به نام خدا

مرکز دانلود رایگان
مهندسی متالورژی و مواد

www.Iran-mavad.com

ASM HANDBOOK

VOLUME

11

Failure Analysis and Prevention

ASM INTERNATIONAL®

The Materials Information Company

Publication Information and Contributors

Introduction

Failure Analysis and Prevention was published in 2002 as Volume 11 of the *ASM Handbook*. The Volume was prepared under the direction of the ASM Handbook Committee.

Volume Coordinator

The Volume Editors were William T. Becker and Roch J. Shipley.

Authors and Contributors

- **Jonh Layer**

  NASA

- **Tom Adler**

  Albany Research Center

- **Rehan Ahmed**

  Heriot-Watt University

- **Debbie Aliya**

  Aliya Analytical

- **Stephen D. Antolovich**

  Washington State University

- **Roy G. Baggerly**
- **Ray Bayer**
- **William T. Becker**
- **M. Belassel**

  Proto Manufacturing Limited

- **Daniel J. Benac**

  Bryant-Lee Associates

- **Jayashree Bijwe**

  Indian Institute of Technology, Delhi

- **R.R. Blackwood**

Tenaxol Inc.

- **John B. Bowles**

  University of South Carolina

- **M.E. Brauss**

  Proto Manufacturing Limited

- **Charlie Brooks**

  University of Tennessee

- **Mohan D. Chaudhari**

  Columbus Metallurgical Service

- **Yan-Ming Chen**

  CETIM

- **Ashok Choudhury**

  Lockheed Martin Energy Research

- **Scott S. Chumbley**

  Ames Laboratory

- **Carl Czajkowski**

  Brookhaven National Laboratory

- **Daniel R. Danks**

  Danks Tribological Services

- **Daniel P. Dennies**

  The Boeing Company

- **Bill Dobson**

  Binary Egnineering Associates, Inc.

- **Don Duvall**

  Engineering Systems Inc.

- **Rob Dwyer-Joyce**

  The University of Sheffield

- **Susan Freeman**
- Millennium Metallurgy, Ltd.
- **Arun M. Gokhale**
- Georgia Institute of Technology
- **Mark Hadfield**

  Bournemouth University

- **Crispin Hales**

  Triodyne Inc.

- **Larry D. Hanke**

  Materials Evaluation and Engineering, Inc.

- **Jeffrey A. Hawk**

  Albany Research Center

- **Thomas R. Jack**

  NOVA Research and Technology Center

- **Jeffrey Jansen**

  Stork Technimet, Inc.

- **L.M. Jarvis**

  Tenaxol Inc.

- **Kent Johnson**

  Engineering Systems Inc.

- **Mitchell Kaplan**

  EAD Aerospace, Inc.

- **Randy K. Kent**

  MDE Engineers Inc.

- **Matthew T. Kiser**

  Caterpillar Inc.

- **John Landes**

  University of Tennessee

- **Allan D. Lang**
- **Iain Le May**

Metallurgical Consulting

- **Roger Lewis**

  The University of Sheffield

- **Cameron Lonsdale**

  Standard Steel

- **Russell A. Lund**
- **Charles R. Manning**

  Accident Reconstruction Analysis Incorporated

- **William McBrine**

  Altran Corporation

- **Robert McCabe**

  AMC

- **Steve McDanels**

  National Aeronautics and Space Administration

- **Dennis McGarry**

  FTI-SEA Consulting

- **Daryll W. McKinley**

  McSwain Engineering, Inc.

- **Richard H. McSwain**

  McSwain Engineering, Inc.

- **Brett A. Miller**

  Stork Technimet Corporation

- **Harry R. Millwater, Jr.**

  University of Texas at San Antonio

- **M. Narazaki**

  Utsunomiya University (Japan)

- **Howard Nelson**
- **Dana E. Netherton**

Athos Corporation

- **John G. Newman**

  Physical Electronics Inc.

- **Ronald J. Parrington**

  IMR Test Labs

- **Cheryl Pattin**

  Triodyne Inc.

- **Ray Pelaez**

  CorrTech Inc.

- **James Pineault**

  Proto Manufacturing Limited

- **Robert B. Pond, Jr.**

  M-Structures Inc.

- **Gordon Powell**

  The Ohio State University

- **Julian J. Raphael**

  Columbus McKinnon Corporation

- **Harold S. Reemsnyder**

  Fracture Technology Associates

- **Roger H. Richman**

  Daedalus Associates Inc.

- **Michel Rigaud**

  École Polytechnique (Montreal)

- **Ashok Saxena**

  Georgia Institute of Technology

- **James J. Scutti**

  Massachusetts Materials Research, Inc.

- **Tom Service**

  Altran Corporation

- **Shahram Sheybany**

  Pacific Metallurgical Company

- **David A. Shifler**

  Naval Surface Warfare Center

- **Roch J. Shipley**

  Packer Engineering Inc.

- **Sujeet Sinha**

  National Institute of Advanced Industrial Science and Technology

- **Charles O. Smith**

  Engineering Consultant

- **V.P. Swaminathan**

  Turbomet International

- **Aaron Tanzer**

  Siemens-Westinghouse

- **R. Bruce Thompson**

  Iowa State University

- **George E. Totten**

  G.E. Totten & Associates LLC

- **George F. Vander Voort**

  Buehler, Ltd.

- **James R. Varner**

  Alfred University

- **Frank W. Veit**

  Siemens Dematic Rapistans Systems Division

- **William R. Warke**
- **Thomas C. Wenzel**

Accident Reconstruction Analysis Incorporated

- **Bryan E. Wilde**
- **Richard Wilson**

Albany Research Center

- **Paul H. Wirsching**

University of Arizona

- **Timothy A. Wolff**

EAD Aerospace, Inc.

Foreword

This 2002 edition of *Failure Analysis and Prevention* is the outcome from the devoted efforts of volunteer editors, authors, and reviewers, who have helped organize and develop this revised Volume 11 of the *ASM Handbook* series. This publication would not have been possible without their vision and dedicated efforts in the ongoing improvement of engineering knowledge and education through the analysis, understanding, and prevention of failure.

As noted in the Preface, the authors and editors assembled this Volume with several broad themes in mind. The nature of failure is complex, varied, and unanticipated. Its prevention can also be multifaceted and varied. In this way, failure analysts are not only specialists, but also educators who help others become aware of the root cause(s) of failure. This requires a clear understanding of the many stages in the life of a part from design and manufacturing to anticipated service, inspection, and maintenance. It also involves a host of tools and techniques for effective planning and implementation of a failure investigation.

Thus, failure analysis and prevention can be a complex multidisciplinary activity that requires broad knowledge in design, manufacturing, mechanics, materials, and testing. The editors and authors have tackled this complex nature of failure analysis and prevention in an updated volume that is, in many respects, an all-new volume. This new edition contains over 50 new articles with expanded coverage on the four basic types of failures (deformation, fracture, corrosion, and wear) and the variety of tools and techniques for effective planning, organization, implementation, and reliable conclusion of a failure investigation through proper interpretation of information.

We would like to extend our thanks to the devoted community of volunteers who have helped organize and develop this 2002 edition of *Failure Analysis and Prevention.* The editors, authors, and reviewers are to be commended for their fine contributions on a vital topic for all engineering disciplines, in the very best of tradition of the Handbook series. We especially thank Bill Becker, Roch Shipley, Debbie Aliya, Dan Benac, Larry Hanke, Jeff Hawk, Steve McDanels, Richard McSwain, Ron Parrington, Jim Scutti, Aaron Tanzer, and Richard Wilson. This publication would not have been possible without their vision, knowledge, and efforts.

Gordon H. Geiger
President
ASM International
Stanley C. Theobald
Managing Director
ASM International

Preface

Welcome to the new edition of *ASM Handbook,* Volume 11, *Failure Analysis and Prevention.*

Theme and Purpose of this Volume. The authors and editors assembled this Volume with several broad themes in mind. First, the most important goal of failure analysis is to decrease the occurrence of component failures through the understanding of the root cause for failure. Experienced failure analysts are often frustrated when,

despite extensive engineering research, investigations, and failure analysis reports, the same types of failures occur again and again. When the root cause has been identified as defective global design rather than abuse or misuse, product quality and reliability is improved.

The failure analyst should strive to uncover the underlying or root (technical) cause of the failure. The fact that a specific component appears to have failed in some way does not automatically mean that the component itself is defective. The problem may lie in the way the component was used, inspected, or maintained. If it is truly defective, then the analysis should determine whether the defect originates in design, manufacture (fabrication and assembly), material selection/processing, or unexpected service environment.

This Volume provides a framework for investigating the above issues. In addition to sections devoted to design and manufacture, there are also sections on failures that occur through fracture, corrosion, and wear, as well as an article on failure through deformation. This Volume is also an attempt to address the principles, tools, techniques, and procedures necessary to plan, organize, and conduct a thorough investigation. Not every failure investigation is the same, and a failure root-cause analysis is more than a microstructural examination, a stress analysis, or a chemical corrosion analysis. All of these disciplines, as well as others, may be required to reach a root cause conclusion.

No single volume, no matter how comprehensive, can present all the information that may potentially be needed. The emphasis of this Volume is on general principles with the widest applicability to situations that the reader is likely to encounter. References and sources of further information are provided throughout. While some common types of components or equipment may be included in some detail, not every type of machine can be treated. The reader is encouraged, and in fact urged, to pursue additional sources of information so as to understand the function and history of the component, machine, or system that is under investigation.

Audience. One of the challenges in preparing a work of this type is the diversity of readership. Some readers are students and other novices who may be confronted with a failed part for the first time. They may be looking to the Handbook for guidance on where to start their analysis. Other readers are experienced practitioners, using the Handbook to verify or clarify a critical detail in their analysis. Thus, the contents of this Volume include the essential basics of failure analysis, as well as more advanced discussions from a research perspective.

The discussions of fracture mechanisms are an example of this approach to Handbook organization. The articles “Overload Failures” and “Fatigue Failures” are good starting points for readers wishing to begin their study of fracture. Examination of the fracture surface (when failure did result in fracture) at both the macro and micro scale provides considerable information pertinent to a failure investigation. This subject is introduced in the article “Overload Failures” with some discussion of the mechanisms that may be involved.

For some readers, these may be sufficient, if all they need is to identify the basic fracture mechanism. However, further study can sometimes allow the analyst to learn more about the circumstances of a fracture. Unfortunately, there are few instances in which a single fractographic feature is definitive in identifying a root cause (and to distinguish between abuse and defective design). Casual examination may not distinguish between fine details caused by different fracture processes. Consequently, a detailed study of the fracture surface at both the macroscale and microscale is helpful and may be critical in obtaining a root cause conclusion. The reader who desires a more detailed appreciation and thorough understanding should continue with the article “Fracture Appearance and Mechanisms of Deformation and Fracture” and the article “Stress Analysis and Fracture Mechanics.” These articles introduce quantitative means to relate the fracture process to material properties and, therefore, are critical to distinguishing between abuse or misuse and inadequate quality. Finally, the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals” provides a still more in-depth treatment on the detailed appearances at both the macroscale and microscale, with the intent of extracting the maximum possible information for root-cause failure analysis.

Differences of Opinion. Controversy is, perhaps, inherent in the very nature of failure analysis. If anything, that is even truer today when real or perceived failures are the subject of litigation. The authors have integrated thoughts on legal considerations into many of the articles. However, nothing here should be taken as legal advice. Those who are concerned regarding legal implications should consult competent counsel.

Furthermore, as every circumstance is somewhat unique, the Handbook should be used with care and should not be the sole source of information when critical decisions are to be made. Most articles include extensive references, which should be reviewed if further information is required.

The authors present analyses and interpretations based on scientific principles and experience. All of the articles have been reviewed and edited. However, there can be and still are differences of opinion among failure analysts regarding some issues. It is up to the reader to determine whether the information presented is

applicable and helpful in a particular situation. Experienced analysts should be consulted if there is any doubt. Despite the best efforts of the authors, reviewers, and editors, the reader might find an area that could be improved. If so, please bring this to the attention of the ASM Editors so that your concern can be reviewed and, depending on the consensus of opinion, can be addressed in subsequent printings.

Collaborative Effort. This Volume reflects the efforts of many people. Except for ASM staff, all are volunteers. Many of the volunteers are fully employed and contributed their personal time to the project. Neither they nor their employers receive any compensation for their efforts, except for the satisfaction that accrues from being able to share what they have learned, prevent failures, and contribute to safer, more reliable products. The names of the authors, editors, reviewers, and ASM staff are acknowledged individually elsewhere in this Volume and are too numerous to list here. However, ASM Editor Steven Lampman does deserve special mention for his commitment, dedication, and patience, without which this Volume would not have become a reality.

It has been most enjoyable and professionally rewarding to work with all who were involved in this effort. On behalf of ASM and the readers of this Handbook, we express our appreciation to all for the time and effort expended and for their willingness to share their knowledge and lessons derived from experience. Many of the contributors have established national and international reputations in their respective fields. More than any words of appreciation in a Preface such as this, however, it is our hope that the Handbook itself will be a most fitting tribute to all participants, both now and into the future.

William T. (Bill) Becker
Consultant (retired, University of Tennessee)
Roch J. Shipley
Packer Engineering Inc.

Officers and Trustees of ASM International (2001–2002)

Gordon H. Geiger
President and Trustee
University of Arizona
Donald R. Muzyka
Vice President and Trustee
Stanley C. Theobald
Secretary and Managing Director
ASM International
John W. Pridgeon
Treasurer
Allvac
Aziz I. Asphahani
Immediate Past President and Trustee
Carus Chemical Company

Trustees

Reza Abbaschian
University of Florida
E. Daniel Albrecht
Advanced Ceramics Research, Inc.
Kathleen B. Alexander
Los Alamos National Laboratory
W. Raymond Cribb
Alloy Products
Brush Wellman Inc.
Subi Dinda
DaimlerChrysler Corporation
R.G. (Gil) Gilliland

Oak Ridge National Laboratory
UT-Battelle, LLC
Walter M. Griffith
Air Force Research Laboratory
Andrew R. Nicoll
Sulzer Metco Europe GmbH
George F. Vander Voort
Buehler Ltd.

Members of the ASM Handbook Committee (2001–2002)

Craig V. Darragh
(Chair 1999–; Member 1989–)
The Timken Company
Henry E. Fairman
(Vice Chair 2001–; Member 1993–)
Cooperheat/MQS Inspection Inc.
Bruce P. Bardes (1993–)
Materials Technology Solutions Company
Larry D. Hanke (1994–)
Materials Evaluation and Engineering Inc.
Jeffrey A. Hawk (1997–)
U.S. Department of Energy
Dennis D. Huffman (1982–)
The Timken Company
Dwight Janoff (1995–)
FMC Corporation
Kent L. Johnson (1999–)
Engineering Systems Inc.
Paul J. Kovach (1995–)
Stress Engineering Services Inc.
Donald R. Lesuer (1999–)
Lawrence Livermore National Laboratory
Huimin Liu (1999–)
Ford Motor Company
William L. Mankins (1989–)
Metallurgical Services Inc.
Srikanth Raghunathan (1999–)
Nanomat Inc.
Mahi Sahoo (1993–)
Natural Resources Canada
Karl P. Staudhammer (1997–)
Los Alamos National Laboratory
Kenneth B. Tator (1991–)
KTA-Tator Inc.
George F. Vander Voort (1997–)
Buehler Ltd.
George A. Wildridge (2000–)
Borg Warner Morse TEC Corporation
Dan Zhao (1996–)
Johnson Controls Inc.

Previous Chairs of the ASM Handbook Committee

R.J. Austin
(1992–1994) (Member 1984–1985)
L.B. Case
(1931–1933) (Member 1927–1933)
T.D. Cooper
(1984–1986) (Member 1981–1986)
E.O. Dixon
(1952–1954) (Member 1947–1955)
R.L. Dowdell
(1938–1939) (Member 1935–1939)
M.M. Gauthier
(1997–1998) (Member 1990–2000)
J.P. Gill
(1937) (Member 1934–1937)
J.D. Graham
(1966–1968) (Member 1961–1970)
J.F. Harper
(1923–1926) (Member 1923–1926)
C.H. Herty, Jr.
(1934–1936) (Member 1930–1936)
D.D. Huffman
(1986–1990) (Member 1982–)
J.B. Johnson
(1948–1951) (Member 1944–1951)
L.J. Korb
(1983) (Member 1978–1983)
R.W.E. Leiter
(1962–1963) (Member 1955–1958, 1960–1964)
G.V. Luerssen
(1943–1947) (Member 1942–1947)
G.N. Maniar
(1979–1980) (Member 1974–1980)
W.L. Mankins
(1994–1997) (Member 1989–)
J.L. McCall
(1982) (Member 1977–1982)
W.J. Merten
(1927–1930) (Member 1923–1933)
D.L. Olson
(1990–1992) (Member 1982–1988, 1989–1992)
N.E. Promisel
(1955–1961) (Member 1954–1963)
G.J. Shubat
(1973–1975) (Member 1966–1975)
W.A. Stadtler
(1969–1972) (Member 1962–1972)
R. Ward
(1976–1978) (Member 1972–1978)
M.G.H. Wells
(1981) (Member 1976–1981)
D.J. Wright
(1964–1965) (Member 1959–1967)

Staff

ASM International staff who contributed to the development of the Volume included Steven R. Lampman, Project Editor; Bonnie R. Sanders, Manager of Production; Gayle J. Anton, Editorial Assitant; Nancy Hrivnak and Carol Terman, Production Editors; and Kathryn Muldoon, Production Assistant. Editorial Assistance was provided by Elizabeth Marquard, Heather Lampman, and Mary Jane Riddlebaugh. The Volume was prepared under the direction of Scott D. Henry, Assistant Director of Technical Publications and William W. Scott, Jr., Director of Technical Publications.

Conversion to Electronic Files

ASM Handbook, Volume 11, *Failure Analysis and Prevention*, was converted to electronic files in 2003. The conversion was based on the First printing (2002). No substantive changes were made to the content of the Volume, but some minor corrections and clarifications were made as needed.
ASM International staff who oversaw the conversion of the Volume to electronic files were Sally Fahrenholz-Mann and Sue Hess. The electronic version was prepared under the direction of Stanley Theobald, Managing Director.

Copyright Information (for Print Volume)

Copyright © 2002
by
ASM International
All rights reserved
No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the written permission of the copyright owner.
First printing, December 2002
This book is a collective effort involving hundreds of technical specialists. It brings together a wealth of information from worldwide sources to help scientists, engineers, and technicians solve current and long-range problems.
Great care is taken in the compilation and production of this Volume, but it should be made clear that NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, ARE GIVEN IN CONNECTION WITH THIS PUBLICATION. Although this information is believed to be accurate by ASM, ASM cannot guarantee that favorable results will be obtained from the use of this publication alone. This publication is intended for use by persons having technical skill, at their sole discretion and risk. Since the conditions of product or material use are outside of ASM's control, ASM assumes no liability or obligation in connection with any use of this information. No claim of any kind, whether as to products or information in this publication, and whether or not based on negligence, shall be greater in amount than the purchase price of this product or publication in respect of which damages are claimed. THE REMEDY HEREBY PROVIDED SHALL BE THE EXCLUSIVE AND SOLE REMEDY OF BUYER, AND IN NO EVENT SHALL EITHER PARTY BE LIABLE FOR SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES WHETHER OR NOT CAUSED BY OR RESULTING FROM THE NEGLIGENCE OF SUCH PARTY. As with any material, evaluation of the material under end-use conditions prior to specification is essential. Therefore, specific testing under actual conditions is recommended.
Nothing contained in this book shall be construed as a grant of any right of manufacture, sale, use, or reproduction, in connection with any method, process, apparatus, product, composition, or system, whether or not covered by letters patent, copyright, or trademark, and nothing contained in this book shall be construed as a defense against any alleged infringement of letters patent, copyright, or trademark, or as a defense against liability for such infringement.
Comments, criticisms, and suggestions are invited, and should be forwarded to ASM International
Library of Congress Cataloging-in-Publication Data
ASM International
ASM Handbook
Includes bibliographical references and indexes

Contents: v.1. Properties and selection—irons, steels, and high-performance alloys—v.2. Properties and selection—nonferrous alloys and special-purpose materials—[etc.]—v.21. Composites
1. Metals—Handbooks, manuals, etc. 2. Metal-work—Handbooks, manuals, etc. I. ASM International. Handbook Committee. II. Metals Handbook.
TA459.M43 1990 620.1′6 90-115
SAN: 204-7586
ISBN: 0-87170-704-7
ASM International
Materials Park, OH 44073-0002
www.asminternational.org
Printed in the United States of America
Multiple copy reprints of individual articles are available from Technical Department, ASM International.

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Introduction

ANALYZING FAILURES is a critical process in determining the physical root causes of problems. The process is complex, draws upon many different technical disciplines, and uses a variety of observation, inspection, and laboratory techniques. One of the key factors in properly performing a failure analysis is keeping an open mind while examining and analyzing the evidence to foster a clear, unbiased perspective of the failure. Collaboration with experts in other disciplines is required in certain circumstances to integrate the analysis of the evidence with a quantitative understanding of the stressors and background information on the design, manufacture, and service history of the failed product or system.

Just as failure analysis is a proven discipline for identifying the physical roots of failures, root-cause analysis (RCA) techniques are effective in exploring some of the other contributors to failures, such as the human and latent root causes. Properly performed, failure analysis and RCA are critical steps in the overall problem-solving process and are key ingredients for correcting and preventing failures, achieving higher levels of quality and reliability, and ultimately enhancing customer satisfaction.

This article briefly introduces the concepts of failure analysis, root-cause analysis, and the role of failure analysis as a general engineering tool for enhancing product quality and failure prevention. The discipline of failure analysis has evolved and matured, as it has been employed and formalized as a means for failure prevention. Consistent with the recent trend toward increased accountability and responsibility, its purpose has been extended to include determining which party may be liable for losses, be they loss of production, property damage, injury, or fatality. The discipline has also been used effectively as a teaching tool for new or less experienced engineers.

The importance and value of failure analysis to safety, reliability, performance, and economy are well documented. For example, the importance of investigating failures is vividly illustrated in the pioneering efforts of the Wright Brothers in developing self-propelled flight. In fact, while Wilbur was traveling in France in 1908, Orville was conducting flight tests for the U.S. Army Signal Corps and was injured when his Wright Flyer crashed (Fig. 1). His passenger sustained fatal injuries (Ref 1). Upon receiving word of the mishap, Wilbur immediately ordered the delivery of the failed flyer to France so that he could conduct a thorough investigation. This was decades before the formal discipline called “failure analysis” was introduced.

Fig. 1 Crash of the Wright Flyer, 1908. Courtesy of the National Air and Space Museum, Smithsonian Institution, Photo A-42555-A

Unfortunately, there are many dramatic examples of catastrophic failures that result in injury, loss of life, and damage to property. For example, a molasses tank failed in Boston in 1919, and *another* molasses tank failed in Bellview, NJ, in 1973 (Ref 2). Were the causes identified in 1919? Were lessons learned as a result of the accident? Were corrective actions developed and implemented to prevent recurrence?

Conversely, failures can also lead to improvements in engineering practices. The spectacular failures of the Liberty ships during World War II were studied extensively in subsequent decades, and the outcome of these efforts was a significantly more thorough understanding of the phenomenon of fracture, culminating in part with the development of the engineering discipline of fracture mechanics (Ref 3). Through these and other efforts, insights into the cause and prevention of failures continue to evolve.

References cited in this section

1. P.L. Jakab, *Visions of a Flying Machine: The Wright Brothers and the Process of Invention,* Smithsonian Institution, 1990, p 226

2. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 229–230

3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

---

Concepts of Failure Analysis and Prevention

Clearly, through the analysis of failures and the implementation of preventive measures, significant improvements have been realized in the quality of products and systems. This requires not only an understanding of the role of failure analysis, but also an appreciation of quality assurance and user expectations.

Quality and User Expectations of Products and Systems. In an era that initially gained global prominence in the 1980s, corporations, plants, government agencies, and other organizations developed new management systems and processes aimed at improving quality and customer satisfaction. Some of these systems include Total Quality Management (TQM), Continuous Improvement (CI), and, more recently prominent, Six Sigma. Historically, these initiatives are founded on the philosophies of the quality visionaries W. Edwards Deming (Ref 4) and Joseph Juran (Ref 5).

In their most basic descriptions, TQM and CI represent full organizational commitment to a system focused on “doing the right thing right the first time” and not merely meeting but exceeding customer requirements (Ref 6, 7). They are focused on process improvements, generally in a production environment. Six Sigma adopts these themes and extends the “reach” of the system to all levels of organizations, with a system to achieve, sustain, and maximize business success (Ref 8). Six

Sigma is founded on the use of measurements, facts, and statistics to move organizations in directions that constantly improve and reinvent business processes (Ref 8). The roots of this business system are in the statistical limits set for the maximum number of defects in a product, as a fraction of the total number of opportunities for such defects to occur. To the practitioners of this system, “six sigma” is a statistical metric referring to six times the statistical standard deviation of a normal distribution, which allows no more than 3.4 defects per million opportunities (equivalent to 99.9997% reliability). This is indeed a lofty goal for any organization (be it a manufacturing company, a petrochemical plant, a service business, or a government agency), but companies committed to Six Sigma have reported significant gains in productivity with simultaneous improvements in organizational culture (Ref 7, 8, 9).

The most positive result of these new management systems is that organizations have responded to the higher expectations of consumers and users and have provided higher-quality products and systems, with attendant increases in customer satisfaction. However, this notion of the *quality* of a product or system is multifaceted. Juran described quality as “fitness for use” (Ref 5). TQM defines quality as the ability to satisfy the needs of a consumer (Ref 10). These characteristics of quality also apply internally to those in organizations, either in the services, or in manufacturing, operating, or administering products, processes, and systems (Ref 10). The intent is to provide not only products and systems that garner high customer satisfaction, but also that increase productivity, reduce costs, and meet delivery requirements.

In general, high quality refers to products and systems manufactured to higher standards, in response to higher expectations of consumers and users. These expectations include such attributes as:

- Greater safety
- Improved reliability
- Higher performance
- Greater efficiency
- Easier maintenance
- Lower life-cycle cost
- Reduced impact on the environment

Some or all of these qualities at one time appeared mutually exclusive. However, customer demands and the aforementioned new business-management systems have provided a means of measuring and quantifying these attributes, creating a new paradigm for business. With the business-culture changes that have occurred through the implementation of one or more of the aforementioned improvement systems, users in recent years have experienced, in general, improvement in all of these areas simultaneously. That translates to reduced product failure and greater likelihood of preventing failures. It is important to recognize that, with all the gains achieved under these management systems, the full potential for maximizing these attributes is yet to be achieved.

Though all of the various improvement systems are unique, they have two aspects in common. They are all *customer focused* and are founded on *problem solving* as a means for improvement.

When addressing customer focus, producers and other organizations have identified that the form, fit, function, and service-life requirements of a product or system are actually defined ultimately by customers. Customer-focused manufacturers strive to meet these requirements in designing, developing, and producing their products or systems. In a broad sense, form, fit, function, and service life represent the technically relevant properties of a product. The form, or physical characteristics of components or products, include the size and shape of a product, as well as the materials of construction and the manufacturing techniques used. The manner in which individual components are assembled into and integrate with the product as a whole describes the fit of components. The function of a product or system is its ability or capability to serve the need for which it was intended. Service life is the duration over which the product or system successfully serves its function. These characteristics define products in the customer's eyes. Arguably the most important characteristics, from a consumer's perspective, are how well a product or system functions and how long it serves a useful life.

Problem Solving, Quality, and Customer Satisfaction. Achieving the levels of quality that meet and exceed customer expectations is paramount to customer satisfaction in a customer-focused management system. Since a customer's perspective of quality is strongly tied to the function and service life of a product or system, it follows that failure to provide adequate measures of function and service life presents problems. One proven technique to improving quality is problem solving. Problems can range broadly, from maintenance training issues, to marginal equipment reliability, to business systems conflicts, to policy inconsistencies, to poor working conditions on the shop floor. When a problem occurs, the responsible organization will analyze the problem to determine the cause and solve it. However, due to various business or cultural pressures, some organizations fall into the following pitfalls when problems arise (Ref 9):

- Do nothing and perhaps hope that the problem will go away.
- Deny that the problem exists, minimize its importance, question the motives of those identifying the problem.
- Troubleshoot in a haphazard fashion (i.e., “shotgun” troubleshooting).

- Chase false leads (i.e., “red herrings”).

In an enlightened organizational culture, products or systems require a systematic approach to problem solving, based on analysis, to achieve the levels of quality and customer satisfaction defined by the new management systems. The cultural aspect is critical, as those who have identified problems must be encouraged to come forward. Furthermore, resources and commitment are required to formulate the solutions and implement necessary changes.
Problem-Solving Models. A wide range of problem-solving methods and models are available in the literature (Ref 4, 5, 6, 8, 9, 10, 11, 12), presenting various details of approaches and processes for solving any of the general types of problems defined previously. All of these methods and models are rooted in the scientific method (summarized as follows) (Ref 6):

1. Define the issue
2. Propose a hypothesis
3. Gather data
4. Test the hypothesis
5. Develop conclusions

A concise problem-solving model, adapted from several of the referenced authors, and that has specific applicability to this Volume, is depicted in Fig. 2. The continuous, circular format in the graphic is significant, indicating that the process reinitiates with the identification of a new problem or problems brought to light as a result of the first problem-solving activity. Note the similarity to the classical scientific method summarized previously.

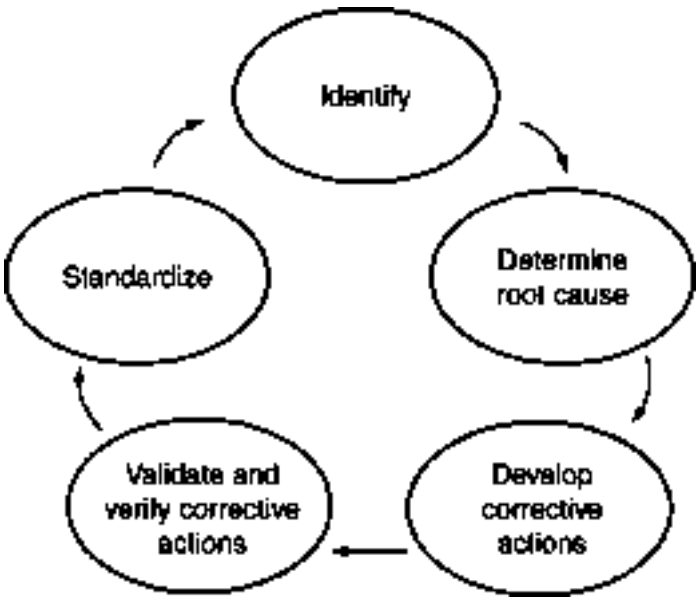


Fig. 2 Problem-solving model

The major steps in the model define the problem-solving process:

1. *Identify:* Describe the current situation. Define the deficiency in terms of the symptoms (or indicators). Determine the impact of the deficiency on the component, product, system, and customer. Set a goal. Collect data to provide a measurement of the deficiency.
2. *Determine root cause:* Analyze the problem to identify the cause(s).
3. *Develop corrective actions:* List possible solutions to mitigate and prevent recurrence of the problem. Generate alternatives. Develop implementation plan.
4. *Validate and verify corrective actions:* Test corrective actions in pilot study. Measure effectiveness of change. Validate improvements. Verify that problem is corrected and improves customer satisfaction.
5. *Standardize:* Incorporate the corrective action into the standards documentation system of the company, organization, or industry to prevent recurrence in similar products or systems. Monitor changes to ensure effectiveness.

The second step in the problem-solving model, *determine root cause,* introduces a very significant process. Solutions to prevent recurrence of problems cannot be developed without identification of the root cause.

Failure Definitions. In the general sense of the word, a failure is defined as an undesirable event or condition. For the purposes of discussion related to failure analysis and prevention, it is a general term used to imply that a component is unable to adequately perform its intended function. The intended function of a component and therefore the definition of failure may range greatly. For instance, discoloration of an architectural feature is a failure of its intended aesthetic function.

Failure can be defined on several different levels. The simplest form of a failure is a system or component that operates, but does not perform its intended function (Ref 13). This is considered a *loss of function.* A jet engine that runs but can only produce partial thrust (insufficient to enable an aircraft to take off) is an example of a loss of function.

The next level of failure involves a system or component that performs its function but is unreliable or unsafe (Ref 13). In this form of failure, the system or component has sustained a *loss of service life.* For example, a wire rope for an elevator has lost its service life when it has sustained fatigue fractures of some of the individual wires, due to irregularities in the wrapping over the sheave. Even though the wire rope continues to function, the presence of fatigue fractures of some of the wires results in an unsafe condition and is therefore considered a failure. Another example of such a failure is the inability of an integrated circuit to function reliably.

In the next level of severity of failure, a system or component is inoperable (Ref 13), such as a pump shaft fracture that causes the impeller to seize or a loss of load-carrying capability of a structural bolt in-service due to fracture.

Failure and Failure Analysis. A logical failure analysis approach first requires a clear understanding of the failure definition and the distinction between an indicator (i.e., symptom), a cause, a failure mechanism, and a consequence. Although it may be considered by some to be an exercise in semantics, a clear understanding of each piece of the situation associated with a failure greatly enhances the ability to understand causes and mitigating options and to specify appropriate corrective action.

Consider the example of a butterfly valve that fails in service in a cooling water system at a manufacturing facility (Table 1). Recognizing the indicators, causes, mechanisms, and consequences helps to focus investigative actions:

- *Indicators(s):* Monitor these as precursors and symptoms of failures.
- *Cause(s):* Focus mitigating actions on these.
- *Failure mechanism(s):* These describe how the material failed according to the engineering textbook definitions. If the analysis is correct, the mechanism will be consistent with the cause(s). If the mechanism is not properly understood, then all true cause(s) will not be identified and corrective action will not be fully effective.
- *Consequence(s):* This is what we are trying to avoid.

Table 1 Example—Failure of a butterfly valve in a manufacturing plant cooling water system

Item	Description	Indicators
Cause	Throttling of valve by the operator outside of the design parameters	Flow gages and records Operator logs
	Low-strength copper nickel alloy construction	Material specifications Laboratory analysis
	Flow-induced cavitation	Rumbling noise in system Vibration of system
Failure mechanism	Erosion-fatigue damage	Laboratory examination of disk, thinning
Consequences	Inability to manufacture at normal production rates	

Life-Cycle Management Concepts. The concept of life-cycle management refers to the idea of managing the service life of a system, structure, or component. There is a cost associated with extending the service life of a component, for example, higher research costs, design costs, material and fabrication costs, and higher maintenance costs. With regard to product failures, it must be understood that failures cannot be totally avoided, but must be better understood, anticipated, and controlled. Nothing lasts and functions forever. For some products, consumers may prefer a shorter life at a more modest cost. In contrast, the useful service life of a product such as an aircraft part may be carefully planned in advance and managed accordingly with routine inspections and maintenance, which may increase in frequency over time. In many cases, avoiding failures beyond a certain predetermined desired life provides no benefit, such as is the case when a surgical implant is designed to far outlive the human recipient. There is also a point of diminishing return on investments related to extending the life of a component. A life-cycle management study of a component would look at these issues as well as other factors such as the issue of obsolescence. How long will it be before the product is obsolete?

Understanding how the typical distribution of failure for a given product must be factored with time is also important when looking at failure patterns (Fig. 3). Early life failures are often associated with fabrication issues, quality-control issues, or initial "shakedown" stresses, while age-related failure rates would increase with time. This is discussed in more detail in the article "Reliability-Centered Maintenance" in this Volume.

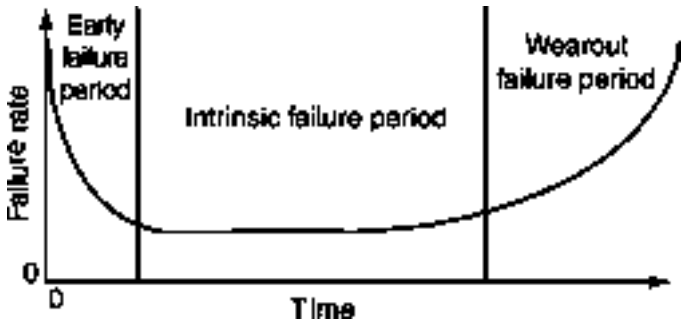


Fig. 3 Typical time distribution of failures ("bathtub curve")

Once the concept of a managed life is prudently adopted over a simple failure prevention concept, design and fabrication costs can be reduced and maintenance and other life-prolonging activities can be optimized.

Diligence in Use of Terminology. Communicating technical information is of paramount importance in all engineering areas, including failure analysis. The choice of technical descriptors, nomenclature, and even what might be considered technical jargon is critical to conveying technical ideas to other engineers, managers, plant personnel, shop personnel, maintenance personnel, attorneys, a jury, and so forth. It is instructive in this introductory article to emphasize that a descriptor can mean something very specific to a technical person and mean something very different to a business manager or an attorney.

For example, the term "flaw" is synonymous with "defect" in general usage. However, to a fracture mechanics specialist, a flaw is a discontinuity such as a crack. Under some circumstances, when the crack is smaller than the critical size (i.e., subcritical), the crack is benign and therefore may not be considered a defect. To the quality-control engineer, flaws are characteristics that are managed continuously on the production line, as every engineered product has flaws, or "deviations from perfection" (Ref 14). On the manufacturing floor, these flaws are measured, compared with the preestablished limits of acceptability, and dispositioned as acceptable or rejectable. A rejectable characteristic is defined as a defect (Ref 14). To the Six Sigma practioner, a defect is considered anything that inhibits a process or, in a broad sense, any condition that fails to meet a customer expectation (Ref 9). To the attorney, a defect refers to many different types of deficiencies, including improper design, inadequate instructions for use, insufficient warnings, and even inappropriate advertising or marketing (Ref 15).

Similar nuances may occur in the basic definitions and interpretations of technical terms used in materials failure analysis. Terms such as ductile and brittle, crack and fracture, and stable and unstable crack growth are pervasive in failure analysis. Even these seemingly basic terms are subject to misuse and misinterpretations, as suggested in Ref 16—for example "brittle cleavage," which is a pleonasm that does not explain anything. Another example noted in Ref 16 is the term "overload fracture," which may be misinterpreted by nonanalysts as a failure caused by a load higher than anticipated by the materials or mechanical engineers. This limited interpretation of overload failure is incomplete, as described in the article "Overload Failures" in this Volume.

Judgmental terminology should be used with prudence when communicating analytical protocols, procedures, findings, and conclusions. Communications during the preliminary stages of an investigation should be factual rather than judgmental. It is important to recognize that some of the terminology used in a failure analysis can be judgmental, and consideration must be given to the implications associated with the use of such terminology. For example, when examining both a failed and an unfailed component returned from service, references to the unfailed sample as "good" and the failed sample as "bad" should be avoided. This is because the investigation may reveal both samples to contain the same defect, and therefore both could be considered "bad." Similarly, *neither* may be "bad" if the analysis actually indicates the failed component met all requirements but was subjected to abuse in service. On completion of the failure analysis, judgmental terminology is often appropriate to use if the evidence supports it, such as in the example of a casting defect that has been confirmed in the example bolt failure analysis.

While discussions of the semantics of terminology may seem pedantic, communicating the intended information gleaned from a failure analysis relies heavily on precision in the use of language.

References cited in this section

4. W.E. Deming, *Out of the Crisis*, MIT Center for Advanced Engineering Study, 1986

5. J.M. Juran and F.M. Gryna, Ed., *Juran's Quality Control Handbook*, 4th ed., McGraw-Hill, 1988

6. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management*, ASQ Quality Press, 1993, p 7

7. F.W. Breyfogle III, *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*, John Wiley & Sons, 1999, p xxvii

8. P.S. Pande, R.P. Neuman, and R.R. Cavanaugh, *The Six Sigma Way*, McGraw-Hill, 2000, p xi

9. M. Harry and R. Schroeder, *Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations*, Doubleday & Co., Inc., 1999

10. G.F. Smith, *Quality Problem Solving*, ASQ Quality Press, 1998, p 7

11. B. Anderson and T. Fagerhaug, *Root Cause Analysis: Simplified Tools and Techniques*, ASQ Quality Press, 2000, p 7, 125

12. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors*, Max Ammerman/Quality Resources, 1998

13. Engineering Aspects of Failure and Failure Analysis, *Failure Analysis and Prevention*, Vol 10, 8th ed., *Metals Handbook*, American Society for Metals, 1975, p 1–9

14. R.K. McLeod, T. Heaslip, and M. Vermij, Defect or Flaw—Legal Implications, *Failure Analysis: Techniques and Applications*, Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 253–261

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause*, 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

16. D. Broek, Fracture Mechanics as an Important Tool in Failure Analysis, *Failure Analysis: Techniques and Applications*, Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 33–44

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

---

Root-Cause Analysis

Failure analysis is considered to be the examination of the characteristics and causes of equipment or component failure. In most cases this involves the consideration of physical evidence and the use of engineering and scientific principles and analytical tools. Often, the reason why one performs a failure analysis is to characterize the causes of failure with the overall objective to avoid repeat of similar failures. However, analysis of the physical evidence alone may not be adequate to reach this goal. The scope of a failure analysis can, but does not necessarily, lead to a correctable root cause of failure. Many times, a failure analysis incorrectly ends at the identification of the failure mechanism and perhaps causal influences. The principles of root-cause analysis (RCA) may be applied to ensure that the root cause is understood and appropriate corrective actions may be identified. An RCA exercise may simply be a momentary mental exercise or an extensive logistical charting analysis.

Many volumes have been written on the process and methods of RCA. The concept of RCA does not apply to failures alone, but is applied in response to an undesirable event or condition (Fig. 4). Root-cause analysis is intended to identify the fundamental cause(s) that if corrected will prevent recurrence.

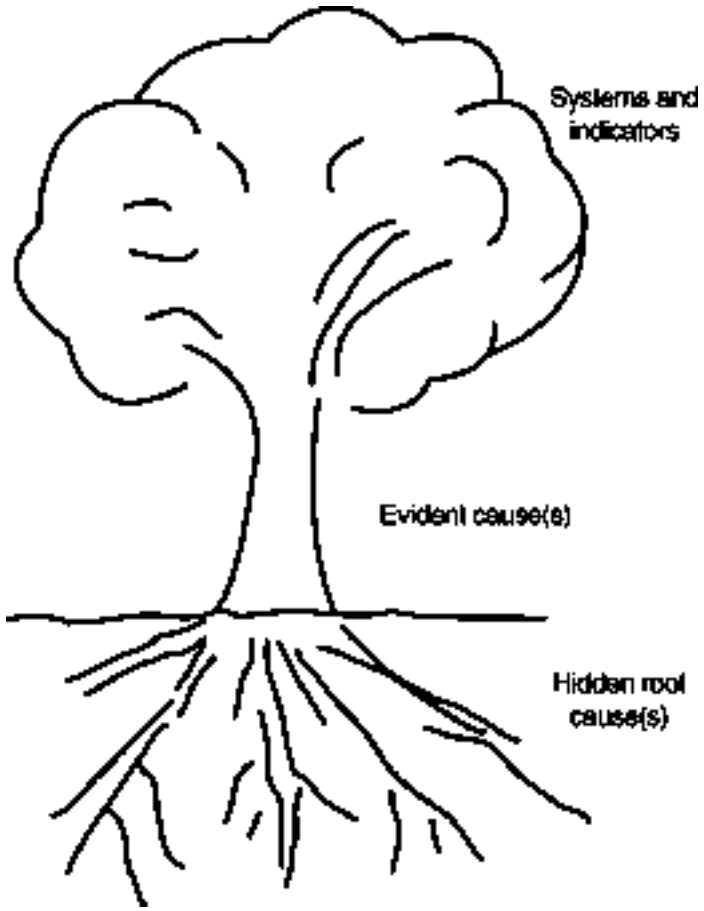


Fig. 4 Root-cause analogy

Levels. The three levels of root-cause analysis are physical roots, human roots, and latent roots (Ref 17, 18, 19, 20, 21). Physical roots, or the roots of equipment problems, are where many failure analyses stop. These roots may be what comes out of a laboratory investigation or engineering analysis and are often component-level or materials-level findings. Human roots (i.e., people issues) involve human factors that caused the failure, an example being an error in human judgment. Latent roots lead us to the causes of the human error and include roots that are organizational or procedural in nature, as well as environmental or other roots that are outside the realm of control. These levels or root cause are best defined by the two examples in Table 2.

Table 2 Examples of root causes of failure of pressure vessel and bolt

Root type	Pressure vessel failure	Bolt failure
Physical roots	Corrosion damage, wall thinning	Fatigue crack; equipment vibration; lack of vibration; isolation
Human roots	Inadequate inspection performed	Improper equipment installed
Latent roots	Inadequate inspector training	Inadequate specification verification process

How deeply one goes into the root causes depends on the objectives of the RCA. These objectives are typically based on the complexity of the situation and the risk associated with additional failures. In most cases, one desires to identify root causes that are reasonably correctable. An example of the variety of possible root causes of an electric motor driven compressor assembly is provided in Table 3 (Ref 22).

Table 3 Possible causes of electric motor driven pump or compressor failures

System design and specification responsibility	Component manufacturer's responsibility	Shipping and storage responsibility	Installation responsibility	Operations and maintenance responsibility	Distress damage or failed components
Application	**Material of construction**	**Preparation for shipment**	**Foundations**	**Shock**	**Distress damages**
Undercapacity	Flaw or defect	Oil system not clean	Settling	Thermal	Vibration
Overcapacity	Improper material	Inadequate	Improper or insufficient grouting	Mechanical	Short/open circuit
Incorrect physical				Improper startup	**Failed components**

System design and specification responsibility	Component manufacturer's responsibility	Shipping and storage responsibility	Installation responsibility	Operations and maintenance responsibility	Distress damage or failed components
condition assumed (temperature, pressure, etc.) Incorrect physical property assumed (molecular weight, etc.) **Specifications** Inadequate lubrication system Insufficient control instrumentation Improper coupling Improper bearing Improper seal Insufficient shutdown devices **Material of construction** Corrosion and/or erosion Rapid wear Fatigue Strength exceeded Galling Wrong hardening method **Design for installation** Unsatisfactory piping support Improper piping flexibility Undersized piping Inadequate foundation	Improper treatment **Design** Improper specification Wrong selection Design error Inadequate or wrong lubrication Inadequate liquid drain Critical speed Inadequate strength Inadequate controls and protective devices **Fabrication** Welding error Improper heat treatment Improper hardness Wrong surface finish Imbalance Lube passages not open **Assembly** Improper fit Improper tolerances Parts omitted Parts in wrong Parts/bolts not	drainage Protective coating not applied Wrong coating used Equipment not cleaned **Protection** Insufficient protection Corrosion by salt Corrosion by rain or humidity Poor packaging Desiccant omitted Contamination with dirt, etc. **Physical damage** Loading damage Transport damage Insufficient support Unloading damage	Cracking or separating **Piping** Misalignment Inadequate cleaning Inadequate support **Assembly** Misalignment Assembly damage Defective material Inadequate bolting Connected wrong Foreign material left in General poor workmanship	**Operating** Slugs of liquid Process surging Control error Controls deactivated/not installed Operating error **Auxiliaries** Utility failure Insufficient instrumentation Electronic control failure Pneumatic control failure **Lubrication** Dirt in oil Insufficient oil Wrong lubricant Water in oil Oil pump failure Low oil pressure Plugged lines Improper filtration Contaminated oil **Craftsmanship after maintenance** Improper tolerances	Sleeve bearing Seal Coupling Shaft Pinion/ball/turning gear Casing Rotor Impeller Shroud Piston Diaphragm Wheel Blades; foil, root, shroud Labyrinth Thrust bearing Pivoted pad bearing Roller/ball bearing Cross-head piston Cylinder Crankshaft

System design and specification responsibility	Component manufacturer's responsibility	Shipping and storage responsibility	Installation responsibility	Operations and maintenance responsibility	Distress damage or failed components
Unsatisfactory soil data Liquid ingestion Inadequate liquid drain Design error	tight Poor alignment Imbalance Inadequate bearing contact Inadequate testing			Welding error Improper surface finish Improper fit General poor workmanship **Assembly after maintenance** Mechanical damage Parts in wrong Parts omitted Misalignment Improper bolting Imbalance Piping stress Foreign material left in Wrong material of construction **Preventive maintenance** Postponed Schedule too long	

Requirements for Effective RCA. Performing an effective RCA requires an interdisciplinary approach in order to ensure that the results are correct and proper corrective actions are identified. In fact, most failures involve factors that spread across many disciplines such as metallurgy, mechanical engineering, hydraulics, electrical engineering, quality control, operations, maintenance, human factors, and others. The analysis team on a complex failure will ideally represent a spectrum of expertise to ensure a very broad perspective.

The best analysis team leader must be a good communicator, have a broad background, be able to integrate factors, and be able to select the best expertise for the project. On less complex failures it is often beneficial to have an individual with a diverse background participate in addition to the specialists, once again to ensure a broader perspective. For example, a metallurgist may be more likely to report a metallurgical deficiency in a product that contributed to the failure, a fabricator is more likely to point to fabrication-related contributors, and a designer is more likely to identify design deficiencies. All of these may be important considerations, but one, all, or none may be a primary root cause. Problems related to people, procedures, environmental concerns, and other issues can also be treated effectively by conducting problem-solving processes and RCAs (although the main focus of this article and this Volume is on materials failure analysis).

References cited in this section

17. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993, p 50

18. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 127

19. M. Paradise, L. Unger, and D. Busch, *TapRoot® Root Cause Tree™ User's Manual,* Systems Improvement, Inc., 1996, p 9–14

20. R.J. Latino and K.C. Latino, *Root Cause Analysis: Improving Performance for Bottom Line Results,* Reliability Center, Inc., 1999, p 79–89

21. C. Nelms, *What You Can Learn From Things That Go Wrong,* 1st ed., Failsafe Network, Richmond, VA, 1994

22. H.P. Bloch and F.K. Geitner, *Practical Machinery Management for Process Plants,* Vol 2, *Machinery Failure Analysis and Troubleshooting,* Gulf Publishing Co., 1983, p 5–6

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Primary Physical Root Causes of Failure

Categorizing schemes for the root causes of equipment failures vary among failure analysis practitioners, quality engineers, other engineers, and managers, as well as legal and insurance professionals (Ref 13, 15, 23, 24, 25, 26, 27). Grouping physical root causes into only a few fundamental categories is advantageous and informative because it defines which aspect of a product or system requires corrective action and prevention strategies. Systematic analysis of equipment failures reveals physical root causes that fall into one of four fundamental categories (Ref 28):

- Design deficiencies
- Material defects
- Manufacturing/installation defects
- Service life anomalies

An effective graphical representation of the impact of defects on the service life of a component or system is provided in the application-life diagram (Fig. 5) (Ref 29, 30). The diagram is constructed by plotting the service lives of components having specific characteristics in the design/configuration, as related to the severity of a specific service condition that is anticipated for the application. Typical characteristics include strength, corrosion resistance, heat treatment condition, flaw size, surface finish, bend radius, void content (i.e., in a casting), degree of sensitization, and so forth. Examples of service conditions include magnitude of stress (either cyclic or static), exposure temperature, aggressiveness of environment, radiation exposure, electrical stress, and so forth.

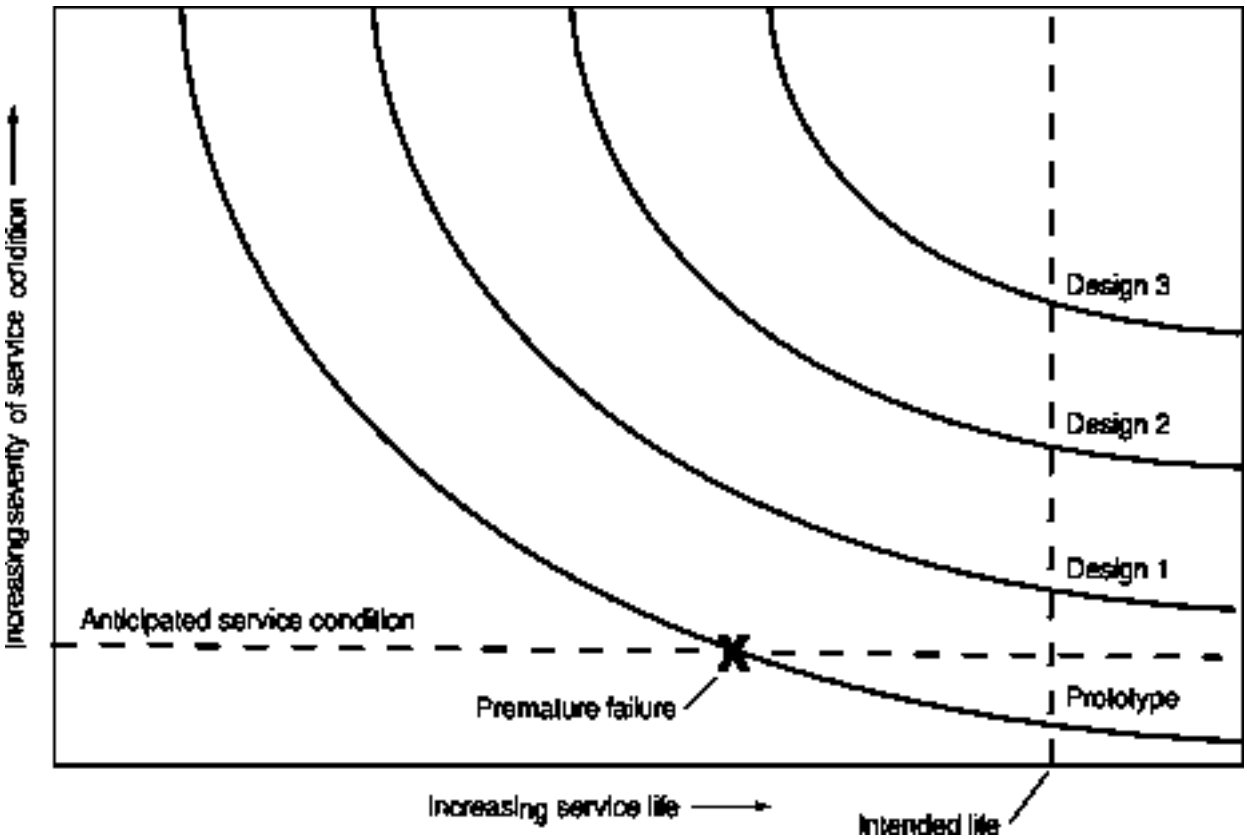


Fig. 5 Application-life diagram comparing the severity of a service condition with the service lives of products having a variable characteristic. This diagram is utilized in specific examples in text.

By varying the characteristics, a family of curves is generated, contrasting the lives of components with the various characteristics and service conditions with the intended service life. Each of the curves represents a different design/configuration characteristic, with increasing degrees of durability as the curves move up the ordinate. Failures can be prevented when the curve for a specific design/configuration lies above the severity of service line, and to the left of the intended service life line. However, if the severity of service conditions increases (either intentionally during operation or as a result of some other change in the system), the propensity for failure may increase, since the characteristics curves intersect the severity of service condition line "to the left," that is, at an earlier point in the service life.

Design Deficiencies

Root causes of failures that stem from design deficiencies refer to unacceptable features of a product or system that are a result of the design process. This process encompasses the original concept development, the general configuration definition, and the detail design, including selection and specification of materials and manufacturing processes. Design involves identifying and defining a need for the product or system, followed by definition of the performance requirements, anticipated service conditions in the application(s), the constraints on the design, and the criticality or risks associated with failure (Ref 31). Discussion of the design process as it relates to failure analysis and prevention is provided in the article "Design Review for Failure Analysis and Prevention" in this Section.

Some examples of design deficiencies include unintended stress raisers due to excessively sharp notches (Ref 32) (e.g., in keyways on shafts) or insufficient radii (e.g., on shafts at bearing journals). Other examples include unanticipated residual stresses associated with heat treating configurations designed with complex geometries, or assembly stresses from configurations that contain unwanted interference. Inappropriate surface treatments could result in failures, such as the use of cadmium plating on an A286 superalloy fastener, subjected to service temperatures above 315 °C (600 °F) (the melting temperature of cadmium is 320 °C, or 610 °F). Two metals specified for use in a wear application could sustain galling if the metals are incompatible, such as sliding wear of components made from 300 series stainless steels.

Selection of a material that is incapable of providing adequate mechanical properties for the application (including strength, fatigue resistance, fracture toughness, corrosion resistance, elevated temperature resistance, etc.) is also a type of design deficiency. Materials can exhibit anisotropy, or variability in properties within a product, such as between the thick

and thin portions of a casting, or between longitudinal and transverse properties in a wrought material. Note that a material can be shown to meet the properties required or specified (i.e., a separately cast tensile bar used to certify a casting, or the longitudinal tensile properties to certify a complex aluminum extrusion), but the specific properties required for the application may rely on the strength, toughness, or stress-corrosion cracking resistance in a direction other than longitudinal.

Design-caused failures include inappropriate geometries (as defined on the engineering drawing), which may lead to a compromise of component or system capabilities. Examples of inappropriate geometries include improper joint preparation for welding or brazing, such as an insufficient or missing groove for a groove weld, insufficient fit-up relief in a socket weld, or inadequate joint overlap in a brazed joint. Other geometry-caused failures can result from insufficient section thickness for a failure based on gross yielding, excessive section thickness in the presence of a flaw for a material of limited fracture toughness, or a fabrication configuration with an excessively sharp forming bend, with the resulting high residual stresses causing a reduction in the fatigue life.

For the example of the excessively tight cold-formed bend radius described previously, an application-life diagram can be constructed as shown in Fig. 6. The service condition considered is stress, and the characteristic that is varied is the radius of the cold-formed bend. Upon examination of the relationship between the characteristic curves and the intended service life, the components having the large and moderate bend radii are found to meet the intended service life at the severity of stress that is anticipated in the specific application. However, in this illustration, the component with the small bend radius sustained a premature failure at the anticipated stress level in the application, since the curve intersects the severity of stress line prior to reaching the intended service life.

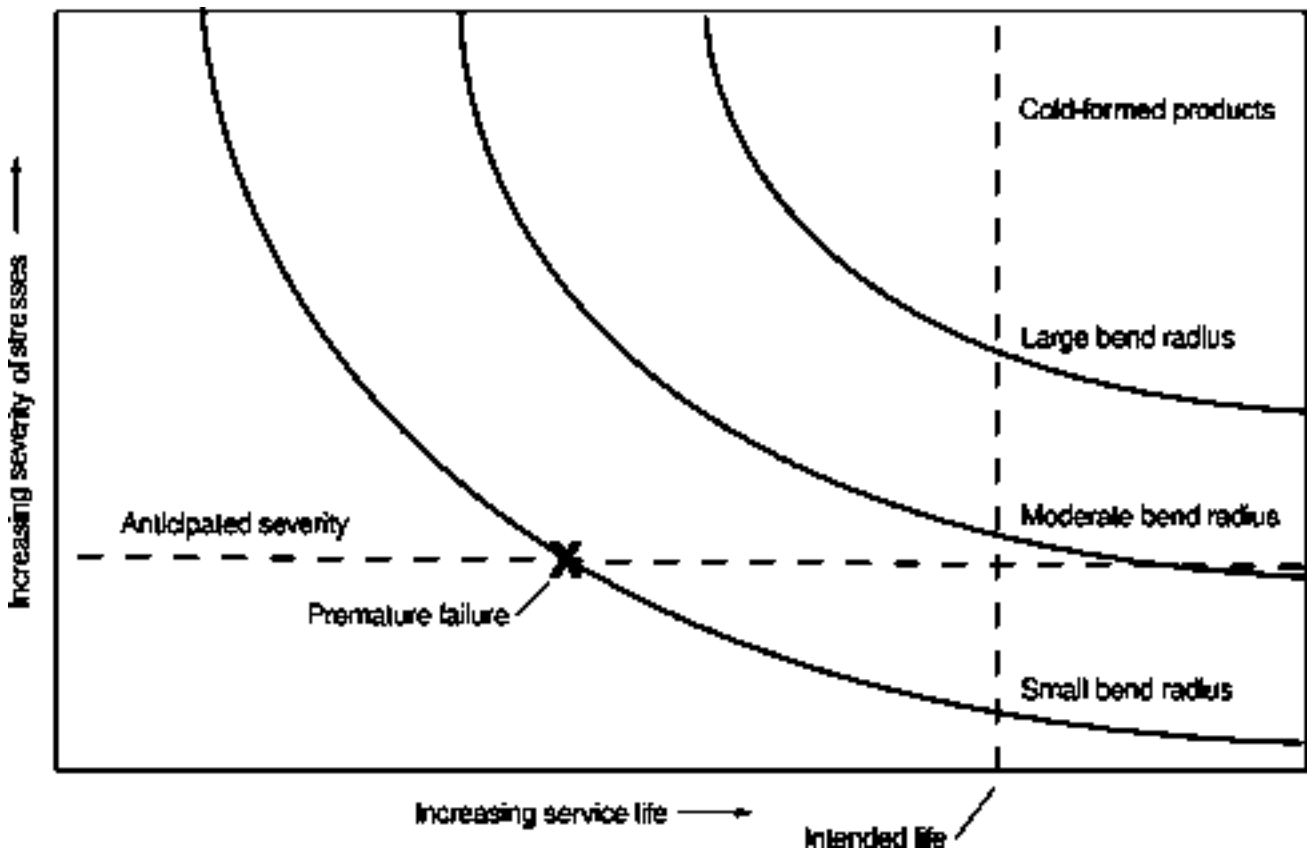


Fig. 6 Application-life diagram for design deficiency

Some of the aforementioned deficiencies in design as well as application-life diagram concepts are illustrated in the following two case histories.

Example 1: Ice Cream Drink Mixer Blade Failures. Excessive assembly stresses and inappropriate detail design caused the premature failures of ice cream drink mixer blades shortly after the mixing machines were introduced into service. A mixer blade as-manufactured is shown on the left side of Fig. 7. As assembled (right side of Fig. 7), the mixer blade is slightly deformed by the contact between the wavy washer at the bottom of the assembly and the bends at the bottom shoulders of the two mixer arms. When properly torqued, the screw that fastens the wavy washer and the mixer blade to the spindle in the center of the assembly places an upward force on the bottoms of the arms (as indicated by the pair of upward facing arrows in Fig. 7). This results in the observed inward deflection of the arms (as indicated by the right and left facing arrows). More significantly, this bending force places the inside radii of the two shoulders of the mixing blade

arms (at the bottom of the blade) in tension. When the mixer is running, the rotational forces further add to the tensile loads on the inside radii of the shoulders.

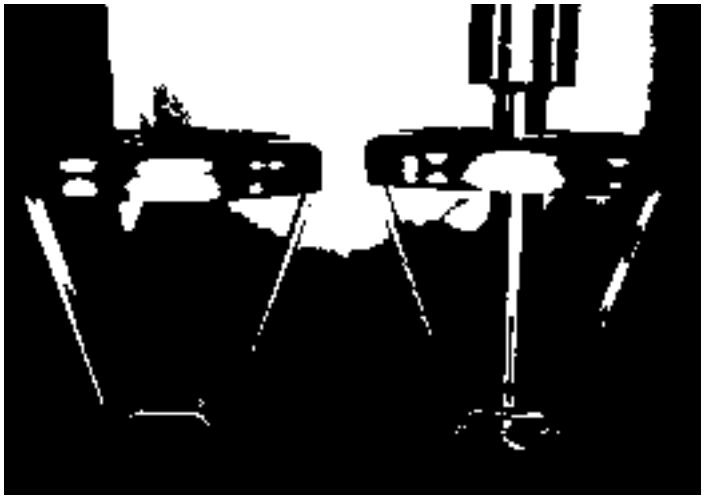


Fig. 7 Ice cream mixer blade as manufactured (left) and assembled to spindle (right)

Analysis of the failed mixer blades revealed multiple fatigue crack origins on the inside radii of the bends at the bottom shoulders (Fig. 8). Metallographic examination of the arm materials revealed additional problems with the configuration: the shoulders on the arms were cold bent, introducing tensile residual stresses on the inside radii of the shoulders and creating a localized area of fatigue susceptibility due to the inherent notch sensitivity of cold-formed 300 series stainless steel.

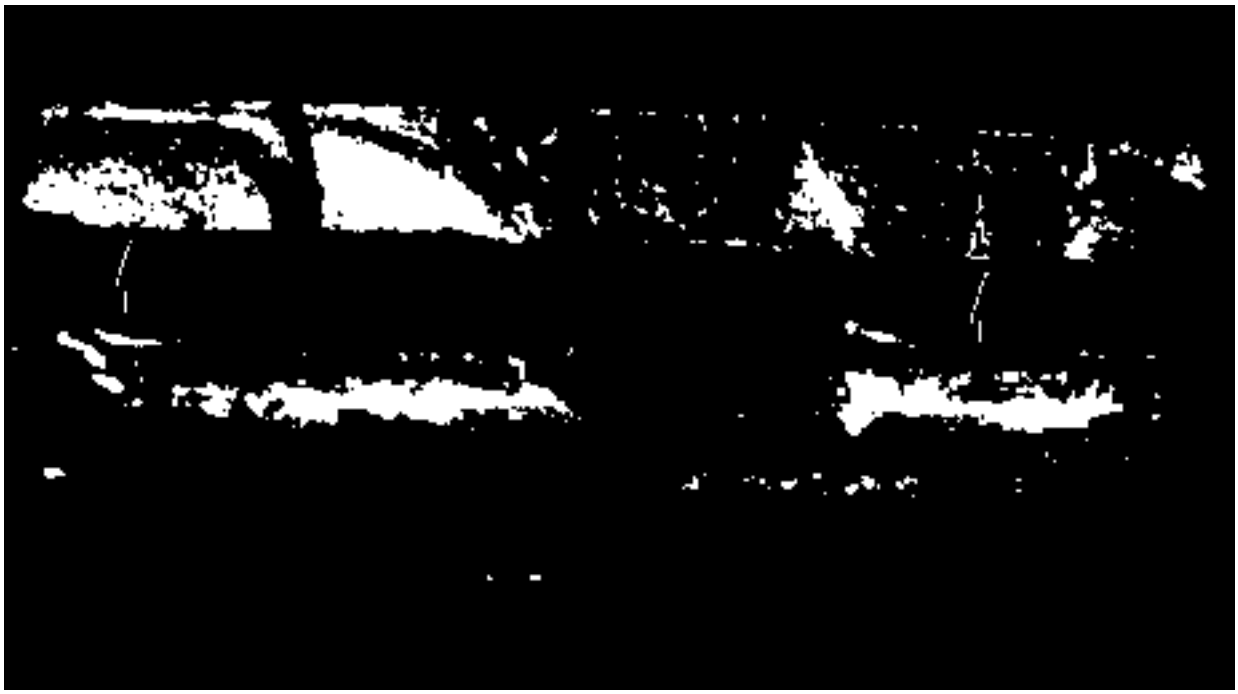


Fig. 8 Fracture surface of failed ice cream mixer blade. Arrows indicate fatigue crack origins. 13×

Clearly, the physical root cause is the design of the mixer blade, which defined two bend areas that contained tensile residual stresses, tensile assembly stresses, and a notch-sensitive microstructure that added to the normal operating rotational and vibratory stresses. The net effect was a reduction in the life of the blade causing loss of function. Corrective-action recommendations included the addition of a stand-off washer between the wavy washer and the bottom shoulders of the blade, or modification of the shape of the wavy washer to prevent contact with the blade shoulders as assembled.

Example 2: Sprocket Locking Device Failure. (Ref 33). A design deficiency involving improper materials selection was revealed through the analysis of a failed tapered-ring sprocket locking device. The device is used to attach a chain sprocket to a shaft without the use of a locking key, enabling the shaft to either drive or be driven anywhere on the shaft (see Fig. 9). The configuration consists of an assembly of four tapered rings (Fig. 10) that are retained by a series of cap screws. As shown in Fig. 11, when the screws are tightened, the middle wedge-shaped rings are pulled closer, forcing the split inner ring to clamp tightly onto the shaft, and the split outer ring to force tightly against the inside diameter of the sprocket. When properly assembled and torqued, the sprocket is fixed to the shaft.

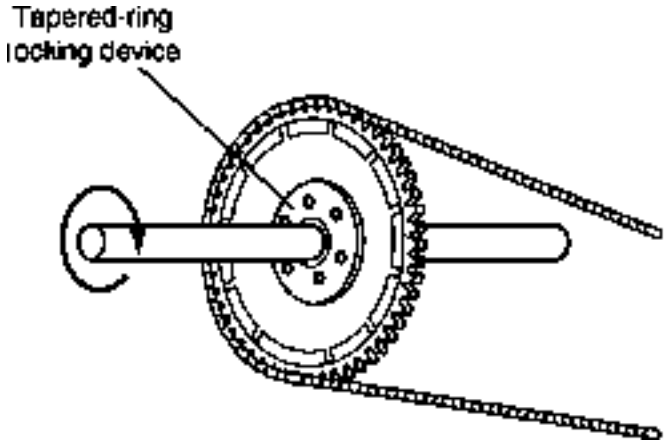


Fig. 9 Sketch of tapered-ring locking device application

Fig. 10 Four tapered rings of locking device. Arrow indicates crack in one of the middle rings.

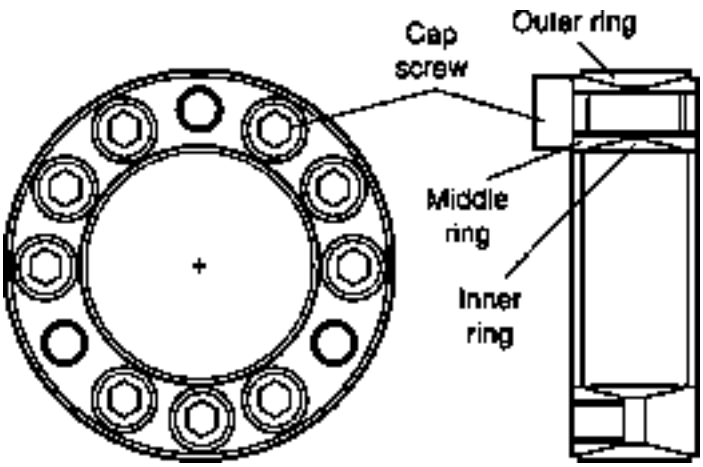


Fig. 11 Plan view (left) and cross section (right) through tapered-ring locking device assembly.

During initial assembly of a new locking device by the manufacturer during a bench test, one of the wedge-shaped middle rings fractured prior to having been fully torqued, preventing the sprocket from being locked to the shaft. The failed assembly was investigated for root cause. One of the middle rings had cracked (Fig. 10, 12a). Examination of the fracture

revealed “woody” fracture features (Fig. 12b), as a result of decohesion between a high volume fraction of manganese sulfide stringers and the matrix (Fig. 13). The matrix fracture features showed ductile dimple rupture.

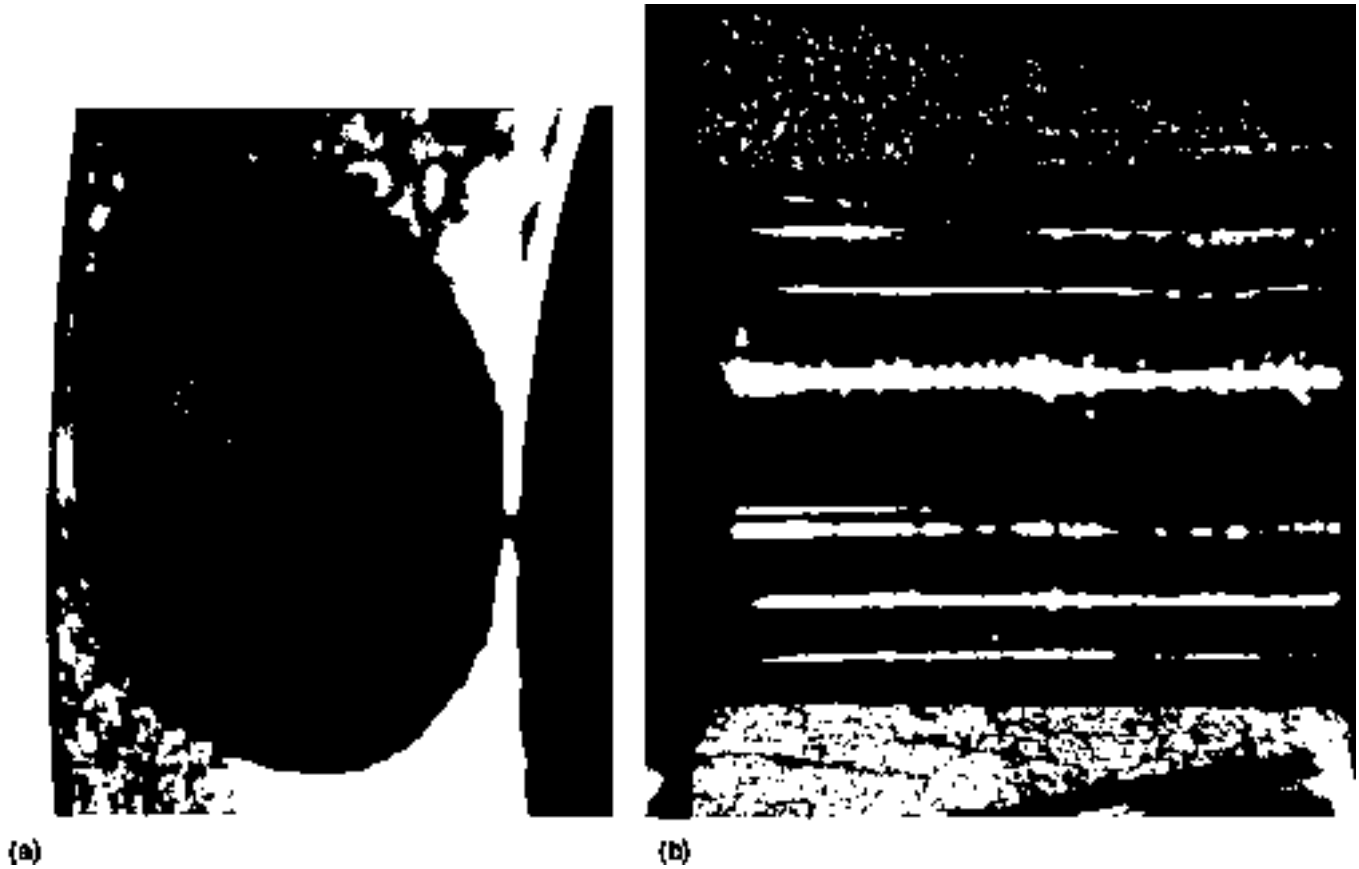


Fig. 12 Crack (a) and broken-open fracture surface (b) of failed wedge-shaped middle tapered ring. 6×

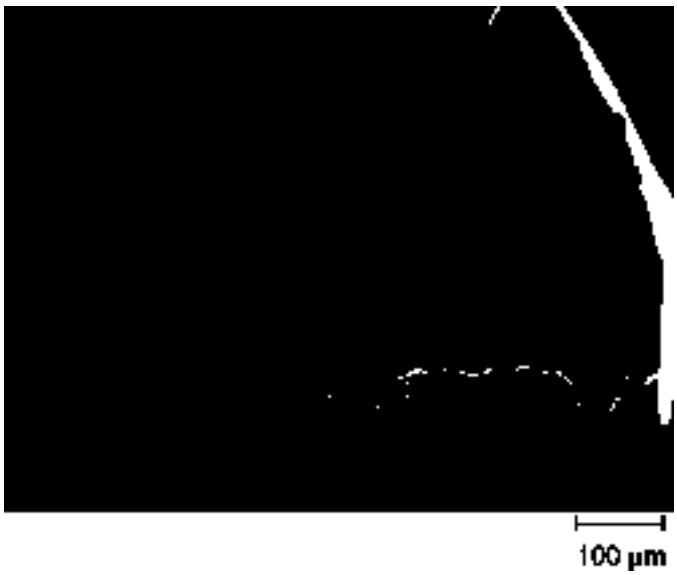


Fig. 13 Higher-magnification view of fracture surface shown in Fig. 12 at origin of cracking. Arrows indicate large manganese sulfide inclusion at origin.

Chemical analysis of the material revealed a resulfurized grade of carbon steel (SAE type 1144, UNS G11440), as required by the manufacturer. This type of steel is marketed as having a rather unusual combination of high strength *and* high machinability. The source of the high strength is in the carbon content and the cold-drawing process used to produce the bar material, giving rise to enhanced *longitudinal* tensile properties. The high volume fraction of manganese sulfide inclusions (Fig. 14) impart the high machinability properties, due to the well-documented enhancement to chipmaking

during machining. The trade-off to this combination of properties, however, is the loss of transverse properties, including strength, ductility, and toughness.

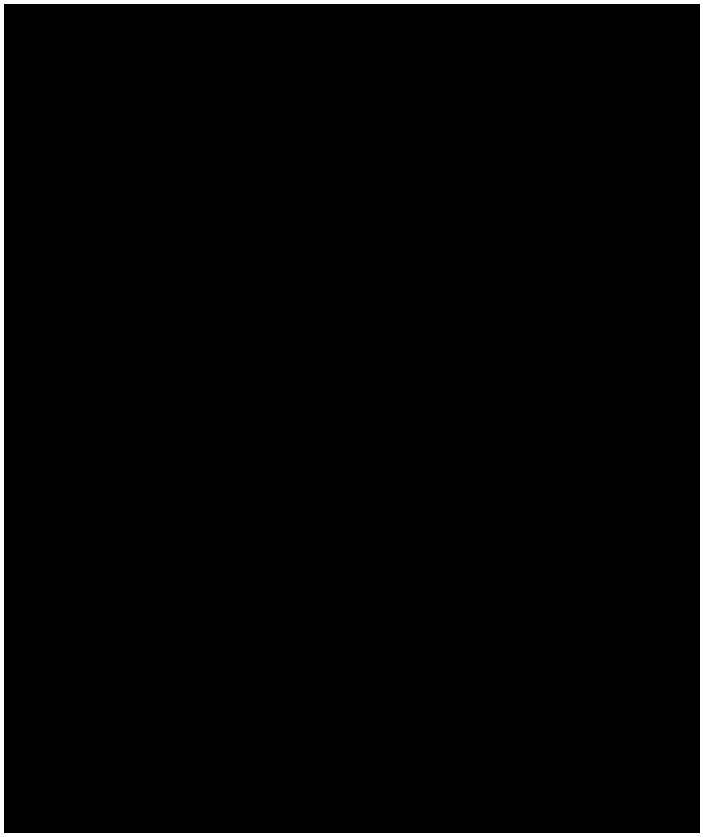


Fig. 14 Significant volume fraction of manganese sulfide inclusions in wedge-shaped tapered ring microstructure. 73×

Analysis of the forces present in the tapered-ring locking device revealed that when the fastening screws were torqued, a significant hoop stress was placed on the middle rings due to the wedging action between the inner and outer rings as well as the relatively small cross section of the middle rings at the fastener holes (see Fig. 11). Since the large inclusion was present at the minimum section thickness zone of the middle ring, the stresses applied to the middle rings during normal torquing caused failure at the inclusion. Since the material contained a high volume fraction of these inclusions, this material choice was not appropriate for this application. The material was weak in an orientation of relatively high stress. Failure prevention recommendations involved specification of a nonresulfurized grade of a low-alloy steel.

Example 2 illustrates some of the complexity and subtlety of RCA. The material was no doubt chosen for its ease of machining. The designer may not have been heavily involved in the material specification or may not have realized the sensitivity of this particular design to material anisotropy. The material itself was not defective or bad, and the part design was reasonable too, except for the material selection, which turned out to be the critical factor in this case.

Material Defects

Unacceptable imperfections or discontinuities in materials are defects, and some types of imperfections may be generally detrimental to the performance or appearance of a product or system. Some of the classical types of material discontinuities that have been identified as causal factor(s) in failures include:

Metal product form	Types of discontinuities
Forgings	Laps Bursts Flakes Segregation

Metal product form	Types of discontinuities
	Cavity shrinkage Centerline pipe Parting line grain flow Inclusions
Castings	Porosity, gas, and microshrinkage Cavity shrinkage Segregation Cold shuts Inclusions
Plate and sheet	Edge cracking Laminations Flakes
Extrusions and drawn products	Edge cracking Seams Steps Central bursts

More detailed descriptions, with physical characteristics and mechanisms for the creation of these defects, are contained in subsequent sections of this Volume. Problems that may develop during subsequent processing, such as heat treating and welding, are discussed in the section “Manufacturing/Installation Defects” in this article.

These material defects can be generally described as discontinuities that degrade the performance of a product in some way. Despite measures taken to control, document, measure, analyze, and improve the processes involved in manufacturing the metal product (such as in TQM and Six Sigma systems), material defects occur. Many defective products are prevented from leaving the mill, foundry, or forge through diligence in adhering to internal procedures and quality-assurance systems. Yet defective materials are sometimes delivered. Depending on the criticality, periodic field inspection may be required and may reveal defects not previously identified. A case study of one such occurrence illustrates the effectiveness of a maintenance plan that includes periodic inspection.

Example 3: Forging Laps in Ski Chair Lift Grip Components. Alloy steel forgings used as structural members of a ski chair lift grip mechanism were identified to have contained forging laps during an annual magnetic particle inspection of all chair lift grip structural members at a mountain resort. A lap in one of the lift grip components (Fig. 15) measured 4.8 mm ($\frac{3}{16}$ in.) long on the surface. An example of the metallurgical cross section through a similar lap is provided in Fig. 16. In accordance with the ASTM standard for magnetic particle inspection, the paint on the forgings was stripped prior to performing the magnetic particle inspection, since the thickness of the paint slightly exceeded the maximum allowable 0.05 mm (0.002 in.) thick paint layer. It should be noted that prior annual inspections, performed at a contracted magnetic particle inspection facility, revealed no significant indications on these forgings. However, the paint was not stripped prior to the magnetic particle inspection at that time.

Fig. 15 Forging lap on ski lift fixed jaw

Fig. 16 Microstructure of forging lap in another ski lift grip component. As-polished. 111×

The presence of the laps, which are rejectable according to the manufacturer's drawings, indicates the forgings were delivered from the manufacturer in this condition. Aside from the obvious procedural roots related to the quality system of the manufacturer, the present issue was whether or not the laps (i.e., sharp-notched discontinuities) had “grown” in a progressive manner, such as by fatigue or stress-corrosion cracking, during the five years that the components had been in service.

The material was confirmed to be 34CrNiMo6 (a European Cr-Ni-Mo alloy steel containing 0.34% C), as required. The broken-open lap (Fig. 17) revealed a darkened area on the fracture surface that was consistent with the dimensions of the lap. The darkened area extended 0.89 mm (0.035 in.) deep. Adjacent to the darkened area, a small area of bright, fibrous fracture features was observed, as well as a transition to a bright, faceted fracture appearance. Scanning electron microscope examination in conjunction with energy-dispersive x-ray spectroscopy (EDS) revealed a heavy oxide on the dark area of the fracture surface (Fig. 18). The bright area adjacent to the dark area contained ductile dimple rupture, which changed to cleavage fracture beyond this area. It was determined through stereomicroscopy, fractography, and metallography that the oxidized portion of the fracture was the preexisting forging lap and that both bright fracture areas were created in the laboratory during the breaking-open process. A cross-sectional view of the broken-open lap is shown in Fig. 19, depicting the field of oxides in the material beneath the lap surface.

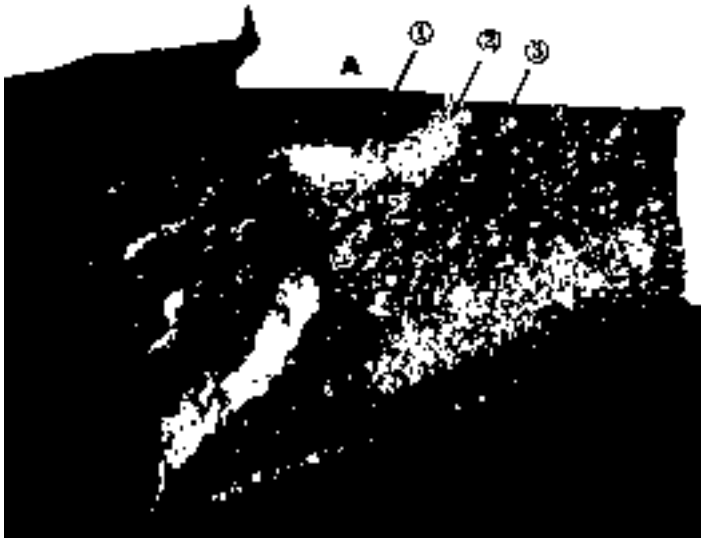


Fig. 17 Broken-open lap. 6×

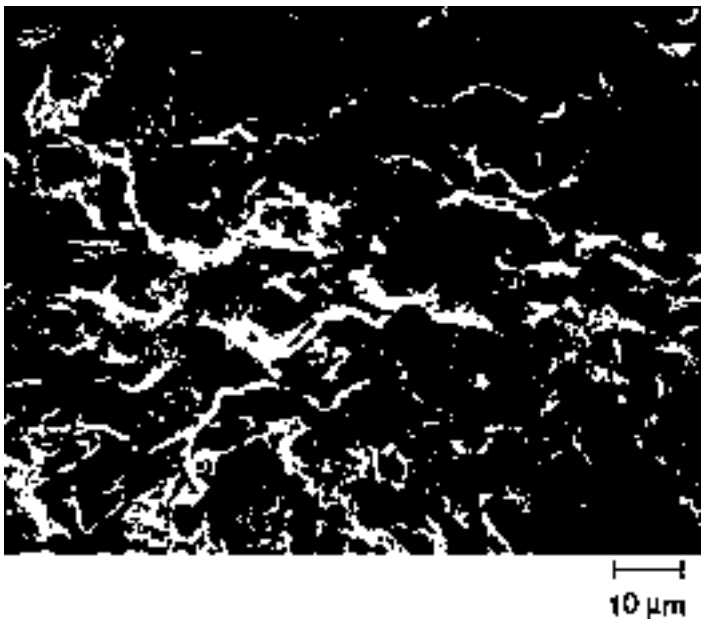


Fig. 18 Scanning electron micrograph of surface features in dark area.

Fig. 19 Micrograph of lap. As polished. 58×

This case is particularly significant in that it is a successful example of *failure prevention* through periodic field inspections. The previously unknown defects were discovered only after magnetic particle inspection procedures adhering

to ASTM standard practices were rigorously followed. Subsequent investigation and analysis of the indications revealed no growth of the laps in service. Nevertheless, the corrective action defined that all forgings showing laps be removed from service. Preventive measures involved critical review and revision of the forging process (so that future lots would be properly forged) and revisions to the nondestructive evaluation (NDE) procedures at the forging supplier.

Building an application-life diagram around this case (Fig. 20) (Ref 29), one can explore the impact of material defects of various sizes on service life. In one possible scenario, the lower curve in Fig. 20 could describe the observed lap, being detectable by NDE and of a size sufficient to sustain growth under the anticipated service conditions *at some time in the future.* However, *at the time of the inspection,* the defect was smaller than that required for crack growth, since the date of the inspection is relatively early in the intended service life of the component. The risk of crack growth and premature failure at some time in the future (as shown by the "X" in Fig. 20) prompted the removal from service of all forgings showing NDE indications.

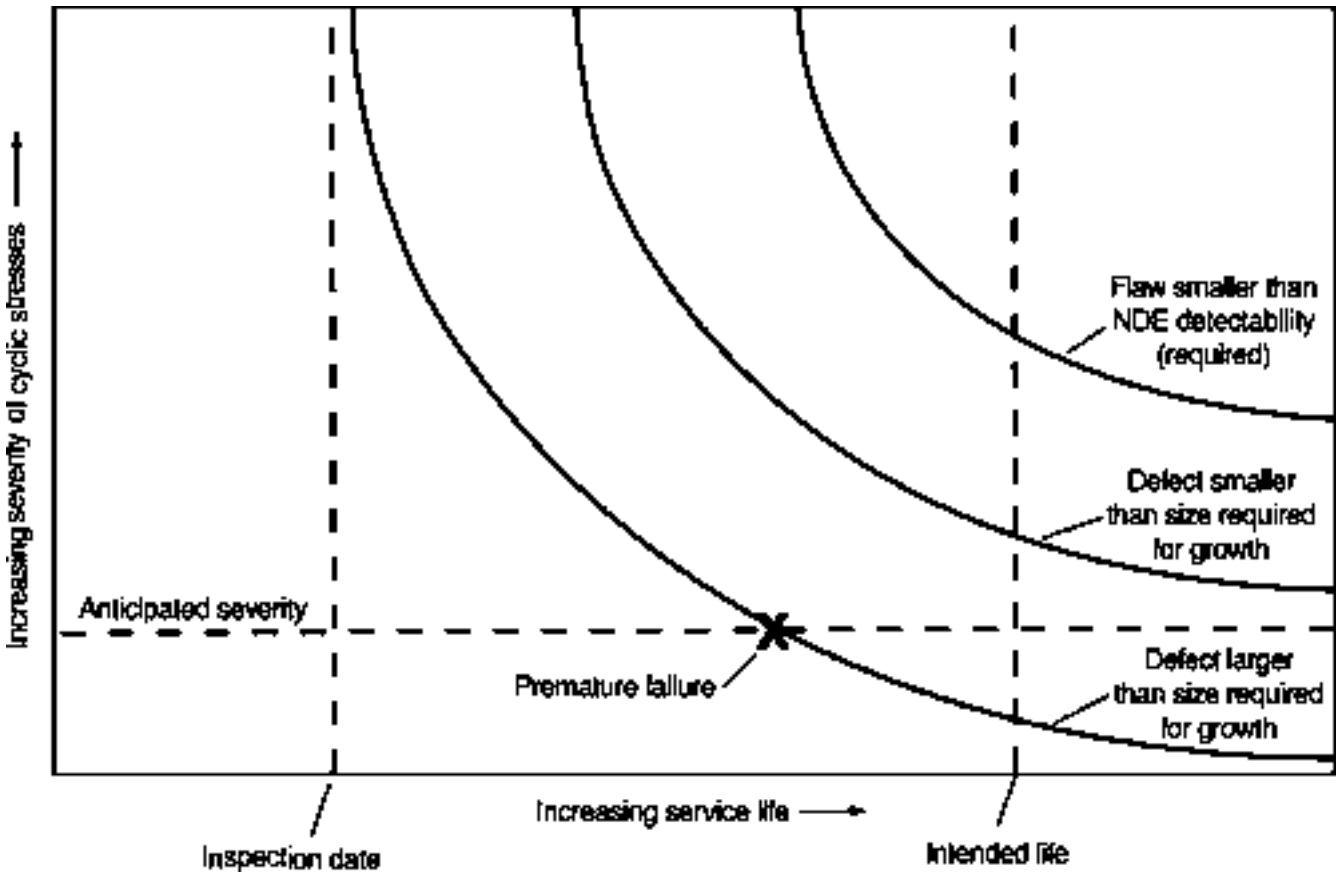


Fig. 20 Application-life diagram showing effects of different sized material discontinuities on service life

Manufacturing/Installation Defects

Manufacture refers to the process of creating a product from technical documentation and raw materials, generally performed at a factory. Installation can be considered manufacturing in-place, such as at a construction site or a new plant. Products can be designed properly using sound materials of construction, yet be defective as delivered from the manufacturer, due to rejectable imperfections (i.e., defects) introduced during the manufacturing process or due to errors in the installation of a system at a site. A wide variety of manufacturing-caused defects exist; each and every manufacturing/installation process has many variables that, when allowed to drift toward or to exceed control limits, can result in a defective product (Ref 34).

Some examples of such manufacturing/installation anomalies are listed below (Ref 35, 36). Failures associated with metalworking, welding, and heat treating operations are also discussed in more detail in other articles in this Volume, and example 4 also illustrates the effects of manufacturing anomalies on the life of a component.

Metal Removal Processes

- Cracks due to abusive machining
- Chatter or checking due to speeds and feeds

- Microstructural damage due to dull tool
- Grinding burn
- Electrical discharge machining recast layer cracking
- Electrochemical machining intergranular attack
- Residual stress cracking due to overheating

Metalworking Processes

- Cracking, tears, or necking due to forming/deep drawing
- Laps due to thread rolling/spinning
- Tool marks and scratches from forming
- Surface tears due to poor surface preparation prior to working
- Residual stress cracking due to flowforming
- Lüders lines due to forming strain rate
- Microstructural damage due to shearing, blanking, piercing
- Overheating damage during spring winding
- Laps and cracks due to shot peening
- Stress-corrosion cracking due to use of improper die lubricants

Heat Treatment

- Grain growth
- Incomplete phase transformation
- Quench cracks
- Decarburization
- Untempered martensite
- Temper embrittlement and similar embrittlement conditions
- Inadequate precipitation
- Sensitized microstructure
- Inhomogeneities in microstructure
- Loss of properties due to overheating during post-plating bake

Welding

- Lack of fusion
- Brittle cracking in heat-affected zone (HAZ)
- Sensitized HAZ
- Residual stress cracking
- Slag inclusions
- Cratering of fusion zone at endpoint
- Filler metal contour out of specification
- Hot cracking
- Cracking at low exposure temperatures
- Hydrogen embrittlement due to moisture contamination
- Liquid metal embrittlement from plating contamination

Cleaning/Finishing

- Corrosion due to inadequate cleaning prior to painting
- Intergranular attack or hydrogen embrittlement due to acid cleaning
- Hydrogen embrittlement due to plating
- Stress corrosion from caustic autoclave core leaching of castings

Assembly at Factory/Installation at Site

- Misalignment
- Missing/wrong parts

- Improper fit-up
- Inappropriate fastening system, improper torque
- Improper tools
- Inappropriate modification
- Inadequate surface preparation

Inspection Techniques

- Arc burn due to magnetic particle inspection
- Intergranular attack or embrittlement due to macroetch
- Fatigue or quench crack from steel stamp mark

Example 4: Forming Process Anomalies in Diesel Fuel Injection Control Sleeve (Ref 28). A user complained of a diesel engine that failed to start in cold weather. Troubleshooting isolated the problem to the diesel fuel control assembly, which was changed out, fixing the problem. Teardown of the fuel control assembly by the manufacturer revealed that a small subcomponent known as the cold start advance solenoid sleeve (Fig. 21) was leaking through the wall. The sleeve operates under relatively high pressure cycles in service. This component is a tubular product with a “bulb” section at one end and threads on the other. The manufacturing method used to create the bulb shape was hydroforming, using a 300 series stainless steel tube in the full-hard condition.

Fig. 21 Cold start advance solenoid sleeve. 0.85×

The leak was attributed to a crack in the sleeve (Fig. 22), in the radius between the bulb area and the cylindrical portion of the sleeve. Scanning electron microscope examination of the broken-open crack revealed fatigue cracks initiated at multiple sites near the outside diameter (OD) of the sleeve (Fig. 23). The crack origins were determined to be extending from shallow (0.013 mm, or 0.0005 in.) zones exhibiting ductile shear (see area between arrows in Fig. 23). Viewing the OD surface of the sleeve adjacent to the fracture plane revealed an extensive network of microcracks on the OD in the radius between the bulb and cylindrical portions (Fig. 24). A cross section through one of the fatigue crack origins revealed slip bands emanating from the microcracks (Fig. 25).

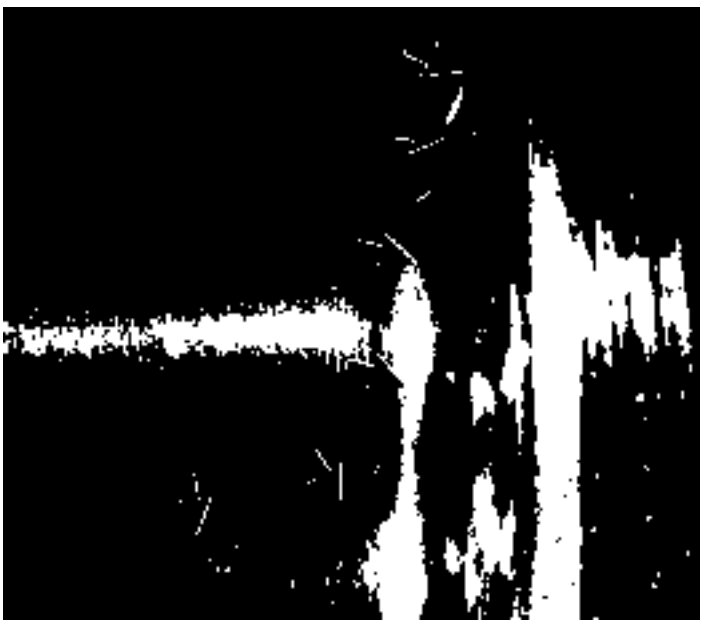


Fig. 22 Crack in sleeve (arrows). 2.5×

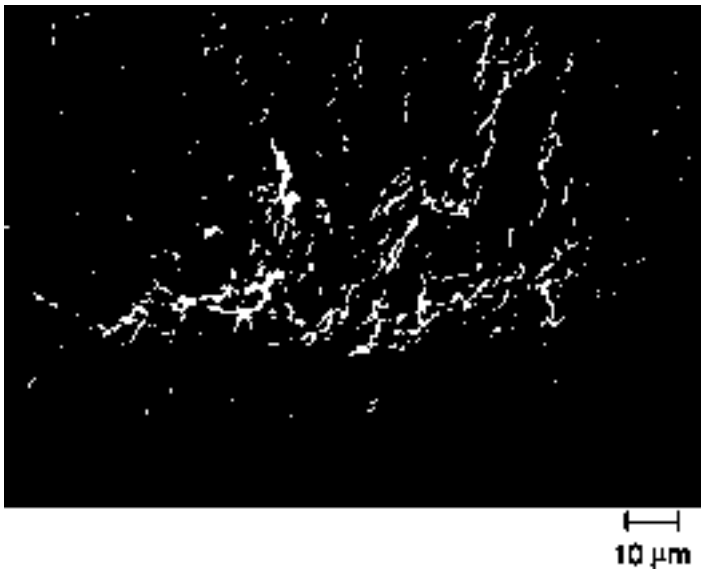


Fig. 23 Fatigue cracking from the outside diameter (OD) of the sleeve (large arrow). Area between small arrows shows evidence of ductile shear at OD surface.

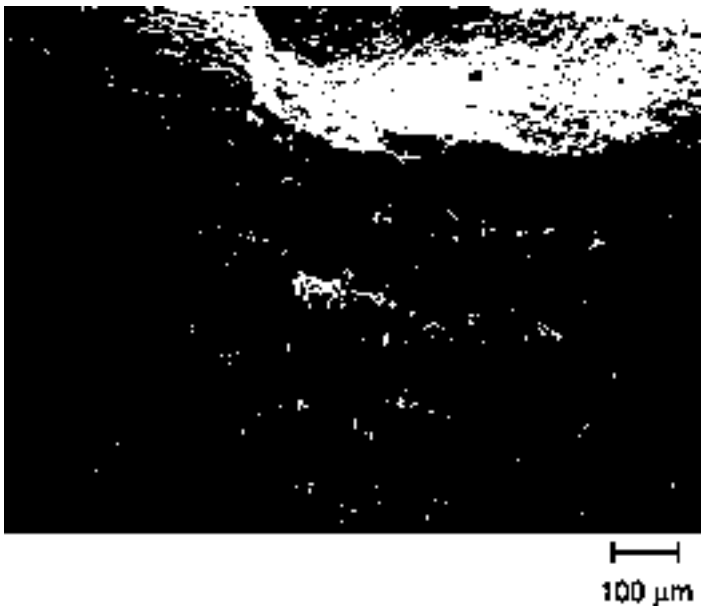


Fig. 24 Network of microcracks (arrows) on the outside diameter surface of the sleeve (lower portion of the micrograph).

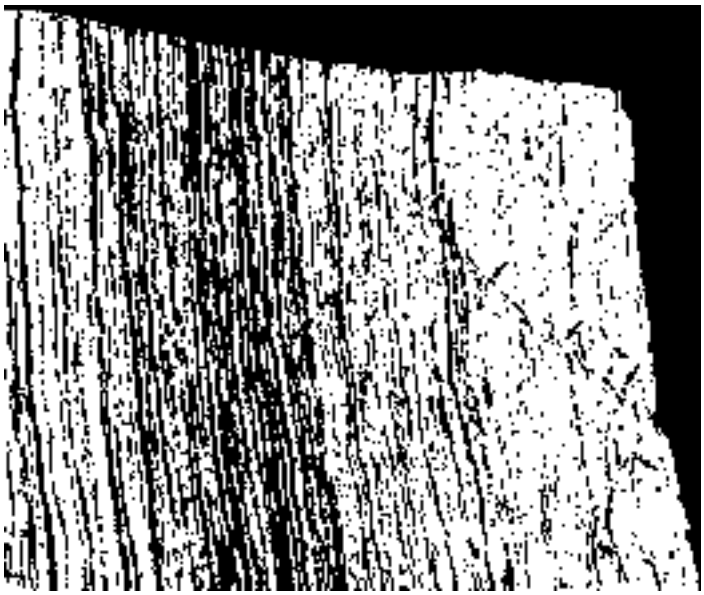


Fig. 25 Microstructure of cross section through outside diameter surface of sleeve adjacent to fracture. Fracture surface is along top of micrograph. Outside diameter surface is along right side of the micrograph. Note slip banding (arrows) emanating from microcrack. 116×

The analysis revealed that during the hydroforming process, heavy biaxial strains were imparted to the sleeve wall, in the radius between the bulb and cylindrical portions of the sleeve. When combined with the heavy strains inherently present in the full-hard 300 series stainless steel, the hydroforming strains in the radius caused the microcracking. The ductile shear areas observed at the origins (see Fig. 23) are microcracks that served to intensify the cyclic service stresses, resulting in fatigue cracks initiating and propagating from these flaws through the wall, causing the leak.

The physical root cause for this failure is a manufacturing process that omitted an intermediate stress relief or annealing treatment prior to hydroforming to the final shape.

Some time later, a similar complaint was received at the factory for a nonstart condition in cold weather. The sleeve was again identified to be leaking due to a through-wall crack. Analysis of the broken-open crack (Fig. 26) revealed fatigue cracks initiated on the *inside diameter* (ID) of the sleeve. This time, the flaw that led to the failure was shallow (approximately 0.005 mm, or 0.0002 in.) intergranular attack on the ID surfaces due to overly aggressive acid cleaning or insufficient rinsing after the acid-cleaning operation. Examination of the OD surfaces revealed no microcracking or evidence of localized strain. Thus a *second* manufacturing defect affecting the same component was identified through failure analysis to have caused the *identical* complaint from the field.

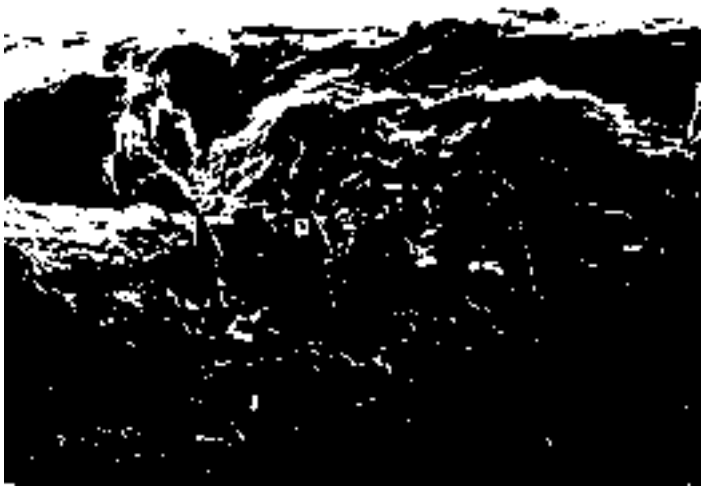


Fig. 26 Multiple fatigue crack origins (arrows) initiating in a network of intergranular attack on the inside diameter of the sleeve. 155×

Using the application-life diagram, the strong effects of minute surface anomalies in this fracture critical component is clearly apparent (Fig. 27). As a result of the severity of the pressure cycles in service, the sleeve cannot tolerate surface flaws.

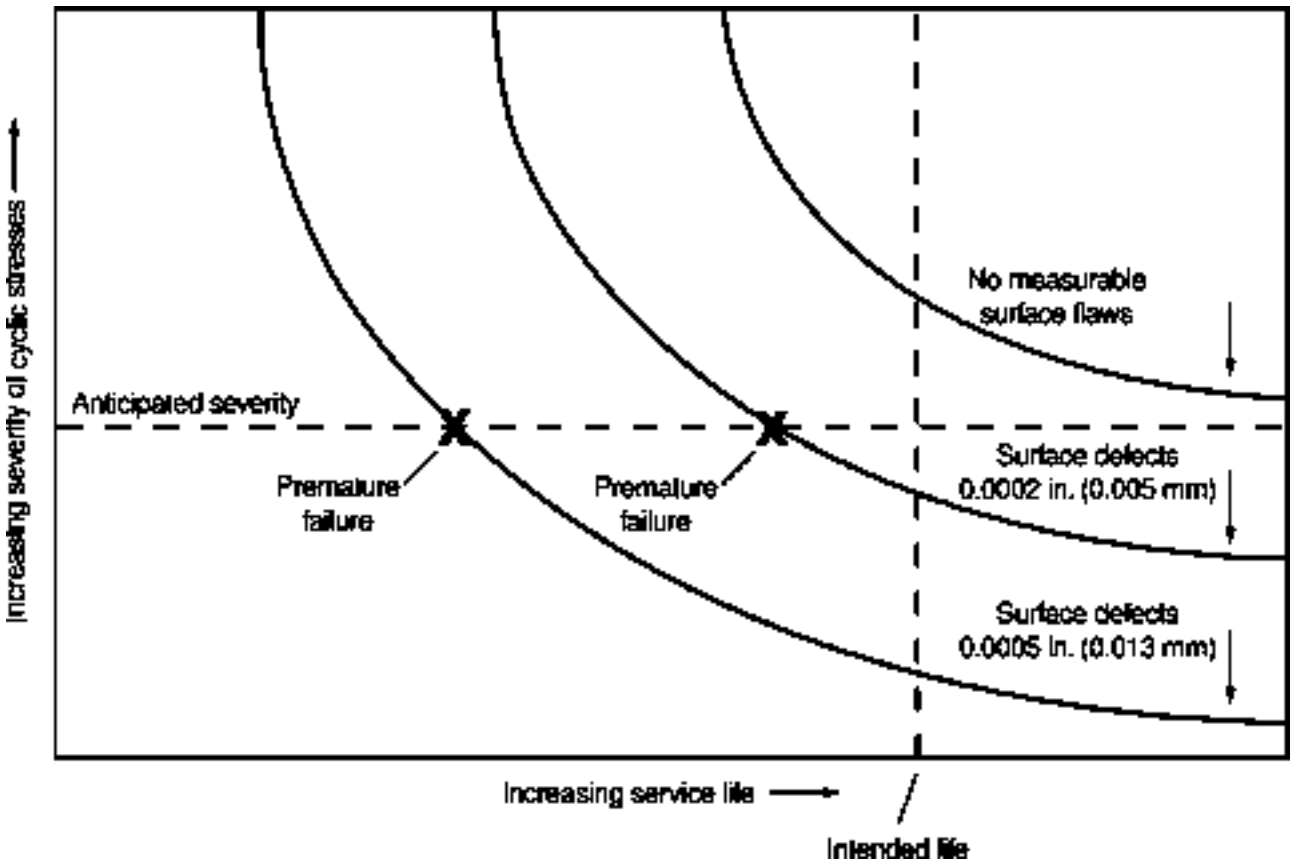


Fig. 27 Application-life diagram showing effects of manufacturing-caused surface discontinuities on service life

Service Life Anomalies

The life of a component or system is heavily dependent on the conditions under which the product operates in service. The service life of a product includes its operation, maintenance, inspection, repair, and modification. Failures due to anomalies in any one of these aspects of service life are unique from those created during the design, procurement of materials, and manufacture of products, as described above. Examples of the types of root causes of failures that result from unanticipated service conditions (Ref 30) are summarized in the following paragraphs.

Operation of the equipment outside of the manufacturer's design parameters would include an example such as a military fighter aircraft in a turn that causes “g” forces that are outside of the operating envelope of the aircraft. Another example is inlet-flow blockage on a high-performance air compressor resulting in excessive cyclic loads applied to the blades causing blade (Fig. 28, 29) and drive shaft (Fig. 30) failures. Failure analysis revealed both the compressor rotor and the shaft sustained fatigue failures.

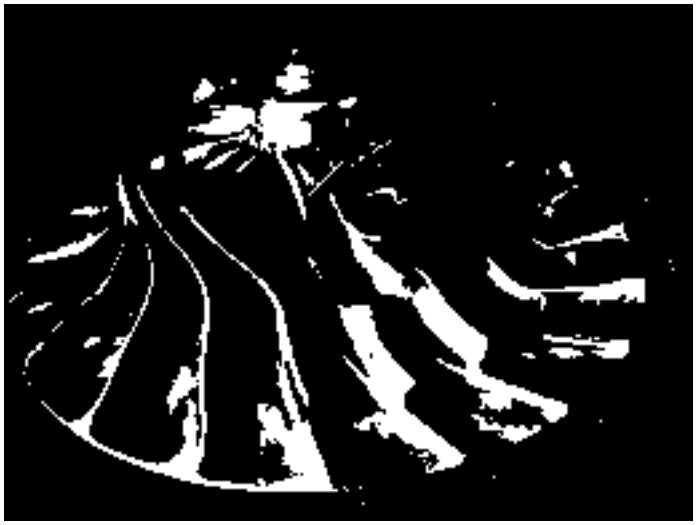


Fig. 28 Failed compressor rotor. Arrows indicate fractured portions of blades. 36×

Fig. 29 Compressor blade fracture surface showing fatigue origins on low pressure (i.e., right) side of blade, as indicated by the arrows. 13×

Fig. 30 Failed compressor rotor shaft. Fracture occurred at radius between large and small diameters. Arrows indicate some of fatigue origins. 1×

Careful fracture analysis revealed fatigue cracks initiated on the low-pressure side of the blades, which are in compression during normal compressor operation. However, when the inlet flow is blocked, particularly when the blockage is only partial, the blades sustain alternating tensile forces, one load cycle per revolution, on the low-pressure side of the blades, resulting in the observed blade fractures. The shaft failed subsequently, due to the severe imbalance and rubbing caused by the blade failures.
Exposure of the product or system to environments more aggressive than forethought would include examples such as:

- Microbiologically influenced corrosion in a cooling-water system using river water in which the ecosystem has changed
- Organic chloride-containing environment exposing a titanium centrifuge bowl, resulting in stress-corrosion cracking
- Faulty sensor cable resulting in an overtemperature condition in a jet engine, which consumes the high-pressure turbine blade life

Improper maintenance would include examples such as:

- Installing a metallic fuel line onto the mating fitting by forcing the tube to align with the mating fitting. Adding the installation stress to the normal cyclic stresses results in a leak due to fatigue cracking.
- Weld repair of a material that is sensitive to high heat cycles, causing brittle cracking and subsequent fatigue failure
- Misalignment of a bearing during rebuild, causing bending loads on the shaft and resulting failure by rotating bending fatigue

Inappropriate Modifications. An example of this would be part-through drill holes in bicycle handlebar stem resulting in fatigue initiation at holes and subsequent fracture (Fig. 31, 32).

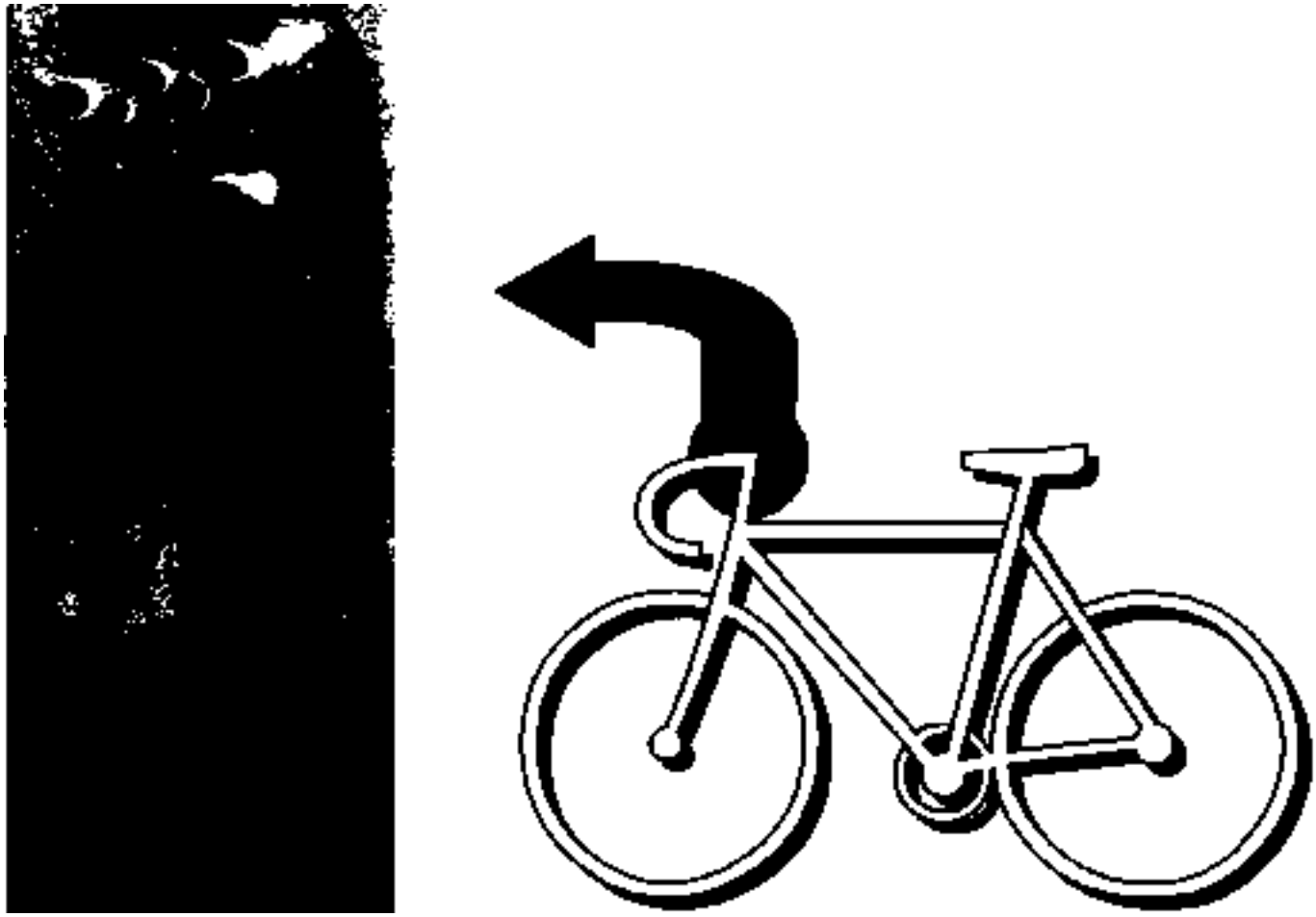


Fig. 31 User-modified bicycle handlebar stem failed in service

Fig. 32 Multiple fatigue initiations at part-through drill holes in user-modified bicycle handlebar stem. 3×

The application-life diagram is useful in exploring the effects of service-life anomalies on the lives of products. For the compressor inlet blockage case described previously, the Fig. 33 depicts the significant loss of service life when the rotor blades sustain the unintended cyclic stresses that occur during an inlet blockage event.

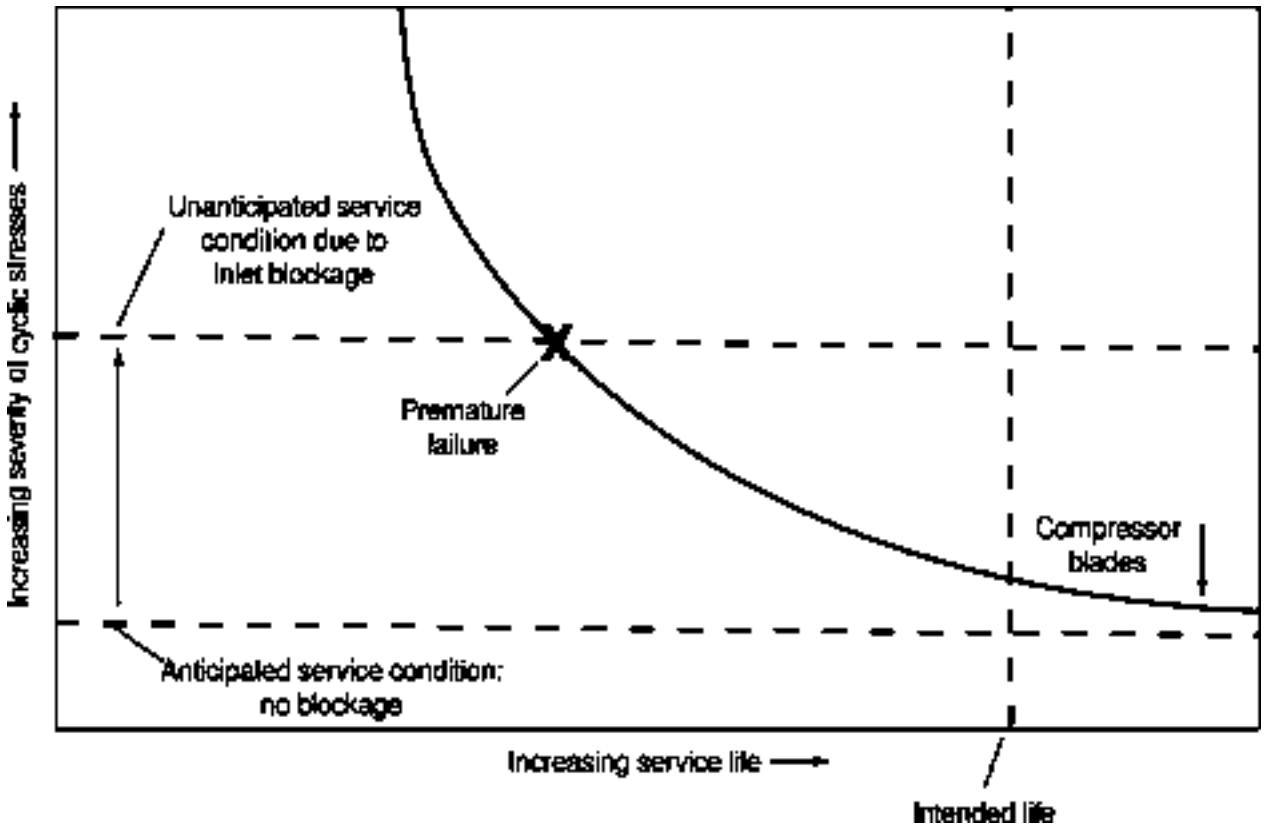


Fig. 33 Application-life diagram showing effects of increasing the severity of the service condition

References cited in this section

13. Engineering Aspects of Failure and Failure Analysis, *Failure Analysis and Prevention,* Vol 10, 8th ed., *Metals Handbook,* American Society for Metals, 1975, p 1–9

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

23. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

24. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998

25. R.K. Mobley, *Root Cause Failure Analysis,* Butterworth-Heinemann, 1999, p 37–39

26. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors,* Max Ammerman/Quality Resources, 1998, p 67

27. *Failure Analysis, The British Engine Technical Reports,* American Society for Metals, 1981

28. F.A. Hossain and J.J. Scutti, Four Fundamental Root Causes of Failure: Case Histories, *Failure Analysis: A Foundation for Diagnostics and Prognostics Development,* Proc. 53rd Meeting of Society for Machinery Failure Prevention Technology, 19–22 April 1999 (Virginia Beach, VA), 1999, p 185–195

29. J.J. Scutti, "Where Things Go Wrong: Where to Look for Failure Prevention Opportunities," Presented to ASM International Materials Solutions Conference, 9–11 Oct 2000

30. J.J. Scutti and F.A. Hossain, Unanticipated Service Conditions, *Proc. Failure Prevention Through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 141–148

31. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Stainless Steels and Special-Purpose Metals,* Vol 3, 9th ed., *Metals Handbook,* American Society for Metals, 1980, p 825–834

32. D. Krashes and J.J. Scutti, Poor Surfaces and Intersections of Surfaces Still Cause Trouble Just Like They Used to Do, *Technology Showcase: Integrated Monitoring, Diagnostics and Failure Prevention,* Proc. Joint Conference, 22–26 April 1996, Society for Machinery Failure Prevention Technology (MFPT), 1996, p 681–690

33. F. Hossain and J.J. Scutti, Failure of Components Although the Causes are Simple & Well Documented, *1998 Technology Showcase,* Proc. Joint International Conference, JOAP International Condition Monitoring Conference, 20–24 April 1998 (Mobile AL), JOAP-TSC, 1998, p 455–464

34. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 53

35. P.F. Wilson, L.D. Dell and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

36. R.P. Baggerly, Preventing Failures Resulting from Machining Issues, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 118–123

Charting Methods for RCA

Many tools exist to assist in performing RCA. The most important element, however, is the preservation of an open mind by the investigator or investigating team. Preconceived ideas or the existence of an investigative bias often obstructs effective root-cause investigations.

A visual representation of an RCA is more easily understood than a long narrative description. Many charting methods have been developed that facilitate the logical organization of information as an aid in performing an RCA. Although such techniques can be invaluable for completeness and logistical analysis, one must not inhibit creativity and an open mind.

The following paragraphs outline a brief and somewhat simplified description of several common charting methods that may be useful in performing an RCA.

A fault-tree analysis is a deductive analysis that identifies a top event, in this case a failure, and then evaluates all credible ways in which this event could have occurred by identifying the interrelationships of basic events or conditions that lead to the failure. The tree is organized by identifying all event strings that lead to the top event and connecting them with a "gate" that depicts the logical relationship. Figure 34 depicts a simplified fault tree.

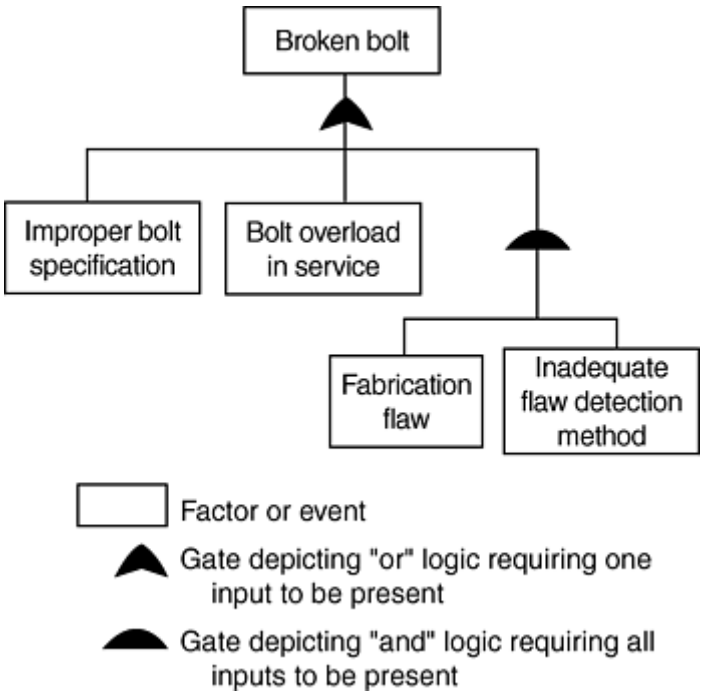


Fig. 34 Simplified fault-tree example

Event and causal factor analysis charting is a very flexible tool that is very useful for performing a logical analysis of the chronological sequence of events and causal factors. The construction starts with a basic timeline with the addition of related conditions, secondary events, and presumptions.

To construct the chart, enclose events in rectangles and connect them in sequence from left to right using solid arrows. The terminal event should be listed at the right-hand end within a circle. In ovals, list conditions, causal factors, and contributing factors and show the relationship between events with dashed arrows.

Barriers may also be added to the chart to identify barriers that failed, allowing events to occur. A barrier can take many forms including a physical barrier such as a locker door or a procedural barrier that was not properly implemented.

The basic elements of the event and causal factor chart (Fig. 35) are primary events, secondary events, and conditions. Events make up the backbone of the chart, while conditions are circumstances pertinent to the situation. The goal of the analysis is to identify the key equipment failures, process failures, or human errors that allowed the loss event to occur. Once the chart is laid out, the causal factors are identified. These are identified as the factors that if eliminated would have prevented the occurrence or lessened the severity of the loss event (Ref 35).

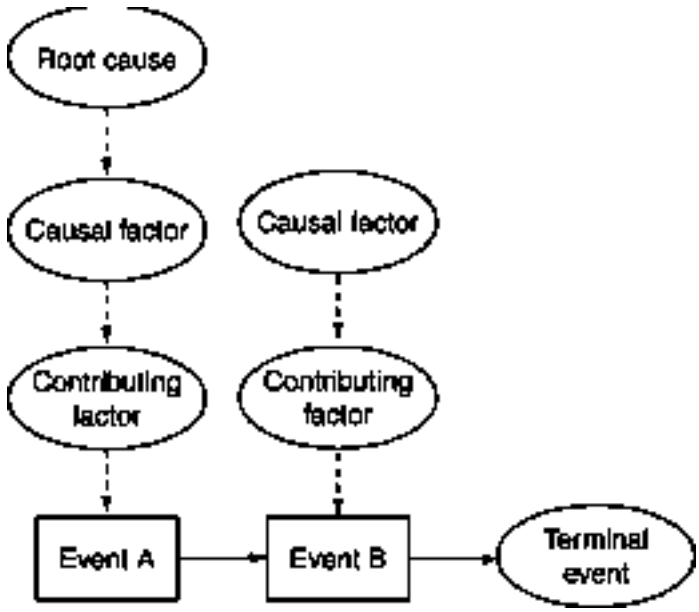


Fig. 35 Simplified event and causal factor chart

Cause-and-Effect Analysis. Failures are always caused to happen. A cause-and-effect analysis is a way to relate causes to a failure in an attempt to find the root cause. Causes can be design problems, human performance, poor fabrication, and so forth. A simple cause-and-effect analysis can take the form of a fishbone diagram (Fig. 36) that can be constructed as follows:

1. Clearly describe the failure at the right side of the diagram.
2. Identify the main cause categories as branches converging on the failure.
3. Brainstorm and list all causes on each branch.
4. Analyze the data until the root cause(s) are identified (Ref 11).

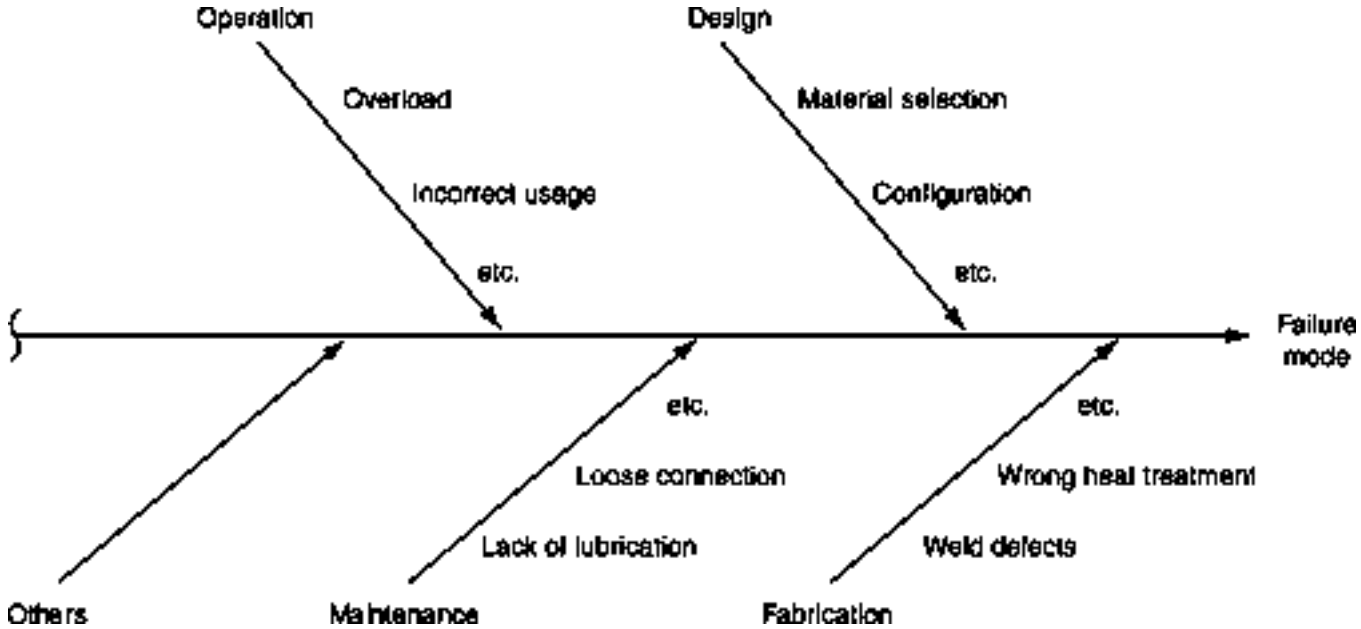


Fig. 36 Simplified fishbone diagram

Five Whys is a simple technique that is intended to lead the user into deeper levels of cause identification, thus leading one further into root cause. The overall objective is to ask "why" after each cause has been identified until true root causes are identified. There actually may be more or less than five "whys" to reach the root-cause level desired (Ref 11). The following example demonstrates this simple concept:

- Event—Highway bridge failure
- Why?—Corrosion damage on structural steel
- Why?—Water collection
- Why?—Debris clogging drainage pipes
- Why?—No maintenance performed to clean pipes
- Why?—Maintenance funding reductions (root cause)

References cited in this section

11. B. Anderson and T. Fagerhaug, *Root Cause Analysis: Simplified Tools and Techniques,* ASQ Quality Press, 2000, p 7, 125

35. P.F. Wilson, L.D. Dell and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Other Failure Analysis Tools

There are many other “tools” that must be considered in performing a failure analysis. In addition to root-cause techniques, tools available to the analyst include:

- Review of all sources of input and information
- People interviews
- Laboratory investigations
- Stress analysis
- Fracture mechanics analysis

Sources of Input

Physical data such as failed parts, samples of environmental influences, photographs, data collection records (pressure, temperature, speed, etc.), and background data are an important part of the investigative process. Forensic analysis of such parts and data is the backbone of any failure investigation. Some of the key elements to an investigation include:

- *Physical evidence:* Broken parts, samples, malfunctioned components, positions, configurations, and so forth. The timely preservation, collection, and recording of physical evidence are essential to any effective failure investigation. The preservation of evidence is done by restricting access to a failure site, preserving of configurations and positions, taking a photographic record of the as-found situation, making sketches, recording of process variables (pressure, temperature, position, etc.), marking and tagging pieces and positions.
- *Background data:* Design data, specifications, technical data, analysis or simulation results, and so forth
- *People:* Witnesses, operators, designers, maintenance personnel, participants, experts

People Interviews

Interviews can provide an essential source of information in any failure investigation. This information, if solicited and documented properly, augments that collected by physical data or research. A very effective way to collect information from people is through the interview process. There are three reasons one would collect data through interviews:

- Firsthand data (witnesses, participants, etc.)
- Background and circumstantial data (historical experiences, related events, situational insights, etc.)
- Expert information (to elicit technical knowledge)

It is essential that those having firsthand data be interviewed as soon as possible after the failure. Important information can be corrupted by losing some of the subtle points over time or by the tendency to have one's firsthand knowledge evolve when discussing the event with other individuals.
Some important points to consider when performing interviews include:

- Explain why the interviews are being performed and maintain confidentiality when possible.
- Interview individually or small groups when possible. Never interview somebody with one's supervisor or manager present or in any other influencing or restricting environment.
- Make the interview environment as comfortable and unintimidating as possible.
- Ask open-ended questions and do not guide the responses.
- Distinguish between firsthand and secondhand knowledge.
- Solicit specific quantitative data, qualitative data, and opinions.
- Get referrals to others who may have pertinent information and other sources of data.
- Recognize biases and paradigms when interpreting answers.

Laboratory Investigations

After pertinent data and samples have been collected, a laboratory investigation is often needed to fully analyze the physical evidence and to identify the failure mechanism. Good procedures in a laboratory begin with good sample collections and handling.

In-Situ Sample Collection and Laboratory Receipt. When collecting samples for laboratory examination, it is a good rule of thumb to collect failed parts, nearby fragments, and lubricant and fluid samples. Collect evidence beyond what is apparent at the time of the initial assessment. Collect undamaged samples of similar components for comparison to the damaged one. Draw diagrams to indicate the position of parts and sample collection locations. Do not be afraid to take many photographs while photodocumenting the scene. Take shots from every angle and always have a scalable object in the photo, preferably a ruled scale. Make in situ markings of fluid levels or other positions that should be recorded prior to disturbing. Having the appropriate documentation and collection tools at a failure site is important to be prepared for activities that may not be anticipated prior to arrival.

Generally, samples should be collected in polyethylene jars or bags using protective gloves and appropriate collection tools. Liquid samples should be collected in glass jars with Teflon-lined covers. Samples for microbiological analysis should be collected in sterile containers and kept cool for prompt analysis. Surfaces should be free from fingerprints or other sources of contamination. Protect samples, particularly delicate items and fracture surfaces, from each other and from other sources of damage.

Tag or label samples in order to indicate when and why it was collected, how it was oriented, who removed it, and what were relevant in situ observations. Generally, it is desirable to collect the largest reasonable sample for laboratory examination prior to sectioning and removal of smaller samples.

Samples received in a laboratory can range from a large component that requires a high-capacity crane to move to something that can only be seen under a microscope. After appropriate collection, receipt, handling, labeling, and appropriate storage of the sample, it is essential to ensure that important evidence is not lost or altered. Samples should always be kept in a dry, secure location and a storage record maintained. A materials safety data sheet (MSDS) should be acquired, and appropriate storage requirement of hazardous material observed. An experienced investigator will also anticipate the disposal of hazardous material after the investigation is completed. For many such materials, disposal in the trash or down the drain is no longer an option. Specialists must be called in to remove and dispose of the material.

Laboratory Analysis. Steps taken in a laboratory after proper receipt may include:

- Initial examination
- Photodocumentation
- Nondestructive examination
- Material verification
- Fractographic examination
- Metallurgical analysis
- Mechanical properties determination
- Analysis of evidence
- Writing of a report

Handling of samples and laboratory techniques employed in a failure analysis are discussed in greater detail in other sections of this Volume (see the article “The Failure Analysis Process: An Overview” for an introduction).

Stress Analysis. Performance of a stress analysis is often a critical part of a structural failure analysis. Stress-analysis techniques are typically used to determine the state of stress as a result of external loadings or other sources of stress such as thermal transients or applied accelerations. Available stress-analysis techniques include hand calculations using theories of strength of materials, approximations derived from reference sources, empirically derived sources and methods, and computerized techniques such as the finite-element analysis (FEA) method.

The FEA method is widely used as both a design tool and a failure analysis investigative tool. Finite-element analysis can be applied to many areas useful in failure analysis, the most common being stress analysis, heat transfer and fluid flow, and electromagnetic properties. Finite-element analysis is able to model complex conditions and handle transient and nonlinear conditions that are typically too complex to perform using hand calculations or other analytical approximations.

The use of FEA in a failure analysis is different from its use in a product design capacity. In a failure analysis, special attention is directed to the failure location. This area of the FEA model may have a finer meshing to capture localized stress concentrations or other localized effects. Applied loadings should include actual load histories that are associated with the failure, including events that are not associated with normal design considerations. This is in contrast to an FEA model used for design that would be used to capture stresses in the entire component as a result of loadings anticipated by design. The results of a failure analysis model would be compared to failure criteria such as shear strength, yield strength, and so forth, or actual observed component deformation. Design models would then be used to qualify the component against the applicable design criteria such as would be published in a code or standard. (See the article "Finite Element Modeling in Failure Analysis" in this Volume for a more in-depth discussion regarding the use of FEA.)

Fracture Mechanics and Failure Analysis

Historically, the discipline of fracture mechanics was developed to understand the relationships among cracklike imperfections, stresses, and crack tolerance for the purpose of fabricating durable structures. As development of this body of knowledge continues, the usefulness of fracture mechanics in failure analysis has been recognized and is appropriately applied as one of the tools for failure analysis (Ref 2, 37, 38).

An in-depth discussion of fracture mechanics as it relates to failure analysis is beyond the scope of this article; more thorough treatment of this subject can be found in *Fatigue and Fracture,* Volume 19 of the *ASM Handbook,* and in the references cited previously. It is instructive to note that the technique is useful in some failure analyses. By performing careful measurements of relevant fracture features, incorporating known material properties (such as tensile strength and fracture toughness), and analyzing the loads and mechanics of the application, relationships can be developed to obtain an estimate of the loads and/or stresses that were operating at the time of fracture or to determine that the material in fact did not have the assumed properties. These can be compared with the loads or stresses either measured or calculated (Ref 37). Note that this is only a very brief summary and an oversimplification of the process. Extreme care must be exercised in performing such a fracture mechanics analysis, since there are uncertainties in failure analysis and in the stress-intensity-factor solutions of the failed component. The results of the stress analysis and fracture mechanics analysis must be consistent with the macroscale and microscale fractographic information and the microstructural information.

When a failure occurs by a progressive form of fracture, such as fatigue or stress-corrosion cracking, fractography can be performed to establish the fatigue striation spacing, or the crack arrest profile, across the fracture surface (as is practicable). These data can be put into appropriate equations to estimate the stress-intensity factors for either fatigue or stress-corrosion cracking, or, under some circumstances, both (Ref 2, 37, 38). Measured fatigue striation densities can be used in fracture mechanics calculations to determine either the stress range or the stress-intensity factor range, when the actual cycle counts for a given length of crack extension are known (Ref 2, 39). The usefulness of fracture mechanics as a tool for failure analysis continues to develop. One goal is to be able to reconstruct the size and growth rate of the crack over time and consider questions such as:

- Was a detectable crack somehow missed during inspection?
- Was the inspection interval appropriate?
- Did a rebuilding, overhaul, or other maintenance operation somehow contribute to the cracking?
- Did a change in service conditions or operating parameters contribute to the cracking?

References cited in this section

2. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 229–230

37. A.R. Rosenfield, Fracture Mechanics in Failure Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 450–456

38. R.M. Pelloux and A.S. Warren, Fatigue Striations and Failure Analysis, *Failure Analysis: Techniques and Applications*, J.I. Dickson, et al., Ed., ASM International, 1992, p 45–49

39. J.C. McMillan and R.M.N. Pelloux, Fatigue Crack Propagation under Program and Random Loads, *Fatigue Crack Propagation*, STP 415, ASTM, 1967, p 505–535

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Categories of Failure

There are many ways to categorize failures and material damage in terms of forms, mechanisms, or cause. No one system is necessarily complete and consistent with the multitude of possibilities. However, categories can help prioritize or identify avenues of investigation, as long as the categories do not limit critical inquiry.

Categories of Material Stressors

To determine the cause of material failure, one must consider the active stressors. A stressor is an external influence that can be a direct or indirect cause of failure. Understanding these influences is important for effective failure analysis and determining root cause. Likewise, mitigation of the stressors is often the most logical solution to reducing susceptibility to failure. The influence of stressors is heavily dependent on the susceptibility of the component, performance criteria, the magnitude of the stressor, exposure, and the material susceptibility.
The six stressors are:

- *Mechanical:* Applied static, dynamic or cyclic loads, pressure, impact, fabrication-induced residual stresses, applied end movements
- *Chemical:* Inadvertent acute or chronic exposure to an aggressive chemical environment, material compatibility issues
- *Electrochemical:* A susceptible metal in a corrosive aqueous environment
- *Thermal:* Exposure to elevated temperatures resulting in materials degradation
- *Radiation:* Ultraviolet lighting, sunlight, ionizing radiation from nuclear power plants, and so forth
- *Electrical:* Applied electrical stress due to the presence of an electric field

Four Categories of Failures

The physical failure of materials can be placed in one of many categories depending on the classification system. The following four categories are a convenient way to descriptively categorize and discuss failures, with the ultimate goal of understanding causes and preventing failures (Ref 3):

- Distortion or undesired deformation
- Fracture
- Corrosion
- Wear

These four categories represent the general forms of failure, and each form of failure may have a variety of different underlying mechanisms (e.g., fatigue crack propagation in the case of fracture or galvanic effects in metal corrosion). It is important to point out that two or more mechanisms can occur simultaneously in some failures. These failure categories integrate with the four fundamental root causes of failures discussed in the section “Primary Physical Root Causes of Failure” in this article. As presented in Table 4, each observed failure category can be associated with any one of the four root causes.

Table 4 Examples of root causes that result in the four failure types

Failure type	Design deficiency	Material defect	Manufacturing defect	Service life anomaly
Distortion	Insufficient section thickness of a tee section results in buckling under normal load.	Cavity shrinkage in a highly stressed area of a complex structural casting used in a gas turbine engine results in permanent deformation in service, and consequential loss of clearances between the rotor and the stator housed by the casting.	Abusive thread rolling causes heavy slip banding in titanium fastener, resulting in localized stretching of the fastener upon torquing as required and an associated inability to adequately clamp joint.	Exposure of an aluminum aircraft structure to excessively high temperatures results in permanent deformation by creep and subsequent buckling.
Fracture	Cold-formed and galvanized carbon steel sheet sustains brittle fracture under normal service loads, due to strain-age embrittlement.	Lap in forging, loaded cyclically in service, grows into a fatigue crack and subsequently fails catastrophically.	Welding of alloy steel with moisture-contaminated filler metal wire results in hydrogen embrittlement and consequential brittle cracking in service.	Inappropriate hole drilling of aluminum structural bicycle component by owner results in fatigue cracks initiating and propagating in service, ending in final failure of the component.
Corrosion	Gray cast iron underground pipes used to transport hazardous materials sustain rupture due to dealloying, or "graphitization."	Iron impurities in wrought aluminum alloy suspension component for railroad car create pitting susceptibility, resulting in loss of structural integrity.	650 °C (1200 °F) stress-relief treatment of a 304L stainless steel formed and welded screen for pulp processing sustained intergranular corrosion, cracking, and failure.	Increased usage of road salt in wintertime in northeastern U.S. results in vehicle electrical problems, traced to corroded electrical contacts
Wear	Incompatible wear couple is specified in the design of an injection mold/ejector pin assembly, resulting in galling and seizure.	Improper melting and hot-working processes lead to poor distribution of primary carbides in tool steel, resulting in rapid tool wear.	Poorly machined surface of a sliding machine element leads to accelerated wear and subsequent mechanical malfunction.	Insufficient lubrication during maintenance results in premature wearout of bearing on pump shaft.

For any of these failure types, materials performance plays a critical role. Just as the performance of a component or system is dependent on the behavior of the materials of construction under the service conditions, the manner in which a component or system sustains a physical failure is strongly affected by materials performance. For example, corrosion failures of dissimilar metals in physical contact in an aggressive environment are associated with the differences in the electrochemical behavior as a result of the chemical compositions of the two metals. This illustrates that one of the most basic tenets in materials science and engineering applies to failures: the interaction of the composition, processing, structure, and properties defines materials performance (Fig. 37), whether satisfactory or unsatisfactory (Ref 40).

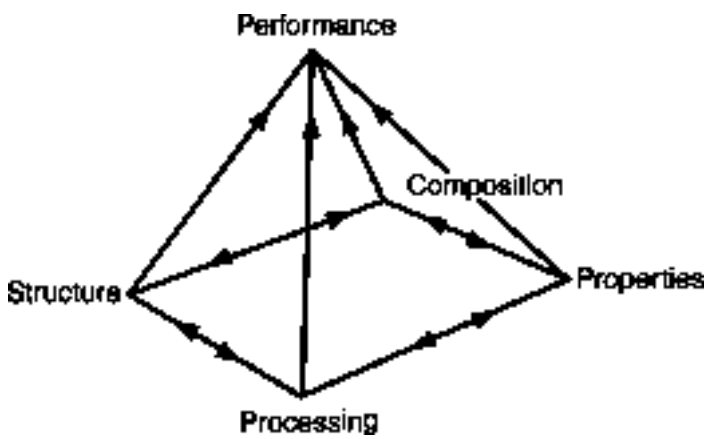


Fig. 37 Materials performance as a result of the interactions among composition, processing, structure, and properties. Source: Ref 40

Distortion. A distortion failure occurs when geometrical changes prevent a component from functioning properly such as a swollen polymer bearing in a pump or bent linkage in a transmission. Geometry change will generally be in the form of volume changes (e.g., swelling or shrinkage) or shape changes (e.g., warping, bending, or buckling).

Common causes of volume-distortion failures include temperature-induced phase changes or thermal expansion in metals, fluid absorption of nonmetallics, and curing shrinkage such as may occur in grouts and adhesives. Common causes of geometry-induced failure include inadequate design, flexural stiffness under load, stress-induced material yielding (Fig. 38), and uneven heating while in service.

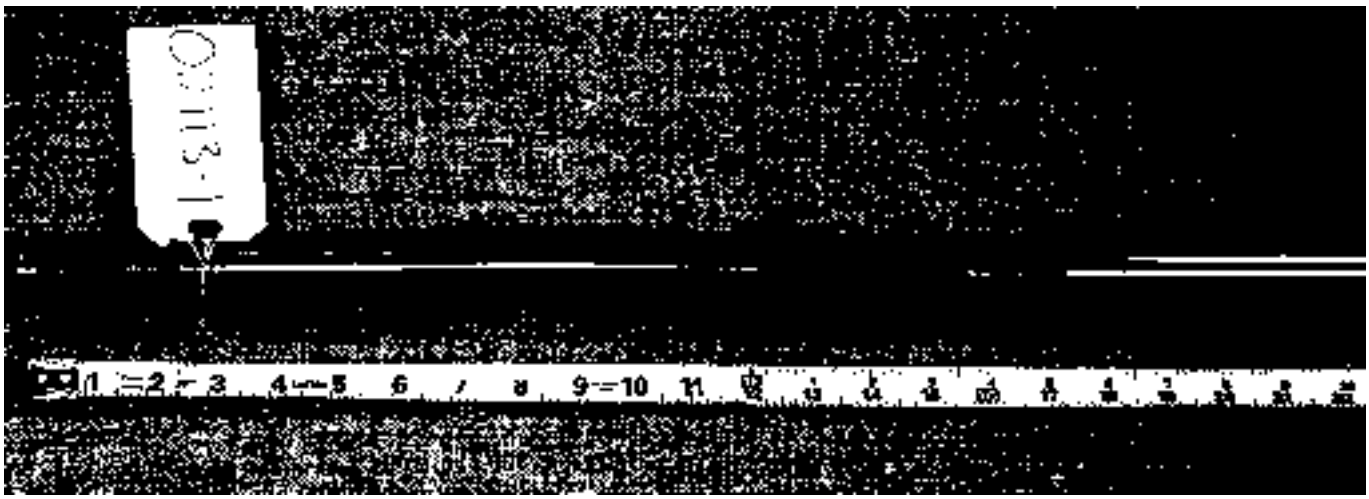


Fig. 38 Example of distortion in an overloaded valve stem

Fracture. A fracture is generally defined as material separation. There are many causes and forms of fracture including brittle fracture (Fig. 39), ductile fracture, and many progressive cracking mechanisms that can lead to final fracture. An understanding of the component design, service loading, environment, and the application of sound laboratory investigative techniques such as interpretation of the fracture surfaces (fractographic examination) are essential to an effective failure analysis in the case of component fracture.

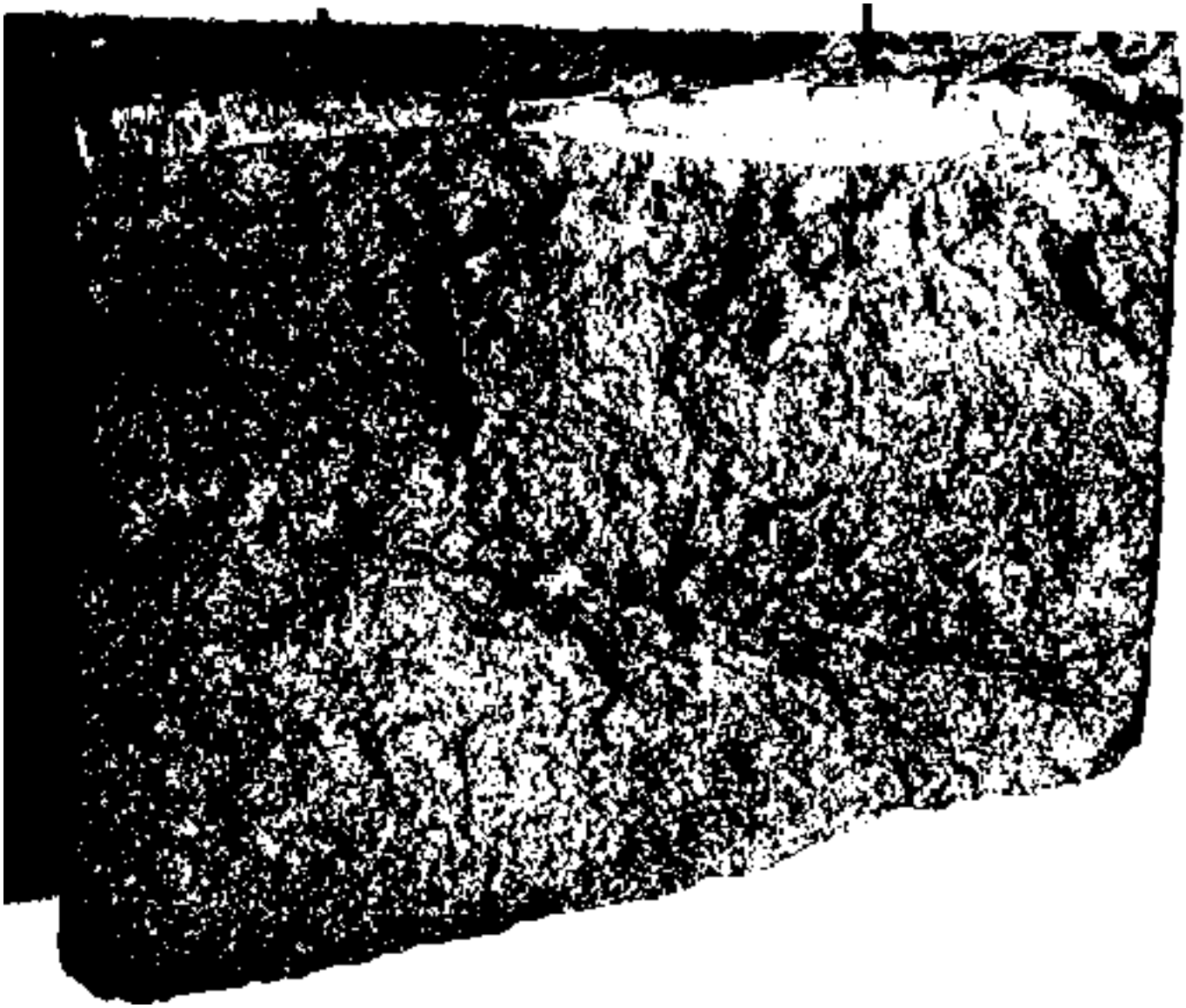


Fig. 39 Example of a brittle fracture of A36 structural steel, after sustaining fatigue cracking initially (at arrows). Source: Ref 41

Material Behavior under Load. Understanding the behavior of materials under load is important to the understanding of fracture modes. The macroscopic behavior of materials under loading is often characterized through tensile testing. It is customary to measure load and elongation during these tests and to plot the results in the form of a stress-strain diagram. Experimentally derived stress-strain diagrams can vary widely between different materials and are influenced greatly by parameters such as the speed of the test and temperature of the specimen during the test. Figure 40 depicts typical stress-strain diagrams. One curve is characteristic of mild steel, and the others are characteristic of other types of materials. However, generally speaking, each material has its own curve. Ductile materials are those that are capable of withstanding relatively large strains prior to fracture as opposed to brittle material to which the converse applies. Nonuniform and unstable transverse contraction referred to as necking in ductile materials indicates a severe overload. It reduces the effective stressed area and results in a distinction between the true stress-strain curve and the engineering stress-strain curve, which considers the original cross section when calculating the stress. A study in dislocation theory further explains the plastic behavior of metals beyond the elastic stress range (Ref 42).

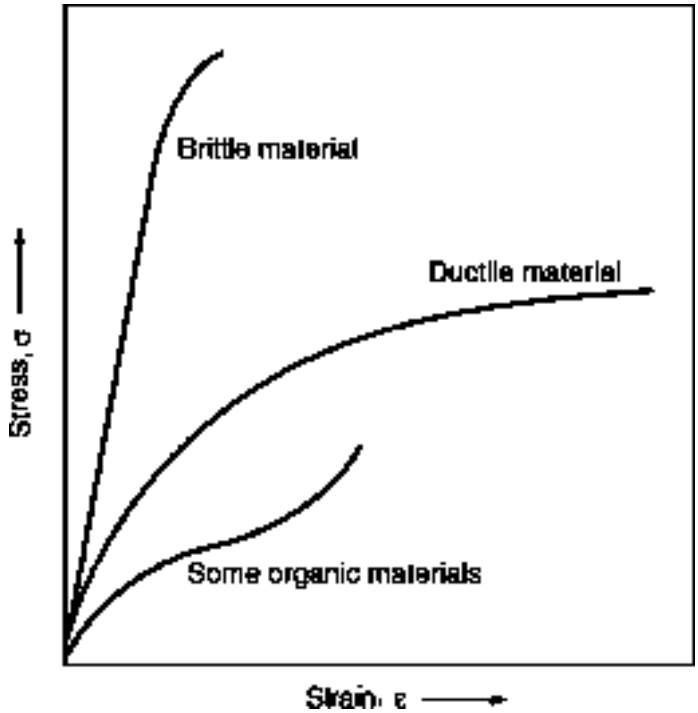


Fig. 40 Typical stress-strain diagrams. Source: Ref 42

Fundamental Fracture Mechanisms. Figure 41 illustrates the three most common fracture mechanisms in metals. Ductile fractures initiate with the nucleation, growth, and coalescence of microscopic voids that often begin at second phase particles or inclusions. Although cleavage fracture is most commonly thought of as a brittle fracture that propagates along crystallographic planes, it can also be preceded by a high degree of plasticity and ductile crack growth (Ref 43).

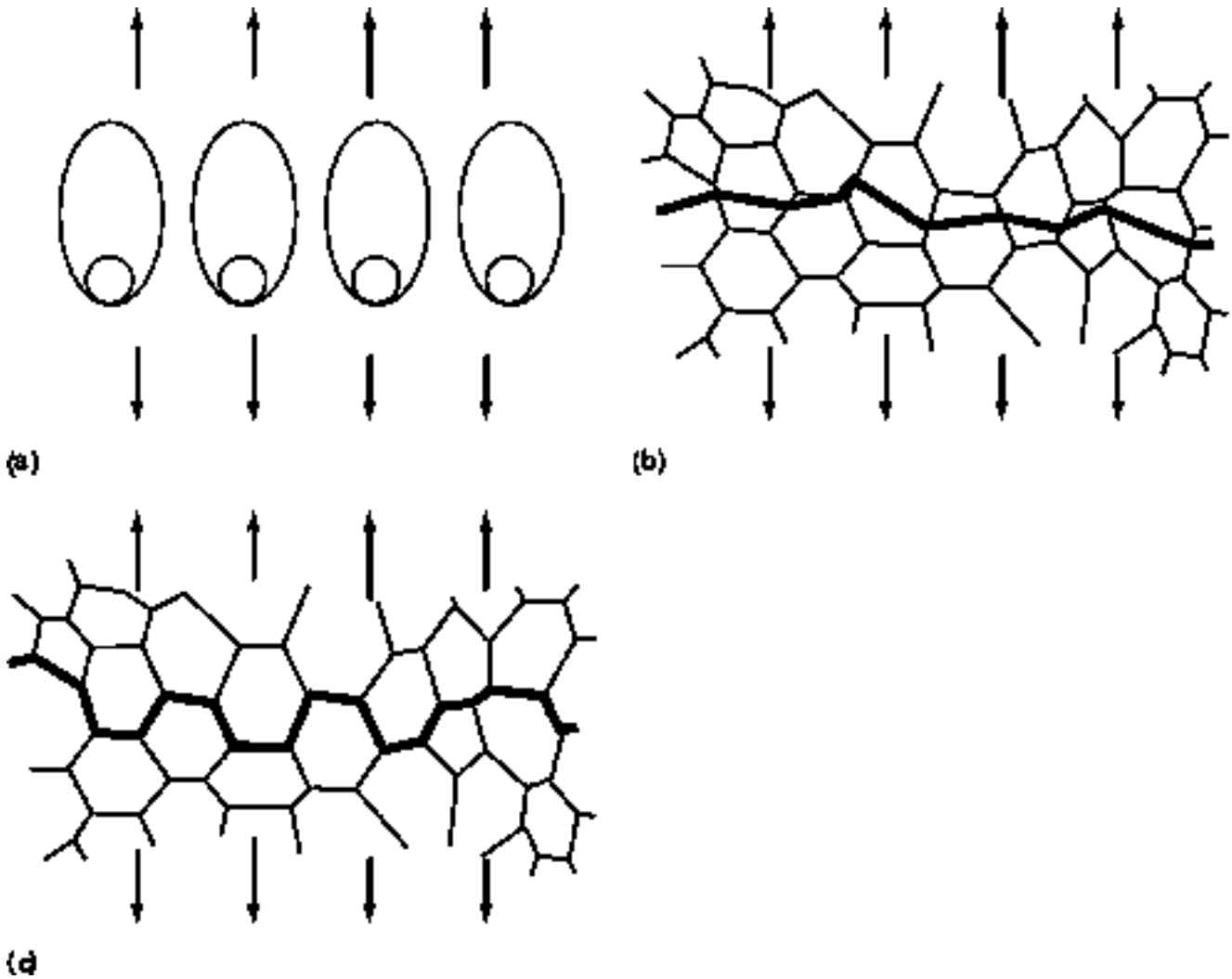


Fig. 41 Three micromechanisms of fracture in metals. (a) Ductile fracture. (b) Cleavage fracture. (c) Intergranular fracture. Source: Ref 43

Corrosion is the environmental degradation of materials. In metals, the most common type of corrosion is an electrochemical phenomenon that occurs on the surface of susceptible metal or metal alloys when exposed to a corrosive aqueous environment. Other forms of corrosion that do not involve electrochemical action include liquid metal embrittlement, corrosion in molten salts, high-temperature oxidation, and so forth. The result of corrosive attack can take the physical form of uniform surface wastage, local wastage, pitting, cracking, embrittlement, and so forth. The loss of material can eventually lead to an overload failure or through-wall penetration. The buildup of oxide scale that has a significantly increased volume when compared to the unoxidized metal can also be a problem by applying wedging load in crevices. Mitigation of corrosive attack involves a change of materials, removal of the corrosive environment, providing a surface barrier such as a coating, or providing cathodic protection. An example of piping system corrosion due to the effects of microbiological activity is shown in Fig. 42.

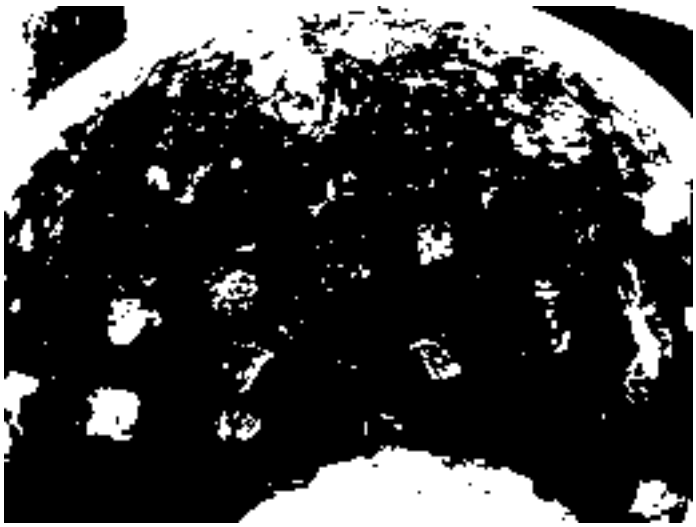


Fig. 42 Microbiologically influenced corrosion in a cooling water piping system

Wear failures result from the removal or displacement of surface material through contact and relative motion with a solid, liquid, or gas. There is a significant influence of friction and lubrication on the rate and severity of wear damage. Wear generally results in loss of material and load-carrying capability, adhesion, increased friction, and debris generation. Whether or not wear damage constitutes failure of a component depends on the performance criteria of the component, such as in a failed diesel engine main bearing that sustained excessive wear and a subsequent loss of control of the crankshaft radial movement (Fig. 43). Slight wear on metal valve seats may result in unacceptable leakage, while severe wear in a less critical application may be anticipated and without consequence and thus be perfectly acceptable. Controlled wear such as is the case with automotive brake pads may be part of the design criteria for a consumable component. The generation of debris could also be a critical consideration if, for example, the contamination of an ultrapure water system is at risk.

Fig. 43 Example of a wear failure in a diesel engine bearing

References cited in this section

3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

40. M. Cohen, MIT Lecture, 1977

41. G.F. Vander Voort, Ductile and Brittle Fractures, *Met. Eng. Quart.,* Vol 16 (No. 3), 1976, p 32–57

42. E. P. Popov, *Introduction to Mechanics of Solids,* Prentice-Hall, 1968, p 101–110

43. T.L. Anderson, *Fracture Mechanics: Fundamentals and Applications,* 2nd ed., CRC Press, 1995

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

Failure Prevention

Failure prevention begins with a state of mind in the specification, design, manufacture/fabrication, installation, operation, and maintenance of any component. However, before failure prevention measures are taken, the degree of reliability required in a specific situation must be determined.

There is a cost associated with failure prevention, and of course there is a cost associated with accepting failures. As shown in Fig. 44, many times it may be reasonable to accept failures should the cost of reliability enhancement outweigh the benefits. For example, the consequence of an aircraft structural failure is very high, thus demanding a high assurance of reliability. In contrast, the failure of a screwdriver may be low cost, although certainly a nuisance.

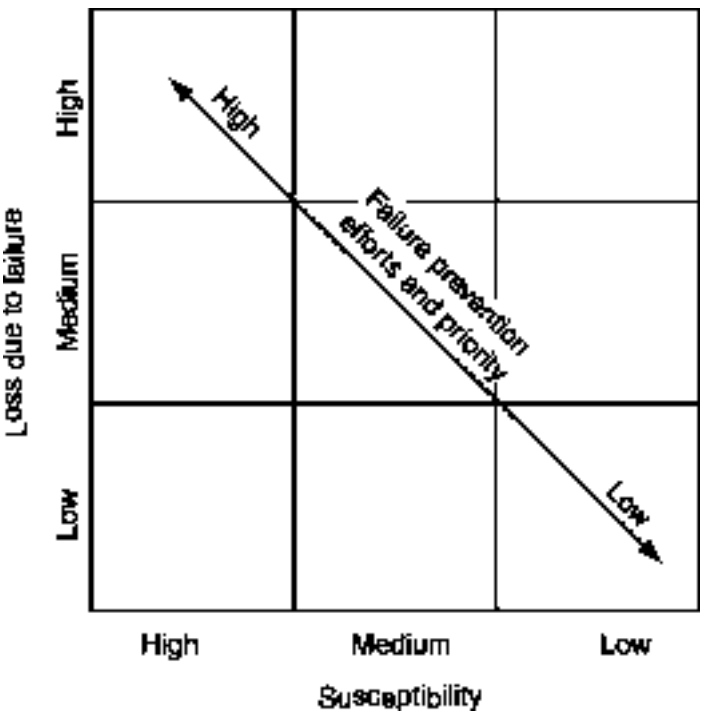


Fig. 44 Failure prevention effort prioritization

Building in reliability from the start is the most efficient way to achieve levels of reliability that will reduce or prevent failures. First, develop a performance specification that establishes the criteria for acceptable performance, answering critical questions. What are the important aspects of form, fit, and function? How long should it last? What are both the

expected and unexpected stressors? The element of life-cycle management becomes an important consideration as previously described.

Failure Modes and Effects Analysis. Just as an effective failure analysis requires a multidisciplinary approach, so does an effective failure-resistant design. Designers, material scientists, engineers, fabricators, and quality-control specialists contribute to failure modes and effects analysis (FMEA).

The FMEA procedure involves examining each item, considering how that item can fail, and then determining how that failure will affect the operation of the entire component or system. The process of identifying possible component failure modes and determining their effects on the system operation helps the analyst to develop a deeper understanding of the relationships among the different system components and to make any necessary changes to either eliminate or mitigate the possible undesirable effects of a failure. The steps involved in performing a FMEA, as identified by J. Bowles in the article "Failure Modes and Effects Analysis" in this Volume include:

1. Identify all item failure modes.
2. Determine the effect of the failure for each failure mode both locally and on the overall system being analyzed.
3. Classify the failure by its effects on the system operation and mission.
4. Determine the failure probability of occurrence.
5. Identify how the failure mode can be detected. (This is especially important for fault-tolerant configurations.)
6. Identify any compensating provisions or design changes to mitigate the failure effects.

Activities that constitute the FMEA complement and add value at every stage of the development cycle.

Applying Codes, Standards, and Regulations. The necessity to ensure interchangeability and compatibility of parts and safety factors in design led to the initial development of codes, standards, and regulations. In general, "codes" are considered to be a collection of laws or regulations that are a result of legislation to control activities. The term *standards* is often considered to be interchangeable with *specifications.* However, specifications are generally considered to refer to a more specialized and specific situation. Standards may be categorized as:

- Government regulations (i.e., requirement mandated by the government such as Occupational Safety and Health Administration, or OSHA, regulations)
- Government standards (federal specifications such as Military Specifications)
- Consensus standards (e.g., ASTM and ANSI standards)
- Technical society, trade association, and industry standards
- Company standards (both the supplier and the purchaser company may have their own standards)
- Standards of good practice
- Standards of consumer expectation

These standards include mandatory standards such as those published by government agencies. An example of these are those specified by the U.S. Code of Federal Regulation (CFR). Also, voluntary standards are often specified for mandatory compliance by a manufacturer or buyer of a product. There are many codes and standards such as those published by the American Society of Mechanical Engineers (ASME), the American Society for Testing and Materials (ASTM), and the American National Standards Institute (ANSI) (Ref 44).

The ASME Boiler and Pressure Vessel (BPV) Code is an example of a code that evolved from the necessity to prevent failures. Typical boiler operating pressures increased gradually from 206 kPa (30 psi) in the mid-19th century to more than 1378 kPa (200 psi) by 1900 and were accompanied by a much more widespread use of steam power. This led to a drastic increase in boiler explosions to a rate of approximately one per day in the United States. The evolution of the code resulted from the need to avoid increasing boiler failures as well as providing a basis for uniformity in the commercial bidding process. Today the ASME BPV Code is widely adopted and specifies acceptable materials and designs as well as fabrication, inspection, and repair methods (Ref 45).

Safety Factors and Reliability. An important element of design is the concept of a safety factor, which is typically a driving influence in the development of failure prevention concepts in codes and standards. A safety factor is generally defined as the ratio of failure load to anticipated load, if the safety factor is applied to stress. However, it can also be applied to fracture toughness, ductility in forming, casting quality, or other failure criteria that are established. Designing and manufacturing a product to adequately perform its intended function is not sufficient. A safety factor must consider an imperfect world, including manufacturing or construction tolerances, material variability, unanticipated stressors, and the effects of aging. An example of such a practice is a code specifying allowable material stresses that are much less than the strength of the material. The advent of modern materials and engineering design has reduced levels of uncertainty and has allowed a reduction of safety factors over the years. Selecting an appropriate safety factor for a given product includes consideration of:

- Degree of uncertainty about loading

- Degree of uncertainty about material strength
- Degree of uncertainty in relating applied loads to material strength
- Consequences of failure in terms of human safety and economics
- Cost of providing a large safety factor

Recommended safety factors for a performance factor (e.g., yield strength or some other failure criterion) may start from a low value of 1.25 to 1.5, where the materials are exceptionally reliable. That is, they are used under controlled conditions with frequent maintenance and inspection and are subjected to loads and stresses that are determined with certainty through testing of statistically significant material populations and/or analysis. Factors such as these are in most cases used where low weight is a particularly important consideration.

More common safety factors are in the range of 3 to 4 or higher when the loads or materials are less certain. Higher safety factors also apply in situations where repeated loads are applied, impact forces exist, materials are brittle, or there is other uncertainty. In the end, the appropriate safety factor is dictated by the applicable code or standard as well as situation specific considerations (Ref 46).

The concept of reliability is closely related to the concept of safety factors, which often incorporate a statistical approach. One must ask: if 1000 “identical” parts are put into service, what is the acceptable failure rate? The usefulness of a reliability approach depends on having adequate information on the statistical distribution of loading applied to parts in service as well as the statistical distribution of strength coming from production runs of manufactured parts. These variables are used in various statistical models with predicted failure rates compared with those considered to be acceptable (Ref 46).

Materials Selection. Design and material selection are fundamentally important in minimizing failures and hence ensuring component reliability. Selecting the most appropriate material for an application is highly product dependent and situation dependent. All functional requirements and environments must be considered in order to satisfy design requirements as well as economic considerations. Significant engineering expertise is required to ensure the material selections are appropriate for the intended function and service (including an understanding of the stressors) because trade-offs are usually required.

One of the common considerations in selecting materials is determining the desired mechanical properties. For instance, having a fracture-tolerant component is often an objective that can be achieved by selecting a material that is ductile and flaw tolerant, reducing the likelihood of brittle fracture. The trade-off is that ductility is often achieved by sacrificing overall strength, wear resistance, and resistance to deformation. In order to achieve ductility and maintain wear resistance, one may select a surface treating process such as a case-hardening process. In metals, the properties that must be considered to both ensure the desired function and reduce the likelihood of failures include:

- Tensile strength
- Yield strength
- Modulus of elasticity
- Ductility (percent elongation)
- Fatigue strength
- Fracture toughness
- Hardness
- Shear strength
- Machinability
- Coefficient of friction
- Impact strength
- Corrosivity
- Density
- Coefficient of thermal expansion
- Thermal conductivity
- Electrical resistivity
- Other physical properties

In the typical application of polymers, there are often other material properties considerations with regard to both performance and failure prevention such as:

- Stiffness
- Chemical, thermal, and ultraviolet resistance
- Electrical resistance
- Dimensional stability
- Resistance to moisture absorption

There are also unique properties associated with other material such as composites and ceramics that must be considered (Ref 47). A more complete treatment of the importance of materials selection in preventing failures is found in the article entitled “Materials Selection for Failure Prevention” in this Volume.

Operation, Maintenance, and Inspection. Achieving the expected service life of a component or system (and consequently preventing failures) requires diligence on the part of the user in operating and maintaining the component or system within established bounds. This process begins, however, with the definition of appropriate operating conditions by the manufacturer during product development and testing. The deliverable product that results from that development effort is an instruction manual that is provided with the product. Information in the manual should include installation instructions, recommended methods for activating, using, and shutting down the component or system, and maintenance recommendations or requirements. Responsibility also lies with the manufacturer to anticipate misuse of a product and adequately warn of the dangers and risks associated with such misuse (Ref 15). Warnings typically are communicated to the user through both the instruction manual, labels prominently affixed to the product, or through public announcements by the manufacturer or a government agency. Such warnings often distinguish risks of personal injury or death versus risks of damaging the product. Design of the warnings also distinguishes hazards that are obvious versus hazards that are not.

Proper maintenance of products is of paramount importance in realizing the expected service life. The user (or the actual owner) is ultimately responsible for proper upkeep of components or systems. However, initially, the responsibility also lies with the manufacturer in developing an appropriate maintenance plan for the anticipated service conditions. Such methodologies as reliability-centered maintenance can be employed to build maintenance plans that optimize requirements as appropriate for specific types of products in specific applications. Emphasis is placed on maintenance of the components or systems with the greatest impact in the event of a failure (see the article “Reliability-Centered Maintenance” in this Volume). Petrochemical and chemical-processing industries also use extensive methods for predictive maintenance for prevention of corrosion failures (see the article “Analysis and Prevention of Corrosion-Related Failures” in this Volume).

Maintenance plans could be as simple as periodic cleaning of, for example, a toaster. Conversely, the plan could be as complex as a comprehensive product-management system involving rigidly defined inspections, servicing, replacements-for-cause, and life-limited component changeouts required at various maintenance levels (that is, sites with specific capabilities). For example, in some aircraft fleet maintenance plans three levels of maintenance are defined, with easily accessible component replacements (including entire engines) allowed at the flight operations sites, partial teardown and rebuilding of more complex components (including parts of engines) at intermediate maintenance sites, and full teardown and rebuilding of all serviceable components (no matter how complex) at maintenance depots. The plan could also include the updating of maintenance manuals, training of maintenance personnel, spare parts procurement, implementation of maintenance directives, and so forth.

Maintenance and repair activities can be provided by:

- The manufacturer, for specialized, complex, and critical systems or for low-volume products
- An approved repair/overhaul company, an approach commonly used for aircraft, and other specialized, complex, and critical systems of moderate to high volume
- Independent repair providers, for less complex systems of high volume, as with automobile service stations
- The end user, as with simple systems such as yard machines

These options are listed in the order of decreasing input from the manufacturer and hence control of the repair processes used. Defining the appropriate service provider requirements can prevent failures and improve service life and reliability.

An important aspect of any maintenance plan for complex or critical products and systems is inspection. As shown previously in Example 3, sound inspection is effective in failure prevention. In general, periodic inspection programs are required for critical systems in which reducing the risk of safety or health issues is desired or required. Other conditions under which inspection programs are implemented include situations where equipment downtime has excessively high cost, such as in a paper mill and where output of a process creates a significant loss in the ability to meet demand, as in the availability of electric power during peak usage.

Inspection programs identify degradation or loss of function of equipment and unanticipated service conditions. In safety and health critical systems, federal regulations often require inspection programs. For example, periodic inspection (or condition assessment) of pressure vessels, tanks, and piping that store or transport hazardous substances is required by federal law under OSHA 1910.119. The inspections are typically performed by visual and nondestructive techniques, documenting internal and external corrosion, corrosion under insulation, poor welds or joint failures that might leak, inappropriate support, lining failures, and so forth, in accordance with the appropriate American Petroleum Institute (API) and ASME codes. Other U.S. government-required inspection programs include aircraft (Federal Aviation Administration, or FAA), transportation of hazardous substances (Department of Transportation), and the manufacture of pharmaceutical products (Food and Drug Administration, or FDA).

Incorporating Lessons Learned. A proper failure analysis and root-causal analysis can provide valuable information into the entire design and production process of a product or system. The implementation of these analyses alone, however, does not ensure that benefit is gained. These results must be communicated into corrective-action recommendations that are routed back to the proper stage of the product life cycle. Implementation of these actions and final verification of these actions should be performed in order to ensure that the desired outcome is obtained in terms of failure prevention. Organizing a multidisciplinary lessons-learned meeting to discuss the outcome of the failure analysis is important from a learning perspective and to ensure proper communications (Fig. 45), particularly in a large company with many departments involved in the evolution of a product (Ref 48).

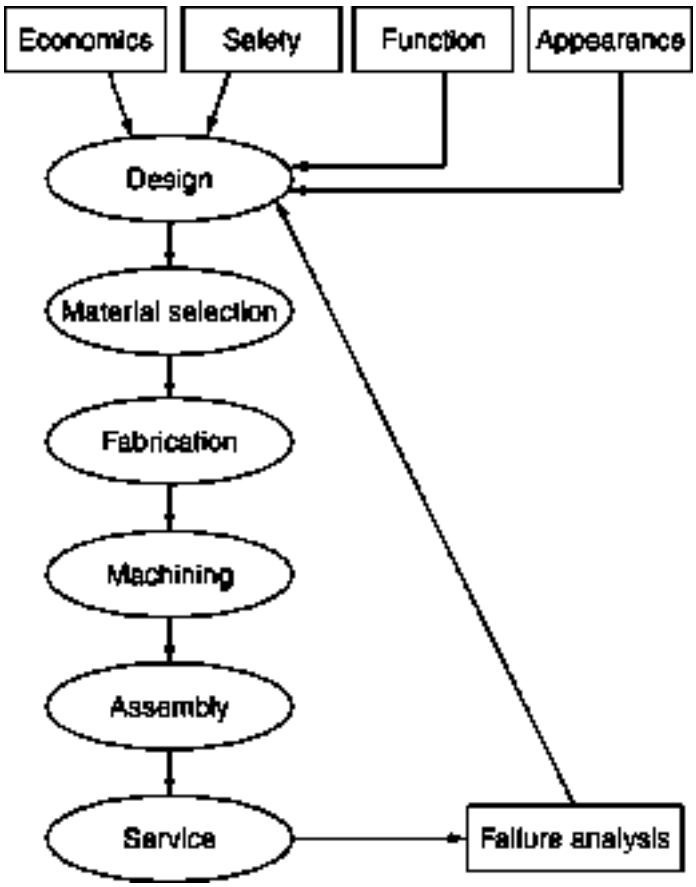


Fig. 45 Feedback of a failure analysis to product evolution. Source: Ref 48

Implementing Corrective Actions. Having completed the failure analysis process and identified the root causes, the next step in preventing future failures involves developing, verifying, and implementing a corrective-action plan. In general, corrective-action plans involve one or more of the three general types of plans: short-term, mid-term, and long-term.

Short-term corrective-action plans involve simple tasks selected to minimize the impact on the operation of the machine or system, such as:

- Repair the inoperative machine or system to get it back up and running
- Identify and manage the suspect population
- Modify the service conditions (make less severe, if possible)
- Issue warnings to other users/maintainers

While these actions may serve to enable continued operation of the machine or system, the amount of service time gained is usually limited. Therefore, mid-term corrective actions can be developed:

- Implement field repair
- Modify design and retrofit in field, at repair facilities, or at the factory

Preventing the failure from recurring in the long run involves implementing long-term “fixes”:

- Redesign
- Implementation of redesign by attrition or retrofit

Standardizing Corrective Actions. Preventing recurrence of failures may require the corrective actions developed through the root-cause failure analysis to be implemented on a much broader basis than for the failed product alone. If similar products are in service that could present risk of failure, those products should be included in the population affected by the corrective actions.

Some corrective actions are standardized internally by the manufacturer, incorporating the actions into division- or corporate-wide standards, such as design guides. Industry associations often step in to standardize corrective actions and lessons learned, in such organizations as ASME in the Boiler and Pressure Vessel Code and the Bridge Welding Code, the API for piping, tanks, and storage vessel inspection programs, the American Welding Society (AWS) for welding standards, the Society of Automotive Engineers (SAE) for fastener standards and material specifications, or the ANSI for safety standards of a wide array of products. The standardization can also be government driven, such as by the FAA, FDA, Consumer Product Safety Commission (CPSC), and so forth in the United States. Clearly, examining failures from a broader perspective enables a much wider impact on preventing failures, through standardization across products, markets, and industries.

References cited in this section

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

44. J.E. Shigley and C.R. Mischke, *Standard Handbook of Machine Design,* McGraw-Hill Inc., 1986

45. M.D. Berstein, Design Criteria for Boilers and Pressure Vessels, *J. Pressure Vessel Technol.,* Vol 110, Nov 1988, p 430–432

46. R.C. Juvinall and K.M. Marshek, *Fundamentals of Machine Component Design,* 2nd ed., John Wiley & Sons, 1991

47. R.L. Mott, *Machine Elements in Mechanical Design,* 3rd ed., Prentice Hall Inc., 1999

48. P.A. Thornton and V.J. Colangelo, *Fundamentals of Engineering Materials,* Prentice Hall Inc., 1995

Acknowledgments

The authors gratefully acknowledge the efforts of Dr. Fahmida Hossain, Veda-Anne Ulcickas, Leonard Norman, and Thomas Painter of Massachusetts Materials Research, Inc. for their enlightening analyses and contributions to the case histories presented. We also express our appreciation to Mr. Randy Fach of Dove Consulting for his insights and resources on the topics of Six Sigma, TQM, CI, problem solving, and quality improvement and to Dr. David French for his review and commentary. We especially acknowledge the inspiration and contributions of Dr. Regis Pelloux, Professor Emeritus, M.I.T. Finally, we extend our appreciation to Dr. Walt Griffith of the Air Force Materials Laboratory for his historical perspective on the Wright Brothers, during his informative presentation on “Materials Usage on the Wright Flyer” to a local chapter of ASM International.

Introduction to Failure Analysis and Prevention

James J. Scutti, Massachusetts Materials Research, Inc.; William J. McBrine, ALTRAN Corporation

References

1. P.L. Jakab, *Visions of a Flying Machine: The Wright Brothers and the Process of Invention,* Smithsonian Institution, 1990, p 226

2. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 229–230

3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

4. W.E. Deming, *Out of the Crisis,* MIT Center for Advanced Engineering Study, 1986

5. J.M. Juran and F.M. Gryna, Ed., *Juran's Quality Control Handbook,* 4th ed., McGraw-Hill, 1988

6. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993, p 7

7. F.W. Breyfogle III, *Implementing Six Sigma: Smarter Solutions Using Statistical Methods,* John Wiley & Sons, 1999, p xxvii

8. P.S. Pande, R.P. Neuman, and R.R. Cavanaugh, *The Six Sigma Way,* McGraw-Hill, 2000, p xi

9. M. Harry and R. Schroeder, *Six Sigma: The Breakthrough Management Strategy Revolutionizing the World's Top Corporations,* Doubleday & Co., Inc., 1999

10. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 7

11. B. Anderson and T. Fagerhaug, *Root Cause Analysis: Simplified Tools and Techniques,* ASQ Quality Press, 2000, p 7, 125

12. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors,* Max Ammerman/Quality Resources, 1998

13. Engineering Aspects of Failure and Failure Analysis, *Failure Analysis and Prevention,* Vol 10, 8th ed., *Metals Handbook,* American Society for Metals, 1975, p 1–9

14. R.K. McLeod, T. Heaslip, and M. Vermij, Defect or Flaw—Legal Implications, *Failure Analysis: Techniques and Applications,* Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 253–261

15. J.J. Asperger, Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 25–29

16. D. Broek, Fracture Mechanics as an Important Tool in Failure Analysis, *Failure Analysis: Techniques and Applications,* Conf. Proc. International Conference and Exhibits on Failure Analysis, 8–11 July 1991 (Montreal, Quebec, Canada), ASM International, 1992, p 33–44

17. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993, p 50

18. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 127

19. M. Paradise, L. Unger, and D. Busch, *TapRoot® Root Cause Tree™ User's Manual,* Systems Improvement, Inc., 1996, p 9–14

20. R.J. Latino and K.C. Latino, *Root Cause Analysis: Improving Performance for Bottom Line Results,* Reliability Center, Inc., 1999, p 79–89

21. C. Nelms, *What You Can Learn From Things That Go Wrong,* 1st ed., Failsafe Network, Richmond, VA, 1994

22. H.P. Bloch and F.K. Geitner, *Practical Machinery Management for Process Plants,* Vol 2, *Machinery Failure Analysis and Troubleshooting,* Gulf Publishing Co., 1983, p 5–6

23. P.F. Wilson, L.D. Dell, and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

24. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998

25. R.K. Mobley, *Root Cause Failure Analysis,* Butterworth-Heinemann, 1999, p 37–39

26. M. Ammerman, *The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors,* Max Ammerman/Quality Resources, 1998, p 67

27. *Failure Analysis, The British Engine Technical Reports,* American Society for Metals, 1981

28. F.A. Hossain and J.J. Scutti, Four Fundamental Root Causes of Failure: Case Histories, *Failure Analysis: A Foundation for Diagnostics and Prognostics Development,* Proc. 53rd Meeting of Society for Machinery Failure Prevention Technology, 19–22 April 1999 (Virginia Beach, VA), 1999, p 185–195

29. J.J. Scutti, "Where Things Go Wrong: Where to Look for Failure Prevention Opportunities," Presented to ASM International Materials Solutions Conference, 9–11 Oct 2000

30. J.J. Scutti and F.A. Hossain, Unanticipated Service Conditions, *Proc. Failure Prevention Through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 141–148

31. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Stainless Steels and Special-Purpose Metals,* Vol 3, 9th ed., *Metals Handbook,* American Society for Metals, 1980, p 825–834

32. D. Krashes and J.J. Scutti, Poor Surfaces and Intersections of Surfaces Still Cause Trouble Just Like They Used to Do, *Technology Showcase: Integrated Monitoring, Diagnostics and Failure Prevention,* Proc. Joint Conference, 22–26 April 1996, Society for Machinery Failure Prevention Technology (MFPT), 1996, p 681–690

33. F. Hossain and J.J. Scutti, Failure of Components Although the Causes are Simple & Well Documented, *1998 Technology Showcase,* Proc. Joint International Conference, JOAP International Condition Monitoring Conference, 20–24 April 1998 (Mobile AL), JOAP-TSC, 1998, p 455–464

34. G.F. Smith, *Quality Problem Solving,* ASQ Quality Press, 1998, p 53

35. P.F. Wilson, L.D. Dell and G.F. Anderson, *Root Cause Analysis: A Tool for Total Quality Management,* ASQ Quality Press, 1993

36. R.P. Baggerly, Preventing Failures Resulting from Machining Issues, *Proc. Failure Prevention through Education: Getting to the Root Cause,* 23–25 May 2000 (Cleveland, OH), ASM International, 2000, p 118–123

37. A.R. Rosenfield, Fracture Mechanics in Failure Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 450–456

38. R.M. Pelloux and A.S. Warren, Fatigue Striations and Failure Analysis, *Failure Analysis: Techniques and Applications,* J.I. Dickson, et al., Ed., ASM International, 1992, p 45–49

39. J.C. McMillan and R.M.N. Pelloux, Fatigue Crack Propagation under Program and Random Loads, *Fatigue Crack Propagation,* STP 415, ASTM, 1967, p 505–535

40. M. Cohen, MIT Lecture, 1977

41. G.F. Vander Voort, Ductile and Brittle Fractures, *Met. Eng. Quart.,* Vol 16 (No. 3), 1976, p 32–57

42. E. P. Popov, *Introduction to Mechanics of Solids,* Prentice-Hall, 1968, p 101–110

43. T.L. Anderson, *Fracture Mechanics: Fundamentals and Applications,* 2nd ed., CRC Press, 1995

44. J.E. Shigley and C.R. Mischke, *Standard Handbook of Machine Design,* McGraw-Hill Inc., 1986

45. M.D. Berstein, Design Criteria for Boilers and Pressure Vessels, *J. Pressure Vessel Technol.,* Vol 110, Nov 1988, p 430–432

46. R.C. Juvinall and K.M. Marshek, *Fundamentals of Machine Component Design,* 2nd ed., John Wiley & Sons, 1991

47. R.L. Mott, *Machine Elements in Mechanical Design,* 3rd ed., Prentice Hall Inc., 1999

48. P.A. Thornton and V.J. Colangelo, *Fundamentals of Engineering Materials,* Prentice Hall Inc., 1995

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Introduction

MATERIALS SELECTION is an important engineering function in both the design and failure analysis of components. In design, materials selection can be a complex, iterative process that solves a particular set of engineering objectives for a given component. Materials selection is just one part of this overall design process, which may involve a complex set of relationships regarding product function, shape, materials, and manufacturing process (Fig. 1, Ref 1). In the past, engineering design was performed as a sequential procedure, with the material decisions made last, sometimes literally as an afterthought. After the dimensions and property requirements were identified, the cheapest material meeting those requirements was sought. This philosophy may have been more justifiable when fewer material choices were available or when less sophistication in design and processing was required. However, with the heightened awareness of efficient design, optimized performance, economic competition, environmental impacts, safety concerns, and legal liability, current methods of materials selection are viewed more and more as a simultaneous and integral procedure of the entire design process, even during the early stages of design.

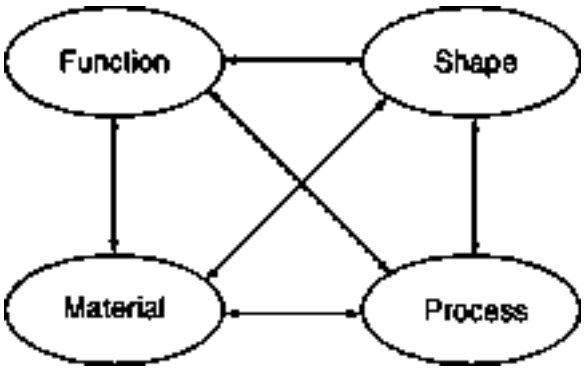


Fig. 1 Interrelated factors involved in the design process. Source: Ref 1

Materials selection and design are also closely related to the objectives of failure analysis and prevention. These processes are inextricably intertwined, probably to an extent that is not readily apparent to most engineers. Failure analysis augments the development process by real-time identification of design inadequacies, providing opportunities for optimization. Finding the root cause of a failure also often takes as much imagination as the original design concept. In fact, failure analysis can be viewed as a figurative reassembly of the component in the original condition, or design in reverse. For example, Fig. 2 is a diagram comparing the general procedures for both engineering design and failure analysis. The basic philosophies of the two processes are reversed. Design is the process of synthesizing and analyzing conditions into the reality of an actual or hypothetical component. In contrast, failure analysis is the dissection of an actual component in order to synthesize and understand the significance of a hypothetical design in a given failure. It is important to note that the analysis and synthesis of engineering factors are prominent in different areas of each process, although the individual steps within the processes contain both.

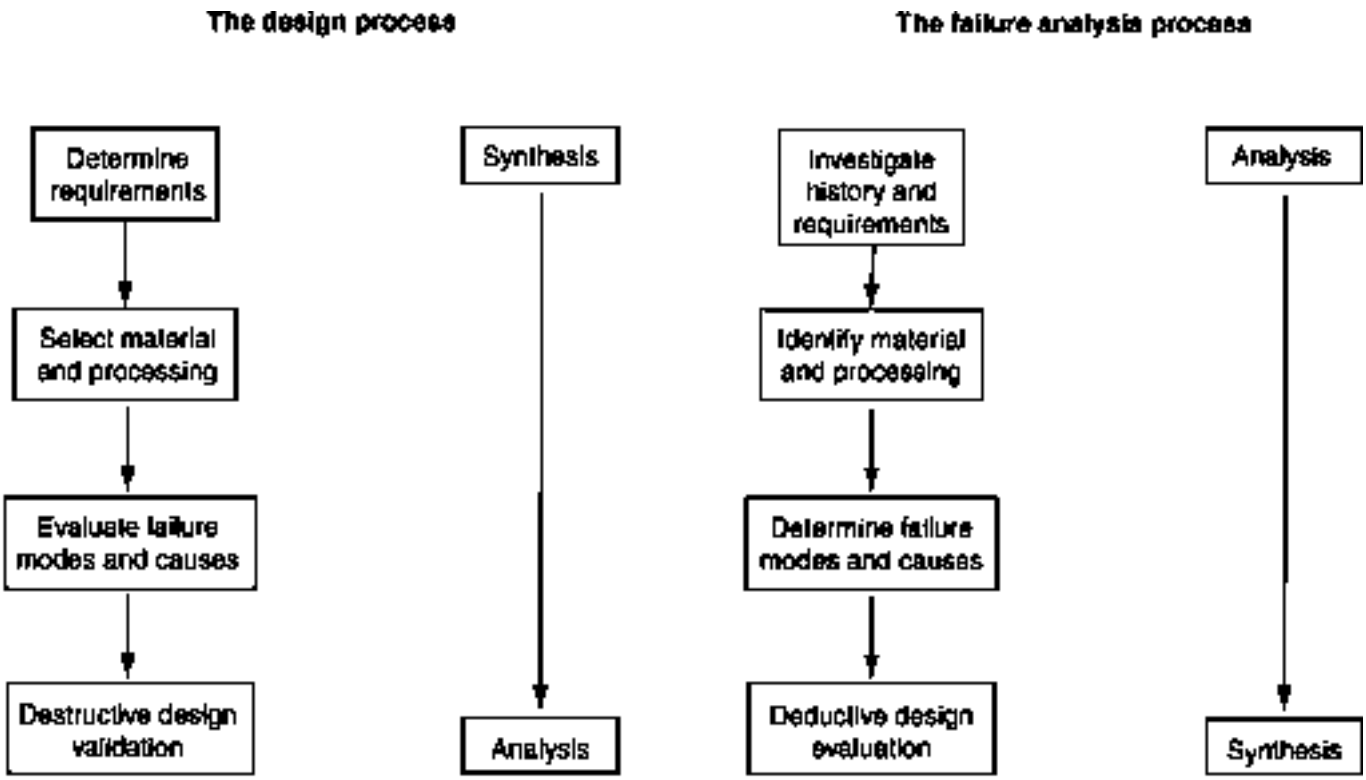


Fig. 2 General steps and the roles of synthesis and analysis in the processes of design and failure analysis

Materials selection augments and is supported by failure analysis in several ways. Of course, a basic objective of any successful design is to prevent failure (which can be defined here in the general context of a part not being able to perform its intended function). Therefore, failure analysts need to understand the underlying principles and practices of design and materials selection as basic tools in failure prevention. However, failure analysts also recognize that the synergistic effects of service conditions, manufacturing effects, and material characteristics are not always captured within the axioms and discrete attributes of a design process. Therefore, the analysis of failed parts can provide important insights for metallurgists and other engineers involved with design in general and materials selection in particular.

Design also sometimes can be an emulative process, whereby a successful design is adapted for a similar but separate service condition. Emulative design assumes that the new intended service is analogous, that the materials and processing are the same, and that the knowledge of the prior design is complete. Using prior design as a pattern involves implicit and possibly unappreciated assumptions, which may not account for the synergistic effects of service conditions, manufacturing effects, and material characteristics. It is easy to imagine, for example, that one structural member in an assembly may appear to be sufficiently strong when, in actuality, its portion of the load may have been displaced and accommodated by the overdesigned, surrounding structure.

This article briefly reviews the general aspects of materials selection as a concern in both proactive failure prevention during design and as a possible root cause of failed parts. This article cannot detail the many particulars of materials selection, because every industry or component application has many specific requirements, guidelines, or procedures, some of which may be mandated by federal or state statute. Therefore, coverage is more conceptual with general discussions on the following topics:

- Design and failure prevention
- Materials selection in design
- Materials selection for failure prevention
- Materials selection and failure analysis

Because materials selection is just one part of the design process, the overall concept of design is discussed first in the section “Design and Failure Prevention.” The next section, “Materials Selection in Design,” then describes the role of the materials engineer in the design and materials selection process. The other sections of this article focus on the significance of materials selection in both the prevention and analysis of failures. Portions of this article contain adapted content from *Materials Selection and Design,* Volume 20, *ASM Handbook,* with citation to more detailed references on materials selection.

Reference cited in this section

1. H.W. Stoll, *Product Design Methods and Practices*, Marcel Dekker, 1999, p 40, 148

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Design and Failure Prevention

The basis of all engineering is design, and the terms are often used synonymously. The primary objective of design is to develop a useful component or structure that performs an intended function in as safe a manner as possible. Therefore, the prevention of failure (generally defined here as any loss of intended function) is a principal concern of any design process. Simply restated, the primary measure of a successful design process is foreseeing and avoiding failure.

Design generally requires specific engineering expertise and is performed by a wide variety of engineering disciplines, such as:

- Civil engineers design large structural forms, such as bridges, highways, buildings, and power-generation and water supply facilities. Codes and standards regulate many of the materials and design features of these structures, due to safety concerns. Civil engineering designs typically use reliable and economical materials that are not particularly exotic.
- Mechanical engineers design a wide variety of components, such as pressure vessels, vehicles, and machinery of all types. These designs often contain moving parts and use materials with highly specific performance requirements. Materials selection can include all materials and processes. Many codes and standards are also applicable to mechanical engineering designs. Mechanical engineering encompasses such a broad range of equipment that designers generally have expertise in specific functions or types of component.
- Chemical engineers typically use materials in various chemical- and petrochemical-processing industries with design requirements involving corrosion resistance and elevated-temperature service.

Many other engineering disciplines also have unique and special requirements for design and materials selection. In electrical engineering design, for example, the physical property requirements (e.g., magnetic, electrical, electronic, or thermal properties) typically supersede the mechanical property requirements. Other design disciplines include industrial, automotive, welding, mining, aerospace, nuclear, and computer engineering.

Each type of engineering discipline requires specialized design expertise that is beyond the scope of this article. However, the general process of engineering design can be described as an iterative procedure that can be roughly divided into two basic stages (Fig. 3, Ref 1):

- *A conceptual design stage* involving the definition of product specification (or functions) and the underlying physical concept and preliminary layout to achieve the intended functions
- *A detailed design stage* involving both the qualitative definition of part configuration and the quantitative analysis of design parameters (e.g., dimensions, tolerances, materials properties, etc.) to achieve a final layout for a product, assembly, or system

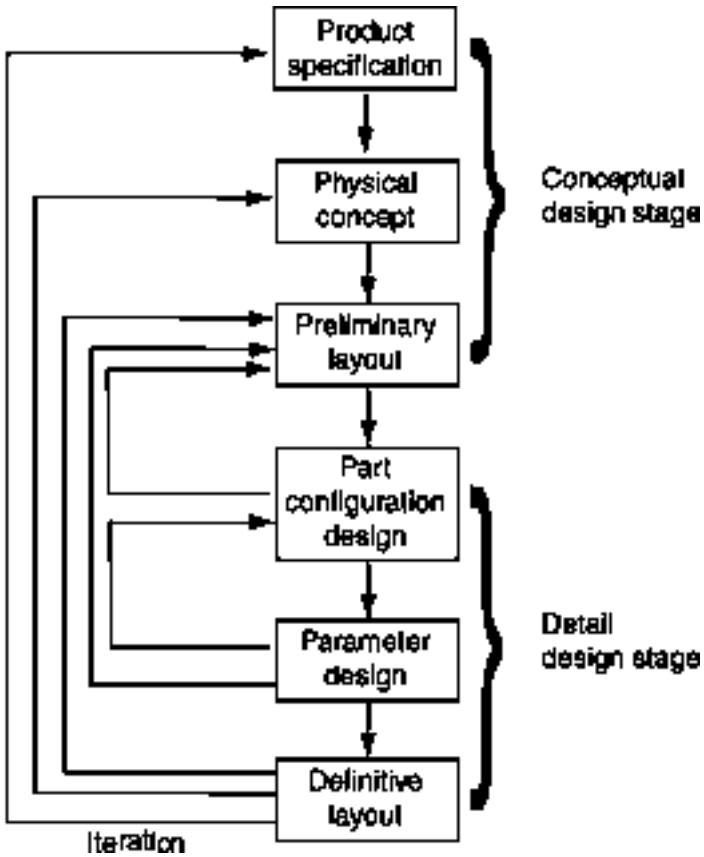


Fig. 3 Stages and steps in the iterative process of design. Source: Ref 1

As shown in Fig. 3, iteration is a key feature for any step in the design process. In fact, engineering design can be thought of as a process of guided iteration (Ref 2), which is a problem-solving methodology that formulates a problem, generates alternative solutions, evaluates the alternatives, and redesigns from the evaluation results if they are unacceptable. This methodology is fundamental to design processes. It is repeated hundreds or thousands of times during a complicated product design. It is used again and again in recursive fashion for the conceptual stage to select materials and processes, to configure parts, and to assign numerical values to dimensions and tolerances (i.e., parametric or parameter design).

Designers and materials engineers are key participants in this iterative process, and failure analysts also need to appreciate and understand the roles and activities of the design process. This provides the basis to identify possible technical root causes of failures and to advise or recommend further design iterations either on a proactive basis or as a lesson learned. Therefore, the basic steps or stages of engineering design are briefly described in the following sections, beginning with the stage of conceptual design and ending with a brief overview on methods of risk assessment in design. The main focus is on the design of part, as opposed to the design of a system or assembly. Systems or assemblies may have many components that involve parameters beyond the typical dimensions, tolerances, and properties of a distinct part. For example, when a system involves human interaction, the design process must address the influence of human factors. This includes not only human factors in equipment design but also a wide range of activities that can include operational methods and procedures, testing and evaluating these methods and procedures, job design, development of job aids and training materials, and selection and training of people (Ref 3).

Conceptual Design

The first stage of design is to determine the physical concept by which the product will function. This includes the physical principles by which the product will work and an abstract physical model that will employ the principles to accomplish the desired functionality. For example, suppose the required function is simply to support a load over an open space. In this case, the physical model could be a beam of uniform cross section or truss. In addition, there is not usually a unique solution for implementing a physical concept, although a concept and its function are inextricably linked.

When a product is more complex, it consists of an assembly of subassemblies and parts. In this case, the physical concept of the system or assembly must be “decomposed” into a set of principal functional subassemblies. For example, an automobile is a set of subassemblies identified as the engine, drivetrain, frame, body, suspension system, and steering system. The physical principles thus require sufficient information about how each of these functional subassemblies will interact with all of the others to accomplish the required product functions. The term “decomposition” is generally used to

describe the part of the design process that identifies the subassemblies comprising a product or larger assembly. That is, in the conceptual design of an automobile, it could be decomposed into the engine, drivetrain, frame, and so forth.

Decomposition in Conceptual Design. Two basic methods of decomposition are used in conceptual design: physical decomposition and functional decomposition. Many design concepts are based on the method of physical decomposition, when existing or emulative designs are used with an implicit functionality. For example, the decomposition of an automobile into engine, drivetrain, frame, body, and so on is an example of physical decomposition. This method is common, but the method of functional decomposition also has benefits. In functional decomposition, the design concept is defined purely in terms of functions, with physical embodiments (or configurations) selected to fulfill the functions. In an automobile, for example, the function of the engine is to convert a source of on-board energy to rotational mechanical power. This function could be achieved with the usual internal combustion engine, but it could also be provided by an electric motor, a turbine powered by compressed gas, human-powered pedals, and many other alternatives. In the case of an automobile, the available alternative sources of power are very familiar. In a new, less-familiar product, however, the advantage of function-first decomposition is that it stimulates designers to consider many ways of fulfilling a given function instead of choosing the most common embodiment that comes to mind.

The whole purpose of decomposition is to provide a systematic description of parts that make up a system or assembly. For an initial concept, it is usually sufficient to perform only one level of functional or physical decomposition, but all subassemblies thus created will ultimately, as a part of their own conceptual design, be decomposed again and again. For example, a lawn mower engine may be decomposed into, among other things, an engine block and a carburetor. Then, in turn, the carburetor may be decomposed into, among other things, a float and a cover. Thus, the process of conceptual decomposition repeats (or recurs) until no new subassemblies are created, that is, until only parts or standard components are obtained. Physical decomposition and functional decomposition are not always mutually exclusive; they also can be used simultaneously in a design.

Conceptual Design of a Part. When the process of conceptual decomposition is completed and a list of parts is obtained, then the next step is conceptual design for each individual part. Conceptual design of a part involves the following steps (Ref 4):

- Determining whether the part is really necessary
- Identifying the required functions of the part
- Selecting the material and a manufacturing process for production

Definition of part function is the essential objective of conceptual design, and examples of common part type or features are listed in Table 1 for various functions. Some parts may have more than one function, and often parts have special features to enhance manufacturing or reduce material costs. Examples of these are described in Table 2. For economic reasons, the process of conceptual design may also address the possibility of eliminating a part or combining functions, because one complex part may be less expensive overall than two or more simpler parts. This step may involve consideration of materials and manufacturing costs and economics, but the reference book by Boothroyd and Dewhurst (Ref 5) provides one relatively easy method for determining whether a proposed part should be assembled from separate components.

Table 1 Functions served by parts

Function	**Examples of part types or features**
Transmit or support force(s) or torque(s)	Brackets, beams, struts, columns, bolts, springs, bosses, knobs
	Levers, wheels, rollers, handles
	Parts that fasten, hold, or clamp, such as bolts, screws, nails
Transmit or convert energy	
Heat	Heat fins, electric resistance heating elements
Mechanical power	Shafts, connecting rods, gears
Electricity	Wires, lightbulb elements, resistors
Provide a barrier (for example: reflect, cover, enclose, or protect)	
Light	Walls, plugs, caps
Heat	Thermal insulators, thermal reflecting surfaces
Electricity	Electrical insulators, magnetic shields
Sound	Walls, sound-absorbing wall surfaces
Control motion	Cams, grooves, slots, gears
Allow passage (of light, rods, shafts, wires, pipes, etc.)	Holes, windows, grooves
Control or regulate the passage of	
Fluids	Nozzles, orifices, pipes, ducts

Function	Examples of part types or features
Light	Shutters, wheels
Indicate	Clock hands, instrument needles, colors, embossing
Locate or guide	Grooves, holes, bosses, tabs, slots

Source: Ref 4

Table 2 Special features designed into parts to aid manufacturing or to reduce material cost

Function	Examples of part features
Aid manufacturing	Fillets, gussets, ribs, slots, holes
Add strength or rigidity (e.g., stiffen)	Ribs, fillets, gussets, rods
Reduce material use	Windows or holes through walls, ribs that allow thinner walls, slots
Provide a connection or contiguity (so the part can be a single part)	Walls, rods, ribs, gussets, tubes

Source: Ref 4

The process of conceptual design may also involve at least a preliminary decision on a material and manufacturing process to be employed. A physical concept of the materials and manufacturing process is generally required here, because most designs can never proceed very far without this information. In a more detailed approach to engineering design, Dixon and Poli (Ref 2) suggest a four-level approach to materials selection:

- *Level I:* Based on critical properties, determine whether the part will be made from metal, plastic, ceramic, or composite.
- *Level II:* Determine whether metal parts will be produced by a deformation process (wrought) or a casting process; for plastics, determine whether they will be thermoplastic or thermosetting polymers.
- *Level III:* Narrow options to a broad category of material. Metals can be subdivided into categories such as carbon steel, stainless steel, and copper alloys. Plastics can be subdivided into specific classes of thermoplastics and thermosets, such as polycarbonates and polyesters.
- *Level IV:* Select a specific material according to a specific grade or specification.

In this approach, materials and process selection is a progressive process of narrowing from a large universe of possibilities to a specific materials and process selection. Level I and level II often may suffice for conceptual design, while level III is needed for embodiment (configuration) design and sometimes for conceptual design. Level IV usually can be postponed until detail (parametric) design.

The four levels of materials selection in the previous list are just a starting point in narrowing options, because the process of materials selection requires the evaluation of many factors, as briefly summarized in more detail in the section "Materials Selection in Design" in this article. However, the key point is that materials selection is an up-front concern with important consequences for processing, product design, cost, availability, recyclability, and performance of the final product. This is why materials and processes selection can be a critical issue in the early stages of design. Moreover, the proliferation of new and specialized engineering materials has changed the complexion of design to the point that no engineer in a design capacity is conversant in all families of potential materials that can be used. The more critical an application is, the more important the materials selection becomes. Specialized materials expertise is mandated by the complexity of critical service, which very often includes extremes of temperature, stress, environment, or all three, as in the case of jet engine components. The services of a materials engineer should be obtained to foresee the complex material and property interactions and synergistic effects that may be attendant to a design. It is a logical conclusion that, particularly for complex and critical engineering designs, a cross-functional approach is best. A materials engineer might specify the materials and associated processes for an engine part but may not be able to design one. Similarly, a mechanical engineer may design an engine part, but may not be able to determine the materials and processes necessary for fabrication.

Integrated Product Development Teams. The integration of diverse engineering disciplines in design and materials selection is important even in the conceptual design. One form of integrating engineering functions to optimize design is the integrated product development (IPD) team concept. One of the strengths of the design team approach is that all disciplines have input early in the process, while decisions are easily changed and inexpensive improvements can be made. Cross-functional IPD teams are formed temporarily in many organizations for a particular product but are also formed somewhat permanently in others for continuous design support.

The IPD approach has been shown to lead to better results faster. For example, the use of an IPD team approach can be useful during configuration design (described subsequently), when designers may inadvertently create parts with geometric features that place severe restrictions on the selection of manufacturing processes, with even less freedom

remaining for materials selection. Similarly, overly restrictive and independent selection of the material will limit the manufacturing processes available. This is all the more reason to use IPD methods. However, until the IPD approach is in more common use, an alternative approach, referred to as a materials-first approach, may be useful. The materials-first approach depends on a thorough understanding of the service environment and advocates choices based on properties that satisfy those performance needs (see the section “Materials Selection in Design” in this article).

Other Specialized Approaches and Tools. Many other specialized design team approaches have been developed to evaluate design, primarily for critical systems where failure can be catastrophic. These methods include failure mode and effects analysis, failure mode, effect, and criticality analysis, fault tree analysis, and fault hazard analysis. These formalized methodologies use systematic evaluation and sophisticated computer programs to predict failure in complicated designs and can be an invaluable aid in materials selection. The section “Risk Assessment in Design” briefly reviews the use of risk and hazard analysis in design.

In addition to multifunctional design approaches, sophisticated design tools have been developed to assist in preventing failures. Fracture mechanics is often used to create flaw-tolerant designs for critical applications. Finite-element analysis (FEA) techniques have been implemented to dynamically evaluate the effects of material characteristics and geometry changes. These tools also can be of great benefit during configuration design, parametric design, design validation, or systematic investigation of a failed component. More details on these methods are discussed in other articles in this Volume.

Configuration Design (Embodiment)

The first step in the detailed design stage of Fig. 3 is configuration design. After the preliminary steps of concept design have been completed, the designers must define the features of the configuration. Ultimately, designers must determine exact numerical values for the dimensions and tolerances of parts during parametric design. However, before this can be done, designers need to define the general configuration of a part in terms of its physical arrangement and connectivity.

Configuration design is a qualitative (i.e., nonnumerical) process that defines the general features of a part in terms of functional interactions with other parts or its surrounding environment. These interactions include forces (loads and available support areas), energy or material flows, and physical matings or other spatial requirements (e.g., certain spaces may be unavailable to the part). The types of dimensional features that are defined during configuration design may include (Ref 4):

- Walls of various kinds, such as flat, curved, and so forth
- Add-ons to walls, such as holes, bosses, notches, grooves, ribs, and so forth
- Solid elements, such as rods, cubes, tubes, spheres, and so forth
- Intersections among the walls, add-ons, and solid elements

Usually, there are several—and sometimes many—ways to configure a part, and the best approach is generating, evaluating, and modifying a number of alternatives. A three-dimensional sketch that shows these interactions to approximate scale is generally a very helpful starting point of configuration design. The sketch shows the essential surroundings of the part and locates loads, possible support points or areas, heat or other energy flows, adjacent parts, forbidden spaces, and so on.

Configuration designs and the various alternatives need to be evaluated before numerical dimensions and tolerances are established. As described in Ref 4, this evaluation process can be guided by qualitative physical reasoning about the functionality of the part configuration and manufacturing. Even when actual dimensions have not been determined in the configuration stage of part design (i.e., when sizes and spatial relationships of the features are still only approximate), knowledge of physical principles and manufacturing processes can still be applied to help create the most effective alternative designs for further evaluation. General physical reasoning involved in the generation of part configuration and manufacturing alternatives is discussed in more detail in Ref 1, 2 and 4.

In addition to qualitative physical reasoning about functionality, effective part configurations also are strongly influenced by manufacturing issues and materials selection. At this point in the part design process, it is necessary to decide on a manufacturing process and at least a class of materials (e.g., aluminum, thermoplastic, steel). However, unless the information is needed for evaluation of the configurations, selection of the exact material (e.g., the particular aluminum alloy or thermoplastic resin) may be postponed until the parametric stage. Consultation with materials and manufacturing experts is, of course, strongly advised, and other factors, such as recycling concerns and existing business relationships, also may be relevant.

Finally, once the set of the most practical part configurations has been generated, a more formal evaluation should be performed. The evaluation can be done by Pugh's method (Ref 6) or by other methods presented in Ref 4. In any method, the comparison criteria for alternative configurations should include the following:

- *Functionality:* Can exact dimensions and tolerances be imposed that will enable the part to perform its function properly and reliably?
- *Use of materials:* When dimensions are imposed, will the configuration provide for efficient use of all the required material?
- *Mechanical failure:* When dimensions are imposed, can the risks of failure from mechanical causes, such as fatigue, excessive stress, buckling, and so forth, be made suitably low?
- *Analyzability:* Does the configuration enable analyses to be performed for stresses, vibrations, heat flow, and so forth?
- *Manufacturability:* Can the selected manufacturing process hold the tolerances that will be needed for the configuration to meet the required functionality? Does the configuration allow for ease of handling and insertion for assembly? Are there special issues that will influence the time required for tooling and production?

Parametric Design and Analysis

Conceptual and configuration designs are based primarily on qualitative reasoning about physical principles and manufacturing processes. In parametric design, however, numerical computations become much more important. The attributes of parts identified at the configuration stage become the design variables, which must be identified and analyzed during the step of parametric design.
Evaluation in parametric design requires computation of performance parameters as well as selection and implementation of a method for evaluating the overall quality of the trial design. During parametric design, the design or process engineer seeks to optimize performance by:

- Identifying design variables and their allowable range
- Identifying performance parameters whose values will be computed or measured to evaluate the performance of trial designs
- Identifying the analysis methods that will be used to compute values for the evaluation

Often, parametric design values and procedures are governed by design codes, standards, or test methods. These may be specific to a particular discipline, industry, or class of material. In addition, design tools such as fracture mechanics and FEA can be used to analyze designs and evaluate the effects of material characteristics and geometry changes. These tools also can be of great benefit during configuration design, parametric design, and design validation.

Risk Assessment in Design

All designs balance expected benefits against potential risks. Therefore, the notion of explicitly looking for technical and manufacturing risks in a design is useful. The idea is to look for any previously unquestioned assumptions that have been made while generating the configuration and while doing the evaluations—that is, to look for issues that may so far have been overlooked.
Risk cannot be avoided completely, even for very conservative designs. Indeed, good designs will have some reasonable risk or they will be too conservative, too costly, too heavy, and so on. However, designers cannot count on good luck; there are many more ways for a design to fail than there are for it to succeed. Thus, risks must be sought out, faced, and evaluated. Table 3 lists questions intended to reveal risks related to part configurations.

Table 3 Questions for revealing part configuration design risks

Factor	Questions
What are the most likely ways the part might fail in service?	
Excessive stress	Can the part be dimensioned to keep stresses below yield failure levels? Add ribs? Use stronger material?
Fatigue	If there will be cyclic loads, can the configuration be dimensioned so as to keep the internal stresses below the fatigue limit?
Stress concentrations	Can the part be dimensioned to keep local stress concentrations low?
Buckling	If buckling is a possibility, can the configuration be dimensioned to prevent it?
Unexpected shocks or loads	What unexpected dynamic loads might be encountered in service or in assembly? Can these be handled by the configuration?
What are the most likely ways the part might not meet its expected functionality?	
Tolerances	Is the configuration such that functionality will be especially sensitive to the actual tolerances that

Factor	Questions
	can be expected in a production situation? Are too many special (tight) tolerances required to make the part work well?
Creep	If creep is a possibility, will it result in loss of functionality?
Strain and deformation	If functional performance is sensitive to retention of size and shape, can the configuration be dimensioned to preserve the required integrity?
Thermal deformations	Might thermal expansion or contraction cause the configuration to deform so that function will be impaired?
Handling and assembly	Might there be unforeseen difficulties with handling and assembly?
Dimensions	Might the part end up being dimensioned so that assumptions about assembleability become invalid?
Tangling	Might the parts tangle if dimensioned in some way?
Will the available production machines be able to make the part?	
Production runs	Are the desired production runs consistent with the machines and expected costs?
Tooling wear	Is tooling wear or maintenance a possible problem that will impact part cost or performance?
Weld lines[(a)]	If the process is a flow process, can weld lines be located appropriately?
Other design and materials factors	
Geometric compatibility	Is the part geometrically compatible with its adjoining parts? What could go wrong is this regard? If there is a small change in this part, or in an adjoining part, can the configuration accommodate the change without major redesign? What about the effects of tolerances of the *adjoining* parts? Or on the assembly as a whole?
Materials	Is the material selected compatible with the configuration and the manufacturing process? Is surface finish properly accounted for? Will standard raw material supplies be of adequate quality? Has the material been thoroughly investigated for its use in *this* particular application? Are there previous uses in similar applications? Have experts on the properties and processing of the material been consulted? Is the material compatible with the rest of the product?
Designer and design team knowledge	Has every possible, unfortunate, unlikely, unlucky, even stupid "What if …" situation been considered? Are there aspects of the part design where the designer or design team is working without adequate knowledge? Where is the design based on insufficient knowledge of materials, forces, flows, temperatures, environment, etc.? Where are there guesses, hopes, fears, and assumptions instead of knowledge: Materials? Stresses? Fastening methods? Manufacturing process? Tolerances? Costs? Adjoining parts? Environmental conditions?

(a) A weld line is formed when a material flow must divide—say around a hole—and then rejoin. The weld lines tend to be weaker and more subject to fatigue failures. Source: Ref 4

Risk and Hazard Analysis. One tool used in the evaluation of risk is the process of risk and hazard analysis, which helps identify the level of risk and to pinpoint the parts of the system that represent the greatest risk for failure. If the analysis is used properly, steps can be taken to eliminate the cause or reduce the risk to an acceptable minimum. Some hardware systems approaching a "failure-free" condition may be produced when actions are taken at all levels that are based on:

- Attention to past experiences with similar systems
- Availability of risk information for all project personnel
- A sound, aggressive risk and hazard analysis during all design phases
- Development of suitable corrective action and safety programs based on the analysis
- A continuous and searching review of all phases of the program efforts

Rigorous applications of risk and hazard analysis have made difficult technological feats, such as landing on the moon, relatively accident-free.

The various analysis techniques of risk assessment have grown out of the search for system reliability. Consequently, the approach is hardware-oriented, with the emphasis on ensuring that hardware is able to perform its intended function. Backup systems and redundancies are also used to reduce such risks. Through cost/benefit analysis, the performance of the system will have a computable value that can be compared to the cost of accomplishing the objectives desired for a product or system.

Risk and hazard analysis tools have been developed to ensure system reliability in critical applications. With the increased emphasis on safety, reliability, and achieving performance objectives, design teams must incorporate risk/hazard considerations in their designs. Figure 4 (Ref 7) is a flow chart that shows the integration of risk and hazard analysis in the overall design process. Even if designers or design managers are not directly responsible for carrying out these analyses, they must be familiar with the methodology, so that they understand how they are carried out and how they can

respond in terms of design or system changes. Most efforts are best carried out during early design phases, and they can be effectively used during design reviews to provide valuable feedback to the design to avoid failures. More information on the principal methods of risk/hazard analysis is presented in Ref 7.

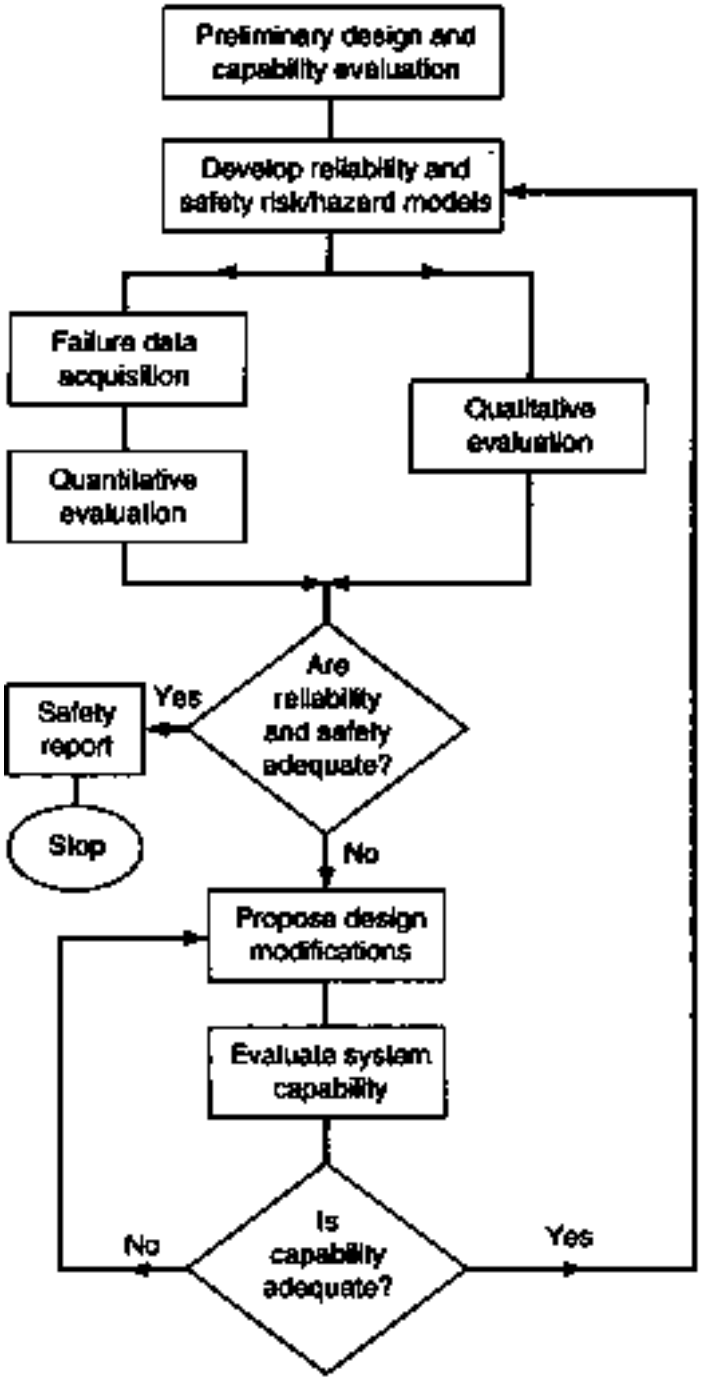


Fig. 4 Flow chart showing the integration of risk and hazard analysis into the design process. Source: Ref 7

References cited in this section

1. H.W. Stoll, *Product Design Methods and Practices,* Marcel Dekker, 1999, p 40, 148

2. J.R. Dixon and C. Poli, *Engineering Design and Design for Manufacturing,* Field Stone Publishers, 1995

3. C.O. Smith, Human Factors in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

4. J.R. Dixon, Conceptual and Configuration Design of Parts, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 33–38

5. G. Boothroyd and P. Dewhurst, *Product Design for Assembly,* Boothroyd Dewhurst, Inc., 1989

6. S. Pugh, *Total Design: Integrating Methods for Successful Product Engineering,* Addison-Wesley, 1991

7. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Materials Selection in Design

One of the chief concerns of any design or engineering effort is materials selection. Materials selection is a process whereby the function and desired final properties of a component are evaluated during all the various stages of design in order to identify suitable materials of construction. During every stage of the design process (i.e., conceptual, configuration, and parametric), some level of materials selection must be made in order to proceed with the design. This is one reason why, as previously noted, integrated product development (IPD) teams have been used.

Moreover, the options in materials selection have proliferated. The number of materials currently available for designers has grown, as shown by the timeline in Fig. 5 (Ref 8). This trend will probably continue, thus making the function of materials selection more difficult than it was many years ago. The manufacturing processes available to designers have also grown substantially. These are additional reasons why IPD teams are used, although perhaps on a limited or temporary basis.

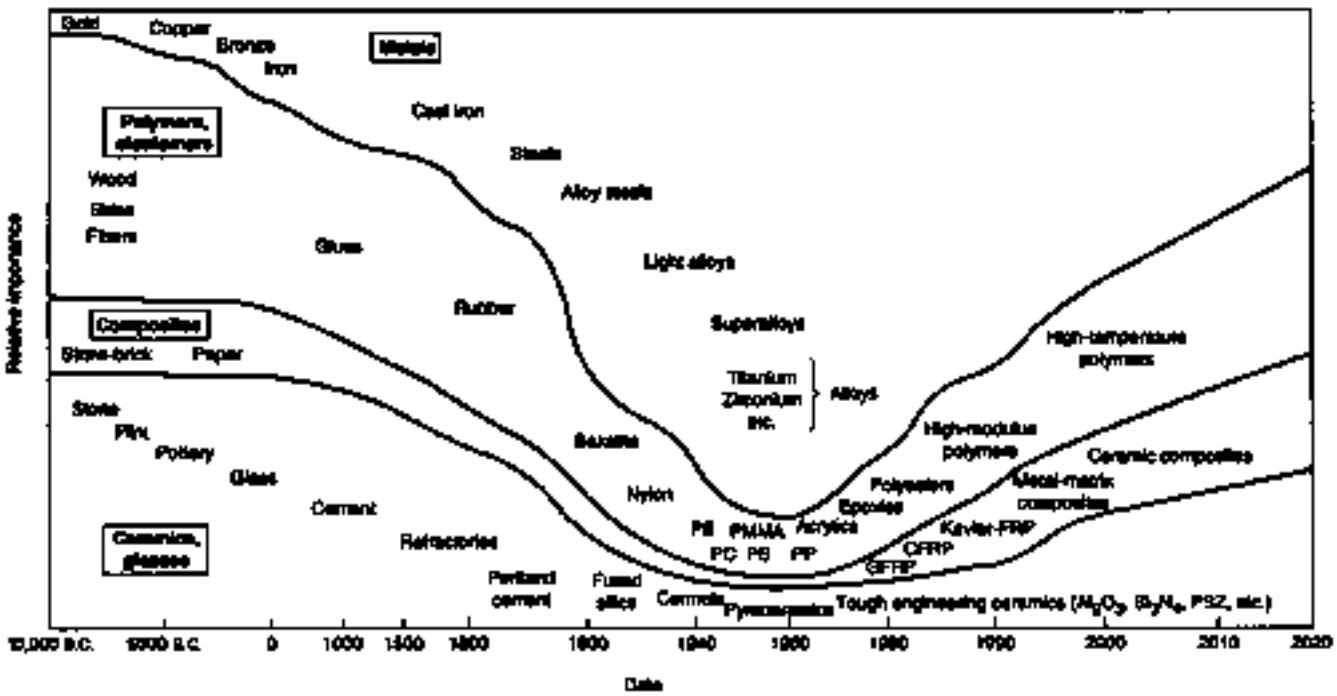


Fig. 5 The evolution of engineering materials through history. PE, polyethylene; PMMA, polymethylmethacrylate; PC, polycarbonate; PS, polystyrene; PP, polypropylene; CFRP, carbon-fiber-reinforced plastic; GFRP, graphite-fiber-reinforced plastic; PSZ, partially stabilized zirconia. Source: Ref 8

Until the IPD approach is in common use, and as an alternative to the traditional “materials-last” approach, another method of materials selection is the “materials-first” approach. The materials-first approach depends on a thorough understanding of the service environment and advocates choices based on properties that satisfy those performance needs. After the component is envisioned, the environment is evaluated, and the constraints are applied, the design concept (including preliminary selection of materials, processes, and product form) can be developed further during configuration design. This materials-first approach involves the selection of a general class of materials during conceptual design and further refines the alternatives through the iterative process of design. The selection process may involve a large set of performance and property criteria, which must be refined and developed during design. However, there are four elementary topics in materials selection that must be addressed during all stages of design. They are posed as simple questions:

- What is it?
- What is its environment?
- What cannot it be?
- What must it be?

These questions can be viewed as conceptual constraints or factors that influence materials selection in the stage of conceptual design. Then, of course, more detailed materials selection criteria and properties are used during the stages of configuration and parametric design.

The process of materials selection varies substantially for different purposes, because steps are typically amended to suit specific applications, and this process is sometimes formalized to ensure all of the steps have been thoroughly performed. This process is dynamic, because changes in design have to be considered in an ongoing fashion. Proper materials selection is a dynamic function that must accompany all design activities while remaining sufficiently flexible to accommodate inevitable engineering design changes. It is an aspect of design that is every bit as crucial as part dimensions and geometry. Proper materials selection is necessary during new design and for the improvement of existing designs that are found to be marginally or completely unsuitable.

Identification of the materials selection as a constant contributor to the design of a component also includes the necessity of incorporating requisite materials information on engineering drawings. Because a complete part drawing is a metric for establishing whether a fabricated component is acceptable, all pertinent characteristics must be identified. The specificity of materials and processing information required for thorough identification on the engineering drawing is a function of the complexity of the component. Newly created drawings are typically electronic computer-aided design constructs. These drawings require less storage space but can be as prone to human errors as the time-honored autographic drafting methods.

Many companies cross-reference stand-alone materials specifications that contain much more detailed purchasing and processing information than could be conveniently placed on the part drawings. This practice has an additional benefit in that a large number of drawings using the same material can be upgraded simultaneously without having to revise each drawing individually. With respect to materials and process information on drawings, some standardized symbols have been created. The symbols in current use include welding joint design and instructions, dimensional tolerances, and finish symbols. These symbols provide a great amount of information without appearing as additional written process descriptions and hence present information more clearly and effectively on drawings.

Materials Selection during Concept Design

At the concept level of design, materials and processes are considered rather broadly. The decision is to determine whether each design concept will be made from metal, plastics, ceramic, composite, or wood and to narrow it to a group of materials. The precision of property data needed is rather low. If an innovative choice of material is to be made, it should be done at the conceptual design step. Materials selection at this stage of design may use tools such as material property charts or general performance indices (e.g., see the articles “Material Property Charts” and “Performance Indices” in *Materials Selection and Design* Volume 20, *ASM Handbook*.)

The four fundamental questions of materials selection should also begin during the stage of conceptual design, where all the functional physical conditions (including any major economic and nontechnical conditions) are imposed. Simply, the component material is defined by what it is, by its intended environment, by what it cannot be, and by what it must do. Often, these criteria, although oversimplified here, can be a good acid test in narrowing the possible alternatives of material classes during conceptual design.

The Design Objective—What It Is. This objective is typically a simple component description before all of the requirements are identified. The designer must ensure separation of what something is from what it does. In other words, the physical embodiment (or configuration) may be related to the intended function, but other configurations may perform the same function. It is relatively easy to lose sight of the very basic utility of a part or structure during design by

inadvertently insinuating arbitrary constraints. This objectivity may be confusing, but incorrect assumptions during design conception can be difficult to surmount later.

The Design Environment—What Its Environment Is. A structure or component can be affected by a wide variety of service environments that may include:

- Temperature extremes
- Temperature fluctuation
- Alkalinity
- Acidity
- Pressure
- Oxygen content
- Flammability
- Flame impingement
- Humidity (wet/dry cycles)
- Galvanic differences
- Moisture
- Liquid metal
- Flow/flow rate
- Erosion
- Cavitation
- Hydrogen content
- Biological agents

This list shows many of the environmental factors that must be kept in mind during the design process, before any thought is given to properties. When the service conditions are not adequately understood (as in the early days of the space program), design can become a costly, iterative process requiring extensive trial-and-error bench testing.

The environment also must be evaluated prior to identifying necessary properties, because mechanical and physical properties can be severely altered by environmental factors. Very specific environments are often considered, including extremely corrosive high-temperature or high-pressure applications.

Design Constraints—What It Cannot Be. Design constraints are industry-specific or self-imposed restrictions on the materials or processes that may be considered in the design process. These constraints come from a variety of sources. Sometimes material constraints are applied by the end user, which may dictate exact materials and processes as a contractual obligation.

Constraints can act as an aid in the design process, because they obviate the consideration of certain prohibited materials or processes, narrowing down the possibilities. It must be kept in mind that in some cases, these constraints may not be realistic or well advised. It is not unusual to encounter over-restrictive or seemingly arbitrary constraints. Constraints can also be indirectly applied. If an entire assembly has a certain constraint, such as total weight, it can become rather complicated to balance the necessary weights and properties of the individual components.

Cost Constraints. Financial constraints accompany each engineering design, except for unusual critical applications when properties are far more important than relative material expense. Design choices can be severely limited by economic factors, particularly in the manufacture of highly competitive consumer items. It is not unusual to be financially constrained to using essentially the same material as the competition, especially in a marketplace without real or perceived product differentiation.

Quantity Constraints. In the case of a single component or structure being designed, it will not be necessary to tool an assembly line or create a manufacturing process capable of making them by the millions. Production of a few components can sometimes be given more personal attention, permitting the inherent labor costs to be a higher percentage of the total price. On the other hand, if a large number of identical or similar items are to be produced, an assembly line approach is mandated, with an accompanying reduction in the relative labor costs. The in-between cases are the most typical, hence the rise of small, medium, and large job shops capable of sufficient flexibility to produce a variety of parts on a short-term contract basis.

Size and Weight Constraints. The rough size and weight of a finished design must be approximated early in the design process, because they may constrain the subsequent design options to a great extent. Maximum sizes and weights can restrict the amount of margin that is possible.

Material Property and Processing Constraints. Many property constraints are placed on materials by the very nature of the item being designed. Restrictions to manufacture can also be present as the willingness to use only those processes and fabrication techniques for which the equipment is already on hand. There are certainly financial advantages to be gained by maximizing utilization of existing facility and equipment capabilities. For example, if a manufacturer has a captive heat treatment department, they may be prone to exploring heat treatment as a preferred processing option. Complicated

and highly technical processing steps are usually best addressed by specialists. Many codes and specifications allow a broad range of materials selections, whereas other codes are very specific and allow few substitutions.

What It Must Be. After the component is envisioned, the environment is evaluated, and the constraints are applied, the design concept, including materials selection, can be developed. A large set of performance and property criteria may be developed in order to define the function and surrounding conditions of a part. Primary and secondary criteria should also be identified. Primary, or absolute, requirements are essential to proper service and cannot be subordinated, whereas secondary requirements are those where judicious compromises can be made. For example, life-cycle considerations (such as recycling or environmental impact) may be a primary or secondary criterion, depending on the product objectives.

Selection Criteria during Detailed Design

The stage of detailed design (Fig. 3) includes the embodiment or configuration level of design and parametric design. During configuration design, the emphasis is on determining the shape and approximate size of a part using engineering methods of analysis, which can be based on methods of qualitative physical reasoning. During parametric design, quantitative methods are used to refine the design further.

Materials selection during configuration design requires the evaluation of a range of material (e.g., a range of carbon steel, low-alloy steel, stainless steel, age-hardening aluminum alloys, etc.), its general product form (e.g., wrought, cast, powder metallurgy, etc.), and the processing method (e.g., forged, die cast, injection molded, etc.). All of these factors must be considered when the shape of a part is defined during configuration design. Material properties during configuration design must also be known to a greater level of precision than in conceptual design, at least to allow qualitative comparison of the alternatives for the possible choices of material type, product, and processing method.

At the detail or parametric design level, the materials selection is narrowed further to a specific grade of material and manufacturing processes. Here, the emphasis will be on quantitative evaluation of allowable variations in material properties, critical tolerances, and any other performance parameters of the design (including the best manufacturing process using quality engineering and cost-modeling methodologies). Depending on the criticality of the part, material properties may need to be known to a high level of precision, with quantitative evaluation of variations in properties or performance. For example, anisotropic variations in the properties of worked products, or the effects of surface finish after machining, are quantitative factors that must be considered during parametric design. At this extreme, the development of a detailed property database or an extensive materials-testing program may be required.

Detailed evaluation of the size, shape, processing, fabrication, and material properties of an engineered part requires communication between designers, materials or manufacturing engineers, quality assurance, and purchasing agents. It can be a relatively simple or complex task, depending on the criteria for materials selection. Examples of materials information required during detailed design are listed in Table 4 (Ref 9). It also includes experience and application history, such as failure analysis reports. During design, it is necessary to identify primary and secondary materials selection criteria, and the following list contains a number of typical materials selection criteria that would be identified during the creation of a new component (Ref 10):

- Size
- Shape
- Weight
- Strength
- Wear resistance
- Environmental resistance
- Loading capabilities
- Life expectancies
- Fabricability
- Quantity
- Availability
- Cost
- Specifications
- Recycling
- Scrap value
- Standardization
- Safety

Table 4 Examples of materials information required during detail design

Material identification
Material class (metal, plastic, ceramic composite)
Material subclass
Material industry designation
Material product form
Material condition designation (temper, heat treatment, etc.)
Material specification
Material alternative names
Material component designations (composite/assembly)
Material production history
Manufacturability strengths and limitations
Material composition(s)
Material condition (fabrication)
Material assembly technology
Constitutive equations relating to properties
Material properties and test procedures
Density
Specific heat
Coefficient of thermal expansion
Thermal conductivity
Tensile strength
Yield strength
Elongation
Reduction of area
Moduli of elasticity
Stress-strain curve or equation
Hardness
Fatigue strength (define test methods, load, and environment)
Temperature (cryogenic-elevated)
Tensile strength, yield strength
Creep rates, rupture life at elevated temperatures
Relaxation at elevated temperatures
Toughness
Damage tolerance (if applicable)
Fracture toughness (define test)
Fatigue crack growth rates (define environment, and load)
Temperature effects
Environmental stability
Compatibility data
General corrosion resistance
Stress-corrosion cracking resistance
Toxicity (at all stages of production and operation)
Recyclability/disposal
Material design properties
Tension
Compression
Shear
Bearing
Controlled strain fatigue life
Processability information
Finishing characteristics
Weldability/joining technologies
Suitability for forging, extrusion, and rolling
Formability (finished product)

Castability
Repairability
Flammability
Joining technology applicable
Fusion
Adhesive bonding
Fasteners
Welding parameters
Finishing technology applicable
Impregnation
Painting
Stability of color
Application history/experience
Successful uses
Unsuccessful uses
Applications to be avoided
Failure analysis reports
Maximum life service
Availability
Multisource? Vendors?
Sizes
Forms
Cost/cost factors
Raw material
Finished product or require added processing
Special finishing/protection
Special tooling/tooling costs
Quality control/assurance issues
Inspectability
Repair
Repeatability

Source: Ref 9

This list contains the most-used criteria for materials selection but is by no means exhaustive (as suggested by the examples in Table 4). Selection criteria can vary as much as the items being designed. These concerns also are not entirely independent, but they are described individually. It is logical to assume that complicated service requirements will result in more stringent selection criteria. More restrictive selection criteria will invariably result in fewer materials that will likely satisfy the design requirements. Even so, engineering materials selection is very rarely a question of a single, suitable material. As in engineering design, materials selection can be an iterative process that compares alternatives during both conceptual and detailed design.

Size Considerations. The size of a designed component can often dictate the form of the material to be used. Very large parts may need to be fabricated from structural shapes, castings, or forgings. Welded fabrication may be necessitated. Extremely small components may need to be created by powder metallurgy, metal injection molding, or other fine forming techniques.

Shape Considerations. The shape and geometrical complexity of a component must also be considered in materials selection. Intricate shapes may not fill during casting or may not be possible to form by other methods, such as extrusion or forging. Many types of material forms are available, such as castings, forgings, extrusions, rolled shapes, wire and rod, plate and sheet, and many hybrid forms. In considering shape during materials selection, it is always best to match the form to the function during configuration design. This practice reduces scrap and promotes the optimal use of material with the desired properties.

Several material forms may be determined to be suitable, and additional factors must be assessed to determine the optimal shape. In the most general sense, increasing complexity narrows the range of processes and increases cost. A cardinal rule of design is, therefore, to keep the shape as simple as possible. This rule may, however, be broken if a more complex shape allows consolidation of several parts and/or elimination of one or more manufacturing steps. Limitations on shape are also imposed by properties of the material and by interactions with the production tooling. For example, minimum wall or section thickness of the web form shown in Fig. 6 (Ref 11) is a function of manufacturing process and material. The aim is, generally, to produce a net shape part ready for assembly. If this is not feasible, a near-net shape part that will need only minor finishing, usually by machining, is desirable.

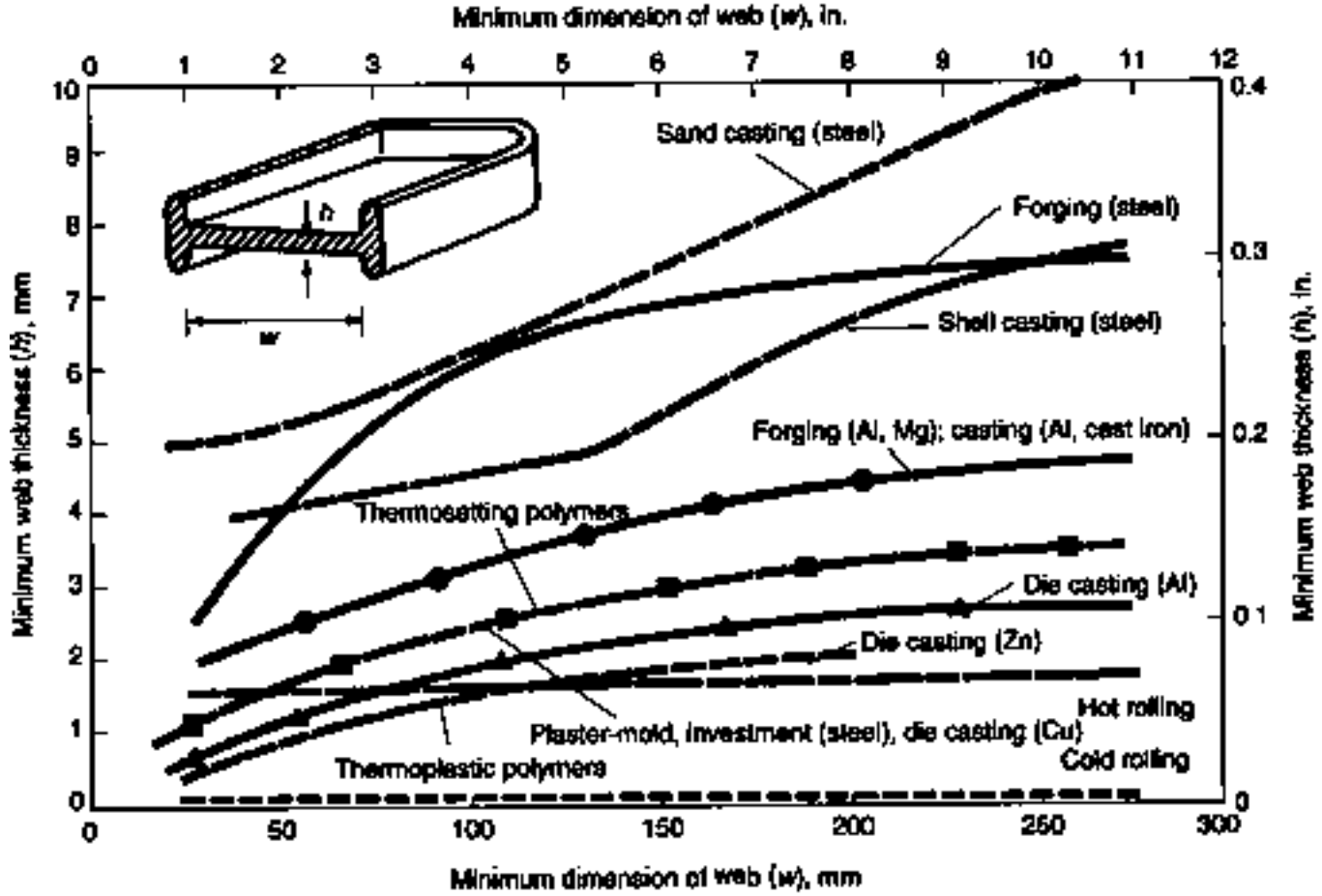


Fig. 6 Example of minimum web thickness for different materials and manufacturing processes. Source: Ref 11

Weight Considerations. There are very few applications of manufactured products where weight is not a consideration. Material weight is of vital importance in aerospace and automotive applications, where incremental weight reduction can be directly measurable in fuel savings and payload capacity. Strength-to-weight ratio, often called specific strength, is a hybrid consideration typically used in vehicle design. Aside from vehicles, weight considerations are also important if the materials are to be transported during manufacture or during service, or if the part moves during service, as in rotating or reciprocating parts of an engine or machinery.

Material Properties. A list of many of the properties to be considered during materials selection is shown in Table 5. Each primary or secondary property attribute must be carefully considered. Probably the most fundamental tenet of materials science is that properties are a function of structure and structure is a function of processing. The properties can never be considered separately from the processing, because processing decisions or steps can affect the nominal value and variability of a property within the geometry of a part. For example, austenitic stainless steel bar can be processed to high strength by drawing but can also be annealed to lower strength and improve ductility. Another example is the variability of properties within the geometry of a part (i.e., anisotropy), depending on the nature of the manufacturing process.

Table 5 Typical material properties used for selection

Tensile strength
Yield strength
Elongation
Compressive strength
Shear strength
Fatigue strength
Fracture toughness
Impact strength
Transition temperature
Modulus of elasticity

Wear resistance
Hardness
Lubricity
Density
Porosity
Melting point
Thermal stability
Thermal expansion
Thermal conductivity
Electrical conductivity
Magnetic characteristics
Galvanic character
Corrosion resistance
Optical characteristics
Fabrication characteristics
Welding characteristics
Finishing characteristics
Hardenability
Aesthetics

It is also important to understand relationships between the mechanical and physical properties. Some of the typical material property relationships are illustrated in Fig. 7 (Ref 12). This diagram shows the general inverse relationships between desirable and possibly undesirable characteristics. Each desired property will likely have attendant properties that may not be desirable. Therefore, materials selection will always contain a level of educated compromise, similar to the design process in its entirety.

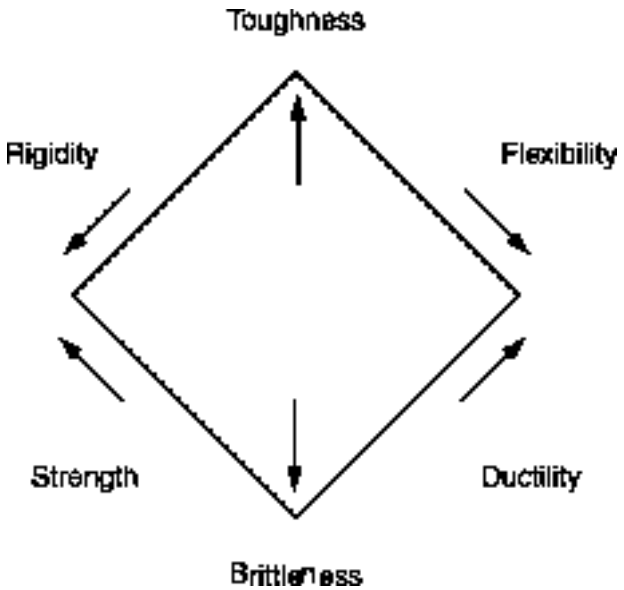


Fig. 7 General relationships of different mechanical behavior. Rigidity and strength are generally inversely related to flexibility and ductility. Source: Ref 12

Wear Resistance is a very important property for all materials that come into repeated or non-stationary contact with other materials. The three principal types of wear are adhesive wear, abrasive wear, and corrosive wear. Unfortunately, the measure of wear resistance is problematic, because there are so many variables involved, including friction factors, lubrication factors, surface finish, and so on. As a result, tables of comparative wear resistance, even ostensibly employing the same standard testing methodology, have limited usefulness in materials selection. Optimal wear properties are usually obtained by prototype testing in an actual intended-service situation.

Knowledge of Operating Environment. This portion of the materials selection differs slightly from the original environmental constraints placed on the design during the conceptual stage. The environmental resistances and behaviors of individual materials are of interest, although very often the conceptual design requirements are the same as those for the configuration of individual parts. However, during configuration design, the surrounding environment may involve galvanic differences and potential deleterious interactions within a complex assembly or structure. Parametric design must account for corrosion rates, material replacement rates, and other life factors.

Types of Loading. The type and magnitude of applied loading are crucial aspects of materials selection. While considering potential materials with other requisite properties, the following general loading types must be addressed:

- Constant, sustained loading
- Cyclic, repetitive loading
- Rapid, shock loading
- Slow loading
- Distributed loading
- Concentrated loading
- Variable loading

The load-carrying attributes of a material can be altered substantially through different types of processing. Any processes that alter mechanical properties, such as heat treatment, rolling, welding, and grinding, can affect the load-bearing characteristics by enhancing or reducing the resistance to specific loading types. Combinations of loading types make the materials selection process more difficult.

Life Requirements. The life requirements influence the materials selection, because longer service duration can often necessitate more sophisticated materials. Some components and assemblies are single use and do not require prolonged capabilities. Many structures are intended to survive fifty or a hundred years before demolition and replacement. Short-term design can often be of a disposable nature, whereas long-term design may permit substantial repair or refurbishment during service.

Fabricability (Design for Manufacturing). Designing for effective and efficient manufacturing can be rather involved but is important in cost-effective designs. Fabrication and manufacturing characteristics may be difficult to quantify, because it may be a composite of many subjective measures, such as formability, machinability, and weldability. Every process that is applied to a material must be evaluated to determine if the process and material are compatible. The first step may be a qualitative comparison during conceptual or configuration design, based on compatibility charts such as the one in Table 6. The necessary processing for a material also might be so cumbersome and costly, or impossible, that an otherwise suitable material would be logically removed from consideration. Materials selections may impose additional inspection, heat treatment, welding, machining, and finishing requirements during manufacture.

Table 6 Compatibility between materials and manufacturing processes

Process	Cast iron	Carbon steel	Alloy steel	Stainless steel	Aluminum and aluminum alloys	Copper and copper alloys	Zinc and zinc alloys	Magnesium and magnesium alloys	Titanium and titanium alloys	Nickel and nickel alloys	Refractory metals	Thermoplastics	Thermoset plastics
Casting/molding													
Sand casting	•	•	•	•	•	•	—	•	—	•	—	X	X
Investment casting	—	•	•	•	•	•	—	—	—	•	—	X	X
Die casting	X	X	X	X	•	—	•	•	X	X	X	X	X
Injection molding	X	X	X	X	X	X	X	X	X	X	X	•	—
Structural foam molding	X	X	X	X	X	X	X	X	X	X	X	•	X
Blow molding (extrusion)	X	X	X	X	X	X	X	X	X	X	X	•	X
Blow molding (injection)	X	X	X	X	X	X	X	X	X	X	X	•	X
Rotational molding	X	X	X	X	X	X	X	X	X	X	X	•	X
Forging/bulk forming													
Impact extrusion	X	•	•	—	•	•	•	—	X	X	X	X	X
Cold heading	X	•	•	•	•	•	—	—	X	—	X	X	X
Closed die forging	X	•	•	•	•	•	X	•	•	—	—	X	X
Pressing and sintering (P/M)	X	•	•	•	•	•	X	•	—	•	•	X	X
Hot extrusion	X	•	—	—	•	•	X	•	—	—	—	X	X
Rotary swaging	X	•	•	•	•	—	—	•	X	•	•	X	X
Machining													
Machining from stock	•	•	•	•	•	•	•	•	—	—	—	—	—
Electrochemical machining	•	•	•	•	—	—	—	—	•	•	—	X	X
Electrical discharge machining (EDM)	X	•	•	•	•	•	—	—	—	•	—	X	X

Wire EDM	X	•	•	•	•	•	—	—	—	•	—	•	X
Forming													
Sheet metal forming	X	•	•	•	•	•	—	—	—	—	X	X	X
Thermoforming	X	X	X	X	X	X	X	X	X	X	X	•	X
Metal spinning	X	•	—	•	•	•	•	—	—	—	—	X	X

•, normal practice; —, less-common practice; X, not applicable; P/M, powder metallurgy. Source: Ref 9

During parametric design, quantitative evaluation of tolerances, tooling, and production costs would be required. It is often necessary to design the manufacturing tooling concurrently with the end product to be made on that tooling. This is especially true for near-net shape processes such as molding, casting, and forging. Sometimes, the designed components must be altered to permit manufacture; hence, the manufacturing functions need to be involved in the design from the beginning.

Quantity Requirements. The quantity of a component to be designed may also influence the material and processing options that are feasible. The manufacture of high volumes of parts may necessitate mass-production methodologies. High volumes may allow use of forming and production techniques that require expensive tooling and dies that would be financially unfeasible when only a few parts are to be produced. These high-production methods can be very cost-intensive, inflexible, and slow to become profitable. Lower quantities of components can often allow more individual attention to the quality and characteristics of each produced part. Many production and processing methods are not applicable to low quantities of parts.

Availability. Materials, as a result of their popularity and relative natural scarcity, may not always be available as production may require. Even abundant materials in unusual forms can become difficult to procure. Design of long-term projects or continuous production includes an implied assumption that the selected material will remain an obtainable and economical choice in the future.

Lower quantities of material types, forms, and shapes can be below that level which mills will supply directly, and these would need to be purchased from a service center or distributor. The uniqueness of the material may be problematic, because small buyers cannot singly affect what mills will produce. Reduced demand from other manufacturers may make desired materials no longer available.

Both raw material and alloying elements are not uniformly available. Foreign sources may be hostile or inconsistent, and general availability may severely restrict supply. Special consideration must be given to using any base materials or processing materials that may not have the requisite availability due to factors that cannot be controlled. These materials are often called strategic materials, and they can become a great concern during wartime, when necessary materials may become of short supply.

Cost. Determination of costs accompanying potential materials selections is not as straightforward as it may seem. In many applications, the material cost was traditionally dictated by the cheapest material that was available that satisfied the previously determined mechanical and physical property requirements. Modern design practices incorporate material and processing costs almost as a property of that material, to be a direct comparison factor. Artificial constraints to using only the subjectively least-expensive material available ignores additional potential benefits of more expensive materials, such as reduced maintenance, longer life, and better reliability. Often a value-in-use approach is employed to better evaluate the costs of potential materials. In this methodology, the additional benefits of better performance can be quantitatively considered along with the basic material cost. In some instances, the additional processing costs for cheaper materials may result in greater total expense than those materials that are traditionally more expensive.

The amount of total component or structure cost that is included in material and subsequent processing varies widely in different industries. Large structures use great volumes of typically lower-cost engineering materials, resulting in the material being a relatively low percentage of the overall project costs. Aerospace and electronic components are typically smaller, specialized items, where the material costs can be higher than the processing and installation costs. In extreme cases, the cost is no object, within reason. The designer must be certain that the accompanying production costs, maintenance costs, potential repair, and downtime costs are considered for prospective materials.

The costs of ordering and warehousing engineering materials are also a consideration. Depending on the amount of material needed, there is likely an economic order quantity that best suits the production requirements and minimizes material costs. This is also a dynamic function, because materials and material forms have unpredictably mutable costs and availability that can alter future purchasing requirements. In general, all costs regarding purchasing, receiving inspection, and storage are reduced, on a per pound basis, by the purchase of large amounts.

Existing Specifications and Codes. In many industries, applicable standards provide materials prohibitions or requirements above those applicable as original design constraints. Standards can restrict material form, heat treatment, welding, and other processing variables. Purchasing material grades and alloys to uniform, popular standards can result in greater availability of materials, due to a greater number of potential suppliers.

Standards can be industry consensus standards, domestic and foreign federal regulations, and customer-supplied engineering specifications. Reliance on these codified requirements is often precarious, because they are sometimes vague and can be interpreted in many ways. Many specifications still require producer and client agreement on crucial processing variables. Some statutes (e.g., the Federal Child Safety Act) may also mandate possible materials and a given design procedure.

Feasibility of Recycling. The potential recycling of manufacturing and process scrap can be an important selection parameter. The expense of a material may be easier to justify if all removed material may be recycled for remuneration via return to the supplier rather than requiring landfill or hazardous material disposal expenditures. Identification of recyclability or other end-of-life considerations as a primary or secondary selection attribute is well advised, even if the

remainder of the design is not part of a life-cycle design effort. This ability considers the ease of sorting and separation and the fluctuating cost of recycled materials compared to newly extracted materials.

Scrap Value. The consideration of whether a designed component may be profitably scrapped at the end of its useful life is an important part of materials selection. Reuse of nondegraded components or rework and refurbishment is often an inexpensive alternative to new purchases.

Standardization of Designs. Standardization of materials selection within organizations that have extensive and continuous design functions must be considered. It is possible to design identical components that will fit in multiple assemblies. This practice of employing analogous designs and materials can prevent costly, redundant design projects or "reinvention of the wheel."

Safety. Perhaps the most important factor in design of a structure or component is safety. A design successfully reaching the manufacturing stage is entirely dependent on critical review and scrutiny about whether the necessary safety factors are satisfied.

Safety also must be included as a selection consideration during manufacture and processing. Materials and processes exhibit potential safety concerns, such as toxicity, flammability, inhalation of fine particles, autoignition, and contact hazards as well as long-term effects such as carcinogenic and pathogenic characteristics. Manufacturers are compelled to make products safer, due to the ethical imperative and economic self-interest.

Regulatory bodies are continually assessing the potential health hazards of relatively newly developed materials. This is due to the greater number of lesser-known and more rare metals and nonmetals that are being incorporated into engineering design. The Occupational Safety and Health Administration (OSHA) and other federal regulatory bodies are constantly collecting greater knowledge of potential health concerns regarding various materials, during both their manufacture and service. Avoiding the use of materials or processes under scrutiny for health concerns would generally be prudent.

References cited in this section

8. R.W. Heckel, Introduction to the Effects of Composition, Processing, and Structure on Materials Properties, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 333

9. G. Dieter, Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 243–254

10. H. Boyer, Introduction: Selection Criteria, *Selection of Materials for Component Design,* ASM International, 1986

11. J.A. Schey, *Introduction to Manufacturing Processes,* 2nd ed., McGraw-Hill, 1987

12. R. Gunia, Ed., *Source Book on Materials Selection,* Vol 2, American Society for Metals, 1977

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

Materials Selection for Failure Prevention

The use of inappropriate materials and processes accounts for a significant number of failed parts. Table 7, for example, itemizes the general causes of failure, with the frequency of occurrence determined from a survey (Ref 13). In this survey, materials selection is the most frequent cause of failure for engineered components. In the case of aircraft components, however, the survey did not identify any failures caused by improper materials. This difference illustrates the important point of how different design methods may influence the process of materials selection. In aerospace, for example, design methods should involve more critical evaluations of material alternatives, because the hazards of failure can be severe.

Table 7 Frequency of causes for failure

Cause	Percentage of failures	
	Engineering components	Aircraft components
Improper materials selection	38	…

Fabrication imperfections	15	17
Faulty heat treatment	15	…
Design errors	11	16
Unanticipated service conditions	8	10
Uncontrolled environmental conditions	6	…
Inadequate inspection/quality control	5	…
Material mix	2	…
Inadequate maintenance	…	44
Defective material	…	7
Unknown	…	6

Source: Ref 13

The selection of materials to prevent failure is typically a structured approach including thorough and diligent research into suitable materials. There is no universal guide that will automatically identify the best material for any service, because the number of interrelated input variables can be difficult to manage by a formal decision structure. Executive decision trees and computer expert systems have been developed to simplify materials selection, and these systems can identify candidate materials from very large databases of engineering materials, with cross-referenced mechanical and physical properties. However, these systems are narrowly applicable and are usually industry or company specific.

Effective materials selection is aided by access to materials property information and acquired engineering knowledge of all engineers participating on the design. No generalizations can be made that will be valid for all materials-selection problems, because each design problem must be considered individually or on the basis of closely related experience. Table 8 (Ref 14), however, provides some general guidance to the criteria that are typically significant in selecting materials in relation to possible failure mechanisms, types of stress, and operating temperatures.

Table 8 Guide to criteria generally useful for selection of material in relation to possible failure mechanisms, types of loading, types of stress, and intended operating temperatures

Failure mechanisms	**Types of loading**			**Types of stress**			**Operating temperatures**			**Criteria generally useful for selection of material**
	Static	**Repeated**	**Impact**	**Tension**	**Compression**	**Shear**	**Low**	**Room**	**High**	
Brittle fracture	X	X	X	X	…	…	X	X	…	Charpy V-notch transition temperature. Notch toughness. K_{Ic} toughness measurements
Ductile fracture[(a)]	X	…	…	X	…	X	…	X	X	Tensile strength. Shearing yield strength
High-cycle fatigue[(b)]	…	X	…	X	…	X	X	X	X	Fatigue strength for expected life, with typical stress raisers present
Low-cycle fatigue	…	X	…	X	…	X	X	X	X	Static ductility available and the peak cyclic plastic strain expected at stress raisers during prescribed life
Corrosion fatigue	…	X	…	X	…	X	…	X	X	Corrosion-fatigue strength

Failure mechanisms	Types of loading			Types of stress			Operating temperatures			Criteria generally useful for selection of material
	Static	Repeated	Impact	Tension	Compression	Shear	Low	Room	High	
										for the metal and contaminant and for similar time[c]
Buckling	X	…	X	…	X	…	X	X	X	Modulus of elasticity and compressive yield strength
Gross yielding[a]	X	…	…	X	X	X	X	X	X	Yield strength
Creep	X	…	…	X	X	X	…	…	X	Creep rate or sustained stress-rupture strength for the temperature and expected life[c]
Caustic or hydrogen embrittlement	X	…	…	X	…	…	…	X	X	Stability under simultaneous stress and hydrogen or other chemical environment[c]
Stress-corrosion cracking	X	…	…	X	…	X	…	X	X	Residual or imposed stress and corrosion resistance to the environment. K_{ISCC} measurements[c]

K_{Ic}, plane-strain fracture toughness; K_{ISCC}, threshold stress intensity to produce stress-corrosion cracking.
(a) Applies to ductile metals only.
(b) Millions of cycles.
(c) Items strongly dependent on elapsed time.
Source: Ref 14

Perhaps one of the most troublesome areas of materials selection relates to the change (or variation) in properties and performance. Property variations can occur within the part geometry from processing and fabrication, or changes in properties can occur over time from factors such as:

- Wear
- Temperature extremes or changes
- Corrosion
- Fatigue

These application conditions require a great deal of judgment in interpreting laboratory test data into design and extrapolating properties and performance over extended periods of time. Often, simulated service testing may be required. An important role of the materials engineer is to assist the designer in making meaningful connections between materials properties and the performance of the part or system being designed. For most mechanical systems, performance is limited not by a single property but by a combination of them. For example, the materials with the best thermal shock resistance are those with the largest values of $\sigma_f/E\alpha$, where σ_f is the failure stress, E is Young's modulus, and α is the thermal coefficient of expansion. These types of performance indices (i.e., groupings of material properties that, when maximized, maximize some aspect of performance) can be useful to compare materials.

Understanding the connection between properties and the failure modes is also important. Figure 8 is a chart of relationships between common failure modes and material properties (Ref 15). For most modes of failure, two or more material properties act to control the material behavior. However, it is also important to understand how property data should be interpreted. For example, even though most standard specifications require tensile-test data, these data are only partially indicative of mechanical performance in specific conditions. The purpose of tensile testing is often to monitor relative quality of different lots, not necessarily for design. Moreover, except in those conditions where ductile fracture or gross yielding may be the limiting condition for failure (Fig. 8), tensile strength and yield strength may be inadequate criteria for avoiding failure. A high tensile strength, for example, might be indicative of lower ductility and toughness, and thus a part with severe stress raisers might be prone to failure.

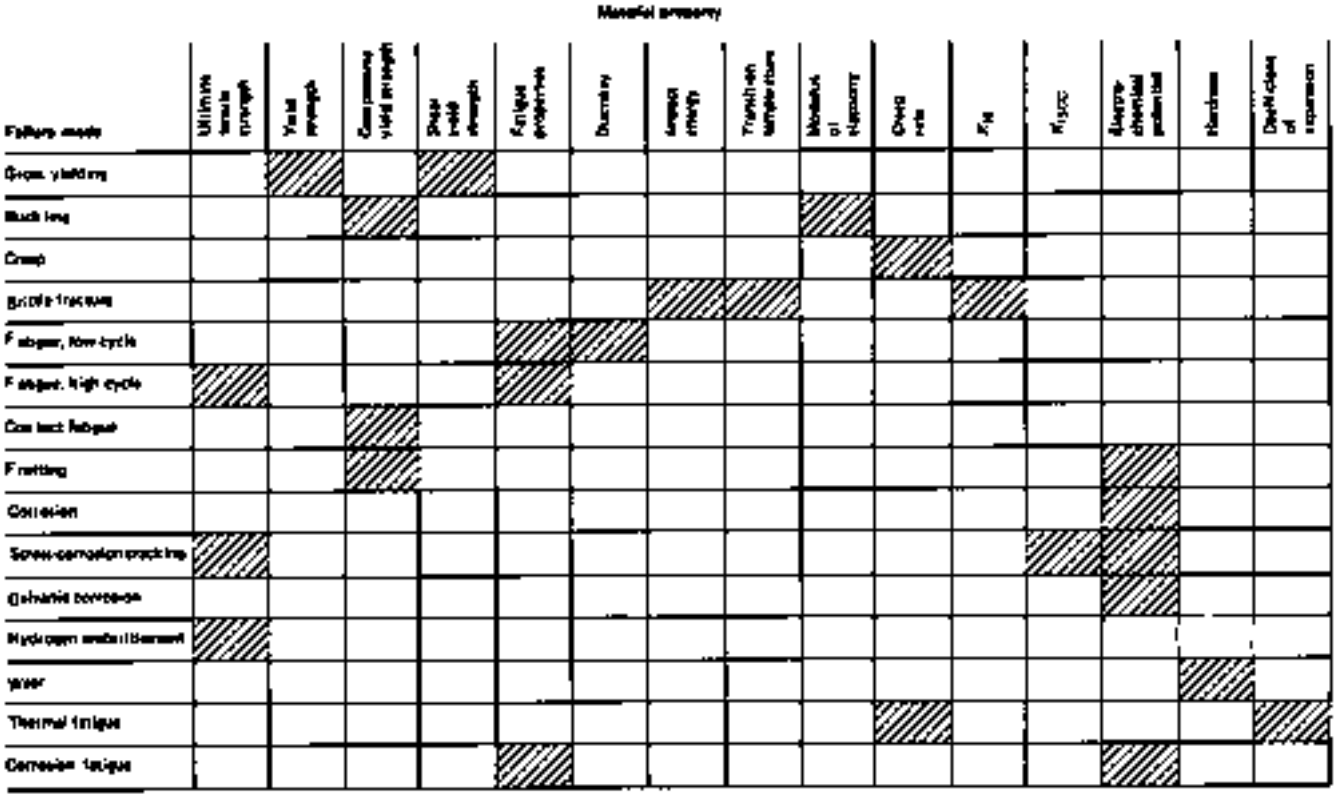


Fig. 8 General relationships between failure modes and material properties. Shaded blocks indicate properties that are influential in controlling a particular failure mode. K_{Ic}, plane-strain fracture toughness; K_{ISCC}, threshold stress intensity for stress-corrosion cracking. Source: Ref 15

References cited in this section

13. G.J. Davies, *Performance in Service, Essential Metallurgy for Engineers,* E.J. Bradbury, Ed., Van Nostrand Reinhold, London, 1985, p 126–155

14. Fundamental Sources of Failure, *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 4

15. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Properties and Selection: Stainless Steels, Tool Materials, and Special Purpose Metals,* Vol 3, *Metals Handbook,* 9th ed., American Society for Metals, 1980, p 825–834

Materials Selection and Failure Analysis

Case histories of failure investigations provide an indispensable tool not only for design but also in the education and training of engineers. Even historical case studies for obsolete materials and technology can still offer insight in identifying root causes and preventing failures in new designs. The investigative process can also identify secondary contributory causes so that they may be accounted for in a preemptive manner rather than by repetitive trial and error. In this way, case studies can be very important to the overall design process and materials selection.

Conversely, the materials selection process is of great importance to the failure analyst. A working knowledge of materials selection is a prerequisite for all engineers performing failure analysis. Throughout the failure analysis process, the investigator needs to consider the appropriateness of the selected material and processing. Inadvertent material substitutions and processing mistakes are often encountered. At other times, the physical evidence suggests that slightly unusual or wholly inappropriate materials and processes were employed. Scientific failure analysis may reveal that the original design was inadequate or had not considered all of the relevant service factors.

The composition and grade identification are typically ascertained during a failure investigation. The results are compared to the specified material or compared to standard grades in cases where no specifications were provided. Subtle deviation from the required composition is not necessarily the cause of a failure, as is sometimes erroneously assumed. The investigator should also endeavor to determine the likely processing methods used on the part or structure. This will be an aid, because certain manufacturing methods exhibit characteristic propensities for certain flaw and defect types. Processes such as heat treatment, welding, and machining can be addressed on a postmortem basis by evaluation of the strength and microstructure. Standard materials laboratory tests and scanning electron microscope fractography provide the remaining observations and data necessary to identify the failure mode and causative factors. The material and processing history thereby gleaned can be compared to the known service conditions to deduce the design concept. This reasoning can then be used to improve the design or address other materials or processing inadequacies.

Similar to design, failure analysis is somewhat influenced by the technical background and experience of the analyst. The basic weaknesses of single-discipline failure analysis can be analogous to design. Different engineering disciplines tend to approach failure analyses primarily within their area of specialization. A failure analysis team approach will likely provide the most beneficial corrective actions.

Some Questions the Failure Analyst Should Ask

A large variety of questions would be asked and answered during the course of a failure investigation. Education and experience will indicate to the analyst what questions should necessarily be addressed. The compound questions listed subsequently are among the many an analyst would address during an investigation:

- Should the complex part be an assembly of several parts rather than one?
- How was the component loaded, and was anisotropy considered?
- Is the material capable of being produced with the required properties, in the form used?
- Can any available material meet the specifications?
- Did the strength requirements preclude toughness or corrosion-resistance needs?
- Was the wear resistance adequate for the materials in contact?
- Were the desired properties compromised by the use of low-cost materials or processes?
- Did the materials and processing comply with the applicable codes and standards?
- Was the product made with unique materials and processes?
- Were proprietary or obsolete materials and processes employed?
- Were the manufacturing processes used to create the desired shape appropriate?
- Did the individual processing methods make sense?
- Should it have been preheated prior to heat treatment or welding?
- Did the fabricability requirements compromise the desired mechanical or physical properties?
- Were the manufacturing methods appropriate for the quantity produced?
- Were the operating conditions and maintenance as intended?
- Were the service conditions easy to anticipate?
- Does the material possess adequate durability in the service environment?
- How did the scrap value contribute to repair and maintenance decisions in service?

Examples of Improper Materials Selection

The consequences of improper materials selection can range from simply aggravating to catastrophic. The causes for failures due to materials and processing are many and varied, because the design process involves the balancing of part

function with manufacturing, cost, and service conditions. Failure analysis will typically indicate whether a material was suitable, marginally unsuitable, or drastically incompatible. The following examples describe some failures that suggest a questionable choice of material.

Example 1: Failure of a Steel Lifting Eye. A steel lifting eye that had fractured during service is shown in Fig. 9. No additional service-related information was provided. The eye was reportedly manufactured from a grade 1144 steel and should exhibit a minimum tensile strength of 689 MPa (100 ksi).

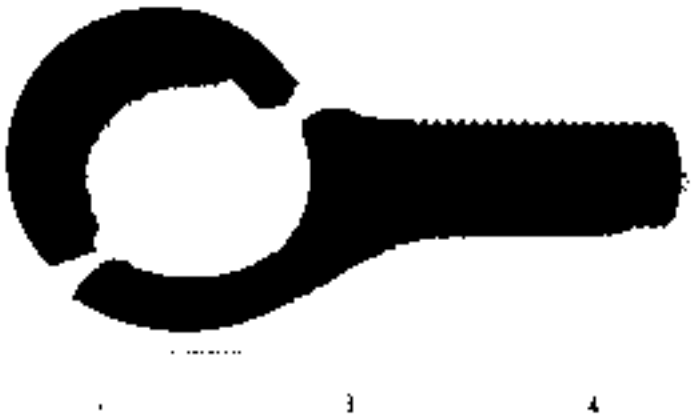


Fig. 9 Steel eye that had fractured in two locations during service

The eye was approximately 70 mm (2.75 in.) long and appeared to be machined. Fracture occurred in two locations: adjacent to the threaded shank and diametrically opposite to this region. The circular eye was deformed longitudinally, and the fracture surfaces exhibited an angular orientation.

Chemical analysis confirmed that the eye was similar to a resulfurized and rephosphorized grade 1144 steel. The sulfur content was slightly below the normal limits, and the phosphorus content was slightly above the typical range.

Scanning electron microscope (SEM) examination of the fracture surfaces revealed woody overload features, typical for resulfurized steels. The morphology was identified as a mixed fracture mode of cleavage and ductile rupture, and the directionality of the features was suggestive of shear overload. A typical region of the fracture nearest the shank, which was the likely origin, is shown in Fig. 10. Fracture preferentially followed the nonmetallic inclusions. Tensile testing could not be performed on the eye, but the hardness was found to be 32 HRC. This is roughly equivalent to 1,000 MPa (145 ksi), which exceeded the drawing specification.

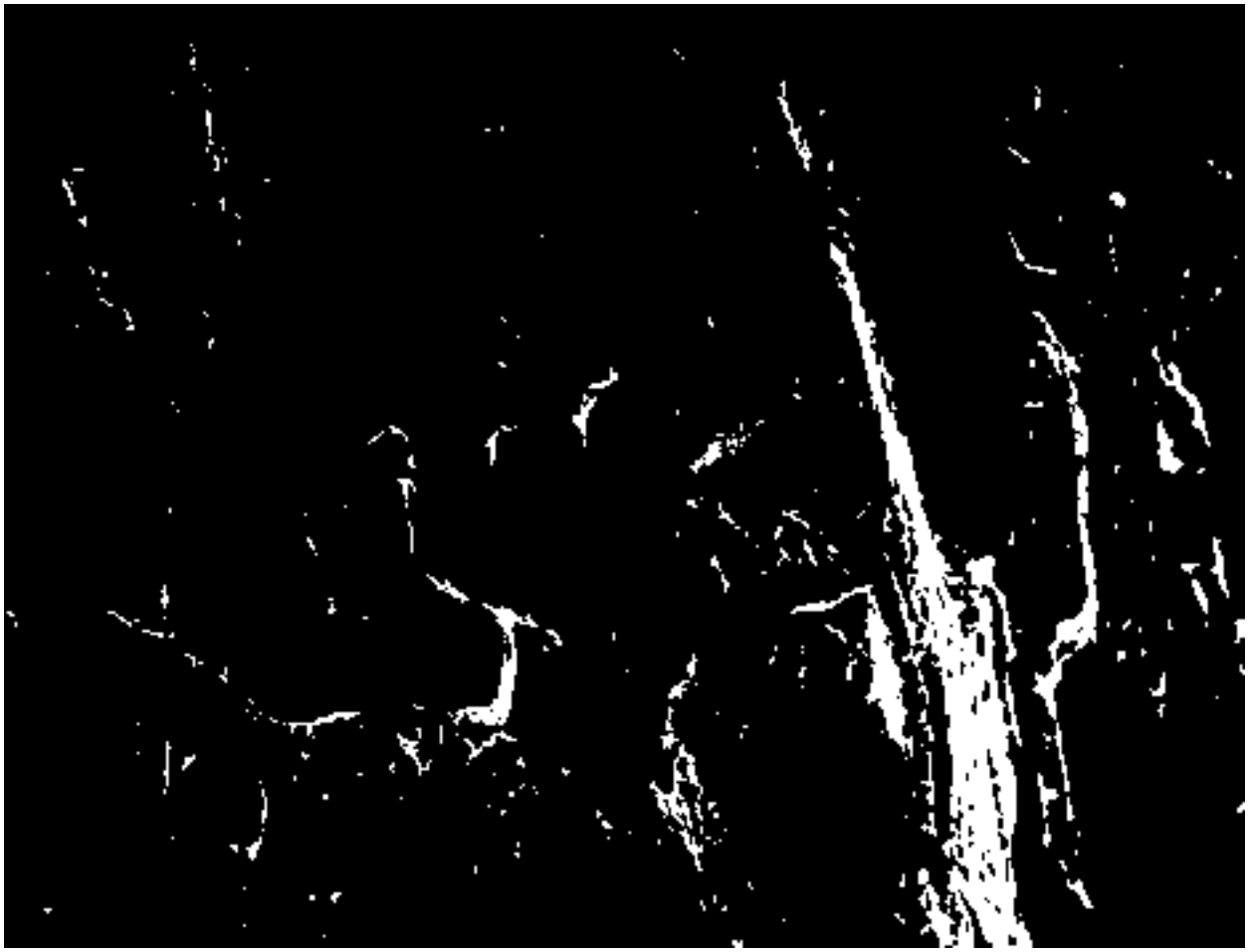


Fig. 10 Scanning electron microscope micrograph of typical eye fracture morphology consisting of woody, ductile features. 500×

Metallographic examination was performed through the fracture surfaces, and the fracture profile of the fracture surface near the shank is shown in Fig. 11 The fracture was parallel to the direction of the manganese sulfide stringer inclusions. Etching revealed the presence of significant banding of the ferrite and pearlite microstructure. The fracture is primarily along the inclusions and through bands of ferrite, as shown in Fig. 12

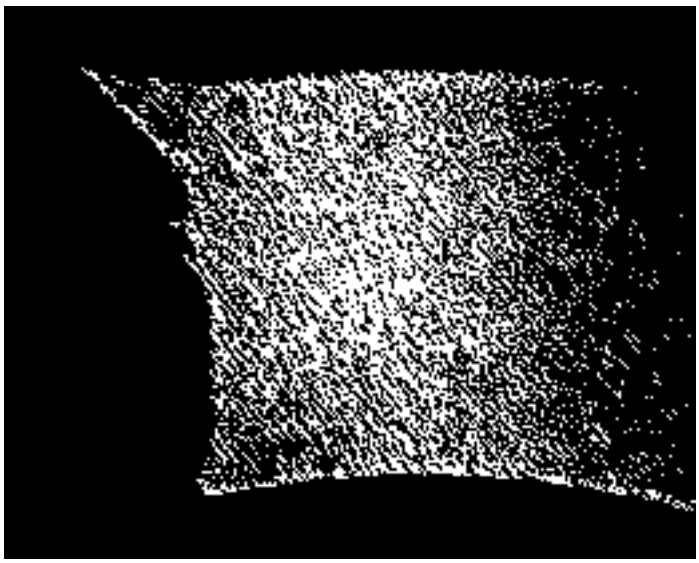


Fig. 11 Cross section through the eye showing cracking through the aligned stringer inclusions. Unetched. 2×

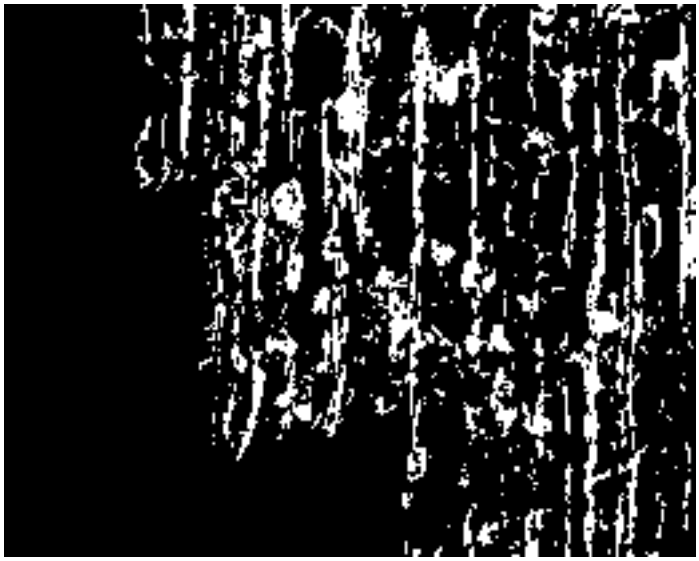


Fig. 12 High-magnification view of the eye fracture surface showing fracture through the sulfide inclusions and a banded microstructure. 2% nital etchant. 28×

It was concluded that the lifting eye failed as a result of overload. Fracture occurred parallel to the rolling direction, through manganese sulfide stringers and ferrite bands in the base metal matrix. The eye was machined from grade 1144 steel that was heavily cold rolled for strength. This material is very anisotropic, exhibiting substantially poorer long and short transverse mechanical properties than the longitudinal properties, which were likely used for design. It is likely that the materials selection process did not properly account for this anisotropy. The selection of a rolled product may also be questionable here. It may be better to use a forged product in this case because of resulting “grain flow” and inclusion orientation.

Example 2: Failure of a Tank Coupling. Leakage was identified around a coupling welded into a stainless steel holding tank. The tank had been in service for several years, storing condensate water with low impurity content. The tank and fitting were manufactured from type 304 stainless steel. The fitting was fillet welded to the tank wall, and the tank was covered with insulation in service.

A diagram of the failed tank section is shown in Fig. 13. The coupling joint consisted of an internal groove weld and an external fillet weld. Cracking was apparent on the tank surface, adjacent to the coupling weld. Some reddish rust was present on the surface, but no gross mechanical damage, yielding, or weld defects were evident.

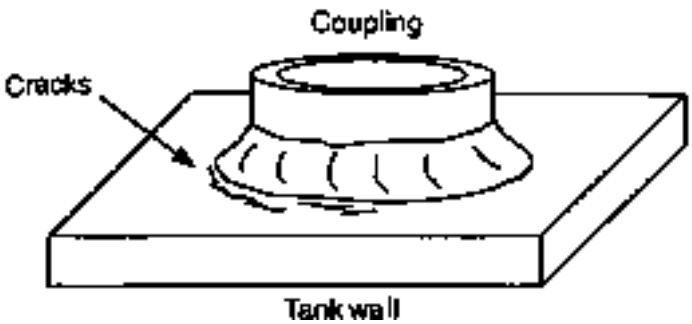


Fig. 13 Diagram of a tank coupling region that leaked during service

The chemical composition of the plate was consistent with a type 304 austenitic stainless steel. No compositional anomalies were detected. Energy-dispersive x-ray spectrometric analysis of the corrosion product on the crack surfaces revealed chlorine, carbon, and oxygen in addition to the base metal elements. The amount of corrosion present at the primary crack prevented SEM examination for morphological identification. Hardness testing of the plate revealed a hardness level considered typical for annealed stainless steel plate.

A metallographic cross section through the most severe cracking is shown in Fig. 14. A great number of secondary, branching cracks are evident in the weld, heat-affected zone (HAZ), and base metal. A typical crack is shown at higher magnification in Fig. 15. Branching, transgranular cracking is evident, emanating primarily from the exterior of the tank. Examination of the HAZ microstructure did not reveal evidence of substantial sensitization.

Fig. 14 Metallographic cross section through the cracked region of the coupling, showing branching cracks from the exterior (top). 10% oxalic acid, electrolytic etch. 1.75×

Fig. 15 High-magnification view at the exterior tank surface showing branching, transgranular stress-corrosion cracking. 10% oxalic acid, electrolytic etch. 14×

The analytical investigation concluded that the tank failed as a result of stress-corrosion cracking (SCC) that initiated at the exterior surface. Contaminant material containing chlorine, which may have leached from the insulation, combined with the inherent susceptibility of the base material and residual stresses from fabrication and welding. These factors combined synergistically to result in cracking. Aqueous chlorides, especially within an acidic environment, have been shown to cause SCC in austenitic stainless steels under tensile stress. The use of a type 304 stainless steel to prevent internal corrosion damage did not adequately foresee the potential for corrosion damage from external contamination.

Example 3: Localized Corrosion of Type 303 Stainless Steel Exposed to Acidic Soft Drinks (Ref 16). This example from *Metals Handbook*, 8th edition, (Ref 16) illustrates how a secondary selection factor (machinability) was not adequately evaluated in terms of a particular service environment and function. In this case, the failure is related to the selection of type 303 stainless steels for ease of machining instead of type 304 stainless steel.

After about two years in service, a valve in contact with soft drink in a vending machine occasionally dispensed discolored drink with a sulfide odor. The soft drink in question was one of the more strongly acidic, containing citric and phosphoric acids with a pH of 2.4 to 2.5, according to the laboratory at the bottling plant.

Manufacturing specifications for the valve called for type 303 stainless steel, a free-machining grade chosen because of the substantial amount of machining required for the part. Other parts in contact with the drink were made from type 304 stainless or inert plastics. In this application, type 303 stainless steel had only marginal corrosion resistance because of the size and distribution of sulfide stringers in some lots. When the machine was unused overnight or over the weekend, there

was occasionally enough attack on exposed sulfide stringers to make the adjacent liquid unpalatable. Specification of type 304, which is suitable for this application, was thus recommended.

Reference cited in this section

16. *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., 1975, p 179

Materials Selection for Failure Prevention

Brett A. Miller, Stork Technimet, Inc.

References

1. H.W. Stoll, *Product Design Methods and Practices,* Marcel Dekker, 1999, p 40, 148

2. J.R. Dixon and C. Poli, *Engineering Design and Design for Manufacturing,* Field Stone Publishers, 1995

3. C.O. Smith, Human Factors in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

4. J.R. Dixon, Conceptual and Configuration Design of Parts, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 33–38

5. G. Boothroyd and P. Dewhurst, *Product Design for Assembly,* Boothroyd Dewhurst, Inc., 1989

6. S. Pugh, *Total Design: Integrating Methods for Successful Product Engineering,* Addison-Wesley, 1991

7. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997

8. R.W. Heckel, Introduction to the Effects of Composition, Processing, and Structure on Materials Properties, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 333

9. G. Dieter, Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 243–254

10. H. Boyer, Introduction: Selection Criteria, *Selection of Materials for Component Design,* ASM International, 1986

11. J.A. Schey, *Introduction to Manufacturing Processes,* 2nd ed., McGraw-Hill, 1987

12. R. Gunia, Ed., *Source Book on Materials Selection,* Vol 2, American Society for Metals, 1977

13. G.J. Davies, *Performance in Service, Essential Metallurgy for Engineers,* E.J. Bradbury, Ed., Van Nostrand Reinhold, London, 1985, p 126–155

14. Fundamental Sources of Failure, *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 4

15. C.O. Smith and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Properties and Selection: Stainless Steels, Tool Materials, and Special Purpose Metals,* Vol 3, *Metals Handbook,* 9th ed., American Society for Metals, 1980, p 825–834

16. *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., 1975, p 179

Brett A. Miller, Stork Technimet, Inc.

Selected References

- D. Askeland, *The Science and Engineering of Materials,* 2nd ed., PWS-KENT Publishing Company, Boston, 1989
- ASM Committee on Material Requirements for Service Conditions, Material Requirements for Service Conditions, *Welding, Brazing, and Soldering,* Vol 6, *ASM Handbook,* ASM International, 1993, p 373–375
- B.P. Bardes and L.J. Korb, Guidelines for Selection of Material, *Properties and Selection: Stainless Steels, Tool Materials, Special Purpose Metals* Vol 3, *ASM Handbook* 9th ed., American Society for Metals, 1980
- J.A. Charles and F.A.A. Crane, *Selection and Use of Engineering Materials,* 2nd ed., Butterworths, London, 1989
- *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986
- M.M. Farag, *Materials and Process Selection in Engineering,* Applied Science Publishers, Ltd., London, 1979
- R. Gunia, Ed., *Source Book on Materials Selection,* Vol 1, American Society for Metals, 1977
- R. Gunia, Ed., *Source Book on Materials Selection,* Vol 2, American Society for Metals, 1977
- R. Kern and M. Suess, *Steel Selection—A Guide for Improving Performance and Profits,* John Wiley and Sons, New York, 1979
- C. MacDermott and A. Shenoy, *Selecting Thermoplastics for Engineering Applications,* 2nd ed., Marcel Dekker, Inc., New York, 1997
- *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997
- F.K. Naumann, *Failure Analysis, Case Histories and Methodology,* Dr. Riederer-Verlag GmbH, Stuttgart, Germany, 1983
- J.B. Park and R.S. Lakes, *Biomaterials: An Introduction,* 2nd ed., Plenum Press, New York, 1992
- H. Petroski, *To Engineer is Human,* Vintage Books/Random House, Inc., New York, 1985
- C.O. Smith, and B.E. Boardman, Concepts and Criteria in Materials Engineering, *Properties and Selection: Stainless Steels, Tool Materials, Special Purpose Metals,* Vol 3, *Metals Handbook,* 9th ed., American Society for Metals, 1980
- G. Vander Voort, Use of Failure Analysis in Materials Selection, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 322–328

Design Review for Failure Analysis and Prevention

Crispin Hales and Cheryl Pattin, Triodyne Inc.

Introduction

THE AIM OF THIS ARTICLE is to assist the failure analyst in broadening the initial scope of the investigation of a physical engineering failure in order to identify the root cause of the problem. Analysis methods for determining the loading pattern and other factors responsible for a service failure are described elsewhere in this Volume. However, for effective measures to be taken in preventing future similar failures, it is essential to identify the root cause of the problem, as distinct from the immediate cause of the physical failure itself. The design process that led up to the physical failure is of great importance when trying to get back to the root cause because it is during the design process that all the original contributing factors are brought together to create the components that are assembled into the complete system. The intention here is simply to set the physical failure within a broader context and then to focus on assessing the design process evidence available within that context. The article purposely does not address issues such as manufacturing, operation, and maintenance, except as they relate to the design process, because these other important aspects are covered in detail within other sections of the current Volume. Neither is it intended to provide prescriptive guidance on carrying out the design process, which is covered in detail within *Materials Selection and Design*, Volume 20 of the *ASM Handbook*. Indeed, if the guidelines and systematic design methods described in Volume 20 are appropriately followed, then the likelihood of a failure due to faulty design is minimized in the first place.

What Is an Engineering Failure?

When a piece of material breaks, cracks, corrodes, or otherwise “fails” in service, it is only natural to look at the “failure” with an initial assumption that it should not have happened. While the assumption may be valid in some cases, in many others it is misleading. The “failure” is simply the physical result of a set of preexisting circumstances, a sequence of events, or a developing situation, and it must be considered in its appropriate context.

It is most important that the initial investigation is approached from a broad perspective, rather than from a specialist viewpoint. Many times the perceived problem is circumscribed too early, and the entire investigation becomes focused on one particular aspect, rather than establishing the root cause. The result is that biased conclusions become accepted, and any actions taken for future improvements are not soundly based. To approach the investigation of an engineering failure from a broad perspective means analyzing the design process that created the product, equipment, or system, as well as the physical failure itself. By using a systematic approach to the analysis, it is possible to review basic design issues and work toward the details in a progressive fashion. This helps to ensure that key points are not missed.

For example, a large number of electric appliance motors failed by seizure when elastomeric components, which had been added to reduce noise and vibration, rapidly disintegrated. A large claim was filed against each of the companies in the component supply chain for providing parts manufactured from substandard material. As it happened, some of the elastomeric material supplied to the component manufacturer was “an equivalent” to the material actually ordered and lacked the oil-resisting capabilities of the material actually ordered by the component manufacturer. To the motor manufacturer the cause of the failure appeared clear: Substandard material had been made into components that disintegrated in service, causing premature motor failures. However, an engineering investigation of the matter revealed quite a different cause of failure. Despite the fact that the company was International Organization for Standardization (ISO) 9001 registered, the additional components had been purchased and fitted to the motors at the sole bidding of an employee engineer whose focus was on reducing vibration and noise in the product and who had seen a similar solution applied by another company. The problem in the motors was never actually defined, no alternative concepts were considered, and the elastomeric material for the additional components was improperly specified on both drawings and purchase orders. Instead of listing an appropriate set of elastomeric material requirements, the

documents simply stated a specific manufacturer's product number. It might have been expected that a design review would have caught these deficiencies before the additional components were made and fitted to the motors on the production line, but there was no such design review. This is because the ISO 9001 requirements for design reviews are essentially customer driven. They are intended to ensure that the customer gets what the customer wants. They do not address the situation in which an employee simply decides that the product could be improved by the adding of components and personally implements the change from within the company. The root cause of the failure in this example was a flawed design process with three major deficiencies. The inevitable failure situation was then exacerbated by a material supplier providing “an equivalent” material without the required notification or approval.

The important point here is that for several years afterward the forensic investigation by each of the companies involved was focused solely on the failure of the specific material in service, and an enormous amount of effort was spent in taking depositions, amassing documents, and life testing motors to try and prove one point or another. However, once the root cause of the failure had been identified, the disputed issues changed and the matter was soon resolved. This is typical of a failure investigation in which the initial focus on the material failure itself leads away from the root cause and masks the true issues. Without addressing the design process issues, the quality of future products would still have been at risk.

The Context of an Engineering Failure

Before windowing in on the design process itself, it is helpful to try to visualize how a particular engineering failure fits into the wider context under the circumstances. This is not easy to do because most likely the failure results from a complex sequence of events taking place at different levels of resolution and seen from different perspectives. Nevertheless, ways of mapping such a context diagrammatically have been developed over the years to help in carrying out design projects, such as shown in Fig. 1. Design processes within this context are discussed in “Overview of The Design Process” by Dixon (Ref 2), and “Conceptual and Configuration Design of Products and Assemblies” by Otto and Wood (Ref 3) in Volume 20 of the *ASM Handbook*. The idea in this article is to set the engineering failure in context so as to help in retracing a path both backward in time and outward in perspective. This is necessary in order to identify possible contributing factors to the engineering failure, which at first may not seem related to the physical failure at all.

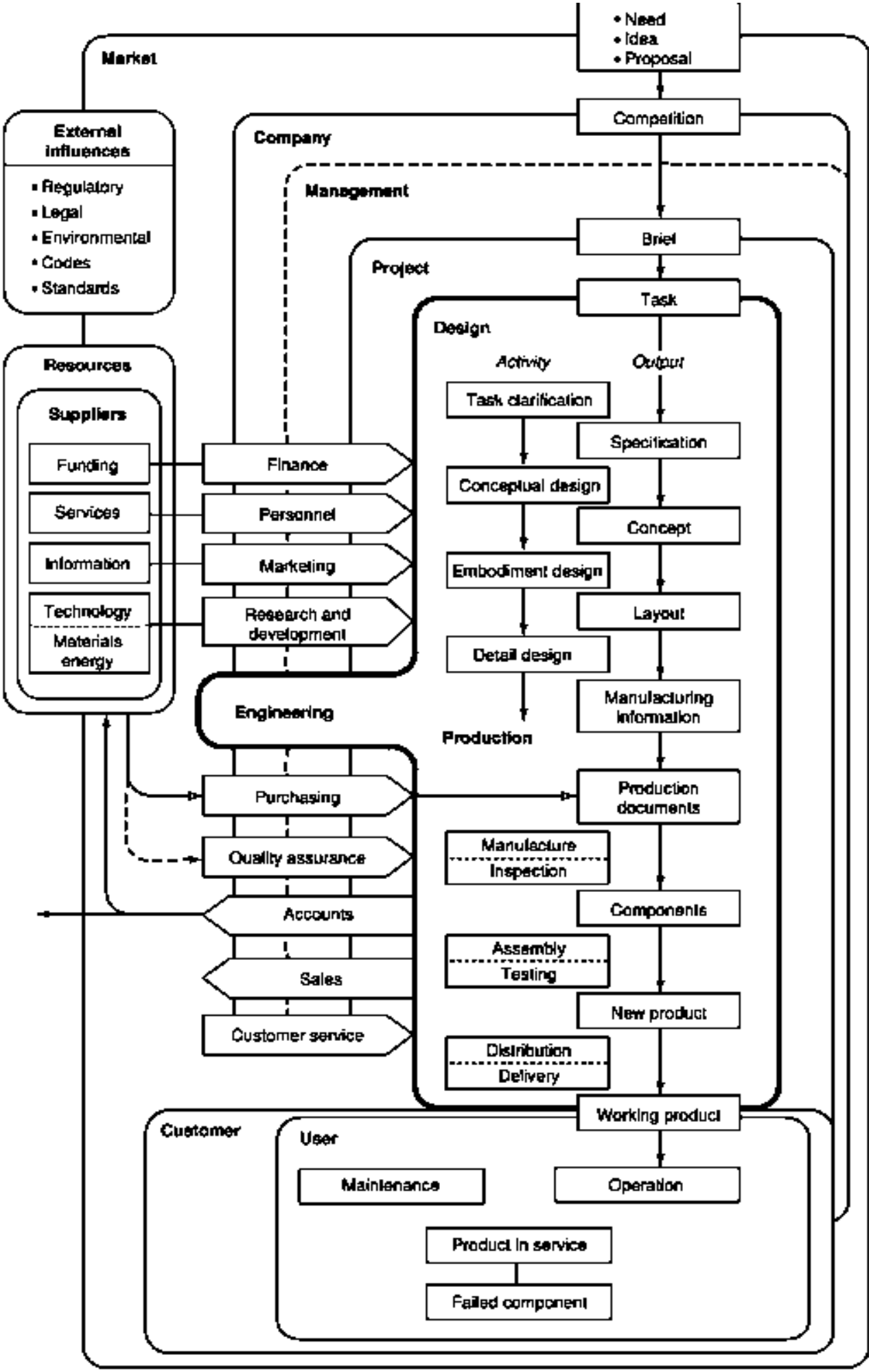

• Need
• Idea
• Proposal
Market
Competition
External influences
• Regulatory
• Legal
• Environmental
• Codes
• Standards
Company
Management
Brief
Project
Task
Design
Activity
Output
Resources
Suppliers
Funding
Services
Information
Technology
Materials energy
Finance
Personnel
Marketing
Research and development
Task clarification
Specification
Conceptual design
Concept
Embodiment design
Layout
Detail design
Manufacturing information
Engineering
Production
Purchasing
Production documents
Quality assurance
Manufacture
Inspection
Accounts
Components
Assembly
Testing
Sales
New product
Customer service
Distribution
Delivery
Working product
Customer
User
Maintenance
Operation
Product in service
Failed component

Fig. 1 Engineering failure set in context. Source: Ref 1

The diagram or map in Fig. 1 shows a component as part of a product (or system or structure) being operated by a user and having been purchased as a working product (or system or structure) by a customer through a distribution and shipping process. Tracking backward up the map reveals that before the sale of the product comes the manufacturing process, which may well include the purchase of component parts or semi-finished products from other suppliers. Prior to that there is some kind of design process, which provides the specifications, instructions, and information for manufacture; and initiating the design process in the first place is a commercial process arising from a need, idea, or market force. All of this is part of what may be considered an economic loop. New developments and products are generated by companies tracking around the full loop or by short-circuiting it and modifying existing products without going through the full design process. Also represented on the map is the notion of *levels of resolution*, shown more specifically in Fig. 2. In this simplified diagram, the design and manufacturing processes, integrated together, can be visualized as forming just one of a number of processes (such as finance or marketing) taking place within a project, within a management system, within a company, within a market, within an external environment. At each level of resolution there are influences that impinge on the levels below and thereby affect the outcome of the project and its resulting product (or system or structure). The remainder of this article is concerned only with the design process at the personal and project levels, but takes into consideration the effects of some higher level influences and interfaces that are often found to contribute to engineering failures.

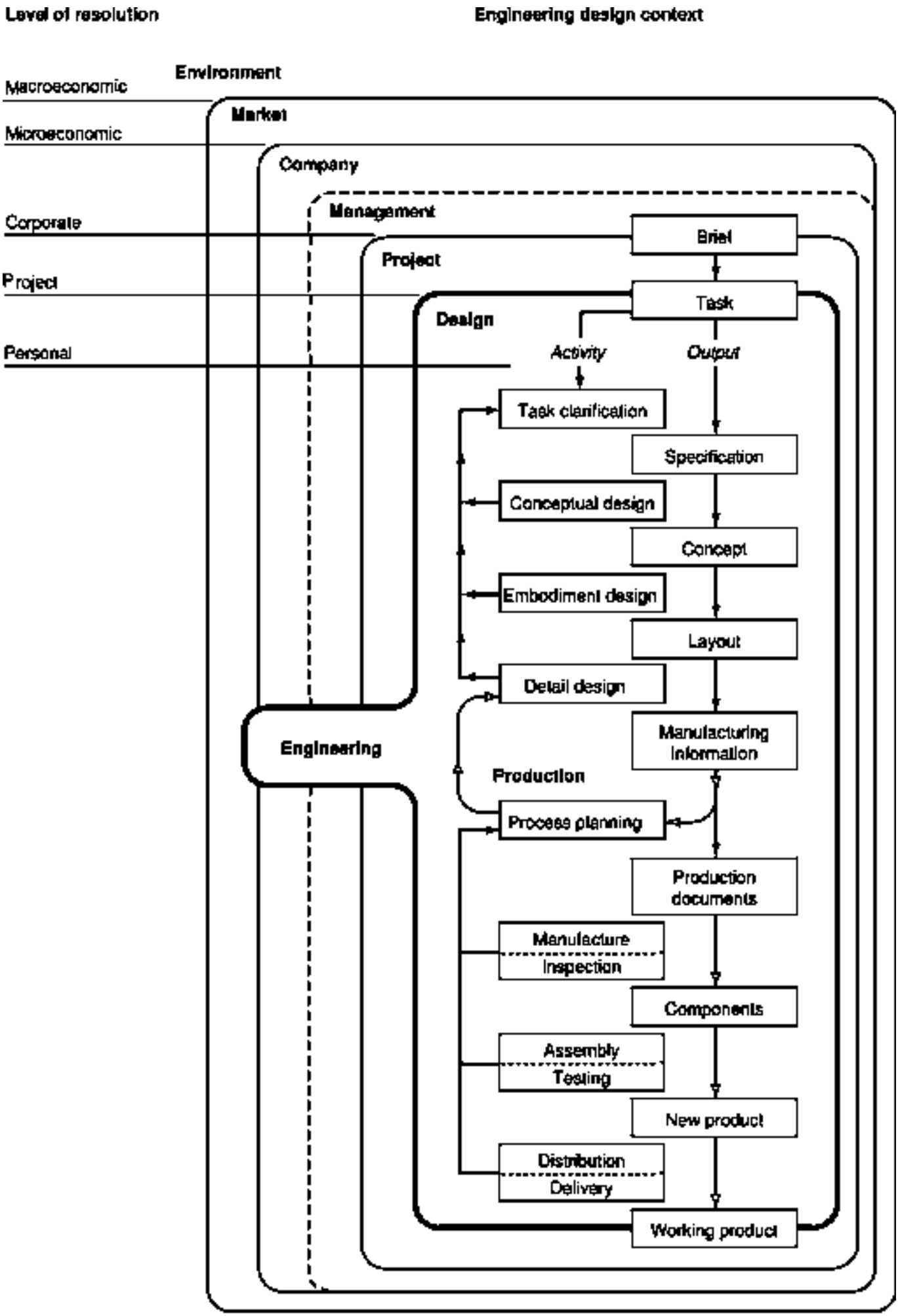

Level of resolution
Engineering design context
Macroeconomic
Environment
Microeconomic
Market
Company
Corporate
Management
Project
Project
Personal
Design
Brief
Task
Activity
Output
Task clarification
Specification
Conceptual design
Concept
Embodiment design
Layout
Detail design
Manufacturing information
Engineering
Production
Process planning
Production documents
Manufacture
Inspection
Components
Assembly
Testing
New product
Distribution
Delivery
Working product

Fig. 2 Levels of resolution related to the engineering design process. Source: Ref 1

References cited in this section

1. C. Hales, *Managing Engineering Design*, Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

2. J.R. Dixon, Overview of the Design Process, *Materials Selection and Design*, Vol 20, *ASM Handbook*, ASM International, 1997, p 7–14

3. K.N. Otto and K.L. Wood, Conceptual and Configuration Design of Products and Assemblies, *Materials Selection and Design*, Vol 20, *ASM Handbook*, ASM International, 1997, p 15–32

The Engineering Design Process

In theory, the basic engineering design process is usually described as a sequence of phases, such as shown in Fig. 2, beginning with a perceived need and finishing with the detailed description of a particular technical system or product (Ref 1, 4, 5, 6, 7). As represented in Fig. 2, each phase involves activities intended to result in particular outcomes. Depending on the product, system, or structure being designed, the phases may be labeled in different ways and will often be carried out in parallel with the design of the manufacturing process (Ref 8). Each phase may be considered as a sub-design process in itself, consisting of an iterative set of steps and commonly summarized as follows:

Step	**Purpose**	**Output**
Task clarification	Definition of the problem	Design specification
Conceptual design	Generation, selection, and evaluation of solutions	Design concept
Embodiment design	Development of the concept	Final layout
Detail design	Definition of every component in shape and form	Manufacturing information

In practice it is unlikely for the design process to progress through the sequential set of project phases exactly as outlined, but the sequencing is less important than the existence, nature, and effectiveness of the actual design activities implied within each phase. For example, there must be some kind of design specification as a starting point, and there must be some kind of concept from which a final design evolved. The concept must be developed to a greater or lesser degree so as to result in a practicable overall design, and the details of every component must be defined to the point at which the product or system can be manufactured. A combination of human activities is required to reach each of these end points or outputs, as described by Smith in his article "Cross-Functional Design Teams" (Ref 9) in Volume 20 of the *ASM Handbook*. The activities of the design team, the influences on the design team, and the consequent output from the design team are the focus when analyzing the design process (Ref 1).

In order to analyze the design process factors that may have contributed to an engineering failure, it is necessary to review the phases of the design process in the context of the original project history, its management, prevailing commercial pressures, and external influencing factors. Figure 3 shows a general sequence for such a review. By starting with the failed component and working backward through its history, collecting evidence on how it was designed, developed, and produced, it is often possible to identify and characterize weaknesses that became contributing factors to the failure.

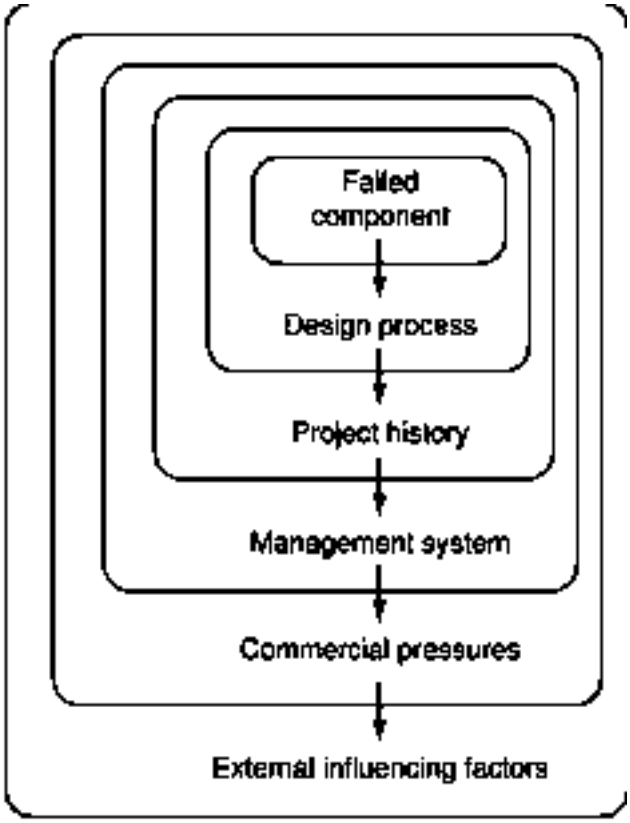


Fig. 3 General sequence of failure investigation from design point of view

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

5. E. Frankenberger, P. Badke-Schaub, and H. Birkhofer, Ed., *Designers—The Key to Successful Product Development,* Springer-Verlag, 1998

6. V. Hubka, *Principles of Engineering Design,* Butterworth Scientific, 1982

7. K.M. Wallace, A Systematic Approach to Engineering Design, *Design Management: A Handbook of Issues and Methods,* M. Oakley, Ed., Basil Blackwell Ltd., 1990

8. M.M. Andreasen and L. Hein, *Integrated Product Development,* IFS (Publications) Ltd. and Springer-Verlag, 1987

9. P.G. Smith, Cross-Functional Design Teams, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 49–53

Preliminary Investigation

When the components in the system that either failed or were associated with the failure are identified, the first step is to request and gather all the available evidence pertaining to the failure. For example, the failed parts themselves, drawings, supply documents, change orders, reports, photographs, videotapes, sales literature, technical information, statements, interview transcripts, deposition transcripts, installation, operating and maintenance manuals, relevant standards and codes—all are potential sources of information that need to be explored. A detailed review of all this material helps to put the failure in perspective and enables a preliminary timeline or outline sequence of events to be compiled.

Analysis of the Engineering Design Process

It is possible to arrive at a design in many different ways, ranging from informal experimentation through to a highly organized formal procedure (Ref 1). When it comes to analyzing any particular case, however, it is essential to have a clear and structured approach to mapping whatever evidence is available, no matter what type of design process is used. The more it can be broken down into measurable components, the less subjective the conclusions will be and the better the chance of developing defensible opinions is. For the purposes of this article, the schematic shown in Fig. 2 represents the basic design and manufacturing process adequately, and the following sections are structured in sequence according to this diagram.

Reference cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

Task Clarification—Defining the Problem

In order for a design project to be carried out, two things need to be established at the outset, each being a complementary result of the "task clarification" phase of the design process:

- A clear statement of the problem to be solved, for which solutions will be sought
- A set of requirements and constraints against which to evaluate the proposed solutions

The first is termed a *problem statement,* or *definition of the problem,* while the second is termed a *specification,* a *target specification,* or more correctly a *design specification.* Both of these are essential if a solution to the problem that satisfies all parties is to be found. Considerable effort (and possibly some preliminary design work) may be needed to help establish what the real problem is, but it must be done. Finding solutions to the wrong problem is unacceptable design practice. Once the problem is defined, the criteria for selecting an appropriate concept must be established in the form of a design specification that lists all the requirements to be met by any solution to the problem. Here again, if the requirements are inaccurate or incomplete, then the design process is flawed from the start.

Frequently it is found that the real design problem is never clearly defined, is incorrectly defined, or the wrong problem is identified, as described in the introductory example. This can be investigated by reference to the original bid documents or equipment quotations and the sequence of communications between supplier and customer that led up to them. The wording of all this needs to be reviewed in detail and with a good understanding of the equipment, as it is easy to miss specification deficiencies that had consequential ill effects on the final design.

During the review it is useful to ask a series of questions concerning the design task, the team that carried out the job, and the details of the team's activities during each phase of the design process. For example:

- Was the overall design problem understood and clearly defined prior to conceptual design work?
- Was a realistic project plan prepared, acceptable to all parties?
- Was a comprehensive design specification compiled?
- Is the design specification independent of solutions, or does it include fictitious constraints?
- Was the design specification circulated to all those involved for comment and approval?
- Was the design specification formally approved before work proceeded on conceptual design?
- Where is the design specification?
- Who developed it?
- Who approved it?
- What changes were made? Who made them? Why? When? How?

Progressing through a checklist of questions for each phase of the design process allows problems and weaknesses to start to emerge. A properly structured design specification provides the maximum design freedom within the given constraints. The degree to which this is carried out may be checked by reference to

standard specifications for the particular equipment involved and checklists such as the general ones offered by Pahl & Beitz (Ref 4) and Hales (Ref 1).

Defining the design problem and listing the requirements in the form of a design specification provides a proper foundation for the project to proceed through the conceptual, embodiment, and detail design phases. Solutions to the defined problem may then be sought by the use of conceptual design methods, and the resulting concepts may be evaluated against the design specification.

The following example illustrates the pitfall of introducing fictitious constraints into the requirements. The result is that the concept selected will be over-constrained and therefore not an optimal solution to the problem.

A large number of custom-designed vertical lift conveyors were required for use in a series of new automated U.S. mail-sorting facilities. These facilities comprise essentially a series of code-reading sorting units at ground level with a large array of horizontal roller conveyors above. Vertical lift conveyors are used to transport plastic trays of sorted mail from ground level up to the horizontal conveyor level, and the reverse. The vertical lifting concept was developed to overcome the slippage problem encountered with inclined roller conveyors when plastic trays superseded the cardboard trays used in earlier facilities.

The general contractor had prepared a voluminous design specification for the post office. This was to apply to all facilities, but for each particular facility there was also a detailed set of special requirements. Within these documents were embedded the requirements for the conveying systems in general, and within those the requirements for the vertical lift units in particular. Prototype vertical lift units that used twin in-running belts to deliver the trays to three-fingered lift platforms already had been developed and tested off-line to the point of acceptable performance for this application. The post office design specification for the lifting units was compiled with the prototype units in mind, even to the point of requiring in-running entry and out-running exit conveyors having two or more belts, and lift platforms with three or more fingers. In such a way the specification unnecessarily constrained the design to a specific concept that was known to have inherent operational problems. The many required units were designed, built, tested, delivered, and installed at a cost of more than $500,000. While some design weaknesses and manufacturing problems were evident, which detracted from the performance of the units, they did pass the acceptance tests laid down in the design specifications, right up to the final production trials. At this point, however, multiple component failures and jams were encountered because the operators were using the vertical lift units on a start-stop basis instead of continuously as intended. It was claimed that the units did not meet the specification, and they were all removed from service and scrapped without payment to the manufacturer. At the same time, another supplier was contracted to provide quite different replacement units. Investigation revealed that a modified design specification had been issued to alternative manufacturers who were invited to bid on supplying the replacement units, and that the units finally selected were based on a concept that could never have met the requirements of the original design specification. In the modified design specification, the requirements for belt conveyors and for the lift platforms to have three or more fingers had been deleted. This removed the fictitious constraints on the design and allowed the use of an articulated slat conveyor, far superior in concept to the suspended tray type of conveyor for this particular application.

The root cause of this failure was a deficient design specification, not the component failures themselves. The result was an enormous waste of effort, money, and materials, as well as the bankruptcy of the original vertical lift unit manufacturer.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

Conceptual Design

Once the problem is defined, it is possible to start generating ideas leading to concepts that will solve the problem (Ref 10, 11). Conceptual design is also discussed in the article “Creative Concept Development” (Ref 12) in Volume 20 of the *ASM Handbook.* This is not the same as invention. What is meant by *conceptual design* is the conscious activity of generating numerous ideas leading to specific concepts that are then selected and

evaluated according to how well they meet the requirements of the design specification. Of course a design concept may become patented as an invention in intellectual property terms, but an invention is not necessarily a design. There is a fundamental difference between an inventor happening to have an inspiration on Thursday and a design engineer producing an acceptable design concept within time, budget, and specification constraints by Thursday. An important part of the necessary concept evaluation activity is searching for weak spots. For example, if an otherwise excellent concept has an inherent reliability problem, it may have to be rejected in favor of perhaps a less innovative concept, but one that is inherently more reliable.

Many times it will be revealed during the investigation that the concept for a design has either simply been assumed, or else has been adapted from a different product, without a genuine search for alternatives or formal assessment against the design specification. Although on the surface it may appear to perform adequately, a non-optimal concept creates secondary problems that then must be addressed by further concepts, which themselves may lead to an unexpected or premature failure. When analyzing the design process, it is necessary to isolate and visualize the principal concept within the context of the project, and to assess how well it meets the requirements of the design specification. Again, the checklist approach may be used for this (Ref 1), but often a simple set of questions will suffice, such as:

- What alternative concepts were produced?
- How were the various alternatives assessed?
- Were all alternatives considered on an equal basis?
- Who decided what concept should be used and how was the decision made?
- Were the weak points of the chosen concept adequately evaluated?
- Why were alternative concepts rejected?

In addition to the example presented in the introduction, the following example illustrates the kind of problem that can occur from the selection of a wrong concept for the application:

A post office in Michigan was fitted with heavy doors that could swing more than 90° from their closed position in either direction. The doors had alloy frames with large glass panels and were fitted with mechanical door closers that always returned the door to its closed position with a “damped spring” action after use. In order to save maintenance and inventory costs, the original architect had specified the same doors throughout the building. Although physically the door could be used as an exterior door, a number of features such as the lack of positive sealing indicated that it was primarily a door for internal use. However, there was no requirement to this effect, and the interchangeability of the doors meant that if, for example, the glass in an external door was ever broken, one of the internal doors could immediately be moved to replace it so that the building security would not be compromised. The external door closest to the public parking area was the most heavily used door in the building. It was exposed to all weathers and the automatic device for damping the door movement and closing it after use required frequent maintenance.

One windy Saturday, a woman was about to enter the post office through this door when a gust of wind blew it first inward and then outward, evidently without any damper control. The door swung outward and then right around, until it hit the edge of the building wall close to its axis of rotation. This caused it to lever itself from its mountings and fall to the ground, killing the woman as it fell. Improper maintenance leading to premature wear of components was claimed.

A detailed investigation revealed that the door closer was not working properly at the time of the accident, allowing the door to swing freely in either direction. In addition, the door had a specially designed hinge system to facilitate the interchange of doors. Mounted on the bottom threshold was a roller that fitted into a fixed socket located within the door, forming the bottom “hinge.” At the top, the door closer unit itself acted as the hinge, mounted on the lintel overhead with its torsion arm connected to the top of the door by means of a spring-loaded quick release mechanism. Heavy usage had resulted in distortion of the door closer torsion arm, sufficient to displace the spring-loaded plunger arrangement almost to the point of release. Uncontrolled swinging, combined with a leverage action when the door hit the wall, generated enough force to separate the door completely from the door closer at the time of the accident.

The design concept of using this particular spring-loaded plunger mechanism as part of the door hinge system was inappropriate in that there was no positive connection between the door and its top “hinge.” A later model of this door design had been installed elsewhere inside the post office, with a bolted connection replacing the quick-release mechanism. If this had been fitted to the door in question, there would have been no accident.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

10. S. Pugh, *Total Design—Integrated Methods for Successful Product Engineering,* Addison-Wesley Publishing Company, 1990

11. M. Khan and D.G. Smith, Overcoming Conceptual Barriers—by Systematic Design, C377/079, WDK-18, *Proc. ICED-89,* I.Mech.E., 1989

12. B.L. Tuttle, Creative Concept Development, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 39–48

Embodiment Design

Embodiment Design Issues. During the phase following conceptual design, the selected concept must be developed into a practical, reliable, and safe design. This phase is often called *embodiment design,* as there are systematic guidelines available for the progressive development of a concept (Ref 1, 4, 5) quite different in nature from those used in the detail design of individual components. These guidelines may be used in the form of a checklist to help identify common deficiencies in developing a particular design. For example, the following issues should be reviewed:

- Simplicity
- Clarity of function
- Safety
- Selection of materials, products, and parts
- Transportation

Simplicity. It is important for a design to be as simple as possible while still meeting the requirements of the design specification. This may be achieved, for example, by reducing the number of components, making components do more than one function, and promoting the use of near net shape manufacturing techniques. It is all too easy to become enamored of exciting new technology or a complex way of doing something, to the detriment of the final result.

For example, some types of vegetables in supermarket display cabinets are misted with water to keep them fresh. This poses design problems such as control of the spray and how to avoid startling customers who are selecting produce as the water jets spray without warning. One innovative design was fitted with an enclosed stainless steel trough, level controls, four ultrasonic misting devices, oscillator circuits, a blower, and electromechanical controls. It was indeed a complex mechatronic marvel, the workings of which could be neither seen nor understood without disassembly. Although it functioned adequately, it was very difficult to clean. The result was that it was left alone, and it began to spray harmful bacteria as well as the misted water. Finally a series of customers fell victim to Legionnaire's Disease, which was traced back to this mechatronic misting device. A survey of designs by other manufacturers showed that the misting could be done just as effectively with simple valves, timers, and sprays, without the risk of harboring bacteria.

Clarity of Function. Clarity in design (Ref 1, 4) means making sure that the design itself explains how it is to be put together, what the load paths are, what the function of each component is, and how the parts move relative to each other. Specific guidelines are available for addressing this issue in a systematic fashion (Ref 4), and in conjunction with other embodiment design guidelines, many types of failure can be avoided. Lack of clarity in design creates ambiguities and design weaknesses that may not be immediately obvious. This is illustrated by the following example.

Heavy trucks usually have a beam axle front suspension with a kingpin mounted near vertically through the eye at each end of the axle. The yoke of each front wheel stub axle fits over the corresponding eye of the beam axle and around the kingpin, thereby forming the axis about which the stub axle can turn in order to steer the truck. Kingpins are normally simple cylindrical hardened shafts extending out of the top and bottom of the axle eye

into the top and bottom bearings of the stub axle yoke. A thrust bearing is fitted between the bottom of the axle eye and the bottom bearings in the yoke to accept the vehicle weight from the axle eye. The forces are thus transmitted from axle eye to yoke without undue friction when steering action takes place. The kingpin is locked in place by means of a cross bolt arrangement through the axle eye.

In Ohio, a heavy dump truck was coming around a right curve in the road when the front left wheel assembly suddenly collapsed and separated from the truck. As the steering box was linked to this particular stub axle, the driver immediately lost all steering control. The vehicle went straight ahead, colliding head-on with a car going the opposite way and killing the passenger in the car.

Investigation revealed that this particular model of dump truck is fitted with a tapered kingpin inserted from the bottom. It is held in place by friction against the tapered hole, together with a nut and washer arrangement at the top. The nut had come loose and the kingpin had dropped progressively as the nut turned. When the nut unscrewed completely, the kingpin fell right out of the axle eye, causing the whole wheel assembly to separate from the truck. In conventional arrangements the vertical component of the truck weight is always transmitted from springs to beam axle and from beam axle to stub axle through the thrust bearing below the axle eye. Even if the thrust bearing were to disintegrate, the load path would still pass through the same components. With the dump truck, however, as wear takes place in the thrust bearing, the clearance between the nut plus washer at the top of the kingpin and the top of the stub axle yoke decreases to the point where the load path changes. The weight of the truck transfers from the thrust bearing to the nut and washer. It is only a matter of time before the torsional friction forces are sufficient to shear the pin that locks the nut, and then to undo the nut in a progressive manner.

In this case the embodiment design was deficient in that it allowed an unacceptable load path change with wear or failure of the thrust bearing. This would have been prevented by the use of accepted embodiment design guidelines, and the accident would not have happened.

Safety. Safety is something to consider throughout the design process, but there are specific embodiment design guidelines such as the safety hierarchy (Ref 13) to follow when developing a design concept. Techniques such as safe-life design, fail-safe design, redundant design, and hazard analysis (Ref 4, 5) must be considered as integral parts of the product development, and not simply as “add-ons” at the end of the project. Safety considerations and hazard analysis in design are discussed further in the articles “Safety in Design” (Ref 14) and “Risk and Hazard Analysis in Design” (Ref 15) in Volume 20 of the *ASM Handbook*. The focus of product liability lawsuits is on the safety aspects of design (Ref 16). It is often claimed that the litigation has the effect of “improving safety” (Ref 17). However, it is probably more accurate to say that such lawsuits are a very expensive way of addressing safety, and that the results are unpredictable (Ref 18). Sometimes safety is improved; sometimes it is unchanged; and sometimes it is compromised. For example, the automobile airbag raises several issues concerning safety in design:

A man was driving at night when a deer leapt into the road right in front of his car. On impact the front of the car knocked the deer's legs from under him and as he rolled through the windshield, the driver's airbag was activated, apparently causing the man's hands to be forced upward into the path of the deer's antlers. An antler caught one of the man's thumbs and tore it off. A lawsuit was filed against the manufacturer, claiming that if the airbag had not deployed, the man would not have lost his thumb.

Would the man have been injured less or more if there had been no airbag? What if the airbag had been there but had not deployed? What if the man had lost control after the airbag deployed and had driven into a tree without any airbag protection? Did the airbag actually protect the man from more serious injuries upon impact by the deer?

If a failure appears to involve safety issues, then a simple starting point in the analysis is to work through the safety hierarchy in the form of the following questions:

- Were the hazards/risks associated with the device eliminated by design as far as practicable?
- Were appropriate protective systems incorporated for remaining hazards and/or risks?
- Were appropriate warnings provided for residual hazards and/or risks?
- Was there provision for instruction and training?
- Was personal protection prescribed (if necessary and as a last resort)?

With regard to large equipment and commercial systems, a hazard analysis, failure modes and effects analysis (FMEA), or other type of formal safety analysis is often required as part of the design process. While such

analyses do help to identify potential problems, and help to reduce the risk of a crippling failure or accident, they are very dependent on the initial assumptions. Many times a catastrophic failure occurs, despite all the elaborate analysis, simply because an assumption was made that *no one would ever do such a thing.* For example, if the absolute minimum allowable brake rotor thickness and the minimum allowable brake pad thickness are both defined in the maintenance manual for an automatic passenger transit system, is it appropriate to assume that the professional maintenance staff, dedicated to ensuring the safe operation of the vehicles, would replace the pads and/or rotors rather than let every pad and every rotor on the vehicle wear to below the absolute minimum? If it is assumed that these specific instructions will not be followed, then what are all the other instructions that must then fall under the same assumption? It may sound ridiculous, but this is exactly what happened, and it was claimed that the ensuing crash caused by pads no longer touching the rotors during braking was caused by a faulty FMEA!

In recent years the legal community has so overreached in cases involving claims over “lack of warnings” that the situation has become untenable for many products. Plastering complex warning notices over every conceivable surface of a product in letters so small that a microscope is required to read them contributes nothing toward safety. However, carefully designed warnings, applied in appropriate circumstances, can be an effective means of preventing accidents and failures.

Selection of Materials, Products, and Parts. Another critical aspect of design, which needs to be addressed formally in the embodiment phase, is the selection of materials, semi-finished products, and standard components. Materials selection in mechanical design is discussed extensively in the article “Overview of the Materials Selection Process” in Volume 20 of the *ASM Handbook* (Ref 19). Detailed methods are provided in Ashby's *Materials Selection in Mechanical Design* (Ref 20). Material selection should be carefully reviewed in light of all intended service conditions to foresee any conflicts, as discussed in the article “Materials Selection for Failure Prevention” in this Volume (Ref 21). For example, a seal might be expected to perform in an operating environment at 150 °C (302 °F). Butyl rubber is often an excellent material choice for seals, but in this case the extensive creep of butyl rubber at high temperature would rule it out.

Sometimes a given part is chosen in order to improve performance in one area while other aspects of the service environment are ignored. In the example described at the beginning of this article, elastomeric components were selected for their capability to reduce noise and vibration in motors, but they degraded rapidly in service due to insufficient oil resistance.

There are numerous aids available for determining material behavior. These resources include graphical materials selection charts (Ref 20), which are appropriate for the selection of a broad class of materials for a given application (such as polymers). Materials selection at this level is appropriate for conceptual design. Detailed web-based software, such as the Cambridge Engineering Selector (Ref 22) or vendor databases, allow the designer to obtain properties of individual materials, often useful in the detail design phase of a project.

It is obvious that if an incorrect selection is made during design, then the risk of failure will be high. What is less obvious is the increasing problem of substandard materials and copied or counterfeit parts. Some industries, such as the aircraft and petroleum industries, have procedures to ensure that supplied components are in fact what they are claimed to be. However, many others have a long way to go in addressing this potential cause for failure, as shown by the following example:

A replacement strainer was fitted into a process steam line in a chicken feed plant. Soon afterward, and luckily during a weekend, the cast iron wall of the strainer failed. As a result of the explosion, the entire plant had to be cleaned and repainted from the steam damage, and of course there were claims and counterclaims over who was at fault. Thickness measurements at the rupture site on the wall of the imported strainer showed the wall thickness to be well below the minimum required by applicable American National Standards. The sales drawing (cut sheet) for the strainer indicated “ISO 9002 Quality Assurance,” yet there was no evidence of any International Organization for Standardization (ISO) 9000 registration nor even that the company staff knew what the term meant. This imported product was held out to be equivalent to those produced by U.S. manufacturers. It demonstrated clearly that it was not.

Transportation Issues. Although at first it may not seem appropriate to review transportation issues when analyzing possible flaws in the engineering design process, it is surprising how many accidents and failures stem from the lack of attention to transportation issues at the design stage. During transportation, a material, component subassembly, or complete product is subjected to conditions completely different from those to be expected during manufacture or subsequent operation in service. The environmental conditions are different; the loading conditions are different; and even the perception of the item being transported is different. For

example, it is common for machines to have a center of gravity that is biased toward one side, or that may be above the midpoint of the height. However, when crated the machine can no longer be seen, and to the shipper it is considered simply a box of a certain size and weight to be conveyed from pickup point to delivery point using agreed modes of transportation. Unless there are clear warnings and instructions with regard to handling of "the box," the shipper may not realize the high risk of overturning and damage due to load shifting. Similarly, when vehicles being shipped are tied down to prevent movement, the rolling elements of axle bearings are held in a fixed position under load and may experience impact loads high enough to cause damage from brinelling. This type of issue should be addressed no later than the embodiment design phase of the design process in order to avoid problems during delivery and afterwards.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

5. E. Frankenberger, P. Badke-Schaub, and H. Birkhofer, Ed., *Designers—The Key to Successful Product Development,* Springer-Verlag, 1998

13. R.L. Barnett and W.G. Switalski, Principles of Human Safety, Safety Brief, Vol 5 (No. 1), Triodyne Inc., 1988

14. C.O. Smith, Safety in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 139–145

15. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 117–125

16. T. Willis, M.P. Kaplan, and M.B. Kane, Safety in Design: an American Experience, C377/110, *Proc. of the Institution of Mechanical Engineers International Conference on Engineering Design (ICED-89),* Institution of Mechanical Engineers, 1989

17. C. Hales, Legal Threats to Innovation in Design, *Proc. ICED 99: International Conference on Engineering Design,* Institution of Mechanical Engineers, 1999

18. P.W. Huber, *Liability: The Legal Revolution and Its Consequences,* Basic Books, Inc., 1988

19. G.E. Dieter (chairperson), Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 241–328

20. M.F. Ashby, *Materials Selection in Mechanical Design,* 2nd ed., Butterworth-Heinmann, 1999

21. B.A. Miller, Materials Selection for Failure Prevention, *Failure Analysis,* Vol 11, *ASM Handbook,* ASM International, 2002

22. M.F. Ashby and D. Cebon, Cambridge Engineering Selector, CES Software Version 3.2, Granta Design Limited (www.grantadesign.com), 2001

Detail Design

Once a design has passed into the detail design phase, it used to be common for it to be *given to the draftsman* to finish. It is a fatal mistake to think that detail design is unimportant and needs less attention than other phases. While it is true that excellent detail design cannot compensate for a bad concept, it is equally true that

poor detail design can ruin a good concept. Detail design is critical and it needs to be understood and investigated when tracking the root cause of an engineering failure. As shown by the following example, this is not just a matter of considering individual components, but can also involve the interaction of various materials in an assembly or with the environment.

An articulated tractor and low-loading trailer was being driven through some hilly country in Missouri after delivery of a bulldozer to a construction site. The driver noticed that his rear trailer brakes had started to smoke, and he found that the brakes were partially applied even when he wasn't using them. He assumed that the brakes were running hotter because he was driving faster through the hills without the load of the bulldozer, and that this heat was causing his brakes to drag. Using a wrench he backed off the brake shoes from each wheel to the point at which he thought they were correctly adjusted. He telephoned his dispatcher from the next rest area, explained what had happened, and was told that he should continue with his journey. A few miles later the weather changed to a misty drizzle, and far ahead he saw that a minor rear-end accident had just occurred at an intersection. He applied his brakes but found that there were now no brakes on his heavy trailer. People started running about on the road and shoulder ahead of him, so he could not steer around. He decided all he could do was to apply his brakes as hard as he could. The tractor-trailer jackknifed completely and collided head-on with a vehicle coming the other way. One person was killed and several were seriously injured.

Large articulated trucks generally have air-operated braking systems, and certain types of low-loading trailers require the use of a particular relay valve because of their physical layout. The relay valve directs and releases air, depending on signals from the driver's brake pedal, trailer brake controls, and the integrity of the air system.

An inspection of the valve from the accident trailer showed that a small steel valve head with a 5 mm threaded stem had come loose. The valve head hangs down from a piston, and when the threaded connection came loose it progressively unscrewed to the point that the valve head blocked an air exhaust port, thus causing the brakes to remain partially applied. Without knowing it, the driver had backed off his brake shoes to the point that they could not come into contact with the drums when the brakes were applied. It was later found that the detail design of this particular threaded connection had been revised and tested numerous times by the manufacturer and that several thread-locking methods had been put into production over the years. The thread involved in the accident had not been locked in place, and the loose thread tolerance allowed complete unscrewing to occur.

The detail design was flawed in that a critical component was hung vertically inside a valve by a loose 5 mm thread. The unscrewing of the thread resulted in the death of one person and serious injuries to several others. It was a known problem, which had been brought to the attention of management, and if it had been addressed properly the accident would not have happened.

Management Influences

Management Control. One of the most frustrating things for design engineers is the way projects are manipulated by those who have very little to do with the design process itself. It is critical for the design manager or team leader not only to be aware of the impact of various influences, but also to exercise control over those that can be controlled and compensate for those that cannot, in the best interests of the customer, the project, and the design team.

Design team activities must be directed and monitored for performance. The design output must be assessed against the specification requirements continually. The effect of influencing factors must be actively predicted, monitored, and controlled when possible. Management involvement in these issues is crucial to the development of high-quality and cost-competitive products (Ref 1, 23, 24, 25). From the design management point of view, the ultimate goal is to produce the highest quality product, meeting the user's expectations for the lowest cost in the shortest time.

A particular challenge in the management of engineering design is to be able to cope with issues that range from “hard” to “soft”—for example, from the dimensional tolerance on a single component to the user's satisfaction with a product in service. Another challenge is that the critical issues must be considered at different levels of resolution and from different points of view. The design manager must be able to see the overall picture while rapidly windowing in on the details and understanding the effect that even tiny details might have on the overall project. A lack of management skill in this area has contributed to many engineering disasters, such as the failure of the solid rocket booster (SRB) on the space shuttle Challenger. Again, simple sets of questions, based on fundamental design principles and asked at the appropriate time by a manager with

adequate technical understanding, can highlight design weaknesses long before a disaster becomes inevitable (Ref 26).

While such issues are not often the focus of a forensic engineer's investigation, it is important to be aware of the influence of the management factor and to be able to recognize situations in which it contributes to the failure. For example, it turned out that a bearing failure in a large industrial dynamometer was the result of the withholding of design information from a new employee by one who felt that his position had been usurped. What started as a routine bearing and material failure analysis ended up as a seminar for the company on project management.

Design Team. Nowadays it is common for design teams to be created and staffed on a project-by-project basis, and specifically for the duration of the project. Increasingly it is common for the work to be carried out in multiple locations by means of electronic communication. Bringing together and orchestrating a team to produce a quality design in a timely fashion is not easy and must be recognized as a critical task. The article "Cross-Functional Design Teams" (Ref 9) in Volume 20 of the *ASM Handbook* addresses this aspect of engineering design.

People have a functional role in the team, using their particular technical expertise and experience, and obviously this has to be matched to the work at hand. They also have a team role, using their particular character traits to help make the team operate effectively as a team (Ref 24). If the set of team roles is not well balanced, then the output suffers badly, no matter how good the balance of functional roles is. A team may be adequate in a functional sense, having the right expertise and experience, yet may not have the right balance of personalities to be productive. Teams need a mix of personalities covering basic "team roles," with the addition of "specialist" roles in technical situations (Ref 1).

The negotiating ability and the negotiating power of the team are critical to the design process. To be successful, a design team needs to be good at negotiating, and it needs to negotiate from a position of power. Often design teams have more power than they realize, for without their input the company would fail. The more incisively the design team can present its case, the better it is able to control the things that matter. This is illustrated by the failure of the space shuttle Challenger. The failure of management to comprehend the importance of detail design, coupled with the failure of the design team to get the message across, set the scene for the failure. If the design team had understood and learned how to use its latent power effectively, Challenger would not have been launched. It is interesting to note that the great engineers of the world, such as Eiffel, Brunel, and Ford, were not only excellent technically, but also were persuasive, entertaining, and politically involved individuals.

Communication Problems. The importance of effectively communicating the results of the design process is discussed in the article "Documenting and Communicating the Design" (Ref 27) in Volume 20 of the *ASM Handbook.* In fact, effective communication has become even more important as the idea of *global design teams* gains in popularity. The recent failure of the Mars Orbiter mission due to mixed imperial and metric units highlights this type of problem, which becomes compounded by differences in language and culture. What used to be a matter of simple transmittal of information is now a matter of ascertaining what someone actually interpreted from the information provided. The following example shows how easy it for a failure to occur when communication is ineffective during the design process.

To upgrade its production facilities, a European manufacturer ordered two huge custom-built metalworking presses, high on the scales of novelty, complexity, and cost, with a short time frame for delivery. The concept involved a unique automatic transfer system for progressing parts longitudinally through a series of operating stations, so the design work was subcontracted to an experienced North American company with special knowledge of the particular type of machine required. As the time frame was so short, the documentation of customer requirements was perfunctory, and there was some confusion as to the responsibilities of the various parties. Under pressure to meet unrealistic deadlines, the design was rushed through to detailing with insufficient time spent either on the concept evaluation or on development of the concept to meet the design specification. Some 4000 drawings were produced by hand, involving translation from one language to another and changes from imperial to metric units. The machines were built according to the translated drawings with almost no communication between the design company and the manufacturer. Apparently it was assumed that the knowledge of the design company would be fully imparted to the manufacturer through the medium of the drawings. The time frame was too short to allow for testing and commissioning, so the machines were built and put into service without normal shakedown procedures. Although slow-speed operation was achieved, the machines could not be run at the speed specified by the customer and agreed to by the manufacturer. Severe

mechanical damage to components was caused in the attempts to reach full speed. The design work was never paid for, and the claims for damages far exceeded the total cost of design and manufacture of the machines. Inadequate communication, a deficient design specification, and the separation of design from manufacture all contributed to the inevitable failure.

Time and Money Constraints. It is an unfortunate fact that the design process takes time and costs money at an early stage in a project, long before there can be any return on investment. Even still, because the outcome of the design process sets all future costs on a project, common sense would dictate that sufficient time and money should be invested in the design process to ensure that the optimal design is produced to minimize downstream costs. In practice it is more often the case that time and money constraints force the design team to take shortcuts, resulting in costly mistakes or design weaknesses. By breaking the design process down into phases for the purpose of analysis, it is possible to check the amount of effort (ideally in hours) that is spent on design activities during the course of a project and thereby come to a conclusion as to whether the time and money constraints contributed to the failure. For example, in a case where a company claimed to have developed its own technology without copying from others, an analysis of the time spent by each different team member between the start and finish dates showed that the job could not have been done in that time except by copying an existing design.

References cited in this section

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

9. P.G. Smith, Cross-Functional Design Teams, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 49–53

23. "Guide to Managing Product Design," *BS 7000:1989,* British Standards Institution, 1989

24. R. Stetter and D.G. Ullman, Team-Roles in Mechanical Design, *8th International Conference on Design Theory and Methodology, 96-DETC/DTM-1508,* J. Cagan and K.L. Wood, Ed., American Society of Mechanical Engineers, 1996, p 1–8

25. E.E. Rothschild, *Product Development Management,* T. Wilson Publishing Company, 1987

26. C. Hales, Analysis of an Engineering Design—the Space Shuttle Challenger, *Engineering Design and Manufacturing Management,* A.E. Samuel, Ed., Elsevier, 1989

27. G. Vrsek, Documenting and Communicating the Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 222–230

External Influences

In coming up with a design that meets the user's needs in the best way possible, trade-offs must be made among the requirements of function, safety, timeliness, cost, ergonomics, the environment, and aesthetics. Often, for example, safety systems can be improved, but at additional cost. In this case the question becomes one of *how safe is safe enough.* Established standards and codes provide the design engineer with a basis for making such judgments in a professionally acceptable manner, and regulations help to enforce accepted levels of practice. The effects of codes and standards on design are discussed in the article "Designing to Codes and Standards" (Ref 28) in Volume 20 of the *ASM Handbook.* In general, standards are more concerned with setting a level of performance, quality, or safety by the definition of criteria, while codes are more concerned with ensuring a level of performance, quality, or safety through adherence to a set of rules or guidelines. The variety of each is enormous; the requirements vary from area to area; and commonly there are inconsistencies that are complicated to resolve. It may be very difficult for the design engineer even to determine which standards or codes apply under particular circumstances, let alone to interpret the details of the fine print. Some are in ISO units; others are in Imperial units. Some are international; others are national; others are regional, and others are

local. Some are specific to a particular product, while others are more generic. Some deal with the minute details of a material composition, while others deal with the testing of whole assemblies or lay down safety procedures. Even the terminology used varies from one document to another, and subtle differences in meaning can sometimes lead to expensive misunderstandings.

When analyzing a failure from the design point of view, it is often important to identify which standards and codes applied to the product at the time of design, as well as at the time of the failure. It is then necessary to determine whether there has been a violation and whether the requirements are legislated and mandatory, voluntary, or just accepted professional engineering practice. There are now standards for carrying out and managing the design process itself, such as VDI Guideline 2221 (Ref 29) and British Standard BS 7000 (Ref 23); with regard to the broader issue of product quality, the series of *ISO 9000 International Standards for Quality Management* is a key reference.

The most controversial standards are often those concerned with safety. They are of concern to the design manager because they strongly influence the design, operation, and maintenance of technical systems and products. A safety standard is a document that attempts to specify components and practices that will result in predictable and acceptable levels of safety. The concept of what is safe needs more careful definition in the design context than in general usage, and its definition is dependent on a set of related ones, as shown by the following definitions adapted from Hebert and Uzgiris (Ref 30):

accident.
: An undesirable event or failure that results in *harm*

harm.
: An adverse effect that occurs in an *accident*

hazard.
: A condition or situation exhibiting the potential for causing *harm*

risk.
: A measure of the probability and severity of *harm:* the potential of a *hazard* to cause *harm*

safe.
: A characterization of a machine, product, process, or practice whose attendant *risks* are judged to be acceptable

safety.
: A state or condition wherein people and property are exposed to a level of *risk* that is judged to be acceptable

safety standard.
: A set of criteria or means for achieving a level of *risk* that is judged to be acceptable by the body formulating the *safety standard.*

In terms of these definitions, the task of the design engineer is to design a safe system by identifying the hazards and controlling the associated risks to within acceptable limits. The criteria for what is acceptable are set, in part, by safety standards. Although compliance with applicable safety standards is generally understood to be a necessary condition for safe design, it may not be a sufficient condition (Ref 31). The standard may not have kept pace with industry or new developments, and it may not address all the hazards involved. It is up to the design engineer to identify hazards, regardless or whether they are described in the standard, and to make sure that the issues are adequately addressed. For example, the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code provides design rules for the pressure-containing components of a boiler, but not the burner system. The burner system is covered by various other codes and standards, depending on the type of fuel and style of burner. This leaves the mechanical interface between burner and boiler not covered by any code or standard. It so happened that an old but serviceable industrial boiler was upgraded by fitting a new and more efficient burner. The frame-mounted fan and ducting to the windbox were replaced by an in-line burner, fan, and motor unit, which was connected directly to the end of the boiler casing by means of a custom-made adapter ring. As there was no code requirement for how to make such a connection, it was simply done on-site by the boilermaker without any engineering review or inspection. Five years later the connection to the boiler casing gave way, allowing the burner to tilt and the flame to impinge directly on the wall of the first pass Morrison tube. This caused overheating of the steel and collapsing of the tube under normal internal steam pressure. The boiler exploded, killing several workers and injuring many more.

References cited in this section

23. "Guide to Managing Product Design," *BS 7000:1989,* British Standards Institution, 1989

28. T.A. Hunter, Designing to Codes and Standards, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 66–71

29. Systematic Approach to the Design of Technical Systems and Products (translation), *VDI Guideline 2221:1987* Dusseldorf: Verein Deutscher Ingenieure, 1987

30. J. Hebert, and S.C. Uzgiris, The Role of Safety Standards in the Design Process, *ASME 89-DE-3* (New York), American Society of Mechanical Engineers, 1989

31. M.A. Dilich and D.F. Rudny, Compliance with Safety Standards: A Necessary but Not Sufficient Condition, *ASME 89-DE-1* (New York), American Society of Mechanical Engineers, April 1989

Design Life-Cycle Issues

When reviewing the design process that led ultimately to a failed component, it is sometimes necessary to consider life-cycle issues such as the expected service conditions, design for maintenance, design life, and design for recycling or disposal. Many types of equipment are now used far beyond their original design life, and often there is no guidance on how to predict what failures might occur, when they might occur, or where they might occur. This has long been an issue with regard to aircraft, especially with their stressed lightweight alloy structures, but it is now becoming critical with less obvious items such as sectional cast iron boilers more than 30 years old. The inspection procedures mandated for such boilers did not envisage the need to check the wall thickness of the cast iron sections in order to determine the annual metal loss due to corrosion, and it is possible to get a catastrophic failure of the wall during operation even though the boiler has been routinely inspected according to the prevailing regulations.

Acknowledgments

The authors acknowledge and thank Robert Koutny, Triodyne Graphic Communications, for his help in preparing the figures.

References

1. C. Hales, *Managing Engineering Design,* Longman Scientific & Technical, 1993; C. Hales and S. Gooch, 2nd ed., Springer-Verlag London Limited, to be published 2003

2. J.R. Dixon, Overview of the Design Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 7–14

3. K.N. Otto and K.L. Wood, Conceptual and Configuration Design of Products and Assemblies, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 15–32

4. G. Pahl and W. Beitz, *Engineering Design,* K.M. Wallace, Ed., The Design Council, 1984

5. E. Frankenberger, P. Badke-Schaub, and H. Birkhofer, Ed., *Designers—The Key to Successful Product Development,* Springer-Verlag, 1998

6. V. Hubka, *Principles of Engineering Design,* Butterworth Scientific, 1982

7. K.M. Wallace, A Systematic Approach to Engineering Design, *Design Management: A Handbook of Issues and Methods,* M. Oakley, Ed., Basil Blackwell Ltd., 1990

8. M.M. Andreasen and L. Hein, *Integrated Product Development,* IFS (Publications) Ltd. and Springer-Verlag, 1987

9. P.G. Smith, Cross-Functional Design Teams, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 49–53

10. S. Pugh, *Total Design—Integrated Methods for Successful Product Engineering,* Addison-Wesley Publishing Company, 1990

11. M. Khan and D.G. Smith, Overcoming Conceptual Barriers—by Systematic Design, C377/079, WDK-18, *Proc. ICED-89,* I.Mech.E., 1989

12. B.L. Tuttle, Creative Concept Development, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 39–48

13. R.L. Barnett and W.G. Switalski, Principles of Human Safety, Safety Brief, Vol 5 (No. 1), Triodyne Inc., 1988

14. C.O. Smith, Safety in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 139–145

15. G. Kardos, Risk and Hazard Analysis in Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 117–125

16. T. Willis, M.P. Kaplan, and M.B. Kane, Safety in Design: an American Experience, C377/110, *Proc. of the Institution of Mechanical Engineers International Conference on Engineering Design (ICED-89),* Institution of Mechanical Engineers, 1989

17. C. Hales, Legal Threats to Innovation in Design, *Proc. ICED 99: International Conference on Engineering Design,* Institution of Mechanical Engineers, 1999

18. P.W. Huber, *Liability: The Legal Revolution and Its Consequences,* Basic Books, Inc., 1988

19. G.E. Dieter (chairperson), Overview of the Materials Selection Process, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 241–328

20. M.F. Ashby, *Materials Selection in Mechanical Design,* 2nd ed., Butterworth-Heinmann, 1999

21. B.A. Miller, Materials Selection for Failure Prevention, *Failure Analysis,* Vol 11, *ASM Handbook,* ASM International, 2002

22. M.F. Ashby and D. Cebon, Cambridge Engineering Selector, CES Software Version 3.2, Granta Design Limited (www.grantadesign.com), 2001

23. "Guide to Managing Product Design," *BS 7000:1989,* British Standards Institution, 1989

24. R. Stetter and D.G. Ullman, Team-Roles in Mechanical Design, *8th International Conference on Design Theory and Methodology, 96-DETC/DTM-1508,* J. Cagan and K.L. Wood, Ed., American Society of Mechanical Engineers, 1996, p 1–8

25. E.E. Rothschild, *Product Development Management,* T. Wilson Publishing Company, 1987

26. C. Hales, Analysis of an Engineering Design—the Space Shuttle Challenger, *Engineering Design and Manufacturing Management,* A.E. Samuel, Ed., Elsevier, 1989

27. G. Vrsek, Documenting and Communicating the Design, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 222–230

28. T.A. Hunter, Designing to Codes and Standards, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 66–71

29. Systematic Approach to the Design of Technical Systems and Products (translation), *VDI Guideline 2221:1987* Dusseldorf: Verein Deutscher Ingenieure, 1987

30. J. Hebert, and S.C. Uzgiris, The Role of Safety Standards in the Design Process, *ASME 89-DE-3* (New York), American Society of Mechanical Engineers, 1989

31. M.A. Dilich and D.F. Rudny, Compliance with Safety Standards: A Necessary but Not Sufficient Condition, *ASME 89-DE-1* (New York), American Society of Mechanical Engineers, April 1989

Failure Modes and Effects Analysis

John B. Bowles, Computer Science and Engineering, University of South Carolina

Introduction

FAILURE MODES AND EFFECTS ANALYSIS (FMEA) has evolved into a powerful tool that can be used by design engineers during all phases of product development to enhance product safety and reliability by eliminating or mitigating the potential effects of item failures. FMEA consists of examining the modes and causes of item failures and determining the product response to the failures. Steps can then be taken to change the design in order to eliminate the failure, mitigate its effects, or develop compensating provisions in case the failure should occur. A structured approach to the FMEA ensures that all appropriate failure modes are analyzed, that the system satisfies its fault mitigation requirements, and that these requirements are properly allocated. The FMEA methodology can be usefully employed throughout the design cycle from conceptual design to production and deployment. Tools have been developed to reduce the amount of labor required for the analysis and to evaluate hardware-, software-, material-, and process-related causes of failure. Significant progress is also being made in automated tools to facilitate the analysis.

This article describes the methodology for performing an FMEA. The overview section describes the process with the specific example of a hot water heater, followed by a discussion of the role of FMEA in the design process. The second section describes the analysis procedures and shows how proper planning, along with functional, interface, and detailed fault analyses, makes FMEA a process that can contribute to the design throughout the product development cycle. The third section describes the use of fault equivalence to reduce the amount of labor required by the analysis. The next section shows how fault trees are used to unify the analysis of failure modes caused by design errors, manufacturing and maintenance processes, materials, and so on, and to assess the probability of the failure mode occurring. The last section describes some of the approaches to automating some of the analysis.

Overview of FMEA

As the name suggests, FMEA is a procedure that examines each item in a system, considers how that item can fail, and then determines how that failure will affect the operation of the system. It is a structured, logical, and systematic analysis. Identifying possible component failure modes and determining their effects on the system operation helps the analyst to develop a deeper understanding of the relationships among the system components and, ultimately, to improve the system design by making changes to either eliminate or mitigate the undesirable effects of a failure.

Although designers have always had to be concerned with the possible effects of item failures, FMEA developed as a formal methodology during the 1950s at Grumman Aircraft Engineering Corporation, where it was used to analyze the safety of flight control systems for naval aircraft (Ref 1, 2). The first article to describe the FMEA process in detail was published in 1960 by Lomas, who was an engineer employed by the U.S. Navy (Ref 2), but the procedure he described was similar to one required by Mil-F-18372 (Aer.) (Ref 3), so some significant earlier work must have been done. Another early description of a procedure for performing a “failure mode effects analysis” was given by Coutinho (also at Grumman Aircraft Engineering Corporation) at the New York Academy of Sciences in 1964 (Ref 4). These first reports described a quite modern approach to doing a FMEA. Lomas included the probability of the failure mode occurrence in the analysis, and he introduced the tabular FMEA worksheet, which initiated the paradigm of doing the analysis by “filling out the form.” Coutinho recommended that FMEA be used as part of a formal design review by nondesigners of the system; he also classified the failures in terms of their consequences. During the 1970s and '80s, various military and professional society standards were written to define the analysis methodology (Ref 5, 6, 7, 8). MIL-STD 1629 (ships), “Procedures For Performing a Failure Mode Effects and Criticality Analysis” (Ref 9) was published in 1974, and, through several revisions, became the basic approach for analyzing a system.

Initially, FMEAs were used primarily as a safety analysis on the system hardware after the design was nearly complete. This application meant that any problems uncovered by the analysis were likely to be extremely expensive to fix. Recent refinements in the methodology have expanded the types of failures that can be analyzed to include functional failures in a functional representation of the system (Ref 10, 11, 12), failures of software components (Ref 11, 13, 14), and failures in the processes through which a product is built or maintained (Ref 15, 16). Along with these extensions in the

methodology, tools have also been developed to reduce the amount of labor required for the analysis (Ref 17), analyze the cause of the failure (Ref 18), and even automate much of the work required for the analysis (Ref 19, 20, 21, 22). Thus, over the last half-century, FMEA has become a “traditional” reliability analysis technique and has evolved into a highly effective tool that can be used throughout the product development process to improve the design of the system to which it is applied.

The term *FMEA* is often used almost interchangeably with the term *FMECA*—failure modes, effects, and criticality analysis. Some authors (Ref 1) take the view that FMEA is limited to an analysis of the effects of item failure modes, while FMECA includes the reliability function of assessing the probability of the occurrence of a failure mode along with its effect on the system. Others (Ref 23, 24) take the view that FMECA extends the analysis to include a ranking of the failure modes based on both a combination of the failure mode's probability of occurrence and the severity of its effect. Still others use the term *FMEA* even when a ranking is included (Ref 15, 16). This article adopts the terminology in Ref 11 and uses *FMEA* as the general term to include assessing the failure mode probability of occurrence under its scope. In the case of software- and process-related failure modes, it often is not possible to determine a meaningful probability of occurrence. This article does not include ranking procedures.

FMEA Example: Analysis of a Domestic Hot Water Heater. Figure 1 shows a schematic of a gas hot water heater. Functionally, the hot water heater takes cold water and gas as inputs and produces hot water as an output; flue gases and heat leakage are also produced as waste outputs. The water temperature is regulated by the controller opening and closing the main gas valve (labeled stop valve) when the temperature of the water in the tank goes outside the preset limits of 60 to 82 °C (140 to 180 °F). The pilot light is always on and the gas valve operates the main burner in full-on/full-off modes. The controller is operated by the temperature measuring and comparing device. The check valve in the water inlet pipe prevents reverse flow due to overpressure in the hot water system, and the relief valve opens if the system pressure exceeds 100 psig.

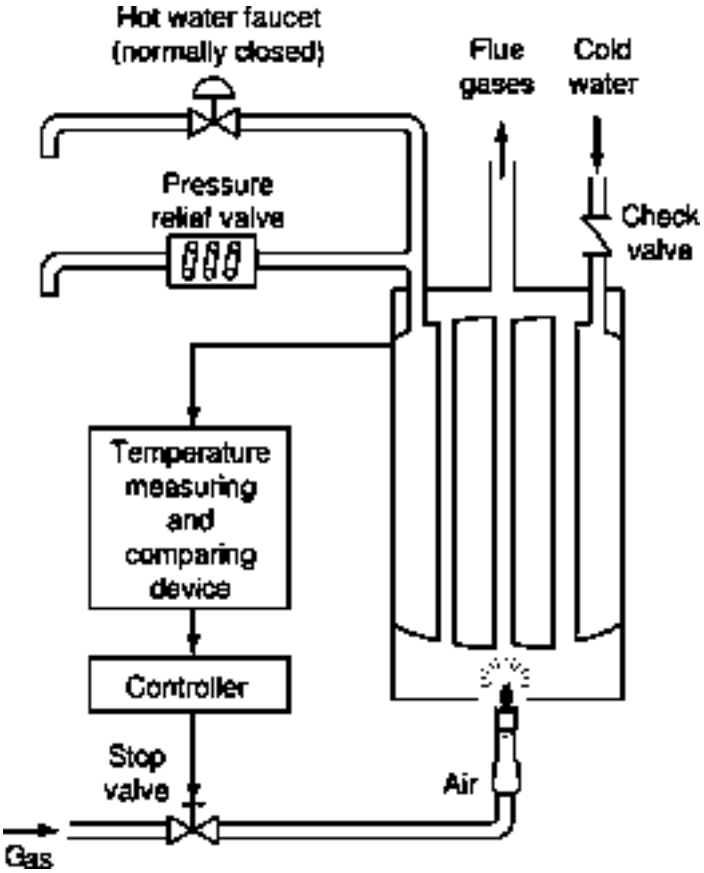


Fig. 1 Schematic for a gas hot water heater (Ref 25)

Several models provide useful representations of the system for analysis. First, the elements of the hot water heater can be represented hierarchically as shown in Fig. 2. This type of hierarchical model reflects the way in which the system design develops by first defining the system functions, then subdividing the system into smaller subsystems. The levels in the model are often called *levels of indenture*—a result of documentation techniques in which the descriptions of subsystem functions were indented relative to the description of the subsystem of which they were a part. Numbering conventions such as 1.23.12, which identify the system (1), subsystem (23), and so on, are often used to identify the system components.

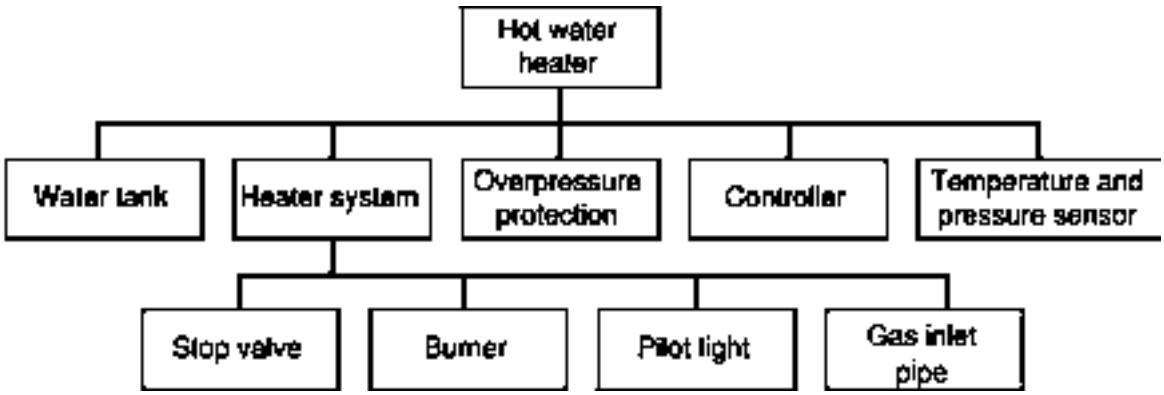


Fig. 2 Hierarchical representation of the hot water heater and its subsystems

The representation in Fig. 2 shows five subsystems, each of which can be further subdivided. For example, the heater subsystem can be subdivided into the stop valve, burner, pilot light, and gas inlet pipe.
The functions of the components shown in Fig 1 are:

Component	**Function**
Stop valve	Controls gas flow; full-on/full-off (controlled by the controller)
Controller	Opens and closes stop valve (responds to temperature sensor)
Temperature/pressure sensor	Senses water temperature and pressure
Check valve	Prevents reverse flow if overpressure
Relief valve	Opens when pressure >100 psig
Pilot light	Lights burner (always on)
Burner	Heats water (operated by stop valve)
Tank	Holds water (safe up to 100 psig)
Faucet	Releases water when needed

The functional relationships between the different system components are shown in a functional block diagram. Figure 3 shows a functional model for the hot water heater. The heater system, water tank, and overpressure protection form the basic system for producing the hot water; the temperature and pressure sensor and the controller provide a feedback loop for monitoring the production of the hot water and regulating its temperature. The heater system can be decomposed into the stop valve, gas inlet pipe, pilot light, and burner.

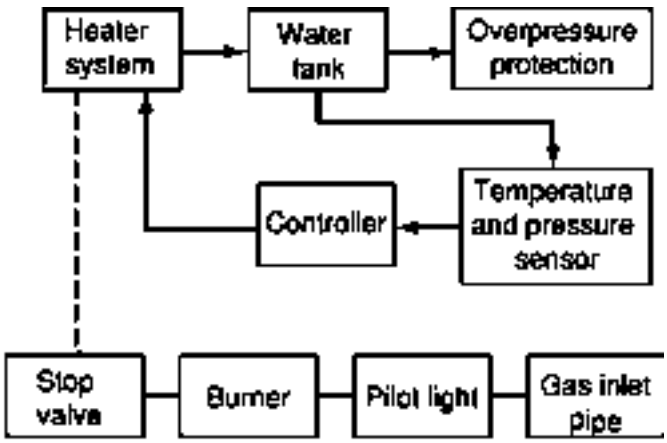


Fig. 3 Functional block diagram of hot water heater

Finally, from a reliability point of view, the hot water heater is a series system in which all components must operate properly for the system to operate. This is illustrated by the series reliability logic diagram in Fig. 4. Again, each system component can be further decomposed into its constituent subcomponents.

Fig. 4 Reliability logic diagram for hot water heater

The hierarchical decomposition of the system, the functional block diagram, and the reliability logic diagram help the analyst to understand the relationships between the system components. The next step is to identify all the ways in which each item can fail and the effect that each of those failures will have on the system. Effects are determined at each level of the system hierarchy—the effect on the module containing the failed component (local), the effect on every subsystem of which the component is a part, and the effect on the system. This analysis is illustrated by considering the stop valve. Table 1 illustrates part of the FMEA for the hot water heater showing the component failure modes and their effects at the local and system levels. Degraded operation (leaks and partial opening or closing of the stop valve), rather than outright failure, is one of the most common types of component failure modes.

Table 1 FMEA analysis of the stop valve for a hot water heater

Failure mode	**Effect**	
	Local	**System**
1. Fails closed	Burner off	No hot water
2. Fails open	Burner will not shut off	Overheats, release valve releases pressure, may get scalded
3. Does not open fully	Burner not fully on	Water heats slowly or does not reach desired temperature
4. Does not respond to controller—stays open	(Same as 2)	(Same as 2)
5. Does not respond to controller—stays closed	(Same as 1)	(Same as 1)
6. Gas leaks through valve	Burner will not shut off; burns at low level	Water overheats (possibly)
7. Gas leaks around valve	Gas leaks into room	Possible explosion, fire, or gas asphyxiation

Source: Ref 23

Generally, the analyst also identifies the possible causes of the failure and attempts to either eliminate them from the design or, if that is not possible, mitigate their effects in some way. For example, the valve could fail to open or close properly due to corrosion. This in turn could be caused by electrolysis between different metals or by a reaction of the valve material with the gas. Careful consideration of the valve materials would eliminate this failure mode. (At least it would eliminate this cause of the failure mode; there could be other causes.)

An analysis of the pilot light would reveal the hazard if it fails (i.e., it is not lit) and the gas is on. This failure mode cannot readily be eliminated, but it can be mitigated by adding an interlock to prevent the gas valve from opening if the pilot light is not on. Many gas hot water heaters have this type of interlock. If the interlock fails, the burner will not be able to heat the water even if the pilot light is on.

Role of FMEA in the Design Process. Figure 5 shows a typical product development cycle, beginning with conceptual design and progressing to deployment in the field. As indicated in the figure, the activities that constitute the FMEA complement and add value at every stage of the development cycle. During the conceptual design and preliminary design phases, the FMEA serves primarily to verify the adequacy of the system requirements; during the detailed design phase it is used to verify compliance with the requirements. During the verification and validation phase, it helps to maintain the integrity of design changes. Finally, during the production, use, and support phase, it serves as a guide for collecting field failure data and for developing maintenance and troubleshooting procedures.

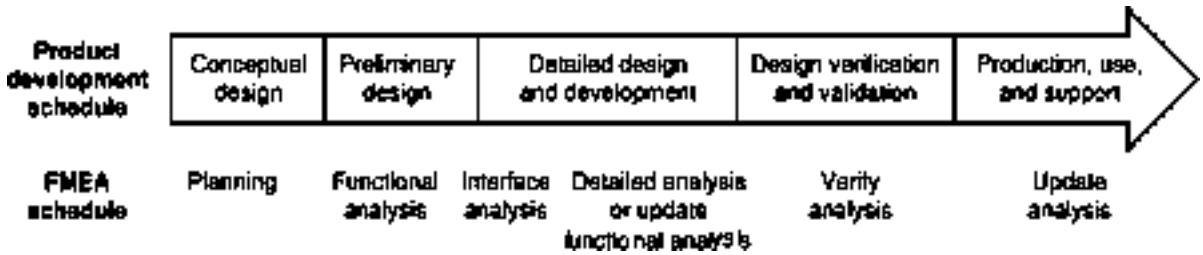


Fig. 5 Typical product development cycle and FMEA schedule (Ref 11)

As the design develops from concept to implementation, the system is subdivided into smaller subsystems. Once the subsystems are defined and their functions and interfaces are specified, the designer can focus on the design details of the subsystem, subdividing it into smaller subsystems until a solution is found. The conceptual subdividing of the system often parallels its subdivision into physical modules, and it is important that the subsystems be defined with a thorough knowledge of the technology (hardware, software, and materials) that will be used to construct each subsystem. The

FMEA helps to keep the designer, who is focusing on subsystem details, aware of the system-level effects of his or her design decisions. Like the design, the FMEA evolves from a functional to a detailed point of view (Fig. 6).

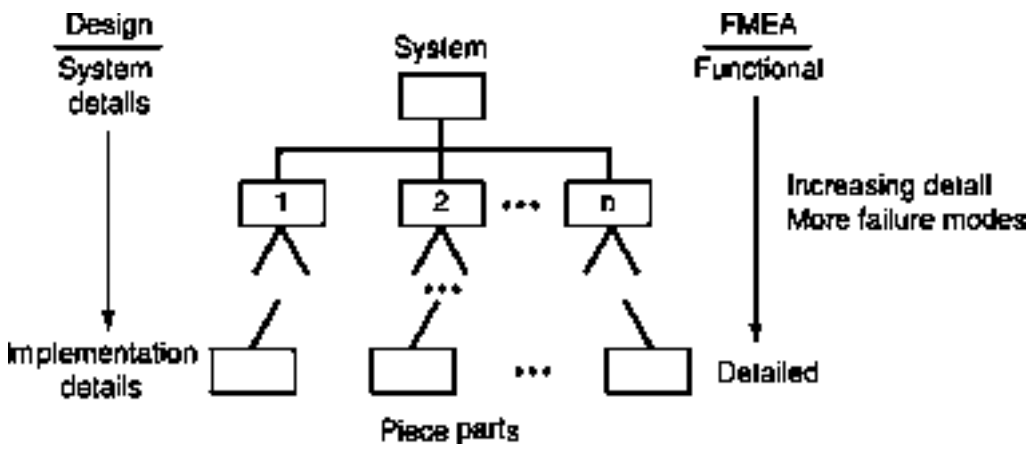


Fig. 6 FMEA tracks the design

Since the FMEA is concerned with the overall system behavior, it can serve as a unifying element for the different engineering groups involved with the product design. Such groups are often discipline specific and focus on their own areas of responsibility while ignoring other aspects of the design. For example, electrical engineers rarely address the design problems associated with vibration or heat transfer, and mechanical engineers sometimes forget effects such as cross talk in cabling. The FMEA provides a common communication tool for all the groups that must be involved with the design of the product: product designers, manufacturing engineers, test engineers, reliability and maintainability engineers, logistic support personnel, users, and others. It identifies potential single-point failure modes of system components and keeps critical items visible throughout the design process. It is useful for identifying the types of tests and testing environments needed to certify whether a design is suitable and as a basis for evaluating the adequacy of changes in product design, manufacturing process, or materials.

During the design process, many different users contribute to the FMEA, and it must meet their needs. The most important of these are discussed subsequently (Ref 23).

System Design. Design has overall responsibility for engineering, analysis, and the detailed design of the product. Because the fundamental reason for doing a FMEA is to improve the product design, the system designer is the most important FMEA user. He or she uses the FMEA to ensure that the results of previous “lessons learned” in the areas of product failure modes, causes of failures, and their effects have been addressed and then designs the product to remove or compensate for any unacceptable failure modes.

The designer contributes to the FMEA by providing detailed product descriptions to the analysis. A meaningful FMEA requires a close working cooperation between the reliability engineer, who has overall responsibility for coordinating all FMEA data, and the system designer, who has the detailed product knowledge to analyze the design and the ability to make the design changes indicated by the analysis.

Reliability. Reliability engineering is responsible for providing analyses to assess the probability that the product will successfully perform its intended function and mission. This organization usually has responsibility for developing the FMEA and maintaining the FMEA database. This entails coordinating the input from design engineering with that from other reliability, maintainability, and support functions. Failure modes developed from the product design specification must be analyzed along with those that are not directly under the designer's control—for example: failures resulting from improper maintenance, operation outside specification limits, or deviations in the manufacturing process.

The completed FMEA is a basis for future design and cost trade-off studies. Documentation supporting changes in design should compare the potential failure modes and compensating provisions of the new design to those of the baseline design.

System Safety. System safety engineers analyze how a product can be safely operated and maintained in its intended operating environment. They develop system hazard analyses identifying catastrophic and severe hazards and the component failures that can cause them. The FMEA is a source for identifying failure modes and causes that can lead to these hazards, and quantifying the hazard probabilities within the failure categories.

Maintainability. Maintainability engineering works with design to ensure that the product can be maintained by the customer in an efficient and cost-effective manner. This function requires analysis of part removal, replacement, teardown and buildup of the product in order to determine the time to do the operation, the skill levels required, the type of support equipment needed, and the documentation required. Maintainability engineers use the FMEA to identify potential maintenance tasks for which detailed task analyses are then developed. Potential preventive maintenance tasks, inspection schedules, and part-replacement schedules, can also be derived from the lists of compensating provisions and estimates of times to failure.

Logistics Support. Logistics support is responsible for ensuring that the product can be supported efficiently by the end user. This includes spare parts provisioning, support equipment, technical publications, and in some cases the provisioning of maintenance facilities, training of maintenance personnel, and packaging, handling, storage, and transportation of the product. The FMEA data help to identify maintenance tasks and other support activities, such as spares provisioning, support-equipment needs, and necessary publications.

The scope of the failure modes required by the logistics support function is considerably more extensive than just those produced by design engineering—they include failure modes induced as a result of the way the product is operated or maintained. (However, because the product design can strongly influence these types of failure modes, the product designer must also be made aware of them.)

Manufacturing. Manufacturing engineers design the sequence of operations through which the product will be produced. They also participate in defining the "critical item" list, considering items that have special producibility concerns. The FMEA provides manufacturing engineers with failure modes and effects which are used to refine the manufacturing processes to prevent manufacturing defects. In some cases, manufacturing may recommend design changes to reduce potential failure modes caused by manufacturing processes.

The manufacturing engineer identifies product-related, process failure modes, and assesses their system effects in what is often called a *process FMEA*. He or she also identifies both the manufacturing or assembly process causes of the failure mode and important process variables to monitor in order to reduce or detect the failure condition. Just as the design engineer has the detailed knowledge and ability to make *design*-related changes to reduce or eliminate unacceptable failure modes found in the (product) FMEA, the manufacturing engineer has the detailed process knowledge and ability to make *process* changes, or impose process monitors, to eliminate unacceptable failure modes identified in the (process) FMEA.

Systems. Systems engineering uses the FMEA to facilitate trade-off studies involving the various disciplines that contribute to product development. In the early design phases, failure modes identified in the FMEA might require system-level design changes to control or mitigate their effects. For example, system-level changes in technology, such as using computer control instead of analog electronic or hydraulic control systems, might eliminate or reduce the system effects of some component failure modes; they may also introduce new failure modes, which must be analyzed. When systems built using "old technology" are upgraded, the potential failure modes and their effects identified in the FMEA are important for ensuring compatibility between the old system and its replacement.

Testability and Quality Assurance. The FMEA provides test engineers with information on anticipated failure modes and causes, against which to match fault detection and isolation capabilities. These are used to ensure that all failures can be detected, that appropriate tests are developed, and that built-in tests can detect and isolate all important failure modes. The FMEA can also provide information needed to develop both built-in and off-line diagnostic routines.

Project Management. The FMEA provides project managers with critical component lists and assurances that potential product reliability problems and safety risks have been identified and addressed in the product design. The FMEA data also indicate maintenance tasks that must be performed, and they substantiate the need for support equipment and other logistic support.

For some products, government agencies, such as the military and those involved in aviation and space exploration, are responsible for ensuring that users receive a system that is reliable, safe, cost-effective, and easy to maintain and use. FMEA is often one of the analyses required for readiness reviews and certification procedures to document that the designer has considered historical failure modes and causes, and has designed the product to minimize the effects of those failures.

In companies that use integrated product development teams to design their products, the design team generally includes members from most of these groups. In this case the FMEA facilitates communication among the team members and helps to ensure that a proper balance is maintained across all competing interests. The FMEA should include all the known component failure modes and their effects, including those related to the product's mechanical design, its electrical characteristics, its material composition, its production, its operation, and its support.

Throughout the design process, each failure mode is treated independently as its effects on the system reliability and safety are analyzed. Possible common mode failures in fault-tolerant systems should also be considered. Where redundant, or backup, subsystems have been proposed, the analysis should be broadened to include the failure conditions that result in the perceived need for such systems and how the failure condition will be detected, even with the redundancy or backup system. Possible design changes should be considered to eliminate or control the effects of all important failure modes. Single-point failures identified during the analysis should be uniquely identified so as to maintain their visibility throughout the design process.

References cited in this section

1. M. Campbell, "History and Background of FMECA," unpublished personal communication

2. W.R. Lomas, “Safety Considerations in the Design of Flight Control Systems for Navy Aircraft,” Paper 60-AV-34, American Society of Mechanical Engineers, 1960

3. Mil-F-18372 (Aer.), “General Specification for Design, Installations, and Test of Aircraft Flight Control Systems,” Bureau of Naval Weapons, Department of the Navy, Washington, D.C.

4. J.S. Coutinho, Failure-Effect Analysis, *Transactions New York Academy of Sciences,* Vol 26, 1964, p 564–585

5. “Fault/Failure Analysis Procedure,” ARP 926, Society of Automotive Engineers Aerospace Recommended Practice, 15 Sept 1967; ARP 926A, 15 Nov 1979

6. “Failure Mode and Effect Analyses,” Electronic Industries Association G-41 Committee on Reliability, Reliability Bulletin No. 9, Nov 1971

7. “Analysis Techniques for System Reliability—Procedure for Failure Mode and Effects Analysis (FMEA),” International Electrotechnical Commission, IEC Standard Publication 812, 1985

8. “Fault/Failure Analysis For Digital Systems and Equipment,” ARP 1834, Society of Automotive Engineers Aerospace Recommended Practice, Aug 1986

9. “Procedures for Performing a Failure Mode Effects and Criticality Analysis,” US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

10. C.S. Spangler, Systems Engineering—The Fault Analysis Process for Commercial Avionics Application, *Proc. of the Third Annual International Symposium of the National Council on Systems Engineering,* 1993, p 729–736

11. “Recommended Failure Mode and Effects Analysis (FMEA) Practices for Non-Automobile Applications,” ARP 5580, Society of Automotive Engineers Aerospace Recommended Practice, July 2001

12. J.B. Bowles, The New SAE FMECA Standard, *Proc. Annual Reliability and Maintainability Symposium,* 1998, p 48–53

13. P.L. Goddard, Validating the Safety of Real Time Control Systems Using FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1993, p 227–230

14. P.L. Goddard, Software FMEA Techniques, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 118–123

15. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and for Manufacturing and Assembly Processes (Process FMEA) Instruction Manual,” Ford Motor Company, Sept 1988

16. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual,” J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

17. C.S. Spangler, Equivalence Relations within the Failure Mode and Effects Analysis, *Proc. Annual Reliability and Maintainability Symposium,* 1999, p 352–357

18. M. Krasich, Use of Fault Tree Analysis for Evaluation of System-Reliability Improvements in Design Phase, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 1–7

19. T.A. Montgomery, D.R. Pugh, S.T. Leedham, and S.R. Twitchett, FMEA Automation for the Complete Design Process, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

20. C.J. Price, Effortless Incremental Design FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

21. C.J. Price, D.R. Pugh, M.S. Wilson, and N. Snooke, The Flame System: Automating Electrical Failure Mode and Effects Analysis (FMEA), *Proc. Annual Reliability and Maintainability Symposium,* 1995, p 90–95

22. D.J. Russomanno, R.D. Bonnell, and J.B. Bowles, Viewing Computer-Aided Failure Modes and Effects Analysis from an Artificial Intelligence Perspective, *Integrated Computer-Aided Engineering,* Vol. 1 (No. 3), Wiley, 1994, p 209–228

23. J.B. Bowles and R.D. Bonnell, Failure Mode, Effects, and Criticality Analysis (What It Is and How to Use It), *Topics in Reliability and Maintainability and Statistics, Annual Reliability and Maintainability Symposium tutorial notes,* 1999

24. "The FMECA Process in the Concurrent Engineering (CE) Environment," AIR 4845, Society of Automotive Engineers Aerospace Information Report, 18 June, 1993

25. H.E. Lambert, Report No. UCID-16238 Lawrence Livermore Laboratory, May 1973

The FMEA Process

The FMEA methodology is based on a hierarchical, inductive approach to analysis; the analyst must determine how every possible failure mode of every system component affects the system operation. The procedure consists of:

1. Identifying all item failure modes
2. Determining the effect of the failure for each failure mode, both locally and on the overall system being analyzed
3. Classifying the failure by its effects on the system operation and mission
4. Determining the failure's probability of occurrence
5. Identifying how the failure mode can be detected. (This is especially important for fault tolerant configurations.)
6. Identifying any compensating provisions or design changes to mitigate the failure effects.

The details of the FMEA analysis are captured on analysis worksheets. These worksheets provide a description of the failure modes and their consequences traceable to diagrams or other design documentation. Generally they include:

- Identification of the component being analyzed
- Its purpose or function
- The component failure mode
- The cause of the failure and how the failure is detected
- The local, subsystem, and system-level effects of the failure mode
- The severity classification and probability of occurrence of the failure mode

A FMEA normally analyzes each item failure as if it were the only failure within the system. When the failure is undetectable or latent or the item is redundant, the analysis may be extended to determine the effects of another failure, which in combination with the first failure could result in an undesirable condition. All single-point failures identified during the analysis that have undesirable consequences must be identified on the FMEA worksheets for proper disposition.

When analyzing failure effects, the analyst must also be concerned about possible failure cascades and common cause failures where a single event can lead to multiple failures. Such failures often result from the physical placement of the components rather than their operational functions. For example, a failure of a disk turbine in which the disk disintegrates and throws off pieces of broken metal could disable several independent hydraulic systems in an aircraft if they are all routed near the turbine.

In Fig. 5, the functional, interface, and detailed analyses provide tools for evaluating the design at each phase of the development cycle. The analysis iterates as the design evolves and expands to include more failure modes as more design details become available, until all the required equipment elements have been completely

defined, analyzed, and documented. Conducting the analysis in this manner enforces a disciplined review of the baseline design and allows timely feedback to the design process.

FMEA Planning. Careful planning for the FMEA tailors the scope of the analysis to the needs of the program and provides a process that efficiently identifies design deficiencies so that corrective actions or compensating provisions can be made in a timely manner. Proper planning requires that the system requirements—including the operating modes and functions of the system, required performance levels, environmental considerations, and safety or regulatory requirements—all be specified. During the planning process, data such as field reports, design rules, checklists, and other guidelines based on lessons learned, technology advances, and the history or analysis of similar systems are collected and studied. Models (often in the form of block or flow diagrams as in Fig. 2 through 4) are developed to illustrate the physical and functional relationships between system components and the interfaces within the system. These models are especially helpful for:

- Identifying the material and component technologies that are being proposed
- Identifying their characteristic failure modes
- Examining the effects of those types of failures on the system safety and operation
- Identifying potential compensating provisions in the design

The failures can then be assessed according to their effects on the operation and mission of the system, how the failures will be detected, and any compensating provisions or design changes needed to mitigate the effects of the failures.

The complexity of the analysis tasks makes it essential to establish a set of ground rules for the analysis as early as possible. These rules help to ensure the completeness, correctness, and consistency of the analysis. Ground rules identify assumptions, limitations, analysis approach, boundary conditions, failure criteria for fault models, and what constitutes a failure (in terms of performance criteria, success/failure criteria, or interface factors) (Ref 11). They also identify requirements to be verified, possible end-item support equipment (e.g., operational or ground support, maintenance support, special test equipment, etc.), lowest indenture level to be analyzed, assumed environmental conditions, possible mission objectives and modes of operation, risk factors defined by system safety analyses, and so on. Some of the ground rules for the analysis of the hot water heater included the assumption of a continuous supply of gas and water; that the water heater was intended for home use, and that it operated between 60 and 80°C (140 and 180°F).

It is also useful to define libraries with descriptions of failure modes and consequences. Such libraries help control the analysis process and ensure consistency in terminology, types of failure modes considered, and so on among all the analysts (including future analysts) contributing to the project. They also provide direction as to the level of detail for the analysis while ensuring a more consistent and uniform documentation. The following libraries should be developed for the FMEA:

- Functional, interface, and detailed failure modes for each item type
- Mission phases and operating modes
- Effects that each failure mode has on the overall system and on the next-higher indenture level above the postulated failure mode
- Descriptions with which to classify the severity of each failure mode's effect on the end-item
- Monitor descriptions that identify how a failure mode is detected

Functional Fault Analysis. A functional fault analysis is performed on the conceptual design to verify that provisions to compensate for component failures are both necessary and sufficient. The relief valve in the hot water heater is an example of such a compensating provision; it is intended to relieve excessive pressure that could potentially rupture the tank if the water overheats. A functional FMEA of the hot water heater would show that it is needed and that it provides the required protection. The relief valve also illustrates the concern about undetected failures: How does one know that the valve is operating properly? Thus, in addition to showing that the valve is needed, the functional FMEA might also result in a requirement for periodic inspections of the relief valve or a requirement for a monitor to detect a failure of the valve.

A functional analysis begins with a functional block diagram or equivalent system representation. The block diagram indicates the input/output transfer function, the flow of information, energy, force, fluid, and so on within the system, and the primary relationship between the items to be covered in the analysis. Functional

failure mode models are assigned to each block resulting in the list of postulated failure modes to be analyzed. Then each function is analytically failed in each of its failure modes to determine the effects and characteristic indications of failure mode in each applicable operating mode.

Typical functional failure modes are of the form “function fails to perform,” “function is continuously performed,” or “function is performed at the wrong time.” In the hot water heater example, the function of the stop valve is to control the flow of gas in full-on/full-off mode. Its functional failure modes are a failure to control the flow of the gas in that:

- Gas is on when it should be off.
- Gas is off when it should be on.
- Gas is not full-on or full-off.

Ideally, a functional fault analysis focuses on the functions that an item or group of items perform rather than the characteristics of the specific components used in their implementation. In practice, the types of failure modes considered for a function may depend on how the function will be implemented. When a functional analysis is applied to manufacturing processes, typical failure mode categories include manufacturing and assembly operations, receiving inspection, and testing. Process failure modes are described by process characteristics that can be corrected. For example, part misorientation, part hole off-center, binding, cracked, and so on. When the analysis is applied to software, typical functional failure modes are: (a) failure to execute; (b) incomplete execution; (c) execution at an incorrect time (early, late, or when it should not have been executed); (d) incorrect result. For some software, the effects of other failure modes may also have to be assessed. For example, the analysis of a real-time system may require an assessment of interrupt timing and priority assignments.

As the design details are developed, the functional block diagrams and analyses are expanded, and the analysis iterates until all the system elements have been completely defined and documented. Any undetected failures that cause loss of system-level functions are corrected by incorporating requirements for compensating provisions into the design and revising the functional fault analysis to reflect the modifications.

A major benefit of a functional FMEA is that the functional failure modes can be identified in the conceptual design before the detailed design has been developed. Thus the analysis is aimed at influencing the design before the construction of any hardware. Typical results of the analysis identify functional failure modes that need to be eliminated or mitigated by changing the functional design of the system (Ref 10, 11).

Interface Fault Analysis. The interface fault analysis focuses on determining the characteristics of failures in the interconnections between subsystem elements. Cables, plumbing, fiber-optic links, mechanical linkages, and other interconnections between subsystem modules provide the basis for the postulated failure modes. Each type of interconnection has its own set of potential failure modes.

The interface fault analysis begins by defining the specific failure modes of the interfaces between subsystem elements. Typical electrical failure modes are “signal fails in the open condition,” “signal fails in the short condition,” and “input or output shorted to ground.” Typical mechanical failure modes are “piping fails in the closed position,” and “hydraulic pressure low.” Software interface failure modes focus on failures affecting the interfaces between disparate software and hardware elements. The four failure modes most often applied to software interfaces are: (a) failure to update an interface value; (b) incomplete update of the interface value; (c) update to interface value occurs at an incorrect time (early or late); and (d) error in the values or message provided at the software interface. Other failure modes specific to the software or the interface hardware may also need to be considered. Process errors that could result in misalignment or improper connection of parts are an important failure mode for a process interface FMEA. These types of errors are often eliminated by designing interconnections that can be made in only one way or ones for which any orientation is valid.

Because the interface analysis involves the interfaces between subsystem elements, it is the responsibility of the system integrator to ensure that the analysis is complete. The subsystem designer is responsible for assessing the effects of all inputs to the subsystem. The integrator uses the results of these analyses to determine the effects of the interface failure modes on the subsystem and system.

The advantage of a separate interface fault analysis is that it can be performed before detailed module designs are available; it can begin as soon as the subsystem inputs, outputs, and their interconnections are defined. Typical results of this analysis are interface failure modes that need to be eliminated or mitigated by interface design changes.

Detailed Fault Analysis. A detailed fault analysis is used to verify that the design complies with the system requirements for: (a) failures that can cause the loss of system functions; (b) single-point failures; (c) fault detection capabilities; and (d) fault isolation. It uses component failure modes postulated from the individual components in the detailed design. This includes the physical devices in the design, software modules, and the processing steps to produce the item.

A detailed fault analysis on the system hardware has traditionally been done in a FMEA. Table 1 is an example of such an analysis. This type of analysis is sometimes called a *piece-part fault analysis* because it is done on the "piece parts" that compose the system. To perform the analysis, an established set of component failure modes and their corresponding occurrence ratios are especially useful. For example, failure modes normally considered for a capacitor are "open," "short," and "leaking"; for an integrated circuit they are "output pin stuck high" and "output pin stuck low." For a bearing, they are "binding or sticking," "excessive play," and "contaminated." The failure mode ratios allow the item failure probability to be apportioned among its failure modes to give the failure mode probability of occurrence. Failure mode ratios are best obtained from field data that are representative of the particular item application but when such data are not available, generic references such as *Failure Mode/Mechanism Distributions 1997 (FMD 97)* (Ref 26) can be used for guidance. Failure mode ratios for a particular component type may vary depending on the operating environment, manufacturer, application, and other factors.

A detailed software fault analysis is intended to assess the effect of failures in the software functionality or in the variables used in the software. Unlike hardware, the software errors found by the FMEA are not due to potential failure modes; rather, they already exist and will be activated under the "right" set of conditions. A software fault analysis is applied to the as-implemented code. The failure modes used in the analysis include errors in the code that implements the software function, such as algorithm singularities and failure modes for each software variable. The variable failure modes must form a logically complete set of the possible error states for the variable type. For example, a Boolean variable with validity flag would have the following failure modes: "value true when it should be false and validity flag set to valid," "value false when it should be true and validity flag set to valid," and "value is correct but validity flag is set to invalid." Extensive lists of failure modes for many common variable types are included in Ref 11, 13, and 14. Due to the usually large number of variables and possible states, and the associated cost of performing a detailed FMEA on even a small piece of software, the analysis is most usefully applied to small embedded systems that have little or no memory protection and computational error checking provided by the hardware.

Process-related failure modes are specific to the manufacturing or maintenance process. For example, a wax coating may be applied too thinly to provide the corrosion protection it is intended to provide.

One problem associated with a detailed fault analysis is that the level of detail required to do the analysis means that it cannot be initiated until the design has matured to the point that detailed schematics and parts lists are available. This means that any major errors found by the analysis are likely to be very expensive to fix. Conversely, even major errors in the design concept are relatively easy to fix in the early design phase when the functional and interface fault analyses are done.

Identify Failure Consequences. The consequences of a failure mode analyzed in a FMEA are its effects, a classification of the severity of the failure mode based on its system-level effects, and the probability of the failure mode occurrence. The analysis is conducted for all phases and modes of system operation including normal operating modes, contingency modes, and test modes, and with respect to the primary and secondary mission objectives. The local, next-higher, and end-level failure effects of each item failure mode must be determined, and corrective actions or compensating provisions must be identified within each applicable operating mode.

Assessment of the failure mode effects must identify the system conditions or operational modes that manifest the anomalous behavior. For example, failures in the landing gear of an airplane that would cause "loss of landing gear extension" will have significantly different effects if the failure occurs on the ground than if it occurs while attempting to land. Likewise, the discovery mechanism for detecting the failure may be different.

The analysis identifies the effects of each postulated failure mode in a bottom-up manner, beginning with the lowest-level items identified. The effects of each failure mode are evaluated with respect to the function of the item being analyzed. Because the item failure under consideration might impact the system at several levels of indenture, the failure effects are then related to the functions at the next-higher indenture level of the design, continuing progressively to the top or system-level functions.

The local effect(s) description gives a detailed accounting of the impact the failure has on the local operation or function of the item being analyzed. The fault condition is described in sufficient detail that it can be used with the next-level effects, end-effects, and detecting monitor(s) to identify and isolate the faulty equipment, thus providing a basis for evaluating compensating provisions and recommending corrective actions.

Next-level effects describe the effect the failure has on the next-higher level operation, function, or status. Descriptions of the next-level effects are normally compiled in a table for consistency of annotation. The failure effect at one level of indenture is the item failure mode of the next higher level of which the item is a component.

End-effects describe the effect the failure has on the ability of the system to operate and properly complete its mission. End-effects also provide a “go/no-go” assessment of system capability to perform its intended mission. The system-level, failure effect descriptions are best derived from the system requirements and compiled in a table for consistency of annotation.

Failure modes are usually classified by an assessment of the significance of the end-effect on the system operation and mission. The FMEA ground rules should provide a ranking and classification system, and appropriate criteria for assessing the severity of failures for the product being analyzed. Often a four-level classification system developed for military equipment with severity classifications ranging from “catastrophic” to “minor” is used (Ref 9). The automobile industry generally uses a ten-level classification system (Ref 16).

Classifying a failure mode and ranking the consequences of failure require knowledge of the system and its phases of operation. For example, in some situations a failed tire might result in nothing more than the inconvenience of having to change the tire. In other cases, a failed tire could lead to loss of control of the vehicle and much more serious consequences.

When items are redundant and there is no warning that a redundant item has failed, the severity should be assessed as if all of the redundant items have failed.

Corrective Action Recommendations. Corrective actions are needed for undetectable faults and for faults having significant consequences—for example, unsafe conditions, mission- or safety-critical single-point failures, adverse effects on operating capability, or high maintenance costs. Corrective actions may not be needed if the risks for the specific consequence(s) of a failure are acceptable based on a low enough probability of occurrence. Corrective actions generally take the form of changes in requirements, design, processes, procedures, or materials to eliminate the design deficiency.

Development of an appropriate corrective action usually requires understanding and eliminating the cause of the specific failure mode; conversely, careful analysis of the failure mode causes may suggest ways to eliminate the failure. Some examples of failure causes are:

- Incorrect material specification
- Overstressing of a component
- Insufficient lubrication
- Inadequate maintenance instructions
- Poor protection from the environment
- Incorrect algorithm (software)
- Software design errors, including software requirements errors

Special attention to the failure mode causes may be needed to ensure that proper materials are used when the operational environment is especially severe due to effects such as extreme temperature cycling, very high or very low operating temperatures, the presence of corrosive chemicals, and so on. Once a corrective action is implemented and validated, the affected fault analyses must be revised to reflect the new baseline configuration.

If a failure results in unsafe system operating conditions, warnings are necessary to alert the user and to ensure satisfactory system status before commencing operations. Monitors must be strategically located to cover all undetected catastrophic, hazardous, and single-point failures, based on the system requirements and intended uses. Once the necessary monitors have been identified, a subsequent FMEA iteration is conducted (in support of maintenance activities) to verify that any remaining undetected failure modes comply with the system fault detection requirements. Requirements for fault detection monitors are then derived to cover the remaining undetected failure modes. These monitoring requirements include operator procedures and human monitoring, as well as built-in test.

When design changes to correct a deficiency are not possible or feasible, compensating provisions must be identified to circumvent or mitigate the effect of the failure when it occurs. Such provisions are often in the form of design provisions or designated operator actions that allow continued safe operation when a failure occurs.

References cited in this section

9. "Procedures for Performing a Failure Mode Effects and Criticality Analysis," US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

10. C.S. Spangler, Systems Engineering—The Fault Analysis Process for Commercial Avionics Application, *Proc. of the Third Annual International Symposium of the National Council on Systems Engineering*, 1993, p 729–736

11. "Recommended Failure Mode and Effects Analysis (FMEA) Practices for Non-Automobile Applications," ARP 5580, Society of Automotive Engineers Aerospace Recommended Practice, July 2001

13. P.L. Goddard, Validating the Safety of Real Time Control Systems Using FMEA, *Proc. Annual Reliability and Maintainability Symposium*, 1993, p 227–230

14. P.L. Goddard, Software FMEA Techniques, *Proc. Annual Reliability and Maintainability Symposium*, 2000, p 118–123

16. "Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual," J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

26. "Failure Mode/Mechanism Distributions 1997," FMD-97, Reliability Analysis Center, IIT Research Institute, 1997

Fault Equivalence

Traditionally, FMEAs have been performed on a component-by-component basis using a tabular format much like that originally developed by Lomas. Each item is listed, followed by a list of its potential failure modes; then the consequences of each failure mode are determined by analytically simulating the operation of the system with the failed item. Worksheets, such as those in MIL-STD-1629 (Ref 9), SAE J1739 (Ref 16), and the one shown in Table 1, are organized to facilitate this type of analysis.

In a manual environment, this type of systematic analysis is important because it ensures that every potential failure mode of every component is analyzed. It also makes the analysis procedure quite laborious, leads to inconsistencies in describing identical failure effects, and has given FMEA its focus on "filling out the form," rather than focusing on what has been learned from the analysis.

When a database or other automated technology is used, much of the duplicative work associated with the FMEA can be eliminated by grouping the failure modes into equivalence groups consisting of all the failure modes that exhibit identical consequences (Ref 17). Such a group is illustrated in Fig. 7 where "no output at A," "open at B," and "open input at C" all have the same effect.

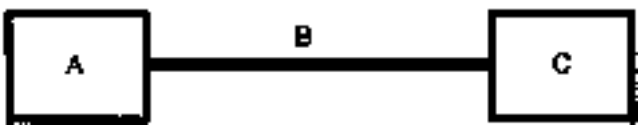


Fig. 7 Fault equivalent failure modes: A, output open; B, open; C, input open

The failure modes that exhibit identical consequences are called "fault equivalent failure modes" and grouped under a single identifying Fault (equivalent) Identifier Number (FIN).

The FMEA process starts with the functional fault analysis, progresses to the interface analysis, and concludes with the detailed analysis. Each analysis represents a more detailed iteration from the previous analyses. The common element

between the analysis types is the subsystem fault, which (along with its associated description of the local effect(s), next-level effect(s), end-effect(s), severity, compensating provisions, and detecting monitor(s)) is identified by the FIN. Because the FIN is generated on a functional basis, it allows previously generated fault information to be traced and used in subsequent analyses. For example, faults generated for the functional fault analysis are used during the interface failure mode analysis when the consequences of the interface failure modes are identical to those of the previously analyzed functional failure modes. In this case, the interface failure modes are assigned the FIN for the already recorded functional fault, thereby maintaining continuity. This process continues for the detailed analysis. Detailed failure modes that have identical consequences to previously analyzed functional and/or interface failure modes are assigned to the same equivalence group. Failure modes whose consequences are not identical to existing faults are assigned new FINs.

In the hot water heater example, the functional failure modes “gas is on when it should be off” and the piece part failure modes “stop valve fails open,” and “stop valve does not respond to controller—stays open” all have identical consequences and hence would belong to the same fault equivalence group.

Organizing the analysis by fault equivalent groups allows the functional fault analysis to be integrated with the interface and piece part fault analyses. Identification of fault equivalence groups permits the analyst to manage failure consequences in place of individual failure modes. This reduces the magnitude of the analysis effort while improving the consistency of the results. Spangler cites one study in which 12,401 failure modes resulted in 1,759 equivalent fault conditions (Ref 17).

References cited in this section

9. “Procedures for Performing a Failure Mode Effects and Criticality Analysis,” US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

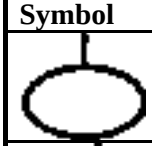
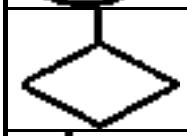
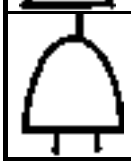
16. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual,” J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

17. C.S. Spangler, Equivalence Relations within the Failure Mode and Effects Analysis, *Proc. Annual Reliability and Maintainability Symposium,* 1999, p 352–357

The Failure Cause Model

As noted in the hot water heater example, faults are postulated at a given level of the hierarchical system model, and their effects are propagated upward to ascertain the overall system effect of each item failure mode. A failure cause model is used to assess the causes of the failure mode and evaluate the failure mode probability of occurrence (Ref 18, 27). The failure cause model often takes the form of a fault tree whose top event is the failure mode of interest. The fault tree for a failure mode at a given level is built up from combinations of subsystem failures at levels lower than that at which the failure mode is postulated. If the failure mode can result from any of several lower-level events, it is represented logically as the OR of those events; if it can result only if all of several lower level events occur, it is the AND of those events. The fault tree is most often represented graphically using logic gates such as those shown in Table 2.

Table 2 Symbols used in the construction of a fault tree

Symbol	Name	Description	Reliability model
	Basic event	Event that is not further decomposed and for which reliability information is available	Component failure mode, or a failure mode cause
	Undeveloped event	A part of the system that has not yet been developed or defined	A contributor to the probability of failure but the structure of that system part has not been defined
	Transfer gate	A gate indicating that the corresponding part of the system fault tree is developed on another page or part of the diagram	A partial reliability block diagram is shown in another location of the overall system block diagram
	AND gate	The output event occurs only if all of the input events occur	Failure occurs if all of the parts of the system fail—redundant system

	OR gate	The output event occurs if any of its input events occur	Failure occurs if any of the parts of the system fail—series system
	Majority OR gate	The output event occurs if *m* of the input events occur	*k*-out-of-*n* module redundancy where $m = n - k + 1$.
	Exclusive OR gate	The output event occurs if one input event but not both input events takes place	Failure occurs only if one, but not both, of the two possible failures occurs
	NOT gate	The output event occurs only if the input events do not occur	Exclusive event or preventive measure does not take place.

From the analysis in Table 1 the probability of the stop valve not working is the logical OR of its failure modes. This is illustrated in Fig. 8, where one should observe that the fault trees for each failure mode are developed separately. The fault tree for the failure mode, "Valve does not respond to controller—stays open" is developed in Fig. 9. A similar fault tree can be developed for each of the other failure modes. Observe that this fault tree includes operational causes of failure such as debris getting in the actuator, manufacturing process failures such as a cold solder connection, and design errors such as using too brittle wire for the connector. Thus, the failure cause model for the failure mode unifies the various types of FMEA for product, process, software, and so on.

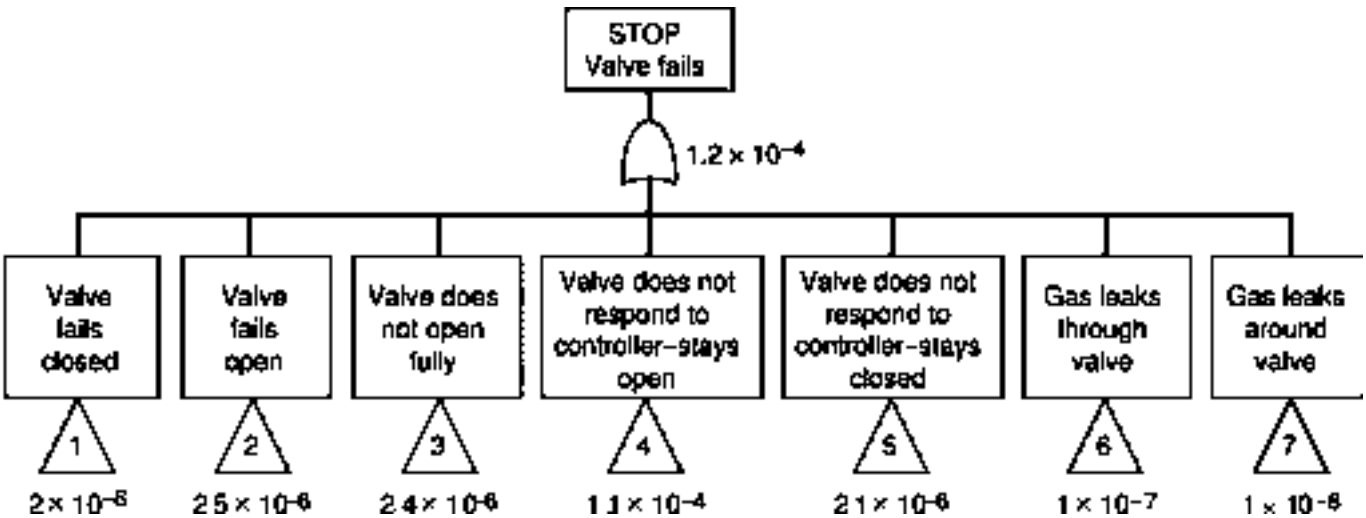


Fig. 8 Fault tree representation of the "stop valve fails" (Ref 27)

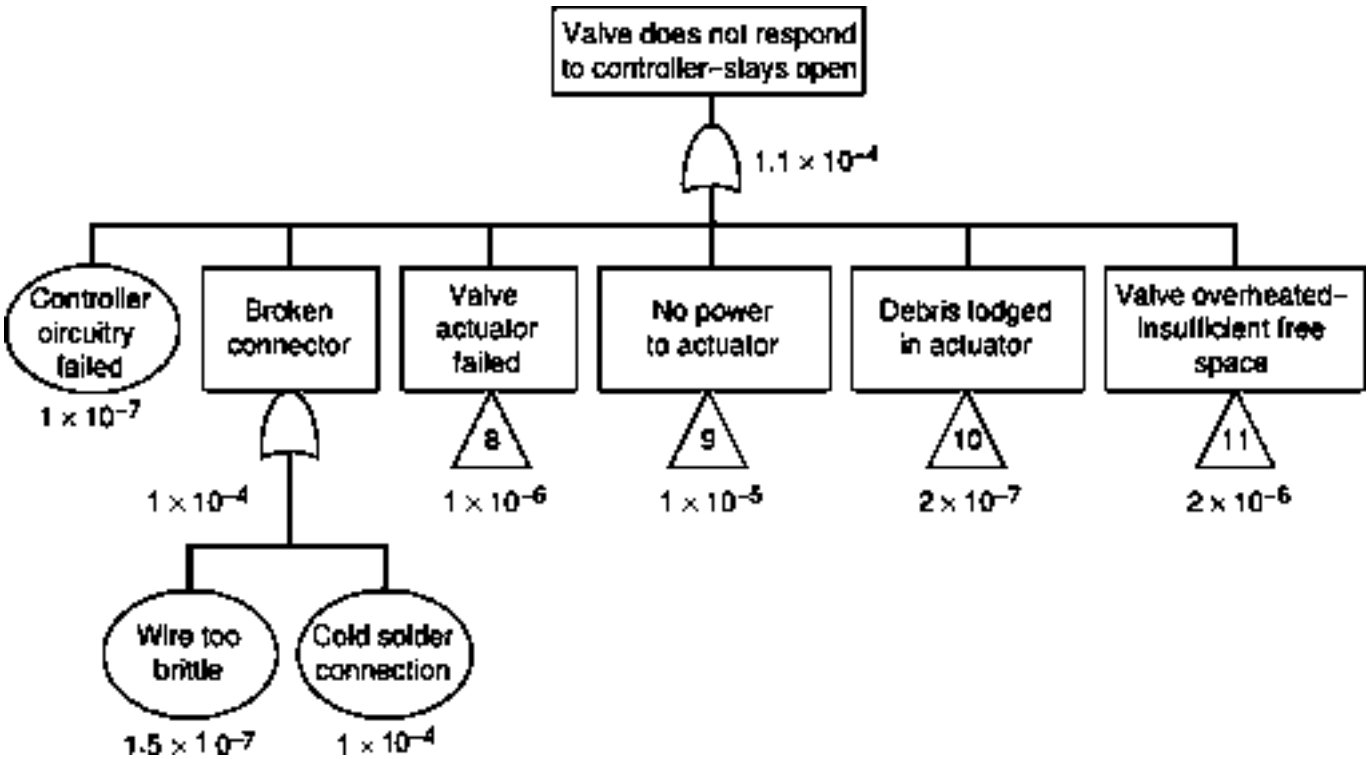


Fig. 9 Fault tree representation of the "valve does not respond to the controller—stays open" failure mode (Ref 27)

The probability of failure for each terminal failure mode can be assessed and then propagated up the fault tree to determine the failure mode probability of occurrence. The failure events in the fault tree are generally assumed to be mutually exclusive and the probabilities of the events at an OR gate are added to obtain the resultant failure probability. If the mutual exclusivity assumption is wrong, the sum must be either reduced by the joint probability or, more commonly, since joint failure probabilities are often difficult to quantify, accepted as an upper bound on the probability. Because the individual event probabilities are usually quite small, this is usually sufficient. If the fault tree includes "repeated events," a more complicated analysis using decomposition is necessary to ascertain the failure mode probability (Ref 28).

The assignment of failure mode probabilities of occurrence, based on the failure cause probabilities, is shown in Fig. 8 and 9. (The numerical values of the probabilities shown are arbitrary and for illustration only.) In Fig. 9 the top event probability of failure is the sum of the terminal event probabilities. From Fig. 9, the most likely cause of the failure mode "Valve does not respond to controller—stays open" is a cold solder connection. Thus, the manufacturing process should be examined, and appropriate controls to prevent this from occurring should be established.

References cited in this section

18. M. Krasich, Use of Fault Tree Analysis for Evaluation of System-Reliability Improvements in Design Phase, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 1–7

27. J.B. Bowles, Designing for Failure: Managing the Failure Response through Analysis, *Process of Failure Prevention through Education: Getting to the Root Cause,* J. Scutti, Ed., ASM International, 23–25 May 2000, Cleveland, p 62–69

28. D. Kececioglu, *Reliability Engineering Handbook,* Vol 2, Prentice-Hall, 1991

Automation

When done manually, FMEAs tend to be tedious and time-consuming tasks. They are error prone, and often are much disliked by product designers who generally focus on the tasks a system must do and how to design the system to do those tasks. Analyzing what will happen when something fails requires viewing the system in a different way, and propagating a failure mode from the component level to the system level is often difficult.

A number of computer-based tools and customized databases have been developed to aid the analysis process. The simplest, sometimes based on commercial spreadsheet software, do little more than provide a consistent format for the analysis. Others have features that help to improve the quality of the analysis. Some useful features of automated tools are:

- Supplying lists of failure mode and failure effect descriptions from which the analyst can choose. This feature encourages the use of common terminology and standardized descriptions in the FMEA report as well as relieving the tedium of writing such descriptions.
- Providing lists of appropriate failure modes and their apportionments for each type of component. This helps to ensure that all failure modes are analyzed.
- Making completeness and consistency checks such as: (a) checking that all appropriate failure modes for a component have been analyzed and that the sum of all failure mode apportionments is 1.0; (b) verifying that failure effects at each level have been given for each failure mode; and (c) checking terminology, grammar, spelling, and mathematical calculations in the report.
- Tailoring the FMEA report for each user's specific needs by sorting, filtering, and reformatting the data in the reports. Data can be sorted by item, failure mode, failure effect, severity class, probability of occurrence, and so on. Other useful features include providing sub-reports such as critical item lists and reformatting the FMEA so that its data can be automatically input to other databases.

The lists of failure mode and failure effect descriptions and the lists of failure modes and their apportionments for each type of item are often maintained in the analysis libraries discussed earlier, or in a database where they can be readily accessed by the FMEA analyst.

Other types of computer programs have been developed to help the engineer with other aspects of the FMEA analysis. In the later design stages, when detailed specifications are available, numerical simulation programs enable the designer to study the behavior and operational parameters of a system or subsystem. Simulation allows the designer to examine the effects of component tolerances on the system operation and, by changing component parameters to their failed state, to evaluate how various component failure modes will affect the system operation and output.

Researchers are also investigating how expert system technology can be used to help perform a FMEA. They hope that this research will yield tools that the designer can use early in the design process when errors and design weaknesses can be easily and cheaply corrected. Current efforts are focusing on the application of sophisticated database technology; the use of functional, causal, qualitative, and case-based reasoning methodologies; and the development of structural, behavioral, and functional system models. A good overview of this work is given in Ref 22 and 24.

The flame system, developed as a prototype application at the University of Wales at Aberystwyth and now marketed commercially as “AutoSteve,” is perhaps the most automated FMEA tool currently available (Ref 19, 20, 21). It is designed for use with automobile electrical systems. Flame uses both qualitative and quantitative device models for its simulation. This allows it to be used over most of the design cycle. Component models, including their possible faults, are stored in a component library, and new components can be added by the design engineer if they do not already exist. Descriptions of the functions performed by automobile circuits are also stored in a database, and the analyst need only link these descriptions to the particular states of the components in the circuit to build the model. Flame simulates the circuit with each failure, for each component, and generates the effect. The simulation is performed by following a list of events provided by the designer. Each event changes the state of the system; for example, the events “turn wiper on,” and “turn wiper to slow” will take the state of the wiper system from off, to intermittent, to slow. Flame analyzes the system by first executing each of the designated events with all components in their operational states. This provides a baseline history of expected system states. It then applies each failure mode to each of the components for each of the system states and compares the results to the correct working of the system. This comparison yields the effect of the failure mode on the system in terms of the functions that fail to operate, states of the system that are different, and components within the circuit that are in a different state than expected. The effects generator can then filter the effects and report only those of interest to the designer. Faults can be made to appear at different times in the simulation, such as during transient analysis, after establishing the direct current (dc) operating point, at a specified time, and so on. The system is also able to assign severity, detection, and occurrence values to each effect automatically (Ref 19, 21).

References cited in this section

19. T.A. Montgomery, D.R. Pugh, S.T. Leedham, and S.R. Twitchett, FMEA Automation for the Complete Design Process, *Proc. Annual Reliability and Maintainability Symposium*, 1996, p 30–36

20. C.J. Price, Effortless Incremental Design FMEA, *Proc. Annual Reliability and Maintainability Symposium*, 1996, p 30–36

21. C.J. Price, D.R. Pugh, M.S. Wilson, and N. Snooke, The Flame System: Automating Electrical Failure Mode and Effects Analysis (FMEA), *Proc. Annual Reliability and Maintainability Symposium*, 1995, p 90–95

22. D.J. Russomanno, R.D. Bonnell, and J.B. Bowles, Viewing Computer-Aided Failure Modes and Effects Analysis from an Artificial Intelligence Perspective, *Integrated Computer-Aided Engineering*, Vol. 1 (No. 3), Wiley, 1994, p 209–228

24. "The FMECA Process in the Concurrent Engineering (CE) Environment," AIR 4845, Society of Automotive Engineers Aerospace Information Report, 18 June, 1993

Conclusions

Failure Modes and Effects Analysis has evolved from an ad hoc technique, dependent on the designer's "experience," to a formal and accepted analysis technique. Failure modes can be described functionally, and the functional system model can be analyzed early in the design phase. Through fault equivalence, results from the functional and interface fault analyses can be used to reduce the amount of labor required to analyze failure modes in the detailed design phase. Fault trees can be used to develop the causes of failure and their associated probabilities of occurrence. The fault tree is able to combine failures due to design errors, material properties, manufacturing errors, service mistakes, and even user errors to give the probability with which the failure mode occurs. Failure modes can be eliminated by removing their causes or at least have their probabilities of failure reduced to acceptable levels. Researchers are making progress in developing tools to assist the designer in performing the analysis.

References

1. M. Campbell, "History and Background of FMECA," unpublished personal communication

2. W.R. Lomas, "Safety Considerations in the Design of Flight Control Systems for Navy Aircraft," Paper 60-AV-34, American Society of Mechanical Engineers, 1960

3. Mil-F-18372 (Aer.), "General Specification for Design, Installations, and Test of Aircraft Flight Control Systems," Bureau of Naval Weapons, Department of the Navy, Washington, D.C.

4. J.S. Coutinho, Failure-Effect Analysis, *Transactions New York Academy of Sciences*, Vol 26, 1964, p 564–585

5. "Fault/Failure Analysis Procedure," ARP 926, Society of Automotive Engineers Aerospace Recommended Practice, 15 Sept 1967; ARP 926A, 15 Nov 1979

6. "Failure Mode and Effect Analyses," Electronic Industries Association G-41 Committee on Reliability, Reliability Bulletin No. 9, Nov 1971

7. "Analysis Techniques for System Reliability—Procedure for Failure Mode and Effects Analysis (FMEA)," International Electrotechnical Commission, IEC Standard Publication 812, 1985

8. “Fault/Failure Analysis For Digital Systems and Equipment,” ARP 1834, Society of Automotive Engineers Aerospace Recommended Practice, Aug 1986

9. “Procedures for Performing a Failure Mode Effects and Criticality Analysis,” US MIL-STD-1629 (ships), 1 Nov 1974, US MIL-STD-1629A, 24 Nov 1980, US MIL-STD-1629A/Notice 2, 28 Nov 1984

10. C.S. Spangler, Systems Engineering—The Fault Analysis Process for Commercial Avionics Application, *Proc. of the Third Annual International Symposium of the National Council on Systems Engineering,* 1993, p 729–736

11. “Recommended Failure Mode and Effects Analysis (FMEA) Practices for Non-Automobile Applications,” ARP 5580, Society of Automotive Engineers Aerospace Recommended Practice, July 2001

12. J.B. Bowles, The New SAE FMECA Standard, *Proc. Annual Reliability and Maintainability Symposium,* 1998, p 48–53

13. P.L. Goddard, Validating the Safety of Real Time Control Systems Using FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1993, p 227–230

14. P.L. Goddard, Software FMEA Techniques, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 118–123

15. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and for Manufacturing and Assembly Processes (Process FMEA) Instruction Manual,” Ford Motor Company, Sept 1988

16. “Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) Reference Manual,” J1739, Society of Automotive Engineers, Surface Vehicle Recommended Practice, July 1994

17. C.S. Spangler, Equivalence Relations within the Failure Mode and Effects Analysis, *Proc. Annual Reliability and Maintainability Symposium,* 1999, p 352–357

18. M. Krasich, Use of Fault Tree Analysis for Evaluation of System-Reliability Improvements in Design Phase, *Proc. Annual Reliability and Maintainability Symposium,* 2000, p 1–7

19. T.A. Montgomery, D.R. Pugh, S.T. Leedham, and S.R. Twitchett, FMEA Automation for the Complete Design Process, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

20. C.J. Price, Effortless Incremental Design FMEA, *Proc. Annual Reliability and Maintainability Symposium,* 1996, p 30–36

21. C.J. Price, D.R. Pugh, M.S. Wilson, and N. Snooke, The Flame System: Automating Electrical Failure Mode and Effects Analysis (FMEA), *Proc. Annual Reliability and Maintainability Symposium,* 1995, p 90–95

22. D.J. Russomanno, R.D. Bonnell, and J.B. Bowles, Viewing Computer-Aided Failure Modes and Effects Analysis from an Artificial Intelligence Perspective, *Integrated Computer-Aided Engineering,* Vol. 1 (No. 3), Wiley, 1994, p 209–228

23. J.B. Bowles and R.D. Bonnell, Failure Mode, Effects, and Criticality Analysis (What It Is and How to Use It), *Topics in Reliability and Maintainability and Statistics, Annual Reliability and Maintainability Symposium tutorial notes,* 1999

24. “The FMECA Process in the Concurrent Engineering (CE) Environment,” AIR 4845, Society of Automotive Engineers Aerospace Information Report, 18 June, 1993

25. H.E. Lambert, Report No. UCID-16238 Lawrence Livermore Laboratory, May 1973

26. “Failure Mode/Mechanism Distributions 1997,” FMD-97, Reliability Analysis Center, IIT Research Institute, 1997

27. J.B. Bowles, Designing for Failure: Managing the Failure Response through Analysis, *Process of Failure Prevention through Education: Getting to the Root Cause,* J. Scutti, Ed., ASM International, 23–25 May 2000, Cleveland, p 62–69

28. D. Kececioglu, *Reliability Engineering Handbook,* Vol 2, Prentice-Hall, 1991

Reliability-Centered Maintenance

Dana Netherton, Athos Corporation

Introduction

RELIABILITY-CENTERED MAINTENANCE (RCM) is a systematic methodology for preventing failures. It is a specific process used to identify the policies that must be implemented to manage the failure modes that could cause the functional failure of any physical asset in a given operating context (Ref 1). In its present form, RCM has been used since the 1980s in dozens of industries at hundreds of sites around the world. Earlier forms were used by commercial aviation in the United States and in Europe starting in the 1960s.

In some ways, the name “RCM” is misleading. Despite the presence of the word “reliability” in its name, RCM is not a field within reliability engineering. Although it can be used with new assets not yet placed in service, RCM is principally intended for use with assets already built, not with assets that are still in the earliest and most fluid stages of the design process.

In addition, despite the presence of the word “maintenance” in its name, RCM is not simply a method for developing maintenance programs; its recommendations reach far beyond the maintenance department. A RCM review may trigger actions for engineering (redesign), for training and technical documentation (addressing human error due to inadequate training or inappropriate procedures), and for operations (addressing errors performed by operators, not just by maintainers).

Reliability-centered maintenance is a process that offers an opportunity for an organization to take a strategic view of its policies for managing the consequences of the failures of its assets. It answers questions such as: Can a facility avoid undesired consequences most effectively by overhauling or replacing everything periodically or by monitoring performance for signs of deterioration? If so, which assets should be monitored, which should be overhauled, and how often? What aspect of performance should be monitored? How much deterioration should trigger repair work?

Most sites have idle assets that are standing by in the event they are needed, such as backup generators or fire safety systems. Being idle, they do not wear out through use and their performance cannot be monitored, so how should they be handled?

Sometimes, no routine cyclic measure can prevent failures with intolerable consequences. What should be done in these cases?

When properly applied, RCM offers rigorous and defensible answers to each of these questions. Its approach to these questions is the result of 40 years of work in the field of physical asset management.

Reference cited in this section

1. “Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes,” JA1011, Society of Automotive Engineers, 1999

History of RCM (Ref 2)

In the late 1950s, the commercial aviation industry of the United States began to put its first jet airliners, the Boeing 707 and the Douglas DC-8, into passenger service. Industry-wide safety statistics began to be published in 1959, and they were troubling. If the jet airliners were flying at today's operating tempos, they would have an accident roughly every day.* Instead, commercial jets operated by U.S. airlines had seven accidents in the whole of the year 2000, that is, on average, about one accident every two months (Ref 3). The story behind this achievement is the story of RCM.

At that time, conventional wisdom said that reliability depended directly on scheduled maintenance; in other words, the more scheduled maintenance, the more reliable the equipment. Conventional wisdom also said that the best form of scheduled maintenance was a thorough overhaul, which was believed to make the asset like new and thus (it was believed) at its highest level of reliability. In accordance with this conventional wisdom,

the initial scheduled maintenance program for the DC-8, published in 1959, established overhaul intervals for 339 items in the aircraft in addition to periodic overhauls for the engines and for the airplane as a whole.

However, there was some evidence that this conventional wisdom had flaws. The airlines had been experiencing persistent problems with some of their reciprocating engines, which they overhauled regularly under the overall supervision of the Federal Aviation Administration (FAA) of the U.S. government. When the airlines shortened the overhaul intervals of these engines, a few became more reliable. A few more were unchanged, but many engines became less reliable.

This discovery stunned most maintenance managers. The airlines responded with a series of reliability programs in close cooperation with the FAA. These programs studied the relationship between reliability and age, first in reciprocating and jet engines and later in other complex items scheduled for routine overhaul. One lesson learned through these programs was that periodic overhauls were seldom effective with complex items. This accounted for the reciprocating engines whose reliability had been unchanged by more frequent overhauls. However, there was still a need to account for those that had become less reliable with more frequent overhauls.

As overhaul intervals grew longer, some items did fail before their scheduled overhaul. Rather than being repaired with a thorough overhaul, they were repaired by replacing or restoring the specific components affected by the failure and then were put back into service. Beginning in 1962, the airlines compared the postrepair reliability of these items with the reliability of items overhauled at the same age. There was only one difference between them: the items that were overhauled were more subject to premature failure. In other words, the overhauls were reducing their reliability. Overhauls did not simply waste resources; in many cases, they harmed the items being overhauled.

Indeed, when the results were assembled from the studies of one airline, it was clear that the relationship between reliability and age did not follow one or two models, such as simple wearout or the bathtub curve well-known in the field of reliability engineering. It followed six models. Three showed a clear, negative relationship between reliability and age (wearout, bathtub, and a linear relationship); they comprised 11% of the several hundred items studied. Two showed, in effect, a constant relationship between reliability and age; these comprised 21% of the items studied. The model that comprised the majority of the items, 68%, showed a positive relationship between reliability and age; the older the item, the more reliable it proved to be. These items did not wear out, they wore in (Ref 2).

By 1972, these results were so conclusive that the FAA permitted the airlines to abandon routine overhauls of engines and other complex items aboard airliners. Meanwhile, the continuing engine reliability problems were being solved by improving the design of the engines, not by overhauling them more often.

With the collapse of the routine overhaul as the maintenance approach of choice, no one in commercial aviation pretended that scheduled maintenance was no longer required. Instead, the airlines looked for an approach to scheduled maintenance that could be demonstrated to work. In 1964, the FAA authorized United Airlines to begin a program called “Test and Replace as Necessary.” The scheduled tasks consisted only of tests to identify reduced resistance to failure. Only those units that failed the test were removed and sent to a shop for repair. By 1969, United had qualified 209 items on various types of aircraft for this program. The program proved remarkably successful and served as a model for on-condition maintenance, that is, maintenance in which repairs are only performed on the condition that a potential failure had been detected.

By the mid-1960s, it was clear to maintenance experts in commercial aviation that scheduled overhauls were on their way out and that on-condition maintenance was on the way in. Experience with the reliability programs was also beginning to develop criteria for identifying the appropriate approach for a given item when installed in a given application. The time was ripe to consolidate the lessons into a comprehensive process for identifying the best approach for handling the failures of items aboard commercial aircraft.

The first public proposal for a decision diagram was presented at a FAA maintenance symposium in 1965. By 1968, the 747 Maintenance Steering Group, or MSG (responsible for creating the initial maintenance program for the new Boeing 747, which was about to enter service, and composed of representatives from the airlines, the aircraft manufacturer, the engine manufacturer, and the FAA), drafted MSG-1, *Handbook: Maintenance Evaluation and Program Development.* Working groups reporting to the MSG “sorted out the potential maintenance tasks and then evaluated them to determine which must be done for operating safety or essential hidden function protection. The remaining potential tasks were evaluated to determine whether they were economically useful. These procedures provide a systematic review of the aircraft design so that, in the absence of real experience, the best (maintenance) process can be utilized for each component or system” (Ref 5).

This experience led to the improved document, *Airline/Manufacturer Maintenance Program Planning Document:* MSG-2, published in 1970. It was used to develop the initial maintenance programs for the Lockheed TriStar 1011 and the Douglas DC-10. A similar document, prepared in Europe and titled "European Maintenance Systems Guide," was used for the Airbus Industrie A-300 and the Concorde. The impact of the new ideas about maintenance is evident from the number of items scheduled for routine overhaul: eight on the Boeing 747 and seven on the Douglas DC-10, compared to 339 on the initial program of the Douglas DC-8, published in 1959.

In 1972, at least one airline used MSG-2 to rewrite the maintenance program of older aircraft (Ref 2). This additional experience showed further areas for improvement; that is, MSG-2 did not make clear the role of failure consequences in establishing maintenance requirements, nor did it address the role of hidden-function failures in a sequence of multiple failures. During the 1970s, proposals emerged for ways to handle those issues. In 1978, responding to a request from the Department of Defense (DoD) of the U.S. government, two United Airlines executives wrote a 500-page technical report that described the state of the art in commercial aviation preventive maintenance. Stan Nowlan and Howard Heap named their technical report "Reliability-Centered Maintenance," and this title gave the final version of the process the name by which it is known today.

Once the DoD published Nowlan and Heap's report, the U.S. military embarked on developing RCM processes for its own use, one for the U.S. Army, one for the U.S. Air Force, and two for the U.S. Navy, because the shipboard and aviation communities of the Navy insisted that a RCM process that worked for one would not work for the other. Support contractors and equipment vendors learned to use these processes when they sold new equipment to the U.S. military. Developed in test programs during the late 1970s, the processes were formally published in military standards and military specifications in the mid-1980s.

In separate but parallel work in the early 1980s, the Electric Power Research Institute, an industry research group for the U.S. electrical power utilities, carried out two pilot applications of RCM in the U.S. nuclear power industry. Their interest arose from a belief that this industry was achieving adequate levels of safety and reliability but was massively overmaintaining its equipment. As a result, their main thrust was simply to reduce maintenance costs rather than to improve reliability, and they modified the RCM process accordingly. (So much so, in fact, that it bears little resemblance to the original RCM process described by Nowlan and Heap; it should be more correctly described as planned maintenance optimization rather than RCM.) This modified process was adopted on an industry-wide basis by the U.S. nuclear power industry in 1987, and variations of this approach were afterward adopted by various other nuclear utilities, other branches of the electricity generation and distribution industry, and parts of the oil industry.

Also in the late 1980s, other maintenance-strategy specialists became interested in RCM. The most widespread and innovative were John Moubray and his associates. Initially, they worked with RCM in mining and manufacturing industries in southern Africa under the mentorship of Stan Nowlan. Later, they relocated to Britain where they expanded their activities to cover the application of RCM in nearly every industrial sector, spanning more than 40 countries. They built on Nowlan's work while retaining its original focus on equipment safety and reliability. For example, they added environmental issues to the safety issues in the decision-making process, clarified the ways in which equipment functions should be defined, developed more precise rules for selecting maintenance tasks and task intervals, and incorporated quantitative risk criteria directly into the setting of failure-finding task intervals. They call their enhanced version of reliability-centered maintenance "RCM2."

In the 1990s, RCM began to be implemented in U.S. industries outside the military and nuclear power. In response, processes emerged that were called RCM by their proponents but that often bore little or no resemblance to the original meticulously researched, highly structured, and thoroughly proven process described by Nowlan and Heap. As a result, if an organization said that it wanted help in using or learning how to use RCM, it could not be sure what process would be offered.

Responding to this need in 1999, the Society of Automotive Engineers (SAE) published the all-industry standard JA1011, "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes." While SAE JA1011 does not present a standard RCM process, it instead presents criteria against which a process may be compared. If the process meets the criteria, it may confidently be called a RCM process; if it does not, it should not.

This section uses SAE JA1011 as its model to describe the key characteristics of a RCM process.

Footnote

* *In commercial aviation statistics, “accident” is defined as “an occurrence associated with the operation of an airplane that takes place between the time any person boards the airplane with the intention of flight and such time as all such persons have disembarked, in which the airplane sustains serious damage, or death or serious injury results from (1) being in or upon the airplane, (2) direct contact with the airplane or anything attached thereto, or (3) direct exposure to jet blast” (Ref 3). In 1959, the accident rate was greater than 30 accidents per million departures (Ref 3). In 2000, U.S. airports saw about 9.6 million departures (Ref 4). At that rate, there would have been at least 288 accidents in 1959, or roughly one per day.

References cited in this section

2. F.S. Nowlan and H.F. Heap, “Reliability-Centered Maintenance,” AD-A066 579, National Technical Information Service, 1978, Appendix B, p 46, 386

3. *Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations, 1959–2000,* Boeing Commercial Airplanes, 2001, p 4, 7, 15

4. *Aviation Industry Overview—Fiscal Year 2000,* Federal Aviation Administration, 2001, p 5

5. *Federal Aviation Administration Certification Procedures,* Federal Aviation Administration, May 12, 1972

Overview of the RCM Process

The RCM process involves asking seven questions, always asked in this order (Ref 1):

1. What are the functions and associated desired standards of performance of the asset in its present operating context? The answers to this question will list the functions of the asset.
2. In what ways can it fail to fulfill its functions? The answers to this question will list the functional failures of the asset.
3. What causes each functional failure? The answers to this question will list what RCM calls the “failure modes” of the asset.
4. What happens when each failure occurs? The answers to this question will list the effects of each failure caused by the failure mode.
5. In what way does each failure matter? The answers to this question will list the consequences of each failure caused by the failure mode.
6. What should be done to predict or prevent each failure? The answers to this question will list the proactive tasks that should be performed to address each failure mode, along with the intervals at which these tasks should be performed.
7. What should be done if a suitable proactive task cannot be found? The answers to this question will list the default actions that should be performed to address each failure mode. In the case of cyclic actions, they will also list the intervals at which these actions should be performed.

The following sections expand on each of these questions, offering definitions when necessary. The first four questions comprise a form of failure modes and effects analysis (FMEA) that is tailored specifically to meet the goals of RCM, so the next section describes the approach of RCM to FMEA. (For another perspective on FMEA, see the article “Failure Modes and Effects Analysis” in this Volume.) The following section describes the failure management policies available under RCM. Then, after a section describing the ways that RCM classifies failure effects in terms of consequences, a section describes how RCM uses failure consequences to identify the best failure management policy for each failure mode.
The closing section, “Managing and Resourcing the RCM Process,” discusses some practical issues pertaining to RCM that lie outside the scope of SAE JA1011.

Reference cited in this section

1. “Evaluation Criteria for Reliability-Centered

Failure Modes and Effects Analysis

As described in the section “Overview of the RCM Process,” RCM involves asking seven questions, of which the first four comprise a form of FMEA that is tailored specifically to meet the goals of RCM. Failure modes and effects analysis of RCM differ from the FMEA used elsewhere in engineering in several important respects.
First, the terminology of RCM (Table 1) differs from that of traditional engineering FMEA. Reliability-centered maintenance uses the term “failure modes” to refer to the events called “failure causes” in traditional FMEA. Reliability-centered maintenance uses the term “functional failure” to refer to the state called “failure mode” in traditional FMEA. These uses are entrenched in the RCM community by decades of familiarity, and it is not likely that they will change to match those of traditional FMEA any time soon. (Table 1 provides a glossary of RCM terms.)

Table 1 Glossary of terms used in reliability-centered maintenance

Term	Definition in RCM
Age	A measure of exposure to stress. Computed from the moment an item or component enters service when new or reenters service after a task designed to restore its initial capability. Can be measured in terms of calendar time, running time, distance traveled, duty cycles, or units of output or throughput
Appropriate task	A task that is both technically feasible and worth doing (applicable and effective)
Conditional probability of failure	The probability that a failure will occur in a specific period, provided that the item concerned has survived to the beginning of that period
Desired performance	The level of performance desired by the owner or user of a physical asset or system
Environmental consequences	A failure mode or multiple failure has environmental consequences if it could breach any corporate, municipal, regional, national, or international environmental standard or regulation that applies to the physical asset or system under consideration.
Evident failure	A failure mode whose effects become apparent to the operating crew under normal circumstances if the failure mode occurs on its own
Evident function	A function whose failure on its own becomes apparent to the operating crew under normal circumstances
Failure consequences	The way(s) in which the effects of a failure mode or a multiple failure matter (evidence of failure, impact on safety, the environment, operational capability, direct and indirect repair costs)
Failure effect	What happens when a failure mode occurs
Failure-finding task	A scheduled task used to determine whether a specific hidden failure has occurred
Failure management policy	A generic term that encompasses on-condition tasks, scheduled restoration, scheduled discard, failure finding, run-to-failure, and one-time changes
Failure mode	A single event that causes a functional failure
Function	What the owner or user of a physical asset or system wants it to do
Functional failure	A state in which a physical asset or system is unable to perform a specific function to a desired level of performance
Hidden failure	A failure mode whose effects do not become apparent to the operating crew under normal circumstances if the failure mode occurs on its own
Hidden function	A function whose failure on its own does not become apparent to the operating crew under normal circumstances
Initial capability	The level of performance that a physical asset or system is capable of achieving at the moment it enters service
Multiple failure	An event that occurs if a protected function fails while its protective device or protective system is in a failed state
Net P-F interval	The minimum interval likely to elapse between the discovery of a potential failure and the occurrence of the functional failure
Nonoperational consequences	A category of failure consequences that do not adversely affect safety, the environment, or operations, but only require repair or replacement of any item(s) that may be affected by the failure
On-condition task	A scheduled task used to detect a potential failure

Term	Definition in RCM
One-time change	Any action taken to change the physical configuration of an asset or system (redesign or modification), to change the method used by an operator or maintainer to perform a specific task, to change the operating context of the system, or to change the capability of an operator or maintainer (training)
Operating context	The circumstances in which a physical asset or system is expected to operate
Operational consequences	A category of failure consequences that adversely affect the operational capability of a physical asset or system (output, product quality, customer service, military capability, or operating costs in addition to the cost of repair)
Owner	A person or organization that may either suffer or be held accountable for the consequences of a failure mode by virtue of ownership of the asset or system
P-F interval	The interval between the point at which a potential failure becomes detectable and the point at which it degrades into a functional failure (also known as “failure development period” and “lead time to failure”)
Potential failure	An identifiable condition that indicates that a functional failure is either about to occur or is in the process of occurring
Proactive maintenance	Maintenance undertaken before a failure occurs in order to prevent the item from getting into a failed state (scheduled restoration, scheduled discard, and on-condition maintenance)
Protective device or protective system	A device or system that is intended to avoid, eliminate, or minimize the consequences of failure of some other system
Primary function(s)	The function(s) that constitute the main reason(s) why a physical asset or system is acquired by its owner or user
Run-to-failure	A failure management policy that permits a specific failure mode to occur without any attempt to anticipate or prevent it
Safety consequences	A failure mode or multiple failure has safety consequences if it could injure or kill a human being.
Scheduled	Performed at fixed, predetermined intervals, including continuous monitoring (where the interval is effectively zero)
Scheduled discard	A scheduled task that entails discarding an item at or before a specified age limit, regardless of its condition at the time
Scheduled restoration	A scheduled task that restores the capability of an item at or before a specified interval (age limit), regardless of its condition at the time, to a level that provides a tolerable probability of survival to the end of another specified interval
Secondary functions	Functions that a physical asset or system has to fulfill apart from its primary function(s), such as those needed to fulfill regulatory requirements and those that concern issues such as protection, control, containment, comfort, appearance, energy efficiency, and structural integrity
User	A person or organization that operates an asset or system and may either suffer or be held accountable for the consequences of a failure mode of that system

Second, RCM and traditional FMEA deliver different outputs, because they are aimed at different goals. Traditional FMEA primarily supports the asset in its initial design. At this early stage, the preferred approach to addressing failure causes/failure modes that have intolerable consequences is to change the design of the asset. Once the reliability of the design has reached the highest level achievable within the practical constraints of the design program, any failure causes/failure modes that still remain will be addressed through compensating provisions such as maintenance. Typically, a traditional FMEA will offer a great deal of information to support the design process but will offer relatively little information about the details of such maintenance (such as the type of maintenance task that should be performed or how often it should be performed).

By contrast, RCM primarily supports the asset after it has been designed and once it is to be put in use. At this stage in the life of the asset, design changes tend to be relatively expensive to research and develop. As a result, they also tend to be slow to be installed, absent a disaster that has galvanized decision makers. Therefore, the preferred approach to addressing failure causes/failure modes that would have intolerable consequences is usually to perform some sort of maintenance. For example, the preferred approach to addressing catastrophic failures in automobile engines that are caused by deteriorated lubricants is usually to replace the lubricants periodically, not to redesign the lubricants or the engine.

Therefore, a RCM review will offer a great deal of information to support the maintenance program but will offer relatively little information about the details of any redesign that might be needed (such as details of the features of the new design).

These different goals are also reflected in the failure causes that each process identifies. The failure process is a chain of events beginning far into the distant past and ending in the inability of the asset to perform its function. Which specific event should be called the cause of the failure? Typically, one selects events that can be easily controlled and that will also have a great effect on the outcome of the process.

During the design phase of the lifecycle of an asset, it is easier to control minute details of the failure process, such as whether the casing of the asset fails by brittle fracture or by ductile deformation. Once the asset is in place, that feature can only be changed by replacing its casing. So, during the operating phase of the lifecycle of the asset, casing failures are more likely to be addressed through inspections or through operational restrictions (such as setting a minimum temperature for the area around the asset).

Thus, the similarities between FMEA of RCM and traditional FMEA are superficial, and traditional FMEA seldom serves as a satisfactory replacement for FMEA of RCM.

Functions. According to SAE JA1011, the description of asset functions in a RCM process has four characteristics (Ref 1):

- It defines the operating context. The operating context is a statement of the relevant features of the context in which the asset operates. For example, it is relevant (to the RCM review) to note whether the asset is operated continuously or in cycles. It is relevant (to the RCM review) to note any production targets that the asset supports. It is relevant (to the RCM review) to note features in the surrounding physical environment that may have an effect on the behavior of the asset.
- The list of functions is a complete list of all functions of the asset, that is, all primary and secondary functions, including all protective functions. It is necessary to say this, because some processes purporting to be RCM processes named only the primary function of the asset or added only a handful of secondary functions.
- Each function statement is composed with a verb, an object, and a performance standard (quantified in every case where this can be done). For example, the primary function of a household drill might be “to drill a hole in wood, metal, or masonry, at diameters in increments of $\frac{1}{16}$ in., plus or minus $\frac{1}{64}$ in, up to the length of the drill bit, in less than 30 s.” (Where no performance standard is given, an absolute standard will be inferred. For example, the function statement “to contain fluid,” without a maximum leak rate, implies that no leaks will be tolerated.)
- The performance standards used in these function statements describe the level of performance desired by the owner or user of the asset in its operating context (not the level of performance defined by the design engineer or the purchasing department).

Readers familiar with traditional FMEA may note the presence of performance standards in the function statement of a RCM review. Traditional FMEA does not always include performance standards, partly (at least) because the design engineer who performs the analysis does not always know the level of performance desired by the owner or user of the asset in its operating context.

However, RCM was not developed for the design engineer but for those responsible for ensuring that the asset is available for productive use. If the maintenance function is to know what it needs to do to ensure that the asset is available, its people must be told what level of performance they are supposed to maintain. If the owners or users of the asset cannot tell the maintenance function what they want it to do, they cannot hold the maintenance function responsible when the asset fails to do it.

Looking at these four characteristics, it should be clear that this question is no easy challenge. It is not unusual to find that about one-third of the time spent in a typical RCM review is spent on defining the functions of the asset under review—one-third of the time is spent on one of the seven questions! It was probably that experience that moved some consultants to speed up RCM by halting after identifying only a few functions.

However, an incomplete list of functions is counterproductive and sometimes dangerous. It is not unusual to find that the solution to a particularly stubborn problem is found in this very step, which is likely to reveal when the owner or user of the asset is unwittingly demanding more of the asset than it is capable of providing. (This happens most often when the asset has been in operation for some time and has been subjected to incremental modifications without perfect consultation with all other people in the organization.) In such a case, the solution is typically very simple (either stop using it so hard or get a model with greater capacity) and is discovered very quickly indeed—long before the official end of the RCM review.

It is also not unusual for an asset to have some protective functions that have been overlooked by its owner or user in favor of its more obvious primary and evident secondary functions. When that happens, for example, with a backup generator or a locker of emergency tools, the site will be unprotected when the emergency eventually does arise. A physical asset management program that neglects its protective functions is likely to be complacent and dangerous.

Functional Failures. According to SAE JA1011, the description of functional failures in a RCM process has one characteristic: it identifies all of the failed states associated with each function (Ref 1).

If the performance standards were defined well, this exercise will be simple, yet precise. If the function of a pump is to provide at least 300 L/min of flow, then the failure of that function is to be unable to provide 300 L/min of flow.

Readers who are familiar with traditional FMEA should note again that this term corresponds to “failure mode” in traditional FMEA, albeit failure modes that always have performance standards.

Failure Modes. In RCM, a failure mode is “a single event that causes a functional failure.” According to SAE JA1011, the list of failure modes in a RCM process has five characteristics (Ref 1).

First, it identifies all failure modes reasonably likely to cause each functional failure. Note that this implies the existence of more than one failure mode for each functional failure. Indeed, there may be several events that might cause a pump to fail to deliver the desired flow rate. A RCM process will identify all of them so that it can find a way to address them all.

A RCM process will not identify all imaginable failure modes. Some people can have active imaginations and can dream up failure modes that are frankly not credible. A RCM process will have a method for weeding out the incredible speculations from the reasonably likely assessments of the RCM review.

Second, the method used to decide what constitutes a reasonably likely failure mode is acceptable to the owner or user of the asset. Different owners may need to use different levels of likelihood when screening the failure modes they should include in the review. For example, should the review include “airplane crashes on site” as a reasonably likely failure mode? While most sites will not, the owners of some nuclear power plants in the United States need to do so.

Third, it identifies failure modes at a level of causation that makes it possible to identify an appropriate failure management policy. Level of causation alludes to the fact that functional failure is the last event in a chain of events. Working backward along that chain can be described as looking at deeper levels of causation. Some failure investigators will look at the events that take place immediately before the failure; others will look at the events that take place long before the failure. In the vocabulary of RCM, those investigators are looking at deeper levels of causation.

Various RCM processes may have different methods of identifying the appropriate level of causation, with each method promising to do so more efficiently and more effectively than the others, but SAE JA1011-compliant processes will all be looking for the level of causation at which the most appropriate failure management policy can be identified.

Appropriate failure management policies are those that are technically feasible and worth doing. In RCM, these phrases are technical terms and are discussed later in this article. At this point, it is enough to remember the observation made earlier when discussing the difference between traditional FMEA and the FMEA of RCM, that design engineers and members of a RCM review group are likely to focus on different events that might be said to cause functional failure.

Design engineers have relatively more control over events that take place early in the design life of an asset, such as the selection of materials or of a specific geometry for controlling stresses in a structure; they have relatively little control over the way that the asset is operated and maintained. A RCM review group has relatively little control over the design of the asset and relatively more control over its operation and maintenance. Although the details of technical feasibility are discussed later, it should go without saying that it is not feasible to expect people to solve a problem by focusing on events over which they have little or no control.

Fourth, the list of failure modes in a RCM process includes those failure modes that have happened before, those that are currently being prevented by existing maintenance programs, and those that have not yet happened “but that are thought to be reasonably likely (credible) in the operating context.”

This is the first of two lists of types of failure modes that are often excluded from reviews. This list sorts failure modes by the type of information used to identify them, such as failure history or the current maintenance program. Some review processes focus on the most recent failure, because they are only intended to solve the problem presented by that failure—after the failure has already occurred. Some review processes examine only the existing maintenance program, because they are intended only to update or optimize that program.

However, the purpose of RCM is “to identify the policies that must be implemented to manage the failure modes that could cause the functional failure of any physical asset in a given operating context.” This purpose is much broader than simply cleaning up after a failure or optimizing a maintenance program. A process that supports this purpose may require changes in work procedures, training, or the design of the asset—and may require them before the failure has actually taken place. In this respect, RCM is truly a process that prevents failures from happening—those failures whose consequences make the effort worth doing, as described subsequently in this article.

Fifth, the list should also include “any event or process that is likely to cause a functional failure, including deterioration, design defects, and human error whether caused by operators or maintainers (unless human error is being actively addressed by analytical processes apart from RCM).”

This, the second list, sorts failure modes by type of failure mechanism. It is human for investigators to limit their examinations to the kinds of failure mechanisms they understand. A metallurgist may prefer to solve materials-related failure modes and may prefer to give human error to someone else to solve. A maintenance worker may prefer to solve the failure of workmanship and may prefer to give design defects to someone else to solve. However, RCM is intended to find solutions to problems that lead to significant failures, wherever the solutions may be found.

Although SAE JA1011 does not address the resources required to perform RCM, this aspect of RCM does suggest that the best setting for conducting a rigorous RCM review is a multi-disciplinary setting, probably in the context of a small

group of technical experts. Indeed, as is discussed later, such a setting has become the norm in most commercial applications of RCM.

Once this step has been completed, the RCM review will have identified all failure modes that are reasonably likely to cause the asset to experience functional failures in its operating context.

Failure Effects. One of the most distinctive features of RCM is its explicit assessment of whether a proposed action is worth doing, that is, whether the benefit gained from the action justifies the effort required to take the action. In order to make this assessment, it is necessary to know what would happen if the failure took place without taking any action. Reliability-centered maintenance collects that information in the form of a description of the effects of the failure.

According to SAE JA1011, the description of failure effects in a RCM process has two characteristics.

First, the description specifies what would happen if no specific task is done to anticipate, prevent, or detect the failure.

There is a temptation to describe what would normally happen if the failure mode occurred, including the normal response of maintenance and casualty personnel to signs that the failure is about to occur—a response that is intended to prevent or mitigate the consequences of the failure.

However, this part of the RCM review is intended to help the reviewers assess whether any such response is needed. After all, not all failures need to be prevented. For example, most homeowners permit their light bulbs to burn out, and they replace them after failure. So, the description of failure effects must not assume the very responses that are to be questioned.

Second, the description includes all information needed to support the evaluation of the consequences of the failure.

Such information includes any evidence that the failure has occurred. Because alarms or other indications sometimes appear after the failure itself occurs (such as low-level alarms in tanks, which may appear some time after the feed pump of the tank has failed), this information can indicate how long the effects of the failure are likely to have been developing without an opportunity for intervention.

It also includes anything the failure does that would kill or injure someone, harm the environment, hinder production or operations, or cause physical damage to other assets. These categories relate directly to consequence categories, which are discussed later.

It also includes any actions that must be taken to restore the function of the system after the failure. Broadly speaking, if a great deal of work is needed after a failure, it is more likely to be worth doing something to prevent having to do it.

Once this list is compiled, the RCM review will have a complete FMEA tailored to meet the goals of RCM.

Reference cited in this section

1. “Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes,” JA1011, Society of Automotive Engineers, 1999

Failure Management Policies and “Technical Feasibility”

Unlike some other processes, RCM considers the full range of options for managing the consequences of the failure of physical assets. This section describes the failure management policies available under RCM and explains the criteria of RCM for deciding when a specific failure management policy is technically feasible.

Determining whether a given policy is technically feasible involves comparing the technical features of the failure mode with the criteria of RCM for technical feasibility for that policy. The criteria of RCM for technical feasibility vary from one type of failure management policy to another, because they are based on the technical characteristics of each policy.

According to SAE JA1011, a RCM process offers two kinds of failure management policies: scheduled tasks and unscheduled policies. In the context of RCM, scheduled tasks consist of tasks that are performed at fixed, predetermined intervals. (Tasks that are performed at irregular intervals but that are planned in advance, such as repairs performed in response to equipment condition, are not called scheduled tasks in the RCM context.)

Scheduled Tasks. According to SAE JA1011, RCM offers four kinds of scheduled tasks: on-condition tasks, scheduled restoration tasks, scheduled discard tasks, and failure-finding tasks (Ref 1).

On-condition tasks look for a potential failure—for an identifiable condition or warning sign that indicates that a functional failure is either about to occur or is in the process of occurring. For example, checking air pressure in tires is an on-condition task. The owner or user of the car only adds air to a tire on the condition that the check revealed low air pressure. It is possible to look for potential failure by using specialized condition monitoring equipment, existing process instrumentation (such as flow rates, pressures, and temperatures), existing quality management techniques (such as statistical process control), or the human senses.

It should be clear that an on-condition task is only technically feasible if there is a potential failure that can be discovered. If failure occurs suddenly and without practical warning (often the case with brittle fractures), then an on-condition task is not technically feasible. According to SAE JA1011, a technically feasible on-condition task has five characteristics (Ref 1):

- A clearly defined potential failure does exist. If the potential failure cannot be readily recognized, it will do little good to look for it.
- There is an identifiable interval between potential failure and functional failure, known as the P-F interval (or failure development period). This and the rest of the characteristics depend on the point that it will do little good to look for the warning sign if it does not give enough time to do something useful.
- The task interval is less than the shortest likely P-F interval. Note that the definition of a "task" in a RCM review includes the definition of the interval at which the task should be performed.
- It is physically possible to do the task at intervals less than the P-F interval.
- The shortest time between the discovery of a potential failure and the occurrence of the functional failure is long enough to take predetermined action to avoid, eliminate, or minimize the consequences of the failure mode. The time between the discovery of a potential failure and the occurrence of the failure mode is shorter than the P-F interval; that is, it is equal to the P-F interval of the asset minus the interval of the task, because a potential failure might emerge soon after the task has been accomplished (with no potential failure in sight), giving the asset nearly the entire time of the task interval to develop toward functional failure. Once the next task does detect the potential failure, there might be little time left until the onset of functional failure.

Human senses often detect the most obvious potential failures and have the shortest P-F intervals and the lowest costs. Specialized inspection technology, which may detect more subtle potential failures, has the longest P-F intervals and the highest costs. Installed instrumentation can also be useful and can lie between these two extremes.

Every inspection technique has some likelihood that a single inspection might fail to detect the potential failure. This likelihood may be increased if the technique is expected to detect extremely subtle potential failures. If a user, exercising such care and skill as it is reasonable to expect in the relevant operating context, is reasonably likely to miss a potential failure with a specific technique, then the potential failure is not clearly defined (see the first characteristic of a technically feasible on-condition task, mentioned previously) and an on-condition task using that technique to focus on that potential failure is not technically feasible.

Depending on local circumstances, people using RCM might decide, instead, to use the same technology to detect more-obvious failure modes that it can detect reliably (such as looking for small cracks rather than extremely tiny cracks), accepting the shorter P-F interval that would result. They might decide to use different technology that is more reliable (if available). They might decide to abandon on-condition tasks altogether and try something else. (In this last case, the technology may still be useful when looking for other degraded areas after a potential failure has been detected, in order to define the scope of the repair work, or when performing postrepair inspections.)

Not all failure modes show potential failures, so on-condition tasks are not always technically feasible. (In fact, experience indicates that tasks using equipment to discover potential failures are technically feasible for no more than 20% of the failure modes encountered in RCM reviews [Ref 6].) Reliability-centered maintenance, therefore, considers other kinds of scheduled tasks as well.

Scheduled discard tasks discard an item at or below a specified age limit regardless of its condition at the time. For example, automobile manufacturers recommend that car owners discard their engine oil at specified intervals of miles driven.

According to SAE JA1011, a technically feasible scheduled discard task has two characteristics (Ref 1).

First, there is a clearly defined (preferably a demonstrable) age at which there is an increase in the conditional probability of the failure mode under consideration.

Conditional probability of failure is the probability that a failure will occur in a specific period on the condition that the item has survived to that period. It is possible for this probability to change over the lifetime of an item. Indeed, people perform scheduled discard tasks because they believe that the conditional probability of failure is higher after the age associated with the task than before that age. If this is true, then we can speak of a wearout age before which the item might be discarded.

However, if the conditional probability of failure remains the same—if it is constant versus age—then a scheduled discard task will offer no improvement. If the conditional probability of failure falls with age, as it did with roughly two-thirds of the items examined by the commercial airlines in their reliability programs of the 1960s and early 1970s, a scheduled discard task could harm the asset by replacing a reliable surviving part with an unreliable new part.

However, the existence of a wearout age is not enough to make a scheduled discard task technically feasible.

Second, a sufficiently large proportion of the occurrences of this failure mode occur after the wearout age to reduce the probability of premature failure to a level that is tolerable to the owner or user of the asset.

This could be called the "barn door" characteristic. An item might have an age at which its conditional probability of failure skyrockets; however, if it is highly unreliable before that age, then waiting until the wearout age to take action solves nothing. By the time such an asset reaches its wearout age, it will probably have already failed—closing the barn door after the horse has already left.

The next type of scheduled task, the scheduled restoration task, is similar to the scheduled discard task in many respects. It restores the capability of an item at or before a specified interval (an age limit), regardless of its condition at the time, to a level that provides a tolerable conditional probability of survival to the end of another specified interval.

According to SAE JA1011, a technically feasible scheduled restoration task has three characteristics (Ref 1). The first two are identical to those of a technically feasible scheduled discard task; the third characteristic is unique to the scheduled restoration task:

- There is a clearly defined (preferably a demonstrable) age at which there is an increase in the conditional probability of the failure mode under consideration.
- A sufficiently large proportion of the occurrences of this failure mode occur after the wearout age to reduce the probability of premature failure to a level that is tolerable to the owner or user of the asset.
- The task restores the resistance to failure (condition) of the component to a level that is tolerable to the owner or user of the asset. Clearly, if the scheduled restoration task is unable to restore the item to a tolerable level, it is not technically feasible.

The last type of scheduled task offered by RCM is the failure-finding task. This task determines whether a specific hidden failure has occurred.

A hidden failure is one whose effects do not become apparent to the operating crew under normal circumstances if the failure mode occurs on its own. It is most frequently associated with protective devices or functions, such as alarms or automatic shutdowns, and with standby assets, such as backup equipment or safety equipment. Such equipment can quietly rust in place or be used as a source of spare parts and then be unavailable for use if needed. The failure-finding task is not intended to prevent such failures; it is intended to detect them before the asset is needed, so that corrective action can be taken before an emergency arises.

Hidden failures are different from other failures, because they have no consequences if they occur on their own. They only have consequences in the event of a multiple failure, that is, only if the protected asset or function fails while the hidden function is in a failed state. Therefore, the evaluation of a failure-finding task takes into account some of the features of a multiple failure.

According to SAE JA1011, a technically feasible failure-finding task has four characteristics (Ref 1):

- The basis on which its task interval is selected takes into account the need to reduce the probability of the multiple failure of the associated protected system to a level that is tolerable to the owner or user of the asset. It should be clear from the previous discussion that the purpose of the failure-finding task is to reduce the probability that the multiple failure may occur and to reduce it to a level that is tolerable to the owner or user of the asset. If the task cannot do this, it is not technically feasible. Note, once again, the close association that RCM makes between the task and its interval.
- The task confirms that all components covered by the failure mode description are functional. Because the description of a hidden failure mode tends to be broad ("pump fails"), a technically feasible failure-finding task tends to be broad in scope ("test-run the pump"). This can be a challenge, because test features do not always cover all components. For example, the test button on a household smoke detector only tests that the battery and the horn are working; it does not test the smoke sensor.

- The task and its interval take into account any probability that the task itself might leave the hidden function in a failed state.
- It is physically possible to do the task at the specified intervals. Sometimes, if the owner or user is to be sufficiently confident that an unreliable protective device will not be in a failed state when the protected device fails, the protective device must be tested at intervals of seconds or minutes. Clearly, such a task is not technically feasible.

Unscheduled Policies. In the event that no scheduled task is an appropriate way to handle the consequences of a failure mode, RCM offers two alternatives: a one-time change and allowing the failure mode to progress to failure.

One-time changes represent a change to the operating context in which the asset works. This change may be applied to the physical system, such as a design change. It may be applied to the methods used to operate or maintain the asset through procedural changes (and any required refresher training). It may be applied directly to the people who operate or maintain the asset through training that increases their understanding of the asset.

Such changes usually require investment in redesign, procedural development, and training development. Investments generally require additional levels of approval before work can begin and thus are generally harder to implement than a scheduled task in a maintenance program. This is especially true for systems already in operation whose funding for initial design of hardware, procedures, and training has already been expended.

Thus, RCM tries to achieve desired levels of performance in the system as it is currently configured and operated by applying appropriate scheduled tasks (Ref 1). A one-time change is only technically feasible if such a task is not available.

In the event that no other failure management policy is appropriate, the only technically feasible alternative is to allow the asset to run to failure. Run-to-failure is an acceptable failure management policy in the RCM context, if it is also worth doing according to the criteria of RCM.

The criteria of RCM for deciding whether a failure management policy is worth doing depend on the consequences of the failure mode, so the next section addresses failure consequences.

References cited in this section

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

6. J.M. Moubray, *Reliability-Centered Maintenance,* Industrial Press, 1999, p 155, 290

Failure Consequences and "Worth Doing"

Technical feasibility is not the only factor to be used when considering whether a given failure management policy is appropriate. The other factor is whether the policy is worth doing.

Reliability-centered maintenance bases its examination of this factor on the consequences of the failure mode, first by identifying the consequences of the failure mode and then by applying the appropriate worth-doing criteria to the failure mode and the proposed failure management policy. This section describes the categories of consequences used in RCM and how RCM uses them.

Reliability-centered maintenance uses the term "consequences" to refer to the reason why a failure mode matters to the owner or user of the asset. According to SAE JA1011, RCM has five kinds of consequences: safety, environmental, operational, nonoperational, and hidden. Failure modes that are not hidden are called evident. Operational and nonoperational consequences are sometimes called economic consequences for reasons that will become apparent soon (Ref 1).

Safety. A failure mode with safety consequences matters, because it could injure or kill a human being. Failure modes that could damage equipment (including secondary damage from the failure mode) are addressed in one of the economic categories (operational, nonoperational).

This use of the term "safety" sometimes surprises newcomers to RCM who are accustomed to thinking of a failure with extensive secondary damage as a safety hazard. However, RCM handles the consequences of secondary damage and the consequences of injury or death in very different ways, as discussed subsequently.

Environmental. A failure mode with environmental consequences matters, because it might breach "any corporate, municipal, regional, national, or international environmental standard or regulation which applies to the physical asset or system under consideration" (Ref 1).

Reliability-centered maintenance defines environmental consequences separately from safety consequences, but, as described subsequently, it uses them in the same way as it uses safety consequences.

Operational. A failure mode with operational consequences matters, because it reduces the operational capability of a physical asset or system, that is, its output, product quality, customer service, military capability, or operating costs in addition to the cost of repair (Ref 1).

Reliability-centered maintenance examines the economic impact of this reduced operational capability in order to determine whether a task that addresses such a failure mode is less expensive than the failure mode itself over a comparable period of time. In some cases, the economic impact can be measured directly (for example, if it reduces production of a product that would certainly have been sold when produced). In other cases, the economic impact must be measured indirectly.

For example, the economic impact of lost military capability cannot usually be measured directly. However, when military services plan how many units to procure, they generally take into account the capability they expect to lose in assets that are unavailable due to scheduled maintenance or to unscheduled repairs.

For example, American fighter squadrons flying on missions to Europe in World War II were routinely accompanied by several spare aircraft that would slip into the formation when the failure of equipment in the high-performance (but mechanically unreliable) fighters would force some pilots to turn back before they reached enemy airspace. In some cases, mechanical breakdowns could have a great effect on military capability. When P-51 Mustangs escorted their first daylight bombing raid all the way to Berlin in March 1944, one of the three 16-plane squadrons reached Berlin with only two fighters. The rest had turned back because of equipment problems (Ref 7).

In a modern example, the United States Navy operates and maintains 12 nuclear aircraft carriers in order to have four carriers on-station around the world. With the rate of equipment failures aboard these aircraft carriers and the amount of scheduled maintenance performed on them, the U.S. Navy has found that if n number of carriers are on-station, then that same n number of carriers will be receiving shore-based repairs at any given time, on average, and that n number (again) will be travelling between their station and the shore-based repair depot. When $n = 4$, the total inventory of carriers becomes 12.

If World War II fighters had been more reliable, the military would have been able to buy fewer of them in order to have the same number of aircraft over the target. If nuclear aircraft carriers were more reliable, the military would be able to buy (and maintain) fewer of them in order to have the same number of carriers on-station. The higher procurement numbers—and the higher costs—reflect the indirect economic consequences of the failure modes that reduce the operational capability of these military assets.

Nonoperational. A failure mode with nonoperational consequences matters only because it requires the repair or replacement of one or more items. The magnitude of the consequences is measured simply in economic terms, that is, the cost of the repair, considering both parts and labor (Ref 1).

Every failure mode without noticeable safety, environmental, or operational consequences will be categorized as having nonoperational consequences.

Hidden failure modes are assigned to the category of hidden consequences. Failure-finding tasks are only considered if the failure mode is hidden, because there is no need to find failures when the failure mode is evident.

Some hidden failure modes also have safety or environmental consequences, because many protective devices or functions are installed to protect lives or the environment. The remaining hidden failure modes apply to protective devices or functions that are installed to protect equipment. These hidden failure modes are coupled with economic (operational or nonoperational) consequences.

Worth Doing. It is common for engineers to judge a technical approach on the grounds of its technical feasibility. Reliability-centered maintenance adds another assessment, that is, a comparison of the value added by the approach to the drawbacks of the approach. If the value outweighs the drawbacks, the approach is worth doing.

According to SAE JA1011, a RCM process declares a scheduled task worth doing if it satisfies the criteria in Table 2 (Ref 1). A one-time change is worth doing if the change meets the appropriate criteria in Table 3 (Ref 1).

Table 2 Worth-doing criteria for scheduled tasks

	Evident failure mode	**Hidden failure mode**
Safety or environmental consequences	The task reduces the probability of the failure mode to a level tolerable to the owner or user of the asset.	The task reduces the probability of the multiple failure to a level tolerable to the owner or user of the asset.
No safety or environmental consequences	The direct and indirect costs of doing the task are less than the direct and indirect costs of the failure mode, when measured over comparable periods of time.	The direct and indirect costs of doing the task are less than the direct and indirect costs of the failure mode, plus the cost of repairing the hidden failure mode, when measured over comparable periods of time.

Table 3 Worth-doing criteria for one-time changes

	Hidden failure mode	Evident failure mode
Safety or environmental consequences	The one-time change must reduce the probability of the multiple failure to a level tolerable to the owner or user of the asset.	The one-time change must reduce the probability of the failure mode to a level tolerable to the owner or user of the asset.
No safety or environmental consequences	The one-time change must be cost-effective, in the opinion of the owner or user of the asset.	The one-time change must be cost-effective, in the opinion of the owner or user of the asset.

Once it has been decided that a failure mode has safety or environmental consequences, it will be handled in terms of probability that it will occur. Then, that probability will be compared to the probability that is tolerable to the owner or user of the asset. It can sometimes be a challenge for owners or users to decide what probability of a dangerous failure they are prepared to tolerate. Still, if they do not know what probability they are prepared to tolerate—until they know how much protection they need—they cannot know whether the protection they will get from the task they are evaluating will be adequate.

When examining a failure mode with economic consequences, it is important to compare the costs of the failure mode against the costs of the task over comparable periods of time. It is misleadingly easy to compare the costs of one failure against the costs of performing the task one time. However, the costs of a task done many times will add up over time, and it is those accumulated costs that must be compared against the cost of an expensive but infrequent failure.

Failure modes and effects analysis, failure management policies and technically feasible, and failure consequences and worth doing are the building blocks of RCM. In the next section, they are put together to create a failure management program.

References cited in this section

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

7. G.C. Hall, Jr., *Death Squadron*, Kensington Publishing Corp., 1946, p 178, 183–84

Failure Management Policy Selection

With the list of functions, functional failures, failure modes, and failure effects as well as an understanding of failure consequences and the characteristics of the failure management policies available, all the pieces are in place to support the selection of the appropriate failure management policies for the asset under review.

According to SAE JA1011, all scheduled tasks selected in a RCM process must be technically feasible and also worth doing (Ref 1). The other options, one-time changes and a policy of run-to-failure, must meet their own specific criteria. If more than one policy meets these criteria, then the more cost-effective policy will be selected (Ref 1).

Also, policies are selected as if no specific task is currently being done to anticipate, prevent, or detect the failure (Ref 1).

The criteria used by RCM at this stage are complex, and it can be hard to keep them (and their results) straight. Most RCM processes guide their users through the process with a decision logic diagram. Indeed, the use of decision logic diagrams is so widespread in RCM that some people mistakenly believe that any process with a decision logic diagram is a RCM process. However, no decision logic diagram is the sole RCM decision logic diagram.

The most complex version—also the most highly optimized version—considers all failure management policies for each failure mode, identifies all those that are technically feasible and worth doing, and selects the one that is least expensive.

Figure 1 shows one such decision logic diagram, in this case developed by the U.S. Naval Air Systems Command (Ref 8). It begins the failure management policy selection process by determining whether the failure mode is hidden or evident. This determination makes failure-finding tasks available as an option later if the failure mode is hidden.

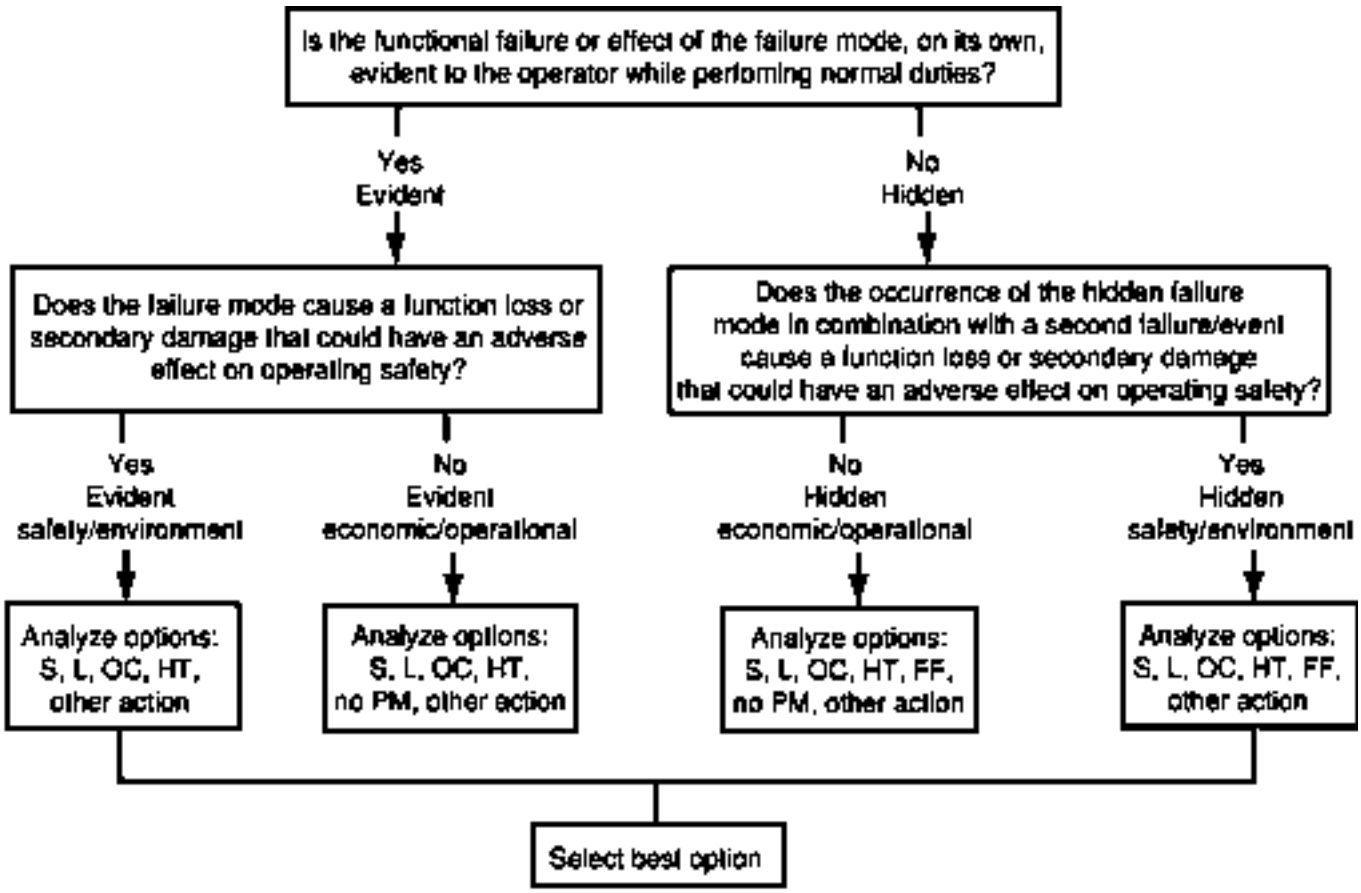


Fig. 1 Typical RCM decision logic diagram. S, servicing; L, lubrication; OC, on-condition; HT, hard-time (comprises scheduled restoration and scheduled discard); FF, failure-finding; PM, preventive maintenance. Note: S, L, and HT are aviation-unique terms and are not standard RCM terms.

It then examines the consequences of the failure mode, determining whether the consequences are safety, the environment, operational, or nonoperational (which U.S. Naval Aviation calls “economic”). This determination makes the specific worth-doing criteria available when considering each failure management policy.

In the next step, the process examines each failure management policy in turn, to determine whether the policy is technically feasible (considering the technical characteristics of the failure mode and the technical characteristics of the policy) and whether it is worth doing (considering the consequences of the failure mode and the consequences of executing the policy). (The abbreviations used to name the failure management policies in Fig. 1 reflect the slightly different vocabulary used in the RCM program of U.S. Naval Aviation.)

In the final step, the process that uses Fig. 1 selects the most cost-effective policy that is both technically feasible and worth doing.

While SAE JA1011 supports such a process, it does not require it. Indeed, most practical RCM processes stop considering failure management policies for a failure mode when the review identifies the first policy that is both technically feasible and worth doing. This is reflected in their use of decision logic diagrams in which types of failure management policies are considered in turn until a suitable policy is identified.

The decision logic diagrams used by different RCM processes show this policy examination in different ways. Figure 1 simply has a box labeled “Analyze options.” Diagrams developed for some other RCM processes spell out the steps of the analysis, that is, which policies to examine first and whether to continue looking after a policy has been found to be both technically feasible and worth doing. Some diagrams even provide a summary of the specific criteria for technically feasible and worth doing for each kind of consequence and each kind of failure management policy. The choice of a specific decision logic diagram is usually part of the choice of a specific RCM process and rests on assumptions about the level of RCM expertise held by those who are expected to use the process, because highly expert people probably need fewer hints.

When this step is concluded, the RCM review will have answered all seven questions listed earlier in “Overview of the RCM Process.” It will have a list of the functions of the asset, its functional failures, its failure modes, and the effects of those failures. It will also have a list of the failure management policies that are technically feasible for those failure modes and that are worth doing, considering the consequences of those failure modes. In many cases, the review also

gathers a wealth of supporting technical information relating to the behavior of the asset—information that is often very useful for other purposes, such as training or troubleshooting.
This concludes the discussion of the RCM process itself. However, the success of a RCM program does not depend only on technical proficiency in using the RCM process. It depends on other matters as well that are outside the scope of SAE JA1011. They are the subject of the next section.

References cited in this section

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

8. "Guidelines for the Naval Aviation Reliability-Centered Maintenance Process," Management Manual NAVAIR 00-25-403, Commander, Naval Air Systems Command, 2001

Managing and Resourcing the RCM Process

In this article, repeated reference has been made to SAE JA1011 and to RCM processes. In practice, several existing processes conform to SAE JA1011 (and are RCM processes) without being identical processes. By and large, they differ from each other in the way that they address issues outside the scope of the SAE standard, which focuses only on the technical issues relating to RCM. These additional issues pertain to the challenge of managing and resourcing the RCM process. Chief among them are:

- Prioritizing assets and establishing objectives
- Planning
- Level of analysis and asset boundaries
- Technical documentation
- Organization
- Training
- Role of computer software
- Data gathering
- Implementation

Each process offers an approach to these issues that is intended to enable an organization to perform the technical process of RCM in the most efficient way. Many organizations find the benefits of RCM so valuable that they wish to apply it to many of their assets, and as the scope of RCM expands, the efficiency of the RCM process becomes increasingly important. If an organization cannot use the process efficiently in its sites, it will find that RCM itself is not worth doing; that is, its benefits do not justify the effort needed to produce them. In practice, this has sometimes led to the collapse of the RCM program in the organization.
Therefore, when setting up a RCM program, it is important to keep in mind the following considerations.
Prioritizing Assets. If RCM is to be applied to a large number of assets at a site, the assets will have to be reviewed in an order or sequence. Some RCM processes incorporate additional steps intended to help assign different priorities to different assets. The usefulness of such additional steps is the subject of debate within the RCM community. Its proponents argue that they help produce clear thinking and clear decisions. Its opponents argue that the additional steps divert time and attention from the most important thing, the RCM reviews themselves. The most important thing, they argue, is to start doing RCM reviews on the assets that are clearly the most important.
Establishing Objectives. There is no universal objective for a RCM review. The objective of a review will depend on the deficiencies currently perceived in the operation and maintenance of the asset. The deficiencies may be related to the performance of the asset itself (safety hazards, interrupted operations, expensive repairs), or they may be related to the performance of the work done to keep the asset running (problems with training, with scheduling work, and so forth).
Because the objectives differ with each RCM review, it is prudent to identify the objectives of each RCM review on a case-by-case basis.

Planning. The RCM process and its supporting processes (discussed in this section) are complex and not likely to be accomplished efficiently without an explicit and focused plan that takes into account all of the processes and resources involved.

Level of Analysis and Asset Boundaries. Failing to find the right level of detail for analysis of failure modes can have a serious effect on the efficiency and effectiveness of a RCM review. Too high a level of detail can prevent the review from identifying all reasonably likely failure modes. Too low a level of detail can force the review to consider the same failure mode many times, thus wasting time and effort. It also makes it more difficult to identify functions and the consequences of failures.

Different RCM processes have different methods for identifying the best level of detail. Some always go straight to the level of the lowest replaceable component. Others examine everything at the system or major equipment level. Still others try to be yet more flexible. Before selecting a process, the owner or user of the asset should examine and approve the method used by that process for identifying the best level of detail.

Sometimes the best level of analysis on a large asset would produce more failure modes than a RCM review can handle in a reasonable amount of time. In such an event, it is appropriate to divide the asset into subsystems and analyze each subsystem separately. When dividing such an asset, it is important to ensure that the boundaries of each review are identified clearly and that no components are overlooked.

Technical Documentation. Existing documentation about the asset under review should certainly be gathered beforehand. Some RCM processes develop specific documentation, if not already available, such as functional block diagrams. This practice is another subject of debate within the RCM community. Its proponents argue that specific documentation always improves the clarity of discussion about the asset. Its opponents argue that it is only needed occasionally and should only be developed when needed.

Organization. An essential part of the planning required to support RCM is the organization of people who will support it. This organization must:

- Ensure that the RCM process to be used complies with SAE JA1011 and plan each RCM review
- Ensure the RCM reviews are executed as planned
- Apply the process during each review
- Provide information and assist in decision-making (by operators, maintainers, management representatives, and (on a case-by-case basis) designer/vendor representatives)
- Provide physical facilities, such as offices, meeting rooms, and computer hardware and software

One fundamental choice of organization is the option of using individual analysts or using a review team with a facilitator. Most U.S. military RCM programs were organized in the late 1970s and early 1980s, and many still rely on individual analysts who are expected to interview technical experts as-needed. Most commercial RCM programs were organized after the shift to teamwork in the 1990s, and most rely on review teams with a facilitator.

The proponents of individual analysts argue that that approach requires less time from technical experts. The proponents of teams argue that team meetings can be a more efficient way to assemble technical information, if the meetings are managed well, and that they also increase the buy-in of the people who are likely to be the leaders of those who will actually perform the recommended actions. Of course, meetings are only efficient if they are well organized and well managed, so the skill of the meeting facilitator is crucial to their success.

Training. It should be clear that RCM is different from other disciplines bearing the word “reliability” in their names. It is different enough to require specialized training. Because RCM is not routinely offered in the curriculum of undergraduate engineering programs or technical training programs, this training usually must be provided as part of any initiative to begin or support a specific RCM program at a specific site.

Some people in a RCM program require more training than others. A short orientation presentation can be enough for senior managers. Experience has shown that those who need to understand the details of the technique need at least three days of formal classroom training in order to absorb the concepts. Those who will take the lead in applying the process, such as meeting facilitators or individual analysts, require still more training, including on-site mentoring, before they will possess the necessary skills.

Role of Computer Software. Computer software can serve a useful role in recording the information and decisions that accumulate during the RCM review. If a large number of assets are to be analyzed, a computer may be the only practical way to store and manipulate these findings for later use.

Computer software can also perform some of the mathematical and statistical calculations in RCM, such as the cost comparisons that support the determination whether a task is worth doing for a failure mode with economic consequences, or the statistical calculations that identify the best task interval for a failure-finding task. However, computer software also brings two risks to a new RCM program.

The first risk is that people's attention might be diverted to the software and the tasks of populating its database and away from the real needs of the asset being reviewed. This approach can reduce RCM to a mechanical process that adds little real value, because the source of the value of RCM is its effectiveness in eliciting information about the asset and organizing people's thoughts about that information. As RCM expert John Moubray puts it, “RCM is thoughtware, not software” (Ref 6).

The second risk is that the ability of the computer to quickly perform calculations will tempt the software designer to incorporate algorithms based on processes other than RCM. Algorithms that have tempted RCM software designers include:

- Best-replacement-interval algorithms, which balance the cost of repairing failures (based on a given failure rate) against the cost of replacing an item periodically
- Condition-monitoring algorithms that combine condition trend data with age data, triggering repairs based either on evidence of a potential failure or on age

The best-replacement-interval algorithm works only if scheduled restoration or scheduled discard is the most appropriate failure management policy. It adds value (compared to simply performing the task shortly before the wearout age) only if there is a dependable historical record of failures (and their ages). Routine reliance on this algorithm rests on the assumption that both conditions are usually met.

Practical RCM programs do not accept this assumption. One of the key insights of the early efforts in the development of RCM was the discovery that scheduled restoration and scheduled discard tasks are almost never appropriate. In addition, a historical record of failures is almost never available, in practice, especially for the most important failures (which tend to be the most infrequent failures), making mathematical predictions of the cost of replacements versus age too imprecise to make such an algorithm practical.

The combination condition-monitoring algorithms are based on a common practice in vibration analysis: the use of a chart with warning and danger levels that vary with the age of the asset. As the asset ages, the levels that trigger warning and danger alerts on these charts get lower. If the asset survives long enough, these charts will force a restoration or discard task even if the condition of the asset is unchanged or improving.

These charts are only appropriate in cases where an on-condition task must be combined with a scheduled restoration or scheduled discard task in order to create an appropriate task. In practice, this happens about as often as cases where a scheduled restoration or discard task is appropriate on its own—again, almost never.

The standard SAE JA1011 has two specific requirements aimed at such risks. First, it says that the failure management policy selection process in a RCM process takes account of the fact that the conditional probability of some failure modes increases with age (or exposure to stress), that the conditional probability of failure of others does not change with age, and that the conditional probability of failure of yet others decreases with age (Ref 1).

Thus, a process that assumes that all or nearly all assets wear out (and that scheduled restoration or scheduled discard is always or nearly always technically feasible) is not a RCM process.

Second, SAE JA1011 states that the formulae used to support a RCM process are “logically robust,” which means that they are consistent with the expectations of RCM about the behavior and deterioration of physical assets (Ref 1). Clearly, algorithms with formulae that expect most physical assets to wear out over time do not comply with this requirement.

Data Gathering. A RCM review gathers and reviews information about the asset and its operating context in order to find the most appropriate ways to manage the consequences of the failure of the asset. The various steps in a RCM review require historical data about failures, the performance of the asset (including how the asset degrades over time), associated maintenance and operating costs, and the performance of scheduled maintenance. It also needs information about existing scheduled maintenance tasks and about the consequences of failure.

Sometimes, adequate information about the failure modes that may occur and how often they might occur is simply not available (especially with assets containing large amounts of new technology). When the consequences of this uncertainty are intolerable, it can be tempting to ignore the technical feasibility of tasks

and “just do something.” A better choice is to take action that changes the consequences so that the uncertainty no longer matters. For example, if the design of the asset makes it difficult to be sure of detecting warning signs that the failure mode is taking place, a different design might eliminate that problem. If the reliability of new technology is not known precisely enough to be able to assess whether a scheduled task is worth doing, the installation of a backup unit (which will change the nature of the consequences to hidden) may make it possible to address failures through a failure-finding task.

Once this information has been gathered for the review, it should be kept current so that the recommendations of the review can be quickly updated later. The owners of the asset should avoid the common mistake of recording only the repair work performed after a failure and should place equal emphasis on recording the causes of functional failure and the associated consequences (such as equipment downtime).

Implementation. After the RCM review has been completed (or updated), its results must be implemented. Organizations that fail to implement results fail to reap any benefits from their review. Successful implementation involves three key steps:

- Obtain management support for the results. This entails auditing the results to ensure that they are technically valid and that they meet the goals of the management. It also entails presenting the results to management and inviting their support for implementation.
- Plan the implementation of the results. This entails writing descriptions of the scheduled tasks and one-time changes in enough detail to ensure that the work will be done correctly by the people who will do it.
- Execute the implementation. Make the one-time changes and plan and execute the scheduled tasks.

These management and planning activities are outside the scope of the RCM process itself, so those who wish to learn more about them should look for references about maintenance management and maintenance planning.

References cited in this section

1. “Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes,” JA1011, Society of Automotive Engineers, 1999

6. J.M. Moubray, *Reliability-Centered Maintenance*, Industrial Press, 1999, p 155, 290

Conclusions

From its origins in the quandaries of airline maintenance managers in the late 1950s and early 1960s to its worldwide use in nearly every industry on the face of the earth today, RCM has grown into a formidably powerful tool for identifying the causes of asset failures and the most appropriate ways to manage their consequences before the failures occur, thus truly preventing the intolerable consequences of failures outright.

In its systematic examination of failures, failure causes (or failure modes), their consequences, and the failure management policies available in a given operating context, RCM keeps in mind these key points:

- Most assets do not wear out at a consistent age; whereas, a considerable number of assets give warning signs before they fail. As a result, RCM always examines the suitability of on-condition maintenance (checking for signs of potential failure), whether or not it examines the suitability of scheduled restoration or scheduled discard tasks.
- Not every failure needs to be prevented. The consequences of preventing the failure may be worse than the consequences of accepting the failure, especially if the asset tends to be less reliable right after it is worked on (a very common experience). As a result, RCM always checks to see whether a scheduled task is worth doing, even if the task has already been shown to be technically feasible.
- The key to solving reliability problems lies in finding the cause of the failure, not merely in finding the mode or manner of failure. A functional failure is the last event in a chain of events. The goal of failure prevention (and of RCM) is to identify the right point at which to intervene and break the chain.

- Further, the key to finding the cause of the failure is understanding the level of detail required in the current context. In the design context, "component A3 breaks" may be a satisfactory "cause of the failure of asset A," and a satisfactory solution may be to specify a more reliable version of component A3. However, in an in-service context, where equipment is out of warranty and fully paid for, a design change may be difficult to justify in any but the most urgent cases. In such a context, a deeper level of detail may be required in order to identify an appropriate scheduled maintenance task that will be just as technically feasible as a design change but more cost-effective (and thus more likely to be approved for implementation). Reliability-centered maintenance always examines the suitability of scheduled tasks, whether or not it examines the suitability of design changes.
- If the consequences of the failure are intolerable and scheduled tasks cannot make them tolerable, the effort must not stop there. One-time changes must be examined until the consequences can be made tolerable. (In rare cases, the only appropriate one-time change may be to cease operating the dangerous asset entirely.) Reliability-centered maintenance is aimed at the reliability of the asset, not simply the maintenance program of the asset.
- Hidden failures matter, too. Backup and safety equipment are often ignored until they are needed, at which point it may be too late to fix problems. As a result, RCM explicitly asks whether a failure is hidden. If it is, RCM adds failure-finding tasks to the list of options available.

Through these questions, along with the supporting criteria that apply these points, RCM enabled U.S. and Canadian airlines to reduce the crash rate of their jet airliners from more than 30 per million take-offs to less than one—a reduction of more than 3,000%. It has enabled international industry to improve the reliability and availability of their industrial assets on sites around the world, in nearly every field that depends on having its assets being reliable and available for use. It remains the most thorough and flexible tool in the world for identifying and addressing failures before they happen—for truly preventing the failures that matter.

References

1. "Evaluation Criteria for Reliability-Centered Maintenance (RCM) Processes," JA1011, Society of Automotive Engineers, 1999

2. F.S. Nowlan and H.F. Heap, "Reliability-Centered Maintenance," AD-A066 579, National Technical Information Service, 1978, Appendix B, p 46, 386

3. *Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations, 1959–2000,* Boeing Commercial Airplanes, 2001, p 4, 7, 15

4. *Aviation Industry Overview—Fiscal Year 2000,* Federal Aviation Administration, 2001, p 5

5. *Federal Aviation Administration Certification Procedures,* Federal Aviation Administration, May 12, 1972

6. J.M. Moubray, *Reliability-Centered Maintenance,* Industrial Press, 1999, p 155, 290

7. G.C. Hall, Jr., *Death Squadron,* Kensington Publishing Corp., 1946, p 178, 183–84

8. "Guidelines for the Naval Aviation Reliability-Centered Maintenance Process," Management Manual NAVAIR 00-25-403, Commander, Naval Air Systems Command, 2001

Products Liability and Design

Charles O. Smith, Engineering Consultant

Introduction

PRODUCTS LIABILITY is a legal term for the action whereby an injured party (plaintiff) seeks to recover damages for personal injury or property loss from a producer and/or seller (defendant) when the plaintiff alleges that a defective product caused the injury or loss.

If a products liability suit is entered against a company, the plaintiff's attorney and technical experts attempt to convince a jury that the manufacturer did not exercise reasonable care in one or more features of design and/or manufacture, and that because the company did not exercise reasonable care, an innocent party was injured. The defendant's team attempts to convince a jury that the manufacturer was not responsible for the injury.

Products liability is not new. The first law code known to be in writing was established by Hammurabi, King of Babylon, about 4000 years ago, and it contained clauses that clearly relate to products liability (Ref 1).

Who may be a plaintiff? Essentially any consumer, user, or bystander may seek to recover for injury or damages caused by a defective and unreasonably dangerous product. Who may be a defendant? Any corporation, business organization, or individual who has some degree of responsibility in the “chain of commerce” for a given product, from its inception as an idea or concept to its purchase and use.

The situation is schematically summarized in Fig. 1. The editors believe it is important for some general background on the legal aspects of failure analysis to be included in this Volume. This article is not intended to provide legal conclusions, and the opinions are those of the author. Because every situation is different, this article should not be applied to any specific product without additional analysis. It is written from an engineering perspective, and the author is not a lawyer. Designers may wish to consult with counsel to better understand what the law requires of them.

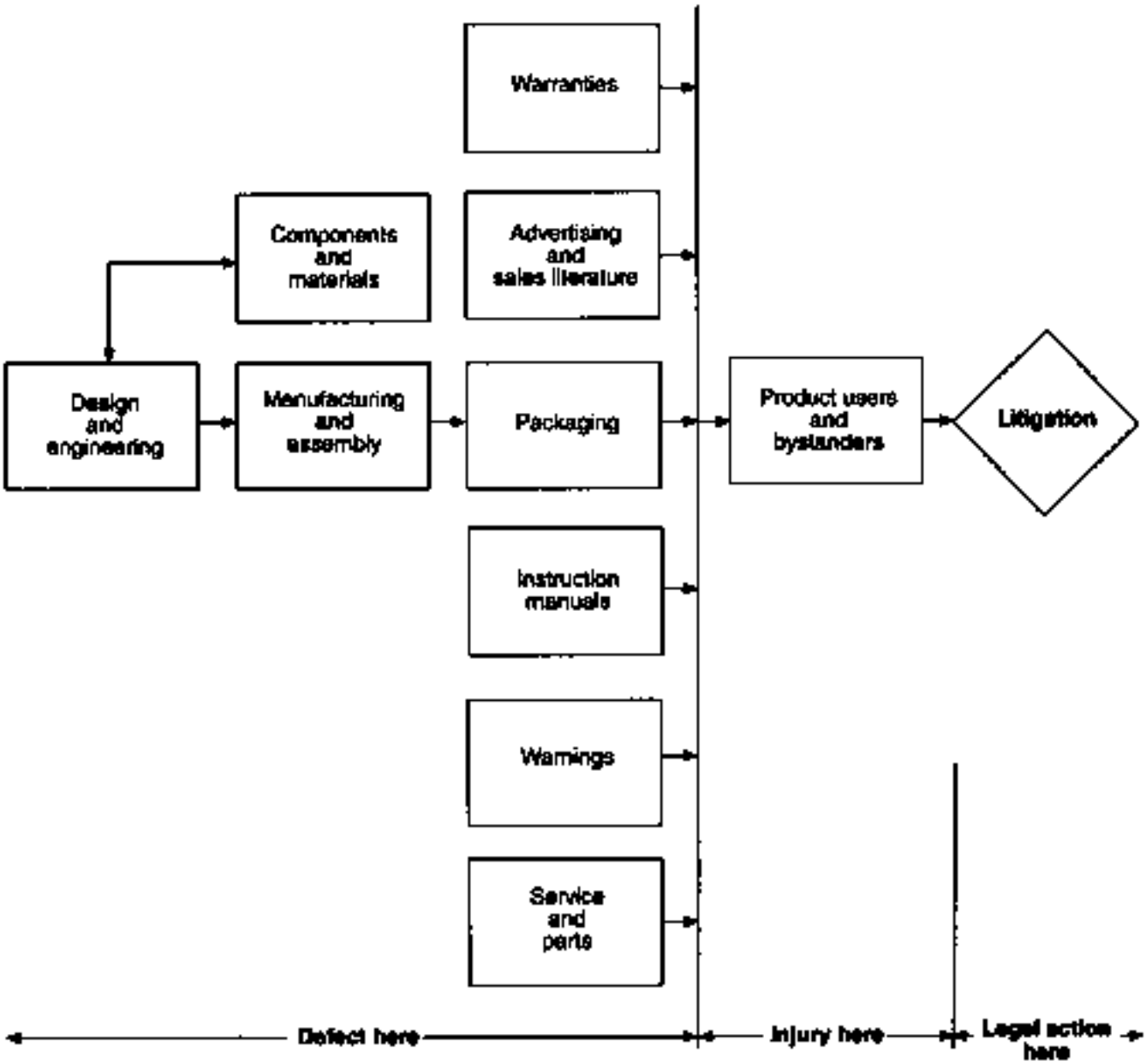


Fig. 1 The essence of products liability

Reference cited in this section

1. *The Code of Hammurabi*, University of Chicago Press, 1904

Legal Bases for Products Liability

The three legal theories on which a products liability lawsuit can be based are negligence, breach of warranty, or strict liability. All three are predicated on the fault system (i.e., a person whose conduct causes injury to another is required to fully and fairly compensate the injured party).

The basic method of imposing liability on a defendant requires the plaintiff to prove that the defendant acted in a negligent manner. Under the negligence theory, the plaintiff must essentially establish proof of specific negligence (i.e., prove that the defendant was almost intentionally negligent). Proof of specific negligence is a difficult task.

A user of a product may, as a result of express oral or written statements, or implication, reasonably rely on the manufacturer's express or implied assurance (including advertising material) as to the quality, condition, and merchantability of goods, and as to the safety of using them for their intended purpose and use. If the user relies on these assurances and is injured, suit can be entered on the basis of breach of warranty.

Both negligence and breach of warranty require proof of some fault on the part of the defendant (i.e., the focus is on the action of an individual). Strict liability, however, focuses on the product itself (Ref 2):

1. One who sells any product in a defective condition unreasonably dangerous to the user or consumer or to his property is subject to liability for physical harm thereby caused to the ultimate user or consumer, or to his property, if (a) the seller is engaged in the business of selling such a product and (b) it is expected to and does reach the user or consumer without substantial change in the condition in which it is sold.
2. The rule stated in subsection (1) applies although (a) the seller has exercised all possible care in the preparation and sale of his product, and (b) the user or consumer has not bought the product from or entered into any contractual relation with the seller.

Although these three bases apply in general, it should be recognized that there are variations. Interpretation and application may vary greatly, depending on the jurisdiction.

Reference cited in this section

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

Hazard, Risk, and Danger

There is substantial confusion about the meaning of words such as hazard, risk, and danger. Webster (Ref 3) defines danger as: “liability to injury, pain, damage or loss; hazard; peril; risk.” Webster makes some distinction by further saying: “Hazard arises from something fortuitous or beyond our control. Risk is doubtful or uncertain danger, often incurred voluntarily.”

One can also consider a hazard as any aspect of technology or activity that produces risk; as the potential for harm or damage to people, property, or the environment; and as including the characteristics of things and the actions (or inactions) of individuals. One can also consider risk as a measure of the probability and severity of adverse effects.

With all the products liability litigation in the United States, there has developed a clear distinction among these three words for legal purposes. In this context, a hazard is a condition or changing set of circumstances that present an injury potential (e.g., a railroad crossing at grade, a toxic chemical, a sharp knife, the jaws of a power press). Risk is the probability of injury and is affected by proximity, exposure, noise, light, experience, attention arresters, intelligence of an involved individual, and so on. Risk (probability of exposure) is obviously much higher with a consumer product than with an industrial product used by trained workers in a shop environment. Danger is the unreasonable or unacceptable combination of hazard and risk. The U.S. courts generally hold that any risk that can be eliminated by reasonable accident-prevention methods is unreasonable and unacceptable. A high risk of injury could be considered reasonable and acceptable if the injury is minimal and the risk is recognized by the individual concerned.

As might be expected, there is extensive and ongoing debate over the meaning of “reasonable” and “unreasonable.” The American Law Institute (Ref 2) says unreasonably dangerous means that: “The article sold must be dangerous to an extent beyond that which would be contemplated by the ordinary consumer who purchases it, with the ordinary knowledge common to the community as to its characteristics. Good whiskey is not unreasonably dangerous merely because it will make some people drunk, and is especially dangerous to alcoholics; but bad whiskey, containing a dangerous amount of fusel oil, is unreasonably dangerous.” The American Law Institute (Ref 2) further says: “There are some products which, in the present state of human knowledge, are quite incapable of being made safe for their intended and ordinary use. … Such a product, properly prepared, and accompanied by proper directions and warnings, is not defective, nor is it unreasonably dangerous.”

References cited in this section

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

3. *Webster's New Twentieth Century Dictionary, Unabridged,* 2nd ed., Simon & Schuster, 1979

Definitions of Defects

The American Law Institute (Ref 2) says that a product is in a defective condition if “it leaves the seller's hands, in a condition not contemplated by the ultimate user, which will be unreasonably dangerous to him.” Peters (Ref 4) indicates that a California Supreme Court decision, Barker v Lull (Ref 5), established a good assessment of “defective condition.” This provides three definitions (or criteria) for manufacturing defects and two for design defects:

Manufacturing defects

- Nonconformance with specifications
- Nonsatisfaction of user requirements
- Deviation from the norm

Design defects

- Less safe than expected by ordinary consumer
- Excessive preventable danger

Manufacturing Defects

A failure to conform with stated specifications is an obvious manufacturing defect and not a new criterion. The aspect of user satisfaction may not be well known, but in the legal context it has long been recognized that a manufacturing defect exists when there is such a departure from some quality characteristic that the product or service does not satisfy user requirements. Under the third criterion (deviation from the norm), added by Barker, a manufacturing defect occurs when a product leaves the assembly line in a substandard condition, differs from the manufacturer's intended result, or differs from other, ostensibly identical units of the same product line.

Design Defects

A product may be considered to have a design defect if it fails to perform as safely as an ordinary consumer would expect. This failure to perform safely is interpreted in the context of intended use (or uses) in a reasonably foreseeable manner, where “foreseeable” has the same meaning as “predicted” in failure modes and effects, fault-tree, or hazard analyses. It appears that many “ordinary” consumers would have no concept of how safe a product should, or could, be without the expectations created by statements in sales material, inferences from mass media, general assumptions regarding modern technology, and faith in corporate enterprise.
A design defect also exists if there is excessive preventable risk. The real question is whether the risk outweighs the benefits. A risk-benefit analysis should include at least five factors:

- Gravity of the danger posed by the design (i.e., severity of the consequences in the event of injury or failure)
- Probability (including frequency and exposure of the failure mode) that such a danger will occur
- Technical feasibility of a safer alternative design, including possible remedies or corrective action
- Economic feasibility of these possible alternatives
- Possible adverse consequences to the product and consumer that would result from alternative designs

Additional relevant factors may be included, but design adequacy is evaluated in terms of a balance between benefits from the product and the probability of danger. Quantification of the risk-benefit analysis is not required but may be desirable.
Proving design adequacy places the burden of proof on the defendant. Once the plaintiff proves that the product is a proximate cause of injury, the defendant must prove that the benefits outweighed the risk. Discussion of manufacturing and design defects of various products is given in Ref 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24.
Note: No paper, book, or handbook relative to products liability can be truly current. In addition, there is substantial variation among jurisdictions (federal, state, and local). All cited publications, however, do have something that is currently pertinent.

Other Defects

The engineer must be alert for other possibilities. Smith and Talbot (Ref 25) point out that a marketing defect exists when there is a failure to provide any warning of hazard and risk involved with use of a product, provide adequate warning of hazard and risk involved with use of a product, or provide appropriate, adequate directions for safe use of a product. In other words, a marketing defect exists when a product, free of design and manufacturing defects, is unreasonably dangerous due to absence of warnings and directions. The designer/manufacturer has control over the directions and warnings provided. The designer/manufacturer is the most knowledgeable about the product and thus presumably the most able to determine the necessary directions and warnings.

Suits against manufacturers often allege a defective product. Careful investigation, however, sometimes shows that the problem is due to improper maintenance (e.g., Ref 26). The designer/manufacturer obviously has no control over the maintenance actually conducted but can try to minimize the possibility of improper practices by providing proper and adequate instructions with the product. In any event, the designer should not overlook the possibility of misuse and improper maintenance on the part of the user.

Example 1: Failure of a High-Speed Steel Twistdrill. A 1.905 cm (0.75 in.) stud broke in the vertical wall of a metalworking machine known as an *upsetter*. A parallel vertical wall left a limited amount of space in which mechanics could work. A pilot hole was drilled with a 0.476 cm (0.19 in.) drill. The drill was held in a Jacobs chuck in a portable drill press that, in turn, was held to the workpiece by an electromagnet. After the pilot hole was finished, the drill press was removed, the pilot drill was replaced by a 1.6 cm (0.63 in.) high-speed steel drill, and the press was repositioned.

One man was doing the drilling while another man was squirting oil into the hole. When the drill was about 1.27 to 1.90 cm (0.50 to 0.75 in.) into the pilot hole, there was a “bang.” The drill shattered, causing a chip to lodge in the right eye of the oiler, ultimately resulting in loss of vision in that eye. Suit was entered against the drill manufacturer alleging a defective drill.

The plaintiff's attorney retained a metallurgist who examined the fragments. An unetched longitudinal section showed a large nonmetallic inclusion parallel to the axis near the center of the drill. After etching, this section showed carbide bands in a martensitic matrix. Hardness measurements indicated 65 to 66 HRC at the edge of the flute with a bulk hardness of 62 to 64 HRC. The drill tip is shown in Fig. 2.

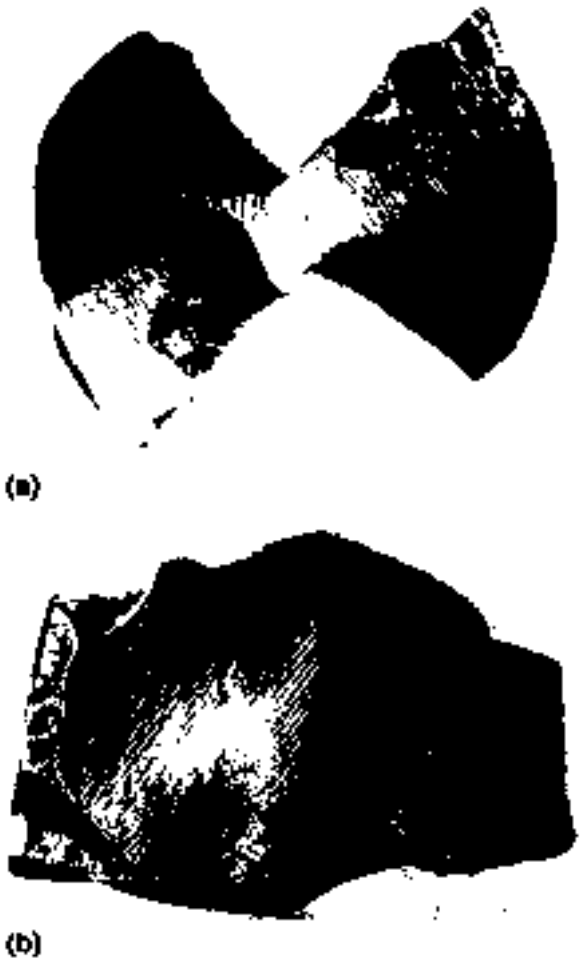


Fig. 2 The tip of the broken twistdrill. (a) End view. (b) Top view. A second drill from the same lot was also examined. There was carbide banding but to a significantly lesser

degree. Hardness was measured as 63 to 64 HRC at both the flute edge and in the bulk of the drill.

Plaintiff's expert concluded that the failed drill was defective while the other drill was satisfactory. He claimed that failure was a cumulative result of the following defective conditions: the steel contained nonmetallic inclusions that were detrimental to the properties of the drill; the carbide segregation was excessive, causing the drill to be brittle; and the cutting edge of the flutes was excessively hard. In his opinion, this high hardness made the edge brittle, so that the edge would chip during drilling. The chips caused the drill to bind and then shatter in a brittle manner because of excessive carbide segregation.

Plaintiff alleged defective design, defective manufacture, unsuitable or defective material, lack of sufficient quality control, and failure to foresee.

The defendant manufacturer believed the twist drill met all specifications for M1 high-speed steel. Both the supplier and manufacturer inspected for carbide segregation, with the poorest rating being “slight to medium.” A “medium” rating was permitted. Heat treatment and nitriding practices were consistent with those published by ASM International. After heat treatment, the drills were within the specified range of 64 to 66 HRC.

Some twenty other inspections for dimensional accuracy, shape, and finish were made after heat treatment. Fourteen drills were given a severe drilling performance test (manufacturer's routine) with no breaking or chipping.

It could be argued that there was nonsatisfaction of user requirements. The counter argument was that there was too high a demand on the drill.

Conclusions by the plaintiff's experts can be viewed as indicating a deviation from the norm. (Of the 5360 drills made from this one lot of steel, the manufacturer received only this one complaint.) The observed nonmetallic inclusion had a maximum width less than 0.0127 mm (0.0005 in.) and a maximum length less than 0.84 mm (0.033 in.). It was located in the shank, more than 6.35 cm (2.5 in.) from the drill tip and along the central axis, which is subjected to essentially none of the bending and twisting loading. While the inclusion is relatively large, it is not likely that it could contribute to the failure.

The largest carbide band was about 8 mm (0.32 in.) long and located about one-fourth of the distance from the central axis to the outer edge. It was also some distance from the drill tip. This location implies relatively light loading.

The manufacturer made hardness measurements on the two drills examined by plaintiff's expert. The results are given in Table 1. These indicate no significant difference between the two drills. The higher hardness at the cutting edge is expected and reasonable for a nitrided M1 steel. The hardness of both drills is within normal specification ranges.

Table 1 Hardness examination of drills

See Example 1 in text.

Measurement	Average hardness, HRC[(a)]	
	Bulk	**Cutting edge**
Plaintiff's measurements		
Broken drill	62–64	65–65
Unbroken drill	63–64	63–64
Defendant's measurements		
Broken drill	64.9, 64.8	65.1[(b)], 66.5[(c)]
Unbroken drill	65.0, 65.1	65.6[(b)], 66.5[(c)]

(a) Tukon readings (100 g) converted to HRC.
(b) At 0.1270 mm (0.005 in.) from the surface.
(c) At 0.254 mm (0.010 in.) from the surface

The manufacturer examined a number of other M1 drills that had satisfactorily met (corporate) standard drilling tests. One of these had a nonmetallic inclusion 1.5 times longer than in the failed drill. Two had edge hardnesses in the 66 to 68 HRC range with carbide banding more pronounced than in the failed drill.

From the viewpoint of design defects, was the drill less safe than expected by the ordinary consumer? Maybe. Presumably the workers did not expect drill failure. It is well known, however, that twistdrills do fail, no matter how well designed and manufactured. Using a drill to remove material after drilling a pilot hole is a common practice and clearly foreseeable. It is clearly more hazardous than drilling without a pilot hole. The drill may have been less safe than expected, but it seems more credible that too much was expected.

Existence of a design defect related to “excessive preventable danger” seems doubtful. The drill design was highly similar to that used by other manufacturers. All dimensions, tolerances, clearances, and so on were consistent with those used by other manufacturers and were based on years of drill use by a great variety of users. There is no question of potentially severe damage and relatively high probability of exposure. But there are no apparent alternatives that are technically and/or economically feasible.

What is your judgment on the validity of the allegations? How should this litigation have been resolved?
Examination of Fig. 2 indicates that one cutting lip is about 0.725 cm (0.286 in.) long while the other is about 0.802 cm (0.316 in.) long, so that the chisel edge is about 2.7 mm (0.106 in.) off center. The shorter lip will contact the work before the longer lip and thus bears all of the initial drilling stresses. The larger of the two chipped areas along the cutting edges in Fig. 2 is on the shorter lip. The broken point also had improper clearance angles (one was close to a negative angle). It was clear that the point of the broken drill was not the original point put on at manufacture but came from regrinding (presumably from “eyeballing” rather than using a jig). The work conditions, including a small pilot hole, a portable drill press, relocation of the press between the two drilling operations, and a questionable supply of coolant, placed abnormal stress on the drill.
The case was eventually settled out of court, with the plaintiff receiving a sum of less than $10,000 at a time when similar injury cases were receiving judgments of $50,000 to $150,000. This was clearly a so-called “nuisance settlement” to get rid of the suit. Much greater detail, both technical and legal, can be obtained by reference to Ref 27, 28, 29. A side aspect of this case relates to the expert witness. It developed that the plaintiff's expert was not sufficiently knowledgeable about high-speed steels, although he was a competent metallurgist.
Finally, it should be noted that this accident could have been avoided if the victim had been wearing safety glasses. The possibility of eye injury when working with tools is well known. Manufacturers usually add a warning to the product itself or its packaging. Employers require both workers and visitors to wear eye protection and provide training on its importance. Eye protection, as well as other personal protective equipment, is also required by both federal and state laws and regulations governing workplace safety.

Human Factors

Two additional examples arising from litigation are given subsequently. Designers certainly must consider the hazards in the design when it is used or operated in the intended manner. A hazard is any aspect of technology or activity that produces risk or danger of injury. The designer must also recognize that the product may be used in unintended but foreseeable ways. Protection must be provided against hazards in all uses that can be foreseen by the designer. Unfortunately, even the most diligent search for foreseeable uses may still leave a mode of use undiscovered. In litigation, a key issue often revolves around the question of whether the specific use was foreseeable by a reasonably diligent designer.
Example 2: A Snowmobile Collision. A female teenager and her boyfriend rented a snowmobile for a few hours to ride around on a lake. He was driving; she was a passenger. They collided with another snowmobile rented from the same rental agency. She was thrown on the ice, injuring her left knee. Surgery was performed but left her with permanent injury.
The owner of the agency admitted that the brakes had been disconnected by the previous owner of the agency (while he was an employee) and that he had left them disconnected. The reason given was that renters became confused about the operation of the throttle and the brakes, with the result that they “tore up the equipment.” Without brakes, the only mechanism for slowing the snowmobile was to release the throttle and let the engine provide braking. The owner claimed that the throttle had been adjusted to limit the speed to between 20 and 25 mph.
The owner's manual read: “The squeeze-type throttle lever, conveniently located on the right side of the steering handle, enables the operator to control the sled and the engine rpm at the same time.” The manual further read: “The hand-operated brake lever is located on the left side of the steering handle. By mounting the brake and throttle levers on the steering handle, the operator is able to maintain complete control of his snowmobile using his hands only.”
The throttle was opened (to increase speed) by closing (i.e., squeezing) the right-hand lever. The brakes were applied (to decrease speed) by closing (i.e., squeezing) the left-hand lever. When the throttle was released, the engine slowed. At some lower (undetermined) speed, the clutch disengaged and the tracks stopped moving, thereby generating significant drag.
As an exercise, this design is analyzed from the human factors viewpoint. If changes would improve the design, what changes would be made?
Most people immediately see this design as very poor from the viewpoint of human factors. Most operators never see the owner's manual, but in this context, that is minor. The natural (automatic) thing for people is that both the left and right hands will do the same thing simultaneously (i.e., both will open or both will squeeze), unless deliberate thought is given to doing otherwise. When a collision is imminent, there is no time for conscious thought, only conditioned reflex. Only well-experienced operators might be expected to have the left hand squeeze while the right hand opens.
It is not necessary, moreover, to have the design that was used. Reversing the action of the brake lever would be a better design (i.e., both hands squeeze to move forward; both hands release to stop). This “deadman” arrangement is common in railway locomotives, power movers, snowblowers, and so on. This arrangement is technically and economically feasible. The argument that keeping the left hand closed is undesirable is weak, because the right hand must be kept closed while the snowmobile is operating.

Another, perhaps superior, alternative is found on many motorcycles and would be familiar to many people. The throttle is operated by rotating a grip on the steering handle. The brake is operated by a squeeze lever on the same side of the steering handle. The two cannot be operated simultaneously. This alternative is technically and economically feasible, with no foreseeable adverse effects.

Example 3: A Ladder Label. Figure 3 shows a full-scale copy of a label “permanently attached” (by some sort of adhesive) to the inside of a rail on a fiberglass ladder. How good is this label? How effective is it? Assuming that users do indeed see the label, how many will read it? Of those who read it, how many will really comprehend what the manufacturer is trying to say? It appears to the author that the label was not well thought out, either in content or in phrasing, which is unclear or ambiguous in many places. The label does not provide clear instructions on use or explicitly warn of the dangers. Thus, in the author's opinion, the label is clearly inferior and essentially ineffective. One might infer that the manufacturer was trying to cover all possibilities to provide “protection” against products liability suits. It may be ineffective in that respect as well. Perhaps the ladder manufacturer is responding to jury awards in which a manufacturer was found negligent for not including a specific warning against some misuse.

FIBERGLASS
SINGLE & EXTENSION LADDER
FOR SAFETY, READ CAREFULLY

INSPECTION

1. Inspect upon receipt and before use.
2. Never climb a damaged ladder. Return for repair or discard.
3. Check all working parts, rivets, bolts, rope and cable for good working order.
4. Never use ladder with missing parts.
5. Discard if exposed to fire or chemicals.

SELECTION

1. Use 300 lb., and 250 lb. Duty-Rated Ladder for maintenance and heavy-duty work. Never use ladder jacks on 200 lb. or 225 lb. Duty-Rated Ladders.
2. Use ladder with correct duty rating to support combined weight of the user and material. Ladders are available with duty ratings of 200, 225, 250, 300 lb.

SET-UP AND USE

1. Set up ladder at 75½° by placing bottom ¼ of length being used out from vertical resting point.
2. Set ladder on firm level ground. Never lean sideways and never use on ice or snow.
3. Use proper size ladder. Never use temporary supports to increase length or to adjust for uneven surfaces.
4. Keep rungs free from wet paint, mud, snow, grease, or other slippery material.
5. Extend only from ground. Never extend from top or by bouncing.
6. Never walk or jog ladder while on it.
7. Securely engage ladder locks before climbing.
8. Erect ladder with fly (upper) section above and resting on base (lower) section.
9. Each section of a multi-section ladder shall overlap the adjacent section by 3 ft. up to and including 36 ft.; by 4 ft. over 36 ft., up to and including 48 ft.; by 5 ft. over 48 ft., up to and including 60 ft.
10. Always have the four ends of the ladder rails firmly supported.
11. Always tie top and base to building.
12. Project ladder minimum of 3 feet above roof edge.
13. Tie down ladder before stepping onto roof.
14. Never over-reach. Move ladder instead. Keep belt buckle inside ladder side rails.
15. Never use in high winds.
16. Never overload. Ladder designed to support one person when properly used.
17. Never use as a horizontal platform, plank or material hoist.
18. Never use on a scaffold.
19. Never fasten different ladders together to increase length.
20. Never apply a side load to ladder to push or pull anything while on ladder.
21. Never drop or apply impact load to ladder.
22. Never sit on end of ladder rails.
23. When reassembling, properly engage all guide brackets and lock prior to use.
24. Never use in front of unlocked doors.
25. Fly section must have safety shoes if used as a single ladder.
26. Hooks may be attached at or near top for added security.
27. To support the top of a ladder at a window opening, a stabilizer should be attached to span the window.
28. Never use ladder when you are in poor health.
29. Never use if taking drugs or alcoholic beverages.
30. Recommend never using if over 65 years of age.

CLIMBING INSTRUCTIONS

1. Never climb onto ladder from the side or from one ladder to another.
2. Face ladder when ascending or descending. Maintain a firm grip and stand on middle of rung.
3. Never stand above 3rd rung from top.
4. Never climb above support point.

STORAGE

1. Support ladder on racks when stored.
2. Never store material on ladder.
3. Properly support ladder in transit.

FOR ADDITIONAL INSTRUCTIONS, SEE ANSI A14.5

Fig. 3 A black-and-white reproduction of a decalcomania label to be placed on the inside of the side rail of a fiberglass ladder. The heading was yellow lettering on black. The text lettering was black on yellow. Reproduction is 100% of original size.

At least two issues deserve comment. One is the content of the text on the label. There are 44 items under five different headings, and many of these items are somewhat ambiguous. For example, in the first three items under "Inspection," it is not clear just what one looks for (i.e., there is no definition of terms). Under "Set-up and Use," the first item is poorly stated. Engineers have no difficulty in understanding what is intended, but many people besides engineers use ladders. In addition, does any user ever actually measure 75° or even "of length being used"? A measurement is not required—a simple estimate of angle or length would be sufficient. Item 10 seems to conflict with item 12. Item 11 says: "Always tie top and base to building." Are ladders used only on buildings? No, but the user might think twice before using it on, for example, a tree. Does anyone ever tie a ladder to a building? If so, how? Item 30 says: "Recommend never using if over 65 years of age." This age limitation presumably refers to the age of the user (not the ladder), but why 65 rather than some other age? One could go on at length about other items. At the bottom, the label says: "For additional instructions, see ANSI A14.5." Will the average user have any concept of what this means or where to find a copy? Will any user, even the intelligent engineer, obtain and read American National Standards Institute (ANSI) A14.5? In any case, the manufacturer is providing a reference for those who feel they might need the additional information that a standard provides. Reference to the standard also tells the user that the manufacturer is conscious of standards and is trying to follow them. This is important to the user even if he or she does not personally obtain and read the standard. It is unclear just what "additional instructions" should be consulted in the standard. Warning of all possible foreseeable hazards and misuses in a very limited space can be difficult, but some further design of the label seems desirable.

The second problem is the type size. It does conform to ANSI Z535.4. The text of the original label had black lettering on a yellow background in keeping with ANSI Z535.1. The three-line heading was yellow on black. The text print size was 6-point type. ("Point" is a printer's measure equivalent to 0.01384 in. or essentially 72 points per in.) ANSI Z535.4 specifies type (in the text) of 1.5 mm high (minimum), or 5 points. However, do you have any difficulty in reading it? Bailey (Ref 30) notes that "type size in books and magazines usually ranges from 7 to 14 points with the majority being about 10 to 11 points. Probably the optimum size is from 9 to 11 points—sizes smaller or larger can slow reading speed."

Anticipating Errors. When humans are involved in the use of a product or system, there will be errors. Some errors are extremely difficult, if not impossible, to anticipate. Also, in many situations, people abuse equipment. This is commonly a result of poor operating practices or lack of maintenance. In other situations, there may be deliberate action by the user (e.g., trying to fit two components together in a manner that is not intended, such as installing thread adapters on pressurized gas containers). There is no question that the designer cannot anticipate all these possibilities and provide protection. Nevertheless, the designer is not relieved of a substantial effort to anticipate such actions and to try to thwart them.

How does one proceed? The designer must be well informed on anthropometrics (physical characteristics), how people tend to behave or perform, and how to combine such data to achieve a suitable, effective, and safe design. A wealth of literature is available.

Hunter (Ref 31) includes enough anthropometric data to give insight into the kind of data to expect. He also provides many examples of sources of information. He comments on Department of Defense documents that provide substantial and significant information. The objectives of these various documents can be applied with equal validity to both civilian and military products.

Human behavior is largely a question of psychology, a topic about which most engineers know little. There seems to be little information readily available that is focused for use by engineers. Possible sources are Ref 32, 33, 34, 35, 36, 37.

Many publications provide varying degrees of insight and help in applying human factors information to design. Some that may be particularly useful are Ref 38, 39, 40.

One of the many objectives of the designer is to minimize the probability of "human error," where human error is any personnel action inconsistent with established behavioral patterns considered to be normal or that differs from prescribed procedures. Predictable errors are those that experience shows will occur repeatedly under similar circumstances. The designer must minimize the possibility of such errors.

People have a strong tendency to follow procedures that require minimum physical and mental effort, discomfort, and/or time. Any task that conflicts with this tendency is highly likely to be modified or ignored by the person who is expected to execute the task.

One of many important considerations in design is to follow common stereotypical expectations as much as possible. Consider a few examples:

- Clockwise rotation of a rotary control (knob) is expected to increase the output.
- Moving a lever forward, upward, or to the right is expected to increase the output.
- On a vertically numbered scale, the higher numbers are expected to be at the top.

- In vehicles, depressing the accelerator is expected to increase speed, and depressing the brake is expected to decrease speed. One expects the right foot to be used to apply force to the accelerator, then moved to the brake pedal. This is true whether one drives on the right or left side of the road.

Smith (Ref 41) tells of a forklift truck that violated the fourth item. The left foot was used to depress a pedal that increased speed, and a brake was applied when the foot was lifted.

References cited in this section

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

4. G.A. Peters, New Product Safety Legal Requirements, *Hazard Prevention,* Sept/Oct 1978, p 21–23

5. *Barker v Lull Engineering Co.,* 20 C. 3d 413

6. C.O. Smith, Manufacturing/Design Defects, Paper 86-WA/DE-14, American Society of Mechanical Engineers

7. C.O. Smith, Mobile Ladder Stand, Paper 87-DE-5, American Society of Mechanical Engineers

8. C.O. Smith, Design of a Saw Table, Paper 87-WA/DE-9, American Society of Mechanical Engineers

9. C.O. Smith, Coffee Grinder: Safe or Not?, Paper 88-WA/DE-6, American Society of Mechanical Engineers

10. C.O. Smith, Collapse of an Office Chair, Paper 89-WA/DE-18, American Society of Mechanical Engineers

11. C.O. Smith, Some Subtle (or Not So Subtle?) Product Defects, Paper 90-WA/DE-23, American Society of Mechanical Engineers

12. C.O. Smith, A Fatal Helicopter Crash, Paper 91-WA/DE-8, American Society of Mechanical Engineers

13. P.D. Beard and T.F. Talbot, What Determines if a Design is Safe, Paper 90-WA/DE-20, American Society of Mechanical Engineers

14. T.F. Talbot and C.S. Hartley, Failure of Fastening Devices in Pump Packing Gland Flange, Paper 89-WA/DE-12, American Society of Mechanical Engineers

15. T.F. Talbot and M. Crawford, Wire Rope Failures and How to Minimize Their Occurrence, Paper 87-DE-7, American Society of Mechanical Engineers

16. T.F. Talbot and J.H. Appleton, Dump Truck Stability, Paper 87-DE-3, American Society of Mechanical Engineers

17. T.F. Talbot, Safety for Special Purpose Machines, Paper 87-WA/DE-8, American Society of Mechanical Engineers

18. T.F. Talbot, Chain Saw Safety Features, Paper 86-WA/DE-16, American Society of Mechanical Engineers

19. T.F. Talbot, Hazards of the Airless Spray Gun, Paper 85-WA/DE-13, American Society of Mechanical Engineers

20. T.F. Talbot, Man-Lift Cable Drum Shaft Failure, Paper 87-WA/DE-19, American Society of Mechanical Engineers

21. T.F. Talbot, Bolt Failure in an Overhead Hoist, Paper 83-WA/DE-20, American Society of Mechanical Engineers

22. W.G. Ovens, Failures in Two Tubular Steel Chairs, Paper 91-WA/DE-9, American Society of Mechanical Engineers

23. J.A. Wilson, Log Loader Collapse: Failure Analysis of the Main Support Stem, Paper 89-WA/DE-13, American Society of Mechanical Engineers

24. T.A. Hunter, Design Errors and Their Consequences, Paper 89-WA/DE-14, American Society of Mechanical Engineers

25. C.O. Smith and T.F. Talbot, Product Design and Warnings, Paper 91-WA/DE-7, American Society of Mechanical Engineers

26. C.O. Smith and J.F. Radavich, Failures from Maintenance Miscues, Paper 84-DE-2, American Society of Mechanical Engineers

27. C.O. Smith, Failure of a Twistdrill, *J. Eng. Mater. Technol.,* Vol 96 (No. 2), April 1974, p 88–90

28. C.O. Smith, Legal Aspects of a Twistdrill Failure, *J. Prod. Liabil.,* Vol 3, 1979, p 247–258

29. C.O. Smith, "ECL 170, Tortured Twist Drill," Center for Case Studies in Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN

30. R.W. Bailey, *Human Performance Engineering: A Guide for System Designers,* Prentice-Hall, 1982

31. T.A. Hunter, *Engineering Design for Safety,* McGraw-Hill, 1992

32. E. Grandjean, *Fitting the Task to the Man,* 4th ed., Taylor and Francis, Washington, D.C. 1988

33. B.A. Sayers, *Human Factors and Decision Making: Their Influence on Safety and Reliability,* Elsevier Science Publishers, 1988

34. K.S. Park, *Human Reliability,* Elsevier Science Publishers, 1987

35. L.S. Mark, J.S. Warren, and R.L. Huston, Ed., *Ergonomics and Human Factors,* Springer-Verlag, 1987

36. B.S. Dhillon, *Human Reliability,* Pergamon Press, 1986

37. C.D. Wickens, *Engineering Psychology and Human Performance,* 2nd ed., Harper-Collins, 1992

38. W.E. Woodson, *Human Factors Design Handbook,* McGraw-Hill, 1981

39. G. Salvendy, Ed., *Handbook of Human Factors,* Wiley, 1987

40. W.G. Ireson, Chapter 12, *Reliability Handbook,* McGraw-Hill, 1966

41. C.O. Smith, Two Industrial Products—Defective Design?, Paper 93-WA/DE-11, American Society of Mechanical Engineers

Preventive Measures

What are the implications of the previously mentioned example for the design engineer? It is necessary to look carefully at the completed design to be sure that it is indeed appropriate and that it does not incorporate problems for which proper technological solutions have existed for some time. (For example, an independent assessment by a design review board, whose members have no parental pride in the design, is highly appropriate.) In addition, there must be recognition that many, perhaps most, consumers have no objective basis to judge how safe a product should be. An engineer making a judgment about safety must understand this lack of appreciation of an appropriate safety level.

Acting as a prudent manufacturer is not enough. The focus should be on the product itself, not the reasonableness of a manufacturer's conduct. Obviously, there will be no viable lawsuits if there are no injuries

or if there are no violations of the law. Undoubtedly the best practice is to sell a well-designed, well-manufactured product. The manufacturer needs to make certain that all reasonable preventive measures have been used in the design and manufacturing process. Much evidence, however, suggests that one of Casey Stengel's comments applies in the area of preventive measures: “In many areas we have too strong a weakness.” While many preventive measures are well known to most design engineers, some comments may be appropriate, even if only in the sense of a checklist of items to be considered.

Design review is an effort, through group examination and discussion, to ensure that a product (and its components) will meet all requirements. In a design of any complexity, there is necessity for a minimum of three reviews: conceptual, interim, and final. Conceptual design reviews have a major impact on the design, while interim and final reviews have relatively less effect as the design becomes more fixed and less time is available for major design changes. It is much easier and much less expensive to include safety in the initial design than to include it retroactively.

A more sophisticated product may require several reviews during the design process. These might be: conceptual, definition, preliminary (review of initial design details), critical (or interim review, perhaps several reviews in sequence—review details of progress, safety analyses, progress in hazard elimination, etc.), prototype (review of design before building a prototype), prototype function review, and preproduction review (final review—last complete review before release of the design to production).

These periodic design reviews should review progress of the design, monitor design and development, ensure that all requirements are met, and provide feedback of information to all concerned.

A design review is conducted by an ad hoc design review board composed of materials engineers, mechanical designers, electrical designers, reliability engineers, safety engineers, packaging engineers, various other design engineers as appropriate, a management representative, a sales representative, an insurance consultant, an attorney in products liability, outside “experts” (be sure they are truly expert!), and so on. Members of the design review board should not be direct participants in day-to-day design and development of the product under review, but the engineers should have technical capability at least equal to that of the actual design team. Vendor participation is highly desirable, especially in conceptual and final design reviews.

Design review checklists should be prepared well in advance of actual board meetings. These checklists should cover all aspects of the design and expected performance, plus all phases of production and distribution. A new checklist should be developed for each new product.

It is good practice for a designer/manufacturer to have some sort of permanent review process in addition to the ad hoc board for each individual product. This permanent group should evaluate all new products, reevaluate old products, and keep current with trends, standards, and safety devices.

If properly conducted, a design review can contribute substantially to avoiding serious problems by getting the job done right the first time. Formal design review processes are effective barriers to “quick and dirty” designs based on intuition (or educated guesses) without adequate analyses.

Some Common Procedures. Many engineers and designers are familiar with such techniques and procedures as hazard analysis; failure modes and effects analysis (FMEA); failure modes, effects, and criticality analysis (FMECA); fault-tree analysis (FTA); fault hazard analysis (FHA); operating hazard analysis (OHA); use of codes, standards, and various regulatory acts; and the Occupational Safety and Health Act (OSHA). These are discussed in the article “Safety in Design” in *Materials Selection and Design*, Volume 20 of *ASM Handbook*. Some other aspects of products liability are perhaps less well known and require some comment.

Prediction methods are necessary in applying FMEA, FTA, and so on. From statistics it is possible to predict performance of a large group of similar products, but it is not possible to predict performance of any one individual item of that group. Various statistical and probabilistic techniques can be used to make predictions, but these are predicated on having good databases.

State of the Art. The meaning of the term “state of the art” should be defined for each specific product. This might be done by comparing the product to those produced by competitors, but this comparison may not be enough. A jury is not bound by negligent practices of a negligent industry, and unfortunately, in some areas, industry practices and standards are low-level consensus practices and standards. Being in step with the state of the art may not be enough—one should strive to be ahead of the state of the art if possible (i.e., better than the competitors). It is not enough to explain what was done, because the plaintiff's expert witnesses may point out what could have been done. Purely economic reasons may not be a compelling defense argument in the courtroom and should be avoided if possible.

Quality Assurance and Testing. A primary function of quality control is to feed back inspection, testing, and other data, showing designers what is happening and revealing any need for design improvement. Manufacturers should test products in various stages of development, including field service, especially if critical components or subassemblies are involved. Final tests are necessary on each individual product or on representative samples of plant output. Care must be taken that quality control is not relaxed, intentionally or unintentionally, for production expediency.

Foreseeability is a factor that requires special attention. It is necessary to determine not only how the product is intended to be used, but also every reasonably conceivable way that it can be used and misused. (Who has never used a flat-tang screwdriver for some other purpose?) All reasonable conditions of use, or misuse, that might lead to an accident should be detailed. The designer must conclusively demonstrate that the product cannot be made safer, even to prevent accidents, during use or misuse. The problem of foreseeability is one that seems especially difficult for engineers to accept.

Consumer Complaints. Data on product failures from test facilities, test laboratories, and service personnel are valuable. Each complaint should be quickly, carefully, and thoroughly investigated. An efficient reporting system can result in product corrections before large numbers of the product reach users, or a product recall before there has been a major exposure of the public to an unsafe product.

Warranties and Disclaimers. Warranties and disclaimers detail the limitations of the manufacturer's liability. When used, they must be written in clear, simple, and easily understood language. Both should be reviewed by highly competent legal counsel knowledgeable in both the industry and products liability. A copy of the warranty and/or disclaimer must be packaged with the product. All practical means must be used to make the buyer aware of the contents. It must be recognized, however, that warranties and disclaimers, no matter how well written, may not be the strongest defense.

Warnings and Directions. Directions are intended to ensure effective use of a product. Warnings are intended to ensure safe use. Both should be written to help the user understand and appreciate the nature of the product and its dangers. If directions and warnings are inadequate, there is potential liability, because it cannot be said that the user had contributory negligence in failing to appreciate and avoid danger.

The burden of full and effective disclosure is on the manufacturer. Directions and warnings, although essential, do not relieve the manufacturer of the duty to design a safe product. The law will not permit a manufacturer, who knowingly markets a product with a danger that could have been eliminated, to evade liability simply because a warning is placed on the product. One must design against misuse.

This topic is discussed in greater detail in the article “Safety in Design” in *Materials Selection and Design,* Volume 20 of *ASM Handbook.* A label is discussed in some detail in the article “Human Factors in Design” in *Materials Selection and Design,* Volume 20 of *ASM Handbook.*

Written Material. All advertising, promotional material, and sales literature must be carefully screened. Warranties can be implied or inferred by the wording on labels, instructions, pamphlets, sales literature, advertising (written and electronic broadcast), and so on, even though no warranty is intended. There must be no exaggeration in such material. The manufacturer must be able to show that the product is properly rated and that the product can safely do what the advertisement says it will do. Additional information on the appropriate level of language is given in the article “Safety in Design” in *Materials Selection and Design,* Volume 20 of *ASM Handbook.*

Human Factors. Many products and systems require operation by a human who thereby becomes an integral part of the system. As such, the human can have a very significant effect on system performance. One must recognize that the human being is the greatest, and least controllable, variable in the system. Some attorneys believe that most products liability suits result because someone (usually the designer) did not thoroughly think through how the product interfaced with society.

Products Recall Planning. It is a fact of life that mistakes are sometimes made even by highly experienced professionals exercising utmost care. When such errors occur, a products recall may be necessary. Unless the specific troublesome part can be readily and uniquely identified as to source, production procedure, time of manufacture, and so on, there will be great difficulty in pinpointing the problem within the producing organization. Placing one advertisement for recall purposes in newspaper and magazines (not including TV) throughout the country is very expensive. An obvious economic need, as well as a regulatory requirement, exists for manufacturers (and importers) to have systems in place for expeditious recall of a faulty product.

Records. Once involved in litigation, one of the most powerful defenses that manufacturers and engineers can have is an effective, extensive, and detailed record. Records should document how the design came about, with

notes of meetings, assembly drawings (including safety features), checklists, the state of the art at the time, and so on. These records, while no barrier to products liability lawsuits, will go a long way toward convincing a jury that prudent and reasonable care has been taken to produce a safe product.

Paramount Questions

No matter how carefully and thoroughly one executes all possible preventive measures, it is necessary to ask:

- What is the probability of injury?
- Who determines the probability of injury?
- What is an acceptable probability of injury?
- Who determines the acceptable probability of injury?

As Lowrance (Ref 42) suggests, determining the probability of injury is an empirical, scientific activity. It follows that engineers are better qualified by education and experience than most people to determine this probability. Presumably, designers will use organized approaches to cope with the complexity. One obvious place for assessing this probability is the design review process. While design review is a most valuable aid for the designer, it is not a substitute for adequate design and engineering.

As Lowrance (Ref 42) further suggests, judging the acceptable probability of injury is a normative, political activity. Obviously, assessing the probability of injury is not a simple matter. Assessing the acceptable probability of injury is far more complex and difficult. Use of the word “acceptable” emphasizes that safety decisions are relativistic and judgmental. It implies three questions:

Acceptable in whose view? Acceptable in what terms? Acceptable for whom? This use of “acceptable probability of injury” avoids any implication or inference that safety is an intrinsic, absolute, measurable property.

In assessing acceptable danger, one major task is determining the distribution of danger, benefits, and costs. This determination is both an empirical matter and a political issue. It involves questions such as:

Who will be paying the costs? Will those who benefit be those who pay? Will those endangered be those who benefit? Answers to these questions may be based on quantifiable data but often must be based on estimates or surveys. A related major task is to determine the equity of distribution of danger, benefits, and costs. This asks a question of fairness and social justice for which answers are a matter of personal and societal value judgment.

Who determines the acceptable level of probability of injury? In terms of ability to judge acceptability, designers/engineers are no better qualified than any other group of people and, in general, are less qualified than many others. It is often alleged that engineers (because of their inherent characteristics, education, and experience) are less sensitive to societal influences of their work and products than others. As for most stereotypes, there is some truth in this view. Clemenceau reportedly said: “War is much too serious a matter to be entrusted to the military.” Perhaps product design is much too serious a matter to be entrusted solely to designers and (especially) business managers.

Jaeger (Ref 43) has summarized the situation thus:

“Nowadays it seems to me that the risk problem in technology has turned out to become one of the most pressing questions concerning the whole of industrial development. This problem is of fundamental as well as of highly practical importance. The answer to the question “How safe is safe enough?” requires a combination of reflective and mathematical thinking as well as the integration of technological, economic, sociological, psychological and ecological knowledge from a superior point of view.”

If the designer cannot adequately make the determination, then who can? Various ideas have been proposed (e.g., Ref 44), but no suggestion yet made is fully satisfactory. The designer/producer must resolve this for each product. References 42 and 45 can be helpful in developing sensitivity to assessing an acceptable probability of injury.

References cited in this section

42. W.W. Lowrance, *Of Acceptable Risk*, William Kaufman, Inc., 1976

43. T.A. Jaeger, Das Risikoproblem in der Technik, *Schweizer Archiv fur Angewandte Wissenshafter und Technik,* Vol 36, 1970, p 201–207

44. C.O. Smith, “How Much Danger? Who Decides?” paper presented at American Society of Mechanical Engineers Conference “The Worker in Transition: Technological Change,” Bethesda, MD, 5–7 April 1989

45. R.A. Schwing and W.A. Albers, *Societal Risk Assessment,* Plenum Press, 1980

Acceptable Level of Risk

Because “danger” has been previously defined as the unreasonable or unacceptable combination of hazard and risk, it is a contradiction to speak of an acceptable level of danger. However, an acceptable level of risk or hazard might be considered, following from the previous section, to be acceptable probability of injury. It should also be mentioned in this section, as it has been in previous sections, that the acceptable level is related to the benefits gained by taking the risk and the risk of alternatives.

It has been suggested that an acceptable level of danger might be 1 in 4000 per year. Currently (2002), 1 in 7000 per year may be a better number; that is, about 40,000 die per year in the United States out of a population of about 275 million. Statistics indicate that this is about the danger of dying from an automobile accident in the United States. One might infer that U.S. citizens consider this an acceptable level in view of the fact that little apparent effort is expended in trying to decrease the accident rate. The National Highway Traffic Safety Administration indicates that about 50% of fatal traffic accidents in the United States are alcohol related. If there were severe penalties for driving under the influence of alcohol (as there are in some other countries), this danger would presumably decrease to about 1 in 8000 per year. Either level of danger may be rational for the public as a whole (obviously debatable), but it probably is not perceived as such by a bereaved family. Such a rate hardly seems acceptable for consumer products. It certainly is unacceptable for nuclear applications. While the majority of manufactured products have a much lower level of danger than this, many of these products are considered to have a level of danger too high to be acceptable. Juries regularly make this decision in products liability actions.

One aspect of a potentially acceptable level of danger is the manner in which it is stated. Engineers might prefer to state the level in terms of probability. The general public, however, might well prefer it otherwise, or even unstated. The general public must be aware of fatalities from automotive accidents. It is possible that if automobile manufacturers were to point out that there is an annual chance of about 1 in 4000 that an individual will be killed, and a much greater chance of being injured (even seriously, such as spinal injuries, which not only incapacitate the victim but require constant attention by others), the attitude of the public might be different.

It must be recognized that while it is possible to reduce the level of danger to a very small number, danger cannot be completely eliminated, no matter how much effort is expended. We do not think there is any one level of acceptable danger. Each situation must be judged independently. The question is not what level of danger the engineer/designer thinks is acceptable for the public but what level the public perceives to be acceptable.

Acknowledgment

This article was adapted from C.O. Smith, Products Liability and Design, *Materials Selection and Design,* Volume 20, *ASM Handbook,* p 146–151.

References

1. *The Code of Hammurabi,* University of Chicago Press, 1904

2. *Restatement of the Law, Second, Torts,* 2d, Vol 2, American Law Institute Publishers, 1965

3. *Webster's New Twentieth Century Dictionary, Unabridged,* 2nd ed., Simon & Schuster, 1979

4. G.A. Peters, New Product Safety Legal Requirements, *Hazard Prevention*, Sept/Oct 1978, p 21–23

5. *Barker v Lull Engineering Co.*, 20 C. 3d 413

6. C.O. Smith, Manufacturing/Design Defects, Paper 86-WA/DE-14, American Society of Mechanical Engineers

7. C.O. Smith, Mobile Ladder Stand, Paper 87-DE-5, American Society of Mechanical Engineers

8. C.O. Smith, Design of a Saw Table, Paper 87-WA/DE-9, American Society of Mechanical Engineers

9. C.O. Smith, Coffee Grinder: Safe or Not?, Paper 88-WA/DE-6, American Society of Mechanical Engineers

10. C.O. Smith, Collapse of an Office Chair, Paper 89-WA/DE-18, American Society of Mechanical Engineers

11. C.O. Smith, Some Subtle (or Not So Subtle?) Product Defects, Paper 90-WA/DE-23, American Society of Mechanical Engineers

12. C.O. Smith, A Fatal Helicopter Crash, Paper 91-WA/DE-8, American Society of Mechanical Engineers

13. P.D. Beard and T.F. Talbot, What Determines if a Design is Safe, Paper 90-WA/DE-20, American Society of Mechanical Engineers

14. T.F. Talbot and C.S. Hartley, Failure of Fastening Devices in Pump Packing Gland Flange, Paper 89-WA/DE-12, American Society of Mechanical Engineers

15. T.F. Talbot and M. Crawford, Wire Rope Failures and How to Minimize Their Occurrence, Paper 87-DE-7, American Society of Mechanical Engineers

16. T.F. Talbot and J.H. Appleton, Dump Truck Stability, Paper 87-DE-3, American Society of Mechanical Engineers

17. T.F. Talbot, Safety for Special Purpose Machines, Paper 87-WA/DE-8, American Society of Mechanical Engineers

18. T.F. Talbot, Chain Saw Safety Features, Paper 86-WA/DE-16, American Society of Mechanical Engineers

19. T.F. Talbot, Hazards of the Airless Spray Gun, Paper 85-WA/DE-13, American Society of Mechanical Engineers

20. T.F. Talbot, Man-Lift Cable Drum Shaft Failure, Paper 87-WA/DE-19, American Society of Mechanical Engineers

21. T.F. Talbot, Bolt Failure in an Overhead Hoist, Paper 83-WA/DE-20, American Society of Mechanical Engineers

22. W.G. Ovens, Failures in Two Tubular Steel Chairs, Paper 91-WA/DE-9, American Society of Mechanical Engineers

23. J.A. Wilson, Log Loader Collapse: Failure Analysis of the Main Support Stem, Paper 89-WA/DE-13, American Society of Mechanical Engineers

24. T.A. Hunter, Design Errors and Their Consequences, Paper 89-WA/DE-14, American Society of Mechanical Engineers

25. C.O. Smith and T.F. Talbot, Product Design and Warnings, Paper 91-WA/DE-7, American Society of Mechanical Engineers

26. C.O. Smith and J.F. Radavich, Failures from Maintenance Miscues, Paper 84-DE-2, American Society of Mechanical Engineers

27. C.O. Smith, Failure of a Twistdrill, *J. Eng. Mater. Technol.*, Vol 96 (No. 2), April 1974, p 88–90

28. C.O. Smith, Legal Aspects of a Twistdrill Failure, *J. Prod. Liabil.*, Vol 3, 1979, p 247–258

29. C.O. Smith, "ECL 170, Tortured Twist Drill," Center for Case Studies in Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN

30. R.W. Bailey, *Human Performance Engineering: A Guide for System Designers*, Prentice-Hall, 1982

31. T.A. Hunter, *Engineering Design for Safety*, McGraw-Hill, 1992

32. E. Grandjean, *Fitting the Task to the Man*, 4th ed., Taylor and Francis, Washington, D.C. 1988

33. B.A. Sayers, *Human Factors and Decision Making: Their Influence on Safety and Reliability*, Elsevier Science Publishers, 1988

34. K.S. Park, *Human Reliability*, Elsevier Science Publishers, 1987

35. L.S. Mark, J.S. Warren, and R.L. Huston, Ed., *Ergonomics and Human Factors*, Springer-Verlag, 1987

36. B.S. Dhillon, *Human Reliability*, Pergamon Press, 1986

37. C.D. Wickens, *Engineering Psychology and Human Performance*, 2nd ed., Harper-Collins, 1992

38. W.E. Woodson, *Human Factors Design Handbook*, McGraw-Hill, 1981

39. G. Salvendy, Ed., *Handbook of Human Factors*, Wiley, 1987

40. W.G. Ireson, Chapter 12, *Reliability Handbook*, McGraw-Hill, 1966

41. C.O. Smith, Two Industrial Products—Defective Design?, Paper 93-WA/DE-11, American Society of Mechanical Engineers

42. W.W. Lowrance, *Of Acceptable Risk*, William Kaufman, Inc., 1976

43. T.A. Jaeger, Das Risikoproblem in der Technik, *Schweizer Archiv fur Angewandte Wissenshafter und Technik*, Vol 36, 1970, p 201–207

44. C.O. Smith, "How Much Danger? Who Decides?" paper presented at American Society of Mechanical Engineers Conference "The Worker in Transition: Technological Change," Bethesda, MD, 5–7 April 1989

45. R.A. Schwing and W.A. Albers, *Societal Risk Assessment*, Plenum Press, 1980

Selected References

- S. Brown, I. LeMay, J. Sweet, and A. Weinstein, Ed., *Product Liability Handbook: Prevention, Risk, Consequence, and Forensics of Product Failure,* Van Nostrand Reinhold, 1990
- V.J. Colangelo and P.A. Thornton, *Engineering Aspects of Product Liability,* American Society for Metals, 1981
- R.A. Epstein, *Modern Products Liability Law: A Legal Revolution,* Quorum Books, Westport, CT, 1980
- R.L. Goodden, *Product Liability Prevention: A Strategic Guide,* ASQ Quality Press, Milwaukee, WI, 2000
- P.W. Huber and R.E. Litan, Ed., *The Liability Maze: The Impact of Liability Law on Safety and Innovation,* The Brookings Institute, 1991
- W. Kimble and R.O. Lesher, *Products Liability,* West Publishing Co., 1979
- J. Kolb and S.S. Ross, *Product Safety and Liability,* McGraw-Hill, 1980
- M.S. Madden, *Products Liability,* Vol 1 and 2, West Publishing Co., 1988
- C.O. Smith, *Products Liability: Are You Vulnerable?,* Prentice-Hall, 1981
- J.F. Thorpe and W.H. Middendorf, *What Every Engineer Should Know about Products Liability,* Dekker, 1979

Failures Related to Metalworking

Introduction

WROUGHT FORMS are produced by a wide variety of metalworking operations that can be roughly divided into bulk-working operations and sheet-forming operations (Ref 1). The general distinction here is that bulk working imposes material flow in all directions, while sheet-forming operations are typically limited to two-dimensional deformation. Metalworking operations are also classified as either primary metalworking (where mill forms such as bar, plate, tube, sheet, and wire are worked from ingot or other cast forms) or secondary metalworking (where mill products are further formed into finished products by hot forging, cold forging, drawing, extrusion, straightening, sizing, etc.). These metalworking operations have a two-fold purpose. First, they are obviously designed to produce parts with the desired configuration. Secondly, metalworking can develop a final shape with internal soundness and improved mechanical properties by:

- Improved internal quality due to compressive deformation
- Grain refinement
- Uniform grain structure
- Elimination of casting porosity and breakup of macrosegregation patterns
- Beneficial grain-flow pattern for improved part performance
- Improved toughness and/or fatigue resistance due to grain flow and fibering
- Burnished surface and controlled surface quality

However, the beneficial factors of deformation processing can become a potential problem if the process is not carefully understood. Potential problems of the deformation process also include:

- *Fracture-related problems:* for example, internal bursts or chevron cracks, cracks on free surfaces, cracks on die-contacted surfaces
- *Metal-flow-related problems:* for example, end grain and poor surface performance; inhomogeneous grain size; shear bands and locally weakened structures; cold shuts, folds, and laps; flow-through defects
- *Control, materials selection, and use problems:* for example, underfill, part distortion, and poor dimensional control; tool overload and breakage; excessive tool wear; high initial investment due to equipment cost; poor material use and high scrap loss

The movement of metal during these processes, whether performed at room temperature or at elevated temperatures, makes them common sources of surface discontinuities, such as laps, seams, and cold shuts. Oxides, slivers or chips of the base material, or foreign material also can be embedded into the surface during working. These surface imperfections produce a notch of unknown severity that acts as a stress raiser, which may adversely affect strength under load.

Subsurface and core discontinuities may also occur. Subsurface flaws often (but not always) originate from the as-cast ingot due to shrinkage, voids, and porosity that form during solidification. For example, this is shown schematically in Fig. 1 for rolled bar that contains ingot porosity and pipe imperfections (discussed later in the section, “Imperfections from the Ingot”). These imperfections can also serve as sites for crack initiation during fabrication or in service.

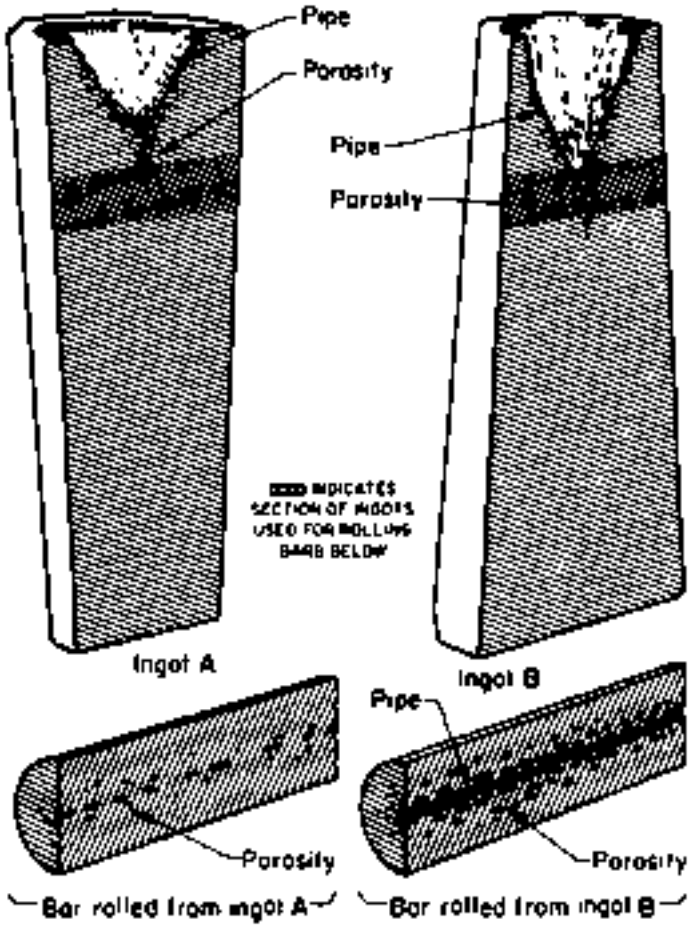


Fig. 1 Longitudinal sections of two types of ingots showing typical pipe and porosity. When the ingots are rolled into bars, these flaws become elongated throughout the center of the bars.

Metalforming and fabrication of wrought forms also involve other manufacturing operations, such as electroplating, heat treatment, machining, or welding. These operations may also introduce possible defects (Table 1) that may be considered in conjunction with possible defects from metalworking. Failures can also occur from a complex series of manufacturing factors. For example, the level of residual hoop and bending stresses can occur in a tube produced by drawing, heat treating, and straightening operations. By varying the severity of these operations, it is possible to produce tubes with very low residual stresses or with very high residual stresses that are near the yield strength of the metal. In other words, parts are made by a rather complex series of operations, which thus requires a broad understanding not only in the context of failure analysis but also in the organizational traditions for failure prevention.

Table 1 Defects that may result from postforming processes

Process	Possible defects
Electroplating	Hydrogen embrittlement, galvanic corrosion
Heat treatment	Excessive grain growth, burning of grain boundaries, brittle structure, carburization, decarburization, quench cracks
Electrolytic cleaning	Pitting
Surface hardening, nitriding, carburizing, anodic hard coating	Excessive case thickness, microcracks, embrittled material at stress raisers
Machining	Tool marks, grinding cracks
Welding	Weld-metal defects, hydrogen-induced cracking, inclusions, improper structure

The primary purpose of this article is to describe general root causes of failure that are associated with wrought metals and metalworking. This includes a brief review of the discontinuities or imperfections that may be

common sources of failure-inducing defects in bulk working of wrought products. This article does not address powder metallurgy (P/M). A good review of powder metallurgy with coverage relevant to failure analysis of P/M forms is in Ref 2.

References cited in this section

1. J. Schey, Manufacturing Processes and Their Selection, *Materials Selection and Design,* Vol. 20, *ASM Handbook,* ASM International, 1997, p 687–704

2. L. Pease III and W. West, *Fundamentals of Powder Metallurgy,* Metal Powder Industries Federation, 2002

Failures Related to Metalworking

Imperfections in Wrought Forms

Various terms are used to describe surface and subsurface imperfections in wrought product. The meaning of terms can vary by location and industry, but Fig. 2 is a schematic illustration of some terms used to describe flaws in rolled bar stock. This schematic is not a complete summary of possible imperfections; for example, die scratches of cold-worked product are not included. However, the schematic of Fig. 2 shows typical terms for surface and subsurface flaws that may occur in wrought products. For example, inclusions (Fig. 2a) and seams (Fig. 2h) are commonly used terms to describe imperfections in wrought form, as discussed in more detail in this section. These imperfections, whether at or the below the surface, can adversely affect performance of a part by creating a notch of unknown severity and serve as a crack-initiation site during fabrication or in service. Corrosion and wear damage can also be assisted by discontinuities, especially at the surface. These flaws may occur from the melting practices and solidification of ingot, the primary or secondary working of the material, or the metallurgical characteristic of a particular alloy system.

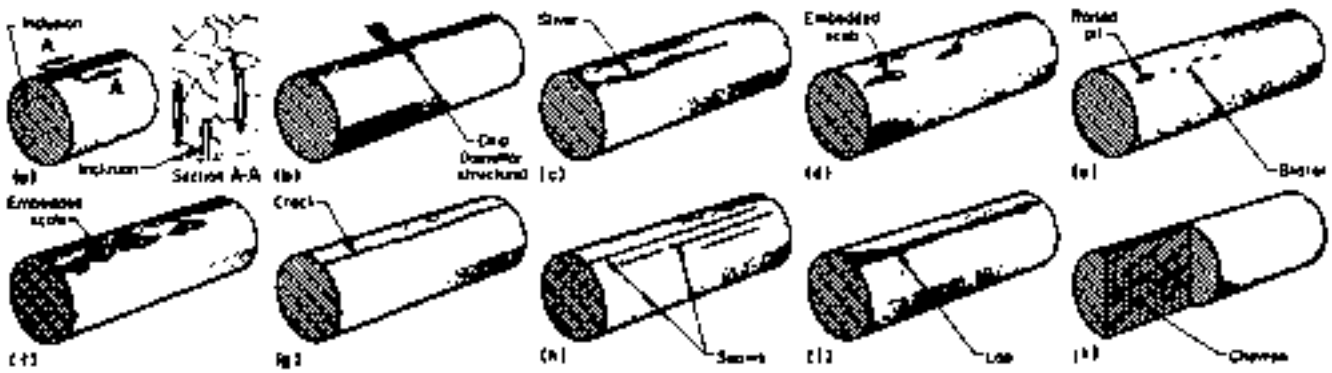


Fig. 2 Ten different types of flaws that may be found in rolled bars. (a) Inclusions. (b) Laminations from spatter (entrapped splashes) during the pouring. (c) Slivers. (d) Scabs are caused by splashing liquid metal in the mold. (e) Pits and blisters caused by gaseous pockets in the ingot. (f) Embedded scale from excessive scaling during prior heating operations. (g) Cracks with little or no oxide present on their edges when the metal cools in the mold, setting up highly stressed areas. (h) Seams that develop from elongated trapped-gas pockets or from cracks during working. (j) Laps when excessive material is squeezed out and turned back into the material. (k) Chevron or internal bursts. See text for additional discussions.

It must be clearly recognized that manufactured materials always have some imperfections or discontinuities that can be acceptable, if they do not interfere with the utility or service of a part. Laps, seams, bursts, hot tears, and thermal cracks are typically considered to be manufacturing defects. However, whether these manufacturing defects are a contributing factor in an in-service failure is a separate question that would need to be confirmed during a failure analysis. A discontinuity or flaw only becomes a service defect when it interferes

with the intended function and expected life of a part. This distinction is important in failure analysis, because a discontinuity or imperfection may be present, even though the failure is attributed to a different root cause.
The distinction between a manufacturing imperfection and a manufacturing flaw is thus critical in determination of root cause. Manufactured components typically contain geometric and material imperfections, but whether the imperfection caused failure and could therefore be a defect should be determined in many situations by analysis. In a cylindrical section under axial load, for example, imperfections along the centerline or at the surface are the most likely locations of crack initiation. In the absence of stress raisers at the surface, crack initiation from manufacturing imperfections is most likely along the centerline of an unnotched bar under tensile loading. In this case, if cracking initiates in another location, it has to do so because the local stress (residual and/or applied) was higher than along the centerline, although there may be exceptions. Some exceptions include the random distribution of fine quench cracks in steels at locations where martensite formed after quenching or the cracking of divorced cementite in the grain boundaries of low-carbon steel after cold working.
Similar arguments can be used to predict initiation sites for various kinds of bending loading and torsion loading. For example, for a three-point loaded beam, cracking is expected to initiate at the location of maximum bending moment. If it is not at that location, the implication is that a geometric or material imperfection has moved the location of the local maximum stress. Another example is rolling-contact fatigue, where maximum stress develops below the surface and is thus expected to cause subsurface crack initiation. Additionally, residual stresses may be distributed in the component as a result of prior mechanical/thermal processing. In this way, the significance of a material imperfection must be carefully evaluated in terms of stresses created by applied loads and part configuration.

Imperfections from the Ingot

Many flaws in wrought products can be traced back to the pouring and solidification of hot metal in molds during production of ingot. Except for forged powder metal components, the starting material in bulk working is a slab, ingot, billet, and so forth produced by casting into stationary molds or by continuous casting techniques. Primary deformation processes, such as hot rolling, tube piercing, extrusion, and open-die forging, are then used for converting the cast structure. Many large open-die forgings are forged directly from ingots, while most closed-die and upset forgings are produced from billets, bar stock, or a preform that has received some previous mechanical working. The product may be suitable for immediate application, but in many cases, it serves as the starting material for another (so-called secondary) deformation process, such as drawing, hot forging, cold forging, and sheet metalworking.
The major problems associated with melting and casting practice are the development of porosity and a condition known as scabs. Porosity is developed in cast ingots when they solidify and is of two types: pipe and blow holes. Factors that work against obtaining a perfect homogeneous product include:

- The fast shrinkage as the molten metals cools (roughly 5% in volume for steel)
- The gaseous products that are trapped by the solidifying metal as they try to escape from the liquid and semisolid metal
- Small crevices in the mold walls, which cause the metal to tear during the stripping operation
- Spatter during pouring, which produces globs of metal frozen on the mold walls because of the great difference in temperature of the mold surfaces and the liquid metal

Scabs are caused by improper ingot pouring, in which metal is splashed against the side of the mold wall. The splashed material, or scab, tends to stick to the wall and become oxidized. Scabs usually show up only after rolling or working and, as can be expected, give poor surface finish.
Ingot imperfections can seriously affect the performance and reliability of wrought products, and the types of the imperfections that can be traced to the original ingot product include:

- Chemical segregation
- Ingot pipe, porosity, and centerline shrinkage
- High hydrogen content

- Nonmetallic inclusions
- Unmelted electrodes and shelf
- Cracks, laminations, seams, pits, blisters, and scabs

The conversion practice must impart sufficient homogenization or healing to produce a product with sound center conditions. Ingot pipe, unhealed center conditions, or voids are melt-related discontinuities, but their occurrence in forgings is often a function of reduction ratio. Macroetching and ultrasonic inspection methods are the most widely used for identifying regions of unsoundness. Preliminary reduction of ingots or billets can also introduce flaws (some of which are similar to flaws that may occur during subsequent bulk working; see the "Forging Imperfections"). This includes internal bursts and various kinds of surface flaws.

Internal bursts occur where the work metal is weak, possibly from pipe, porosity, segregation, or inclusions. The tensile stresses can be sufficiently high to tear the material apart internally, particularly if the hot working temperature is too high. Such internal tears are known as forging bursts or ruptures. Similarly, if the metal contains low-melting phases resulting from segregation, these phases may cause bursts during hot working of the ingot or billet. Bursts can also occur during subsequent bulk-working operations.

Surface flaws from preliminary reduction may include:

- Laps appear as linear defects caused by the folding over of hot metal at the surface. These folds are worked into the surface but are not metallurgically bonded (welded) because of the oxide present between the surfaces. This creates a sharp discontinuity.
- Seams are a surface defect that also appears as a linear discontinuity. They occur from a crack, a heavy cluster of nonmetallic inclusions, or a deep lap (a lap that intersects the surface at a large angle). A seam can also result from a defect in the ingot surface, such as a hole, that becomes oxidized and is prevented from healing during working. In this case, the hole simply stretches out during forging or rolling, producing a linear cracklike seam in the workpiece surface.
- Slivers are loose or torn pieces of steel rolled into the surface. Rolled-in scale is scale formed during rolling.
- Ferrite fingers are surface cracks that have been welded shut but still contain the oxides and decarburization.
- Fins and overfills are protrusions formed by incorrect reduction during hot working.
- Underfills are the result of incomplete working of the section during reduction.
- Rolled-in scale

Chemical Segregation

The elements in a cast alloy are seldom distributed uniformly. Even unalloyed metals contain random amounts of various types of impurities in the form of tramp elements or dissolved gases; these impurities are also seldom distributed uniformly. Therefore, the composition of the metal or alloy varies from location to location. Unfortunately, such variation in chemical composition can often be significant and produce deleterious material conditions. This deviation from the mean composition at a particular location in a cast or wrought product is an imperfection termed segregation.

Chemical segregation originates in alloys during the solidification stage. Such deviations from the nominal composition are due to convection currents in the liquid, gravity effects, and redistribution of the solute during the formation of dendrites. Solute rejection at the solid-liquid interface during dendrite formation typically occurs during solidification, and thus a compositional gradient typically exists from the cores of dendrites to the interdendritic regions, with the latter enriched in alloying elements (solute) and low-melting contaminants. Dendrite arms also are generally lower in impurities, such as sulfur and phosphorus in steel, than the interdendritic regions. Consequently, the dendrite arms are stronger and, on working, do not deform and flow as readily as the matrix in which they are incorporated.

Microsegregation characterizes concentrations of elements in interdendritic regions that range in size from a few to several hundred microns. By contrast, macrosegregation is the gradient difference, measurable on a macroscale, in alloying elements from the surface to the center of an ingot or casting. Macrosegregation becomes more pronounced with increasing section size. Microsegregation, particularly within secondary arm branches, can be eliminated by homogenization. However, macrosegregation is harder to eliminate, because

complete homogenization would require longer times than are economically acceptable under production conditions. Therefore, in very large sections, gross differences in alloy concentration sometimes persist and are carried into the final product.

One function of hot working is to break up the cast (dendritic) structure and promote chemical homogeneity, and a minimum amount of cross-sectional reduction is usually required from the cast ingot to the billet. Hot working can partially correct the results of segregation by recrystallizing or breaking up the grain structure to promote a more homogeneous substructure. Initial working first causes flow in the weaker matrix (interdendritic) regions and tends to reorient the stronger dendrites in the direction of working. With increased mechanical working, the dendrites deform and fracture, thus becoming increasingly elongated.

A certain degree of alloy segregation occurs in all wrought products, and hot working can alleviate some of the inhomogeneity. However, if the ingot is badly segregated, hot working just tends to alter the shape of the segregation region into a banded structure. Figure 3 shows banding from a carbon-rich centerline condition in a hot-rolled 1041 steel. Figure 4 shows an extreme example of banding in a hot-rolled plain carbon steel (1022) in which alternate layers of ferrite and pearlite have formed along the rolling direction. The relationship between increasing percentages of reduction by hot rolling and the intensity of banding in type 430 stainless steel is demonstrated by Fig. 5.

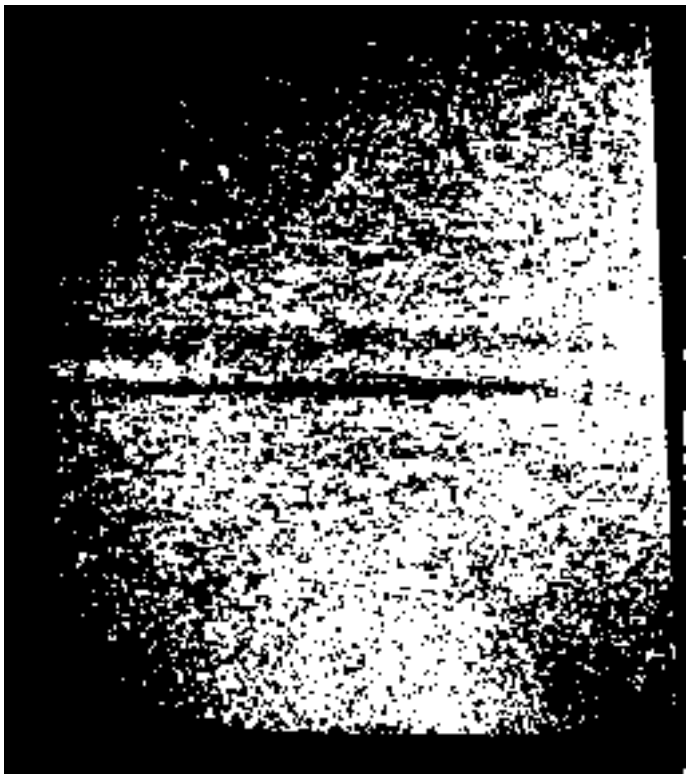


Fig. 3 Longitudinal section through a hot-rolled 1041 steel bar showing a carbon-rich centerline (dark horizontal bands) that resulted from segregation in the ingot. Picral. 3×. Courtesy of J.R. Kilpatrick

Fig. 4 Hot-rolled 1022 steel showing severe banding. Bands of pearlite (dark) and ferrite were caused by segregation of carbon and other elements during solidification and later decomposition of austenite. Nital. 250×. Courtesy of J.R. Kilpatrick

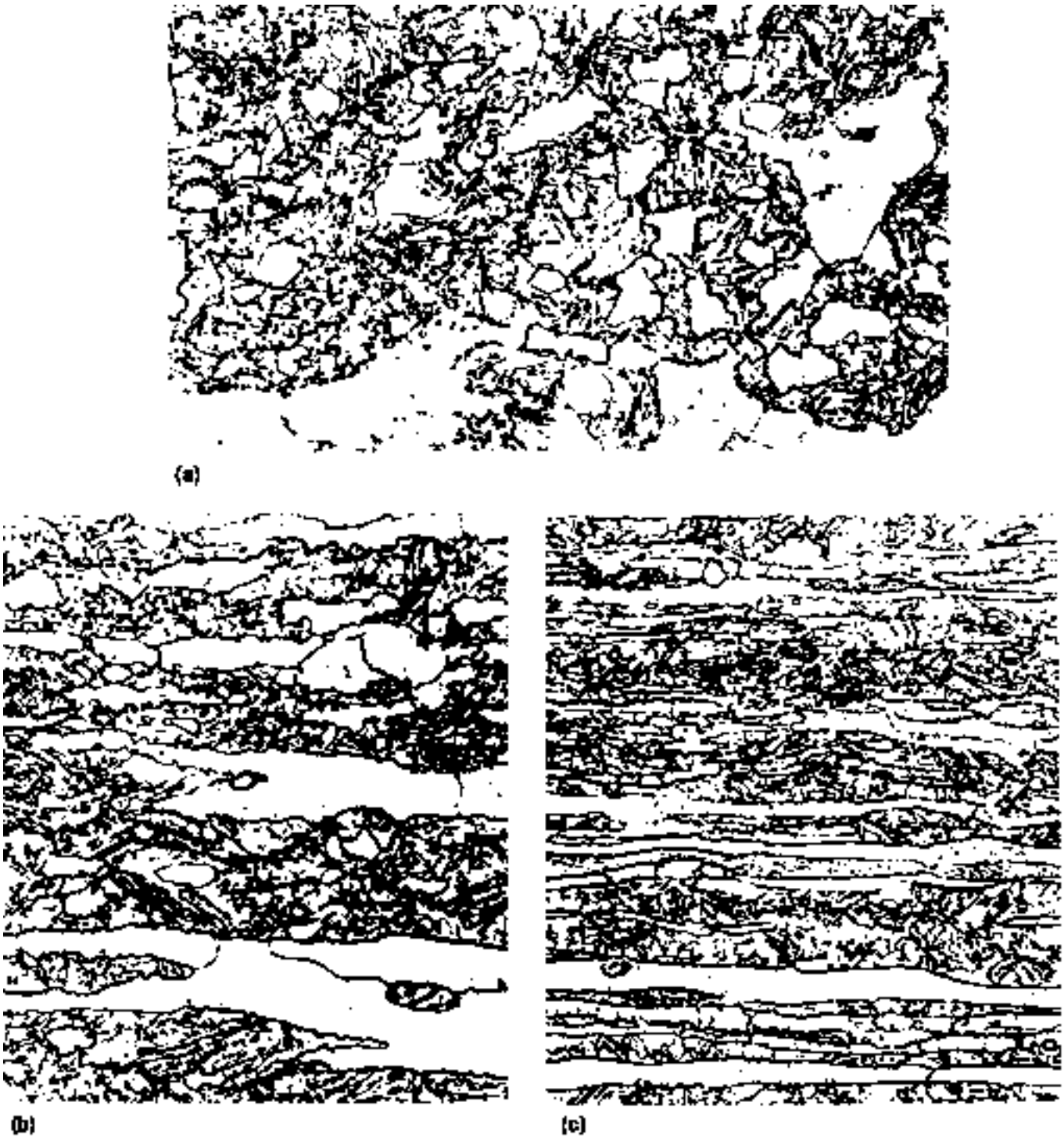


Fig. 5 Type 430 stainless steel hot rolled to various percentages of reduction showing development of a banded structure consisting of alternate layers of ferrite (light) and martensite (dark) as the amount of hot work is increased. (a) 63% reduction. (b) 81% reduction. (c) 94% reduction. 55 mL 35% HCl, 1 to 2 g potassium metabisulfite, 275 mL H_2O (Beraha's tint reagent No. 2). 500×

Depending on the kind and degree of segregation that develops during solidification, some degree of banding carries over to the wrought form. If banding is severe, it can lead to discontinuities that cause premature failure. For example, Fig. 6 shows the fatigue fracture of a carburized and hardened steel roller. Banded alloy segregation in the metal used for the rollers resulted in heavy, banded retained austenite, particularly in the carburized case, after heat treatment. When the roller was subjected to service loads, the delayed transformation of the retained austenite to martensite caused microcracks near the case-core interface. These internal microcracks nucleated a fatigue fracture that progressed around the circumference of the roller, following the interface between case and core.

Fig. 6 Fracture surface of a carburized and hardened steel roller. As a result of banded alloy segregation, circumferential fatigue fracture initiated at a subsurface origin near the case-core interface (arrow).

Excessive segregation also can have an adverse effect on subsequent fabrication and heat treatment. In heat treatable alloys, variations in composition can produce unexpected responses to heat treatments, which result in hard or soft spots, quench cracks, or other flaws. Excessive segregation that leads to significant variations in hardness can lead to premature failure and extreme difficulties during cold working or forming. In this case, one of the simplest and most effective tests for incoming material is a simple standard upset test. The details of such a test can be worked out between the supplier and the cold forger.

The methods to reveal the presence of segregation may depend on the alloy and expected impurities. Macroetching is commonly used, and the American Society for Testing and Materials (ASTM) has established a graded series (ASTM E 381) of macroetchings for center segregation in steel product. Segregations are revealed by differences in the severity of the etchant attack; segregations at the center may appear as a pipe or may be grouped in some fairly regular form about the center, depending on the shape of the ingot and the mechanical work done on it. Segregation as revealed by macroetching does not always indicate defective metal. A polished specimen should also be examined under the microscope to determine whether the revealed segregation is metallic or a concentration of nonmetallic impurities.

Sulfur Print Test. The microscopic identification of segregation may be supplemented by chemical or other means of testing. For regions with expected regions of sulfide sulfur-rich segregation, the sulfur print test (Ref 3) can be used. An example of a failure of a steel I-beam with high levels of carbon, sulfur, and phosphorus segregation in the middle of the section is given in Ref 4. The beam was lying flat on the ground near the seacoast under normal weather conditions. It was flame cut into two sections, as required for construction, and approximately $\frac{1}{2}$h after cutting, a violent sound was heard. The shorter section of the cut beam had split catastrophically into two portions along the entire length and approximately through the middle of the web.

Various samples were taken from both the broken and unbroken sections of the beam for analysis (chemistry, metallography, macroetching, impact testing, tensile testing) and sulfur printing. Sulfur prints taken at various locations indicated segregation of sulfides within a central zone approximately 2 mm (0.08 in.) wide in the thickness direction of the web that extended throughout the length of the beam. The breadth of the segregation zone varied from 60 mm (2.4 in.) at the end face of the unfractured section of the I-beam to almost the total width of the web in most of the fractured section. Sulfide segregation was not found in the flanges of the beam.

Failures similar to the one investigated have occasionally occurred in structural beams in the shop under no load, and a contributing factor was the presence of residual stresses in the material. Flame cutting caused a change in the distribution of the residual stresses, which, aided by low fracture toughness due to the poor quality of the beam, resulted in failure. The failure of the I-beam was probably caused by segregation of carbon, sulfur, and phosphorus within its web section, which resulted in decreased notch sensitivity and low fracture toughness with respect to crack propagation through the web. The detailed investigation (Ref 4) revealed segregation of high levels of carbon, sulfur, and phosphorus in the middle of the web and high residual stresses attributed to rolling during fabrication.

Example 1: Fracture of a Forging Die Caused by Segregation (Ref 5). A cross-recessed die of D5 tool steel fractured in service. The die face was subjected to shear and tensile stresses as a result of the forging pressures from the material being worked. Figure 7(a) illustrates the fractured die.

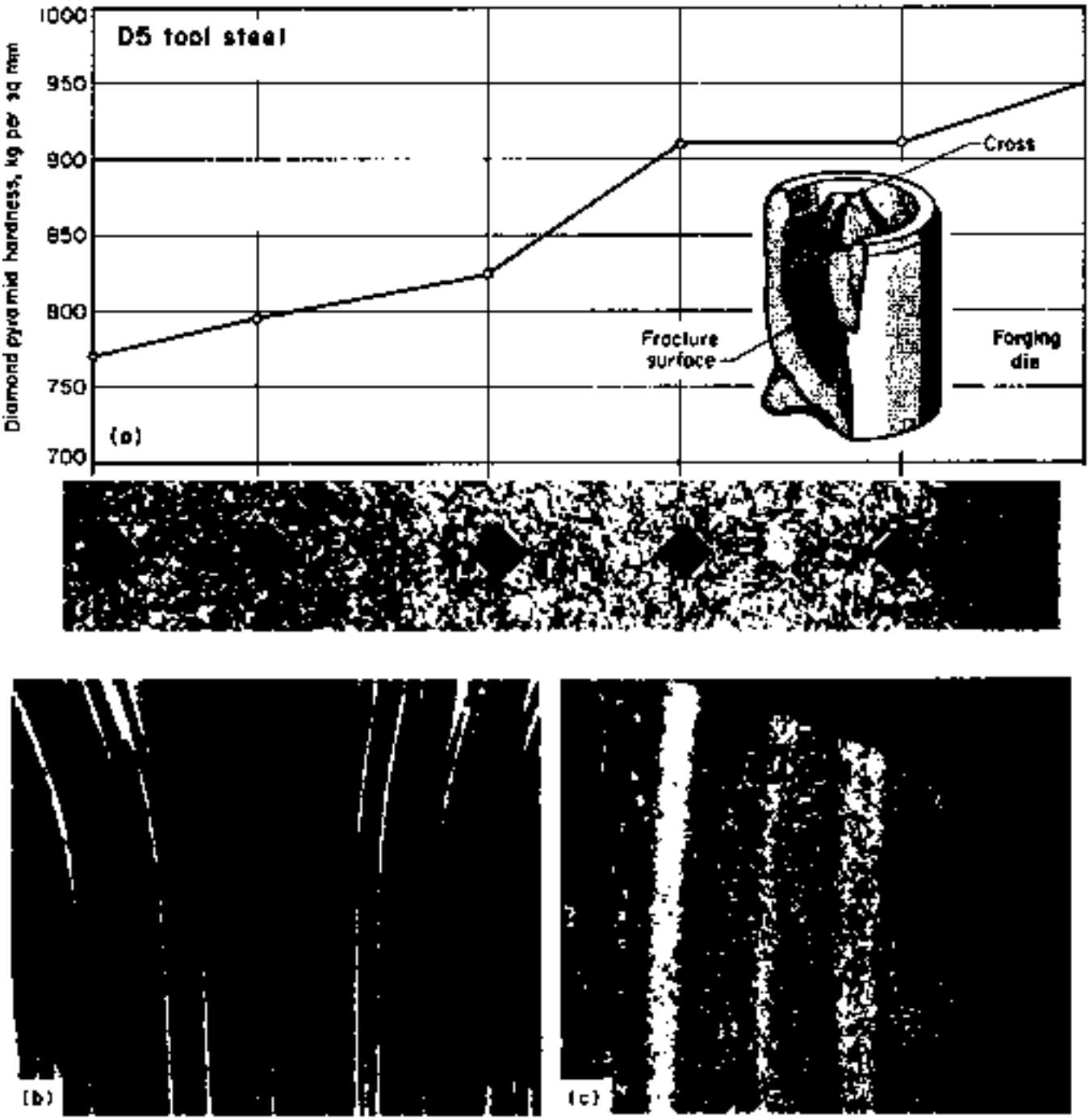


Fig. 7 A D5 tool steel forging die that failed in service because of segregation. (a) Hardness traverse correlated with the microstructure of the die. (b) Section through one arm of the cross on the recessed die face showing a severely segregated (banded) structure. Etched with 5% nital. (c) Micrograph of the segregated area. Etched with 5% nital. 200×

Investigation. A longitudinal section was taken through the die to include one arm of the cross on the recessed die face. The specimen was polished and examined in the unetched condition. Examination revealed the presence of numerous slag stringers.

The polished specimen was then etched with 5% nital. A marked banded structure was evident even macroscopically (Fig. 7b). Microscopic examination revealed that the pattern was due to severe chemical segregation or banding (Fig. 7c).

Hardness measurements were then conducted across the face of the specimen in locations corresponding to the banded and nonbanded regions. These results (Fig. 7a) showed that the segregated region is considerably harder than the neighboring material. The reason is that the increased carbon content of the segregated region, together with its higher alloy content, makes the region more responsive to what would have been normal heat treatment for this grade of tool steel.

The high-hardness material is also subject to microcracking on quenching; microcracks can act as nuclei for subsequent fatigue cracks. Examination of the fracture surface revealed that the fracture originated near the high-stress region of the die face; however, no indications of fatigue marks were found on either a macroscale or a microscale.

Conclusions. Failure of the die was the result of fracture that originated in an area of abnormally high hardness. Although fatigue marks were not observed, the fact was that the fracture did not occur in a single cycle but required several cycles to cause failure.

Example 2: Fatigue Fracture that Originated on the Ground Surface of a Medium- Carbon Steel Forging with a Notch-Sensitive Band Structure. The broken connecting end of a forged medium-carbon steel rod used in an application in which it was subjected to severe low-frequency loading is illustrated in Fig. 8(a); the part shown is from one of two identical rods that failed in service by fracture. In each instance, fracture extended completely through the connecting end in two places. The two fractured rods, together with two similar unused forged rods, were examined to determine the mechanism and cause of fracture.

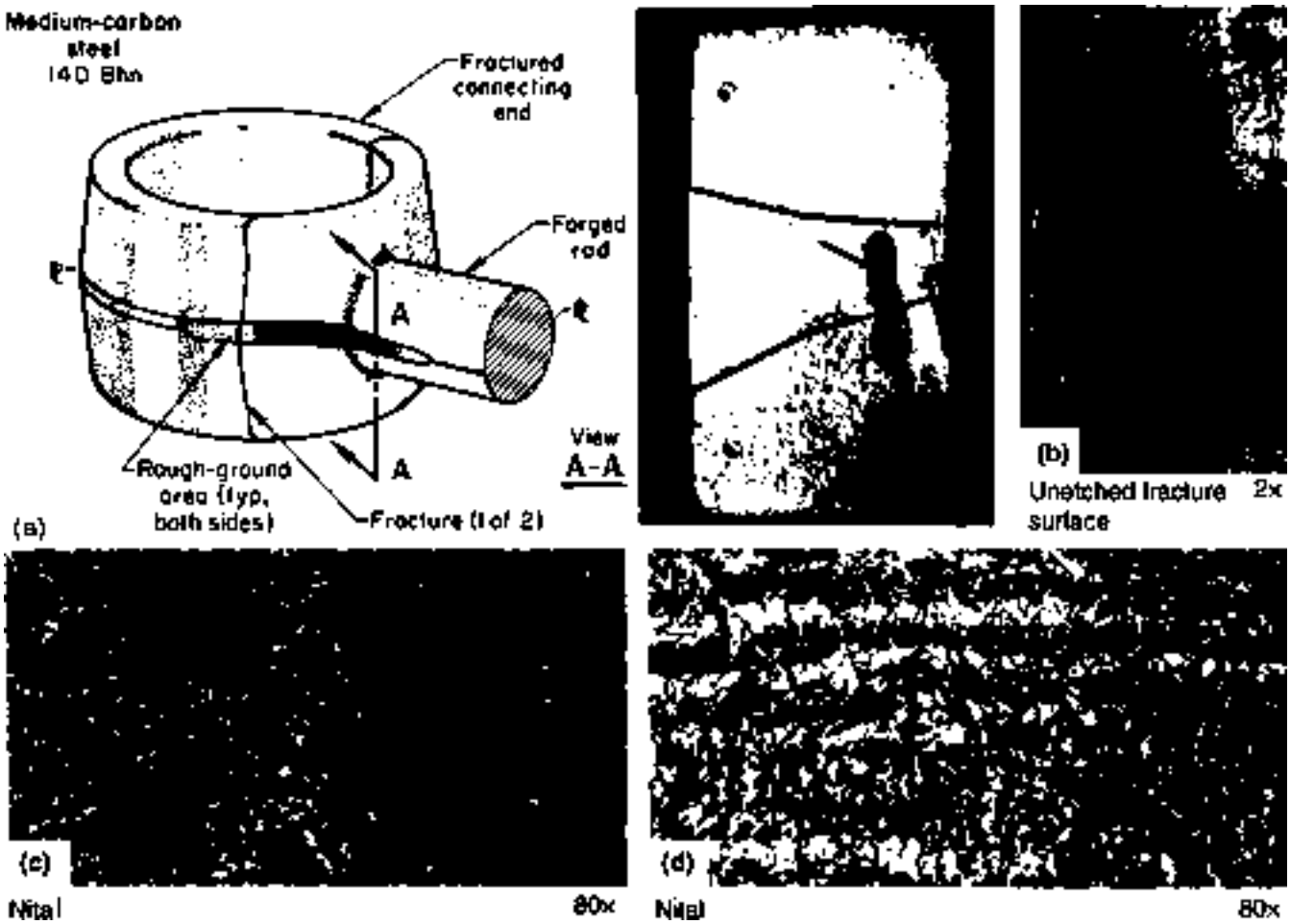


Fig. 8 Connecting end of forged rod, with banded structure from excessive segregation in the billet (Example 2). (a) Rod end showing locations of fractures at rough-ground areas at the parting line; in view A-A, dashed lines denote a rough-ground area, arrow points to a liquid-penetrant indication of an incipient crack. (b) Fracture surface, with beach marks indicating fracture origin at rough-ground surface. (c) Normal, homogeneous structure of an unused rod examined for comparison; this structure contains equal amounts of ferrite (light) and pearlite (dark). (d) Unsatisfactory structure of the fractured rod, which contains alternating bands of ferrite and pearlite.

Preliminary Examination. The material of the four forged rods was found by spectrographic analysis to be within the normal limits for the specified medium-carbon steel. Except for the fractures on the connecting ends of the two failed rods, no significant imperfections or evidence of damage were found by visual examination.

Surface hardness of the four rods, as measured at various points with a Rockwell tester, was equivalent to 140 HB—substantially lower than the specified hardness of 160 to 205 HB.

The fractures in the two failed rods were in areas of the transition regions that had been rough ground to remove flash along the parting line (see Fig. 8a). The fracture surfaces were fairly flat and were radial with respect to the annular connecting ends.

Low-magnification examination of the fracture surfaces revealed the presence of beach marks, indicating that the fractures had originated and propagated by fatigue. The location and curvature of the beach marks on the fracture surfaces of the two broken rods (see Fig. 8b) established that the fracture origin in each rod was at the rough-ground surface. The fatigue region of the fracture surface, which was quite smooth, extended approximately halfway through the thickness of the rod end.

The remainder of the area of the fracture surfaces, corresponding to final, fast fracture, had the typical appearance of brittle fracture, exhibiting little or no evidence of plastic deformation.

Liquid-Penetrant Examination. The four rods were checked for the presence of imperfections in the general vicinity of rough-ground areas, using liquid-penetrant inspection. This examination revealed an incipient crack 9.5 mm approximately ($\frac{3}{8}$ in.) long (see view A-A in Fig. 8) on one of the fractured rods. The crack was apparently a fatigue crack in the initiation stage and was located at a rough-ground area. Liquid-penetrant indications of several other incipient cracks that were smaller in size were also detected on the fractured rods in this examination; several were on rough-ground areas.

Metallographic Examination. Cross sections for metallographic examination were taken through the annular connecting ends of the two fractured rods and of the two unused rods and were polished and then etched with nital. At a magnification of 80 diameters, the microstructure of the unused rods (Fig. 8c) and that of the fractured rods (Fig. 8d) appeared greatly different.

The two unused rods has a fairly fine-grained, homogeneous structure containing approximately equal amounts of ferrite and pearlite—a normal structure for good-quality medium-carbon steel forgings of this composition that had been properly heat treated after forging.

The two fractured rods had a banded structure consisting of zones of ferrite (light) and pearlite (dark). Microhardness of the banded region was 140 HB in the light areas and 145 HB in the dark areas (converted from Vickers hardness readings, 100 g load).

Examination at 320 diameters of sections through the rough-ground areas of the fractured rods established that the incipient cracks found in liquid-penetrant inspection had originated at the surface in the banded region, in areas of ferrite where this constituent had been visibly deformed by grinding.

Discussion. The loads applied to the annular connecting ends of the rods were apparently complex, consisting of torsional, bending, and axial loads from the forged rods; these loads caused cyclic tensile circumferential stresses and bending stresses in the connecting ends.

The microhardness readings on the ferrite and pearlite bands were not considered to represent true hardness values because of the small size and shallowness of the bands. However, the following observations were considered to be significant with respect to the failures:

- The forged rods that had a normal microstructure (see Fig. 8c) did not fail.
- The forged rods that had a banded microstructure (see Fig. 8d) did fail.
- The hardness of all four rods, as measured on the surface using a Rockwell hardness tester, was the same and was equivalent to 140 HB—substantially below the specified range of 160 to 205 HB.
- The fractures originated in rough-ground areas, and incipient cracks were found in an area of ferrite that had been visibly deformed by grinding.

Conclusions. The rod-end fractures had originated and propagated by fatigue during continued exposure of the rods to severe cyclic loads in service. The loads on the rods developed tensile and bending stresses in the connecting ends. Contributory factors were:

- The presence of a notch-sensitive banded structure containing alternately soft (ferrite) and hard (pearlite) layers
- The presence of stress raisers produced by rough grinding to remove the forging flash along the highly stressed transition area between the rods and the connecting end
- Hardness (and, accordingly, strength) below the range specified for the part

The banded microstructure of the two fractured rods apparently resulted from severe segregation in the billet from which they were forged.
Corrective Action. It was recommended that:

- Closer control be exercised over the microstructure and hardness of the forgings
- The connecting end be finished more smoothly in the critical area
- Consideration be given to increasing the thickness (diametral with respect to the ring) of the connecting end in the transition area

Ingot Pipe, Porosity, and Centerline Shrinkage

A common imperfection in ingots is pipe, which is an internal shrinkage cavity formed during solidification of ingots. It occurs in the upper central portion of the ingot (Fig. 1) during solidification and contraction of the metal, when there may eventually be insufficient liquid metal to feed the last remaining portions as they contract. A concave cavity thus forms at the top of the ingot because of metal shrinkage during solidification. The cavity usually forms in the shape of a cone, hence the term pipe.
If not completely cropped before subsequent working, pipe may be carried through the various manufacturing processes to the finished product. The pipe becomes elongated and is found in the center of the final product, as shown in ingot B in Fig. 1. The pipe also can form internal lamination in rolled products (Fig. 9). These laminations may not be immediately evident following rolling but may become apparent during a subsequent forming operation. The occurrence of laminations is most prevalent in flat-rolled sheet products.

Fig. 9 Laminations in rolled steel sheet resulting from insufficient cropping of the pipe from the top of a conventionally cast ingot. Courtesy of V. Demski, Teledyne Rodney Metals

In addition to the primary pipe near the top of the ingot, secondary regions of piping and centerline shrinkage may extend deeper into an ingot (Fig. 10). Primary piping is generally an economic concern, but if it extends sufficiently deep into the ingot body and goes undetected, it can eventually result in a defective product. Detection of the pipe can be obscured in some cases, if bridging has occurred.

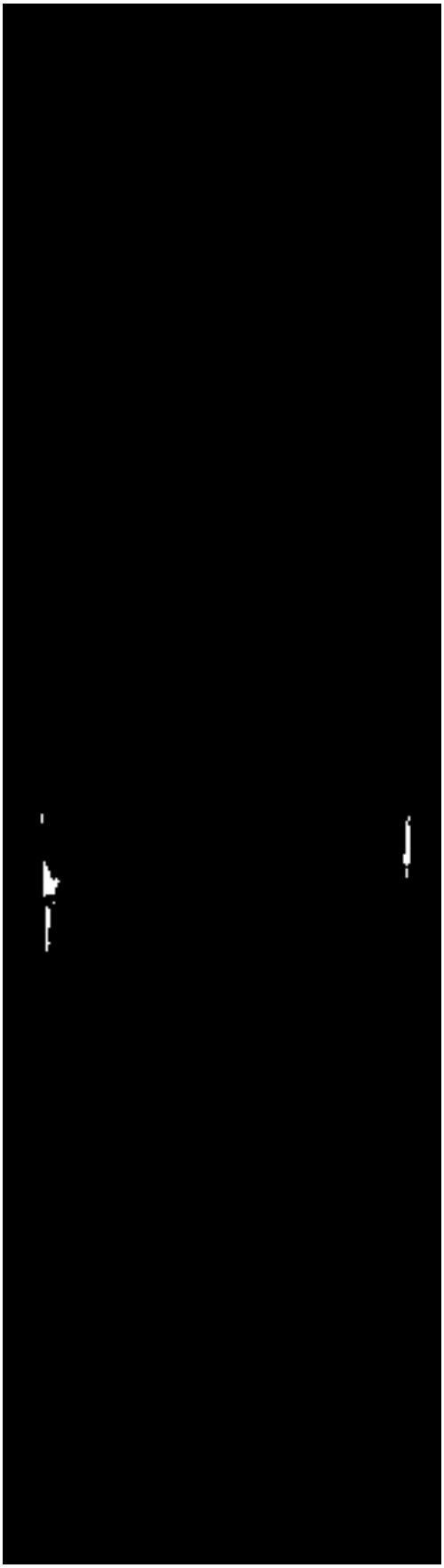


Fig. 10 Longitudinal section through an ingot showing extensive centerline shrinkage

Piping can be minimized by pouring ingots with the big end up, providing risers in the ingot top, and applying sufficient hot-top material (insulating refractories or exothermic materials) immediately after pouring. These techniques extend the time that the metal in the top regions of the ingot remains liquid, thereby minimizing the shrinkage cavity produced in this portion of the ingot. On the other hand, secondary piping and centerline

shrinkage can be very detrimental, because they are harder to detect in the mill and may subsequently produce centerline defects in bar and wrought products. Such a material condition may indeed provide the flaw or stress concentrator for a forging burst in some later processing operation or for a future product failure.

Pipe in steel products invariably is associated with segregated impurities, which are deeply attacked by the acid during macroetching. Cavities in the center that are not associated with deeply etched impurities often are mistaken for pipe, but such cavities usually can be traced to bursts caused by incorrect processing of the steel during forging or rolling. Pipe should be visible after deep etching; it usually can be distinguished from bursts by the degree of sponginess surrounding the defect. Piped metal usually exhibits considerably more sponginess than burst metal. The micrograph of Fig. 11 is an example of secondary pipe, and the micrograph in Fig. 12 is an example of a cavity that appears to be a pipelike flaw.

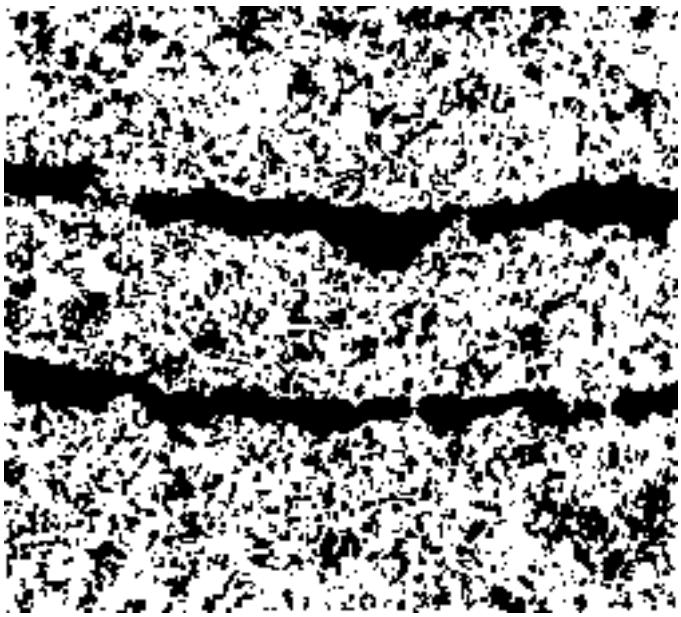


Fig. 11 1038 steel bar, as-forged. Longitudinal section displays secondary pipe (black areas) that was carried along from the original bar stock into the forged piece. Gray areas are pearlite; white areas, ferrite. 2% nital. 50×

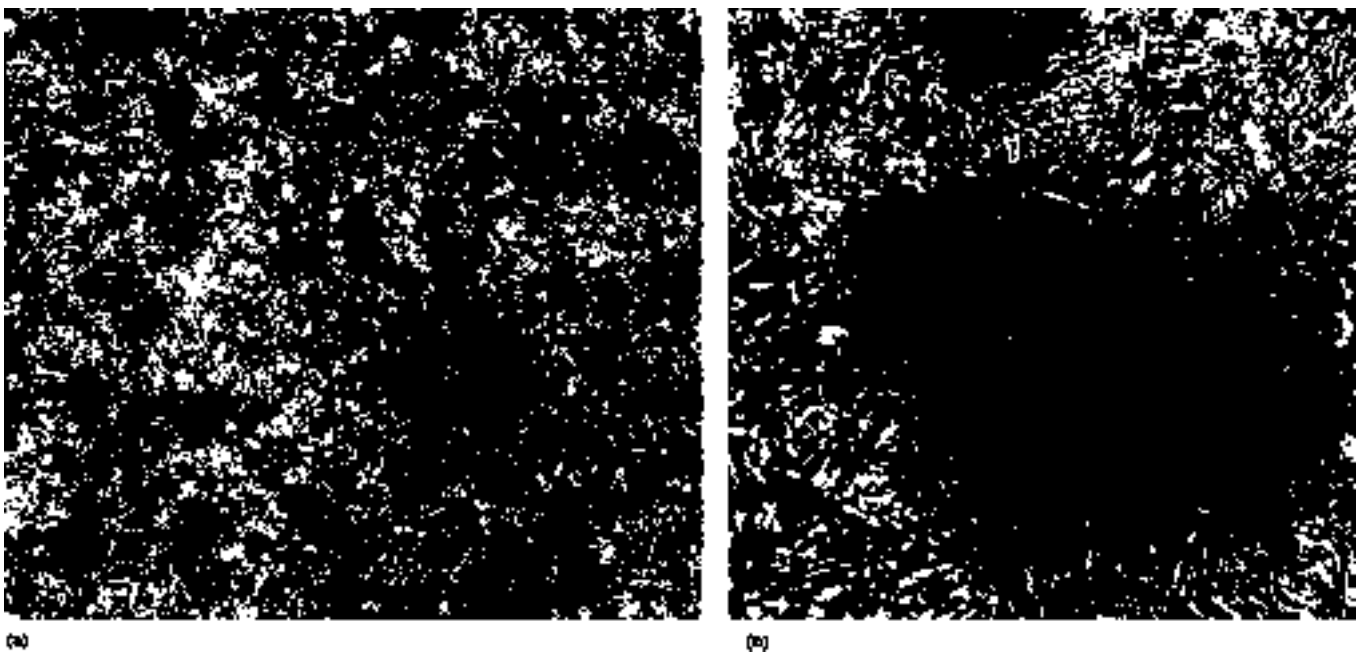


Fig. 12 Inclusions and a pipelike cavity in tempered martensite of AISI E4340 steel (Example 4). (a) 100×. (b) 600×. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

Example 3: Fatigue Fracture of Alloy Steel Valve Springs Because of Pipe. Two outer valve springs broke during production engine testing and were submitted for laboratory analysis. The springs were from a current production lot and had been made from air-melted 6150 pretempered steel wire. The springs were 50 mm (2 in.) in outside diameter and 64 mm (2.5 in.) in free length, had five coils and squared-and-ground ends, and were made of 5.5 mm (0.2 in.) diameter wire. Because both failures were similar, the analysis of only one is discussed.

Investigation. The spring (Fig. 13a) broke approximately one turn from the end. Fracture was nucleated by an apparent longitudinal subsurface defect. Magnetic-particle inspection did not reveal any additional cracks or defects.

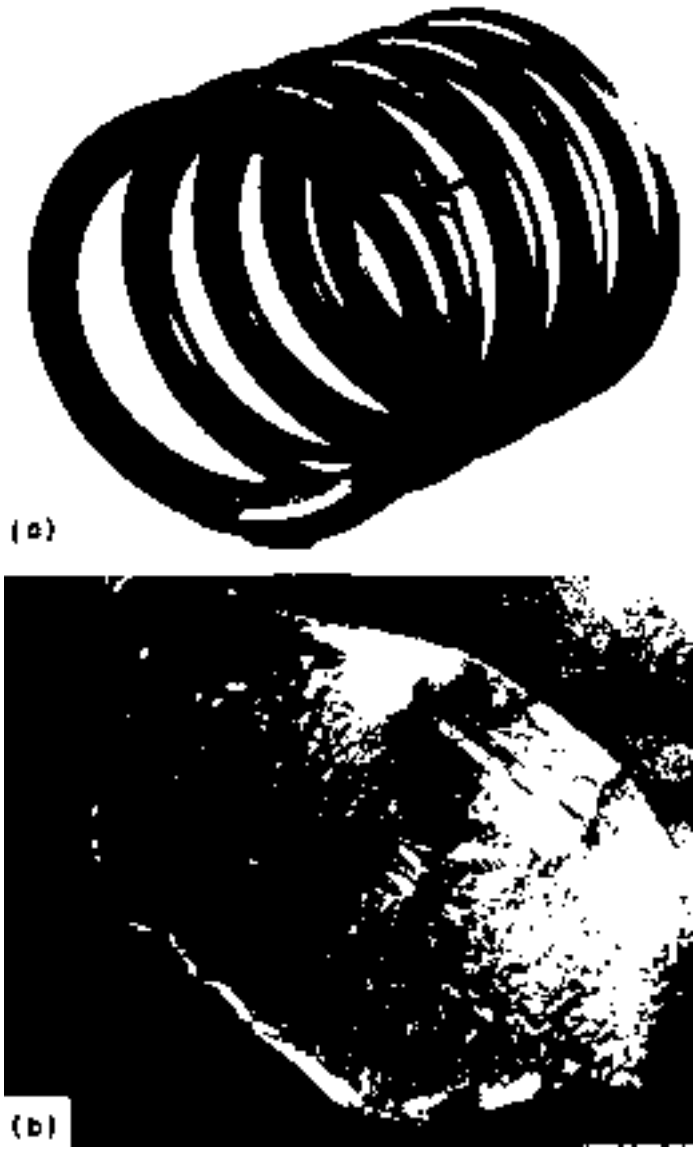


Fig. 13 Valve-spring failure due to residual shrinkage pipe. (a) Macrograph showing fracture, as indicated by arrow. (b) Fracture surface; pipe is indicated by arrow.

Microscopic examination of transverse sections of the spring adjacent to the fracture surfaces revealed that the defect was a large pocket of nonmetallic inclusions at the origin of the fracture. The inclusions were alumina and silicate particles. Partial decarburization of the steel was evident at the periphery of the pocket of inclusions. The composition and appearance of the inclusions and the presence of the partial decarburization indicated that the inclusions could have been associated with ingot defects. The defect was 0.8 mm (0.03 in.) in diameter (max), 25 mm (1 in.) long, and 1.3 mm (0.05 in.) below the surface (Fig. 13b).

The steel had a hardness of 45 to 46 HRC, a microstructure of tempered martensite, and a grain size of ASTM 6 or 7, all of which were satisfactory for the application. The fracture surface contained beach marks typical of fatigue and was at a 45° angle to the wire axis, which indicated torsional fracture.

Conclusions. The spring fractured by fatigue; fatigue cracking was nucleated at a subsurface defect that was longitudinal to the wire axis. The stress-raising effect of the defect was responsible for the fracture.

High Hydrogen Content

A major source of hydrogen in certain metals and alloys is the reaction of water vapor with the liquid metal at high temperatures. The water vapor may originate from the charge materials, slag ingredients and alloy additions, refractory linings, ingot molds, or even the atmosphere itself, if steps are not taken to prevent such contamination. The resulting hydrogen goes into solution at elevated temperatures; but as the metal solidifies after pouring, the solubility of hydrogen decreases, and it becomes entrapped in the metal lattice.

Hydrogen can be absorbed many places during primary and secondary manufacturing. It can be picked up during melting, teeming, welding, pickling, or electroplating. If hydrogen is absorbed into the base metal, it must be removed by a bake-out heat treatment. Otherwise, severe embrittlement of the base metal may occur, especially in steels with hardnesses above approximately 35 HRC. High-strength, highly stressed parts can crack and fracture as a result of hydrogen embrittlement. Failure by hydrogen embrittlement is even more likely to occur if high residual tensile stresses are present.

The deleterious effects of hydrogen depend on the alloy and the concentration of hydrogen. Hydrogen concentrations in excess of only 1 ppm have been related to the degradation of mechanical properties in high-strength steels, especially ductility, impact behavior, and fracture toughness. Hydrogen concentration in excess of approximately 5 ppm has been associated with hydrogen flakes, which are small cracks produced by hydrogen that has diffused to grain boundaries and other preferred sites (such as inclusions or matrix interfaces). Flaking or hairline cracking in steels, particularly in forgings, heavy-section alloy plate steels, and railroad rails, has been an ongoing problem since the beginning of steelmaking and is reviewed in Ref 6. At still higher concentrations of hydrogen in metal, blistering can occur (e.g., see Fig. 18 of “Visual Examination and Light Microscopy” in *Fractography*, Volume 12 of *ASM Handbook*).

Excessive gas contents in ingots are most serious when oxide inclusions are present. Inclusions provide sites for the agglomeration of gas, and thus molten alloys are often treated to reduce the hydrogen content and alkali element concentration as well as to remove nonmetallic inclusions. For example, vacuum degassing is used to remove dissolved oxygen and hydrogen from steel, thus reducing the number and size of indigenous nonmetallic inclusions and the likelihood of internal fissures or flakes caused when hydrogen content is higher than desired.

Despite these methods of producing high-quality metal, problems sometimes arise, particularly if there are no methods available for continuously monitoring metal quality. Hydrogen flaking also is a potential problem for carbon tool steels and for some of the medium-carbon low-alloy steels, such as those used as prehardened plastic-mold alloys. Figure 14(a) shows a die made from AISI O1 tool steel that was found to be cracked after heat treatment. When opened (Fig. 14b), these cracks exhibited a coarse, shiny, faceted appearance. The cracks in this part were longitudinally oriented and were confined to the center of the section. Flakes are not observed in the outer region of sections, because there is apparently more time for the hydrogen in the outer region to diffuse to a safe level. Flakes that were close to the surface of the die exhibited temper color, but some of the deeper flakes did not. The temper-colored flakes apparently opened during quenching.

Fig. 14 Die made from AISI O1 tool steel that was found to be cracked after heat treatment. (a) Longitudinal cracks after the surface was swabbed with 5% nital. (b) One of the cracks opened, revealing features typical of hydrogen flakes. 6.5×

Nonmetallic Inclusions

Nonmetallic inclusions are an inevitable consequence of commercial alloys and their respective melting or casting practices. Inclusions that originate in the ingot are carried onto wrought products, even though their shapes may be appreciably altered. Furthermore, additional nonmetallic matter, such as oxides, may develop during intermediate hot working stages and also end up in the finished form.
Two categories of nonmetallic inclusions in metals are:

- Those that are entrapped inadvertently and originate almost exclusively from foreign matter, such as refractory linings, that is occluded in the metal while it is molten or being cast
- Those that form in the metal because of a change in temperature or composition

Inclusions of the latter type are produced by separation from the metal when it is in either the liquid or the solid state. Some nonmetallic inclusions form in the liquid before solidification; others form during solidification. In steels, for example, aluminates and silicates generally form before solidification, while sulfides form during solidification. Manganese sulfide inclusions frequently form in the interdendritic regions and primary grain boundaries of steel, where the last of the liquid freezes.
Depending on the alloy, oxides, sulfides, nitrides, or other nonmetallic compounds may occur when solubility limits in the matrix are exceeded. Because these compounds are products of reactions within the metal, they are normal constituents of the metal, and conventional melting practices cannot completely eliminate such inclusions. However, it is desirable to keep the type and amount of inclusions to a minimum, so that the metal is relatively free from those inclusions that may cause unwanted discontinuities. Air-melted alloys commonly contain inclusions. Vacuum- or electroslag-remelted alloys more commonly contain nonmetallic inclusions, such as titanium carbonitrides or carbides, when carbon or the hardening element form precipitate during stabilization and aging cycles. Homogenizing cycles prior to conversion or at an early stage of conversion can be done for re-solution of some precipitates, but not all nonmetallic inclusions are affected by homogenization. In steels, manganese sulfides can be homogenized.
Nonmetallic inclusions are unquestionably one of the most common imperfections involved in problems or failures. Nonmetallic inclusions can easily become stress concentrators because of their discontinuous nature and incompatibility with the surrounding composition. This combination may very well yield flaws of critical size that, under appropriate loading conditions, result in fracture. Inclusions that are very brittle and tend to fracture and fragment during working operations also can be very detrimental. In fact, with certain refractory and slag inclusions, fragmentation frequently occurs when the amount of cross-sectional reduction in the workpiece is large.
The deleterious nature of nonmetallic inclusions depends on several factors, including chemical composition of the inclusion, volume percentage, shape, orientation, and the mechanical/physical properties of the inclusion as compared to its surrounding matrix (Ref 7). One study has demonstrated that the local constraint of the matrix by hard inclusions during tensile loading can produce severe stress concentrations that depend on the elastic moduli, size, shape, and orientation of the inclusions (Ref 8). For example, inclusions such as manganese sulfide, which are very deformable at hot working temperatures, elongate in the direction of working. This results in an anisotropic material condition, which, depending on its severity, can produce substantial decreases in transverse mechanical properties, such as ductility, fatigue life, and fracture toughness.
The metal-producing industry attempts to control nonmetallic inclusions, but inclusions cannot be removed with complete assurance. In demanding applications, inclusion shape-control methods are used to produce inclusions predominately with a spherical shape, which is less of a stress concentrator with the surrounding matrix and less damaging than inclusions with more angular shapes. The principle of inclusion shape control is to reduce the surface energy between the inclusion and the metal—normally through suitable chemistry modification of the metal. Sulfide inclusions, for example, have been shape controlled by the addition of calcium or rare earth metal treatments. These inclusions are not plastic or deformable at typical hot working temperatures and retain their globular shape. Therefore, they are less injurious with respect to ductility, toughness, and fatigue life in finished wrought products.
With the number and size of inclusions held roughly constant, inclusion shape control reduces the likelihood of crack initiation from an inclusion. This is shown in Fig. 15 (Ref 9) for a rolled constructional steel with calcium treatment, which provides longer life under cyclic loading due to inclusion shape modification. The advantage of a spherical inclusion shape is also evident in Charpy V-notch impact data for constructional steels (Fig. 16)(Ref 10). Scanning electron microscope (SEM) fractographs also show a pronounced difference in

deformation mode with an inclusion shape modification (Fig. 17). Local ductility increases adjacent to the inclusions when the shape is spherical.

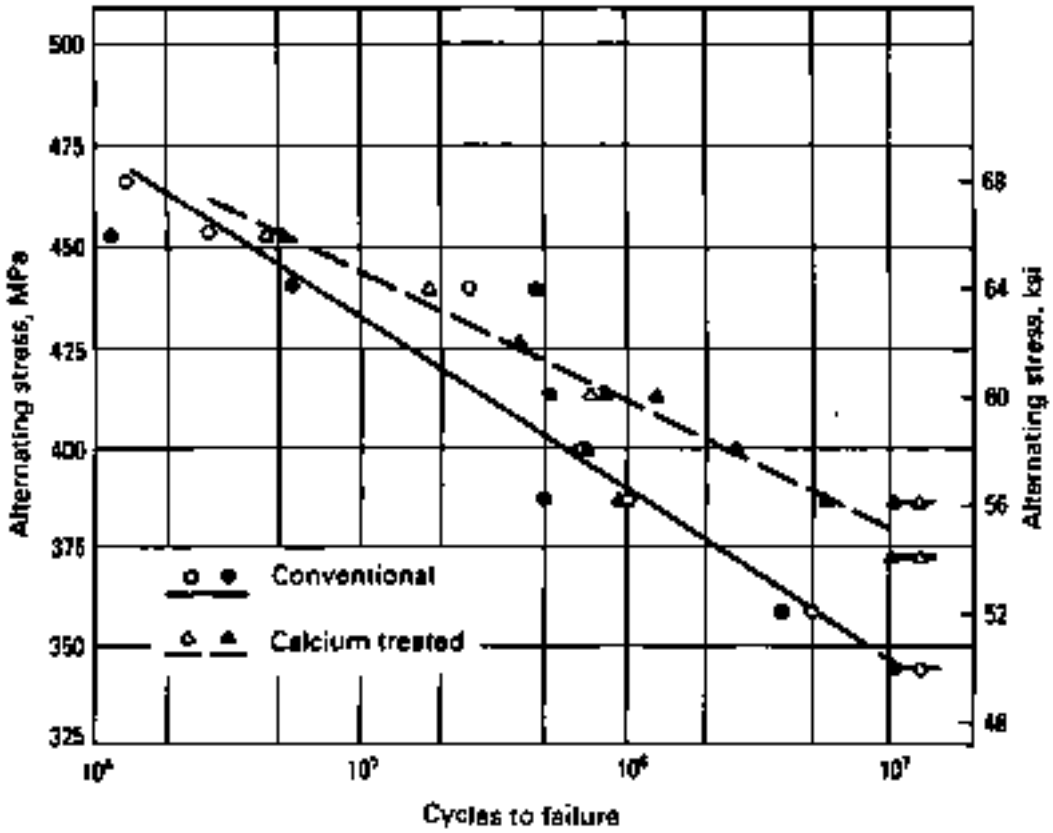


Fig. 15 Comparison of axial fatigue data for untreated and calcium-treated rolled ASTM A516 steel. 51 mm (2 in.) thick plates tested with alternating stress ratio of 0.1. Source: Ref 9

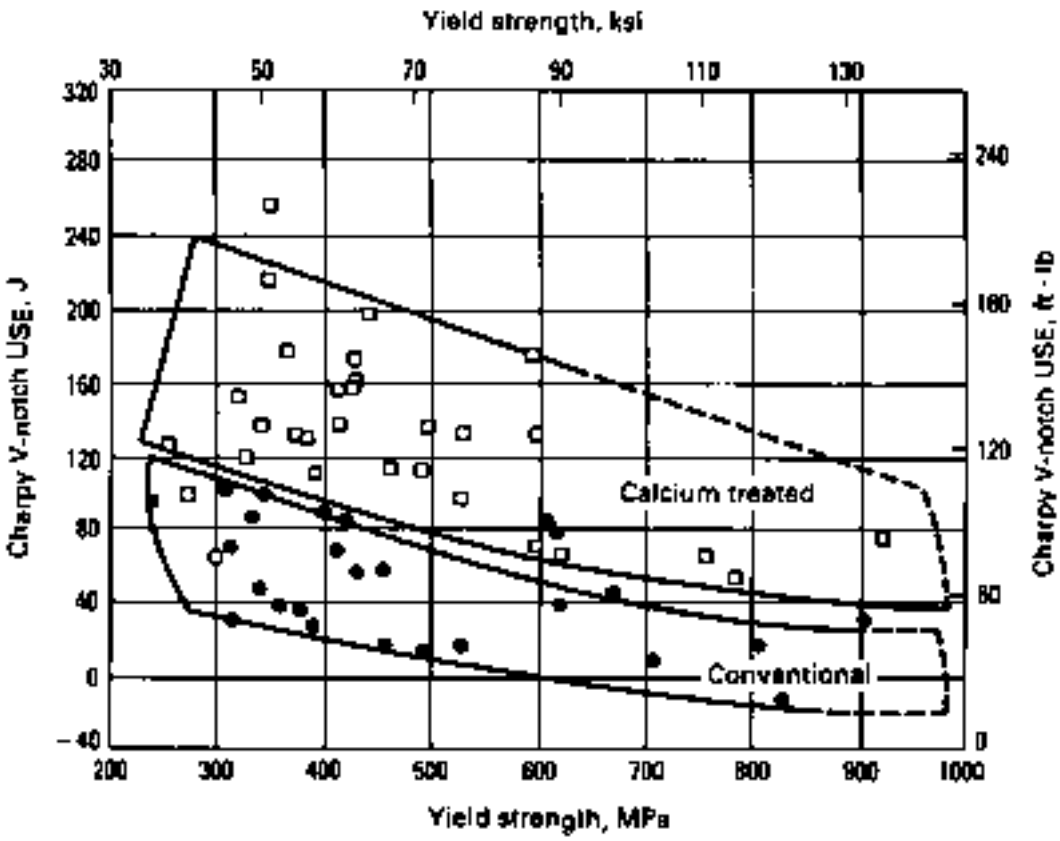


Fig. 16 Comparison of Charpy V-notch upper-shelf energies (USEs) for several grades and thicknesses of untreated and calcium-treated steel. Source: Ref 10

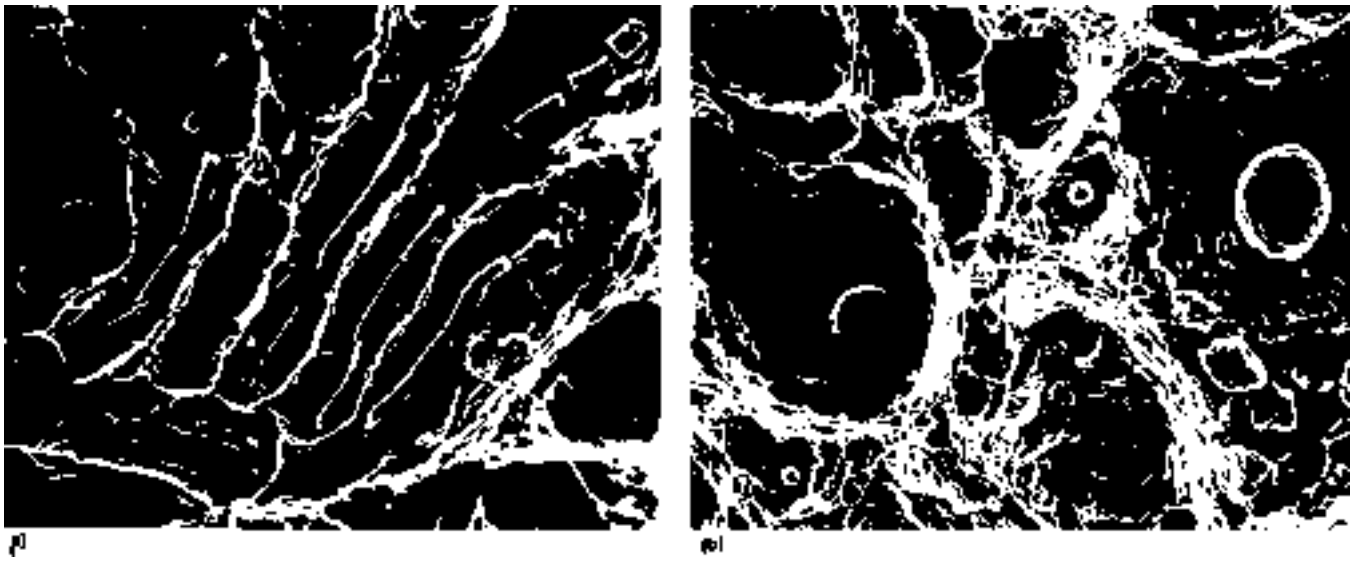


Fig. 17 Scanning electron microscope fractographs showing the effect of calcium treatment on the fracture morphology of ASTM A633C steel impact specimens. (a) Untreated steel with type II manganese sulfide inclusions showing evidence of brittle fracture. (b) Calcium-treated steel with spherical inclusions; the fracture is ductile. Courtesy of A.D. Wilson, Lukens Steel Company

Macroetching can provide a good indication of cleanliness, but further metallography is desirable if information is needed on the character of nonmetallic inclusions. Nonmetallic inclusions, although they usually appear as pits or pinholes after macroetching, must not be confused with pits that result from etching out metallic segregations or from variations in etching procedure. When nonmetallic inclusions are suspected in highly alloyed steels that may contain metallic segregations, an annealed specimen should be compared with a hardened and tempered specimen etched in the same way. If the etching pits are the result of nonmetallic inclusions, they appear similar in the annealed and the hardened specimens. If they are the result of metallic segregation, they differ in appearance, being more prominent in the hardened specimen.

Example 4: Failure Analysis of a Helicopter Main Rotor Bolt (Ref 11). During assembly, a helicopter main rotor made of AISI E4340 steel failed while being torqued. The torque being applied reached approximately 100 N · m (900 lbf · in.) when the pitch horn bolt sheared and shot across the hangar. No other bolts were reported to have failed during assembly.

The rotor bolt fractured in the black-coated region between the threads and the taper section of the shank. The head, the straight portion of the shank, and the nut were gold colored. The fracture surfaces were first cleaned with soap and water and were then ultrasonically cleaned to remove surface debris. Longitudinal and transverse sections were prepared for optical microscopy. The bolt surface was also inspected for identification of coatings and surface treatments.

Figure 18 shows the two mating halves of the fracture. The bolt had been heated to obtain a tempered martensitic microstructure. The metallographic specimen showed a significant amount of nonmetallic inclusions (Fig. 12). Figure 12 also shows a pore, probably a pipe defect in the original bar stock.

Fig. 18 Macrograph of failed bolt

Figure 19 shows a SEM view of the fracture initiation region. The white inclusions were nonmetallic particles that were high in calcium. Various spectra and x-ray maps were collected. The x-ray maps illustrated the presence of calcium in the white spots. These spots were thus interpreted to be slag inclusions in the steel. Areas containing the inclusions were cleaned with a soft brush several times to ascertain that the x-ray indications were not debris that collected on the fracture surface during storage and handling after the failure. The gold-colored areas of the bolt were found to have been cadmium plated, followed by a chromate treatment. The black-coated surfaces contained molybdenum and sulfur, indicating a MoS_2 treatment. Specimen composition, as determined by optical emission spectrometry, was in agreement with nominal specifications for AISI E4340 steel.

Fig. 19 Scanning electron microscope micrographs showing segregation of inclusions along the edge where fracture apears to have initiated.

Conclusion and Remedial Action. It was initially suspected that hydrogen embrittlement caused crack initiation. It was concluded, however, that the failure was caused by a high concentration of nonmetallic inclusions. Inspection of other bolts from the same shipment was recommended.

Other Ingot-Related Imperfections

Laminations in wrought products may occur from various types of discontinuities introduced during the ingot stage. As previously noted, pipe can form into laminations in rolled products (Fig. 9). Laminations may also form spatter (entrapped splashes) during the pouring of the molten metal into the ingot mold. These imperfections are elongated during rolling or other working and are usually subsurface, as shown in Fig. 2(b) for bar stock. Figure 2(b) illustrates a lamellar structure opened up by a chipping tool.

Slivers are most often caused by a rough mold surface, overheating prior to rolling, or abrasion during rolling. Very often, slivers are found with seams. Slivers usually have raised edges, as shown in Fig. 2(c).

Scabs are caused by improper ingot pouring, in which metal is splashed against the side of the mold wall. The splashed material, or scab, tends to stick to the wall and become oxidized. The metal first freezes to the wall of the mold, then becomes attached to the ingot, and finally becomes embedded in the surface of the wrought product (e.g., Fig. 2d). Scabs usually show up only after rolling and, as can be expected, give poor surface finish. Scabs thus bear some similarity to laminations.

Pits and Blisters. Gaseous pockets in the ingot often become pits or blisters on the surface or slightly below the surface of bar products. Other pits may be caused by overpickling to remove scale or rust. Pits and blisters are both illustrated in Fig. 2(e). Embedded scale may result from the rolling or drawing of bars that have become excessively scaled during prior heating operations. The pattern illustrated in Fig. 2(f) is typical.

Cracks and seams are often confused with each other. Cracks with little or no oxide present on their edges may occur when the metal cools in the mold, setting up highly stressed areas. Seams develop from these cracks during rolling as the reheated outer skin of the billet becomes heavily oxidized, transforms into scale, and flakes off the part during further rolling operations. Cracks also result from highly stressed planes in cold-drawn bars or from improper quenching during heat treatment. Cracks created from these latter two causes show no evidence of oxidized surfaces. A typical crack in a bar is shown in Fig. 2(g).

Unmelted electrodes and shelf are two other types of ingot flaws that can impair forgeability. Unmelted electrodes are caused by chunks of electrodes being eroded away during consumable melting and dropping down into the molten material as a solid. Shelf is a condition resulting from uneven solidification or cooling rates at the ingot surfaces. The consumable-melting operation has occasionally been continued to a point where a portion of the stinger rod is melted into the ingot, which may be undesirable, because the composition of the stinger rod may differ from that of the alloy being melted. To prevent this occurrence, one practice is to weld a wire to the stinger rod, bend the wire down in tension, and weld the other end of the wire to the surface of the electrode a few inches below the junction of the stinger rod and the electrode. When the electrode has been consumed to where the wire is attached to it, the wire is released and springs out against the side of the crucible, thus serving as an alarm to stop the melting. A disadvantage of this practice is that the wire may become detached and contaminate the melt.

Unmelted pieces of electrode or shelf conditions appear infrequently in vacuum-melted alloys. Either of these conditions can seriously degrade forgeability. Macroetching or another appropriate type of surface inspection of the machined forging or the billet is the most effective method of detecting unmelted electrodes or shelf conditions.

References cited in this section

3. G.F. Vander Voort, *Metallography: Principles and Practice*, McGraw-Hill, 1984 (out of print), reprinted 1999 ASM International

4. R.K. Dayal, J.B. Gnanamoorthy, and P. Rodriguez, Failure of an I-Beam, *Handbook of Case Histories in Failure Analysis*, Vol 1, ASM International, 1992, p 378–381

5. E. Kauczor, *Prakt. Metallogr.* Vol 8, July 1971, p 443–446

6. G. Vander Voort, Embrittlement of Steels, *Properties and Selection: Iron, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, p 689–736

7. P.A. Thornton, *J. Mater. Sci.,* Vol 6, 1971, p 347

8. D.V. Edmonds and C.J. Beevers, *J. Mater. Sci.,* Vol 3, 1968, p 457

9. A.D. Wilson, Calcium Treatment of Plate Steels and Its Effect on Fatigue and Toughness Properties, *Proc. 11th Annual Offshore Technology Conference,* OTC 3465, May 1979, p 939–948

10. A.D. Wilson, Effect of Calcium Treatment on Inclusions in Constructional Steels, *Met. Prog.,* Vol 121 (No. 5), April 1982, p 41–46

11. M.D. Chaudhari, Failure Analysis of a Helicopter Main Rotor Bolt, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 392

Forging Imperfections

In many cases, the internal and surface imperfections that occur during forging are the same as, or at least similar to, those that may occur during hot working of the ingots or billets, as briefly introduced in the previous section. The inspection of forgings and the types of flaws for specific alloy forgings (steel, nickel, aluminum, and titanium alloys) are described in more detail in the article "Nondestructive Inspection of Forgings" in *Nondestructive Evaluation and Quality Control,* Volume 17 of *ASM Handbook.* A brief review for nonferrous alloys follows:

- *Nickel-base heat-resistant alloys:* Most of the flaws found in forgings of heat-resistant alloys can be categorized as those related to scrap selection, melting, or primary conversion to bar or billet or those that occur during forging or heat treatment. Nickel-base superalloys are highly susceptible to surface contamination during heating for forging. Fuel oils containing sulfur induce a grain-boundary attack, which causes subsequent rupturing during forging. Paint or marking crayons with high levels of similar contaminants cause similar areas of grain-boundary contamination.
- *Aluminum alloys:* The internal discontinuities that occur in aluminum alloy forgings are ruptures, cracks, inclusions, segregation, and occasionally porosity. Ruptures and cracks are associated with temperature control during preheating or forging or with excessive reduction during a single forging operation. Cracks can also occur in stock that has been excessively reduced in one operation. Inclusions, segregation, and porosity result from forging stock that contains these types of discontinuities.
- *Titanium alloys:* Discontinuities that are most likely to occur in titanium alloy forgings are usually carried over in the bar or billet. Typical discontinuities in titanium alloy forgings are alpha-stabilized voids, macrostructural defects, unsealed center conditions, clean voids, and forging imperfections. Alpha-stabilized voids are among the most common discontinuities found in forgings of titanium alloys. Three principal macrodefects are commonly found in ingot, forged billet, or other semifinished product forms. These include high-aluminum defects (type II defects), high-interstitial defects (type I defects or low-density interstitial defects), and beta flecks.

The most common internal imperfections found in steel forgings are pipe, segregation, nonmetallic inclusions, and stringers. Internal flaws caused by forgings also include cracks or tears, which may result either from forging with too light a hammer or from continuing forging after the metal has cooled down below a safe forging temperature. Other flaws in steel forgings that can be produced by improper die design or maintenance are internal cracks and splits. If the material is moved abnormally during forging, these flaws may be formed without any evidence on the surface of the forging. Bursts, as described in the preceding section, may also occur.

A number of surface flaws also can be produced by the forging operation. These flaws are often caused by the movement of metal over or on another surface without actual welding or fusing of the surfaces. The most common surface flaws in steel forgings are seams, laps, and slivers. Laps and seams are surface discontinuities

that are caused by folding over of metal without fusion. They are usually filled with scale and, on steel components, are enclosed by a layer of decarburized metal. Laps, seams, and other surface defects formed during manufacture of the blanks can also lead to major problems during cold forming of the actual parts. Once any surface defect is exposed to the atmosphere, its surfaces become oxidized, providing an initiation point for failure during future operations.

Other surface flaws include rolled-in scale, ferrite fingers, fins, overfills, and underfills. Surface flaws weaken forgings and can usually be eliminated by correct die design, proper heating, and correct sequencing and positioning of the workpieces in the dies. When such cracks form early in the steel manufacturing process, the surfaces of such breaks can become completely decarburized, forming what is known as a ferrite finger. Even if the original crack about which they formed is then removed by conditioning, the remaining ferrite is extremely low in tensile strength and often splits during the metal movement necessary during cold forming.

Cold shuts often occur in closed-die forgings. They are junctures of two adjoining surfaces caused by incomplete metal fill and incomplete fusion of the surfaces. Shear cracks often occur in steel forgings; they are diagonal cracks occurring on the trimmed edges and are caused by shear stresses. Proper design and condition of trimming dies to remove forging flash are required for the prevention of shear cracks.

Control of Heating

In addition to flow-related imperfections, proper control of heating in hot forging is necessary to prevent excessive scale, decarburization, overheating, or burning. Excessive scale, in addition to causing excessive metal loss, can result in forgings with pitted surfaces. The pitted surfaces are caused by the scale being hammered into the surface and may result in unacceptable forgings. The development of scale during preheating of ingots, slabs, or blooms is almost inevitable, particularly for steels. Sometimes descaling operations involving hydraulic sprays or preliminary light passes in rolling operations are not totally successful; scale may get rolled into the metal surface and become elongated into streaks during subsequent rolling.

Severe overheating causes burning, which is the melting of the lower-melting-point constituents. This melting action severely reduces the mechanical properties of the metal, and the damage is irreparable. Detection and sorting of forgings that have been burned during heating can be extremely difficult. Another defect related to heating practice is blistering, which is a raised spot on the surface caused by expansion of subsurface gas during heating. Blisters may break open and produce a defect that looks similar to a gouge or surface lamination.

Hot tears in forgings are surface cracks that are often ragged in appearance. They result from rupture of the material during forging and are often caused by the presence of low-melting or brittle phases.

Thermal cracks occur as a result of nonuniform temperatures in the forging. Quench cracks are one example of such thermal cracks (see the section “Quench Cracks” in this article). Internal cracks, another type of thermal crack, may occur when forgings are heated too rapidly. These occur as a result of unequal temperatures of the surface relative to the center of the mass, and the resulting differences in the degree of thermal expansion produce tensile stresses near the center. The formation of such cracks depends on both the section size and the thermal conductivity of the material. Large section sizes and poor thermal conductivity promote thermal gradients and favor crack formation.

Flow-Related Factors

Several kinds of forging imperfections are related to metal flow, which is influenced by the workability of the material (discussed later) and the details of component and die design. This section discusses some of the flow-related imperfections that might constitute a manufacturing flaw during inspection or a potential problem during service. It is also important to keep in mind the effects of design on the overall process. Forging load increases very rapidly as the sharpness of details is increased, and die wear increases directly with die pressure. Flash geometry also directly influences the amount of back pressure experienced by the metal in the die cavity.

These defects described in this section are related to the distribution of metal. They can be avoided by proper die design, preform design, and choice of lubrication system. Strictly speaking, these defects are not fundamental to the workability of the material. However, knowledge of these common forging defects is necessary for a practical understanding of workability. These are the defects that commonly limit deformation in the forging process.

Most of the defects occur in hot forging, which is most common for impression-die forging. Therefore, defect formation may also involve entrapment of oxides and lubricant. When this occurs, the metal is incapable of rewelding under the high forging pressures; the term *cold shut* is frequently applied in conjunction with laps, flow-through defects, and so on to describe the flaws generated.

Underfill may not seem like a flow-related defect, but aside from simple insufficient starting mass, the reasons for underfill are flow related. These include improper fill sequence, insufficient forging pressure, insufficient preheat temperature, lubricant build up in die corners, poor or uneven lubrication, and excessive die chill. An improper fill sequence may result in excessive flash loss, or it may be the result of extraordinary pressure requirements to fill a particular section. Sometimes, venting may eliminate the problem; more often than not, a change in the incoming workpiece shape or a change in the deformation sequence is required.

Laps and Folds. Laps are surface irregularities that appear as linear defects and are caused by the folding over of hot metal at the surface. These folds are forged into the surface but are not metallurgically bonded (welded) because of the oxide present between the surfaces (Fig. 20). Thus, a discontinuity with a sharp notch is created.

Fig. 20 Micrograph of a forging lap. Note the included oxide material in the lap. 20×

A lap or fold results from an improper progression in fill sequence. Normally, a lap or fold is associated with flow around a die corner, as in the case of an upper rib or lower rib, or with a reversal in metal-flow direction. A general rule of thumb is to keep metal moving in the same direction. The die corner radius is a critical tool dimension, and it should be as generous as possible. In progressing through a forging sequence, the die corners should become tighter, so that the workpiece fillets are initially large and progressively become smaller as the forging steps are completed. Figure 21 shows schematically a lap forming as metal flows around a die corner.

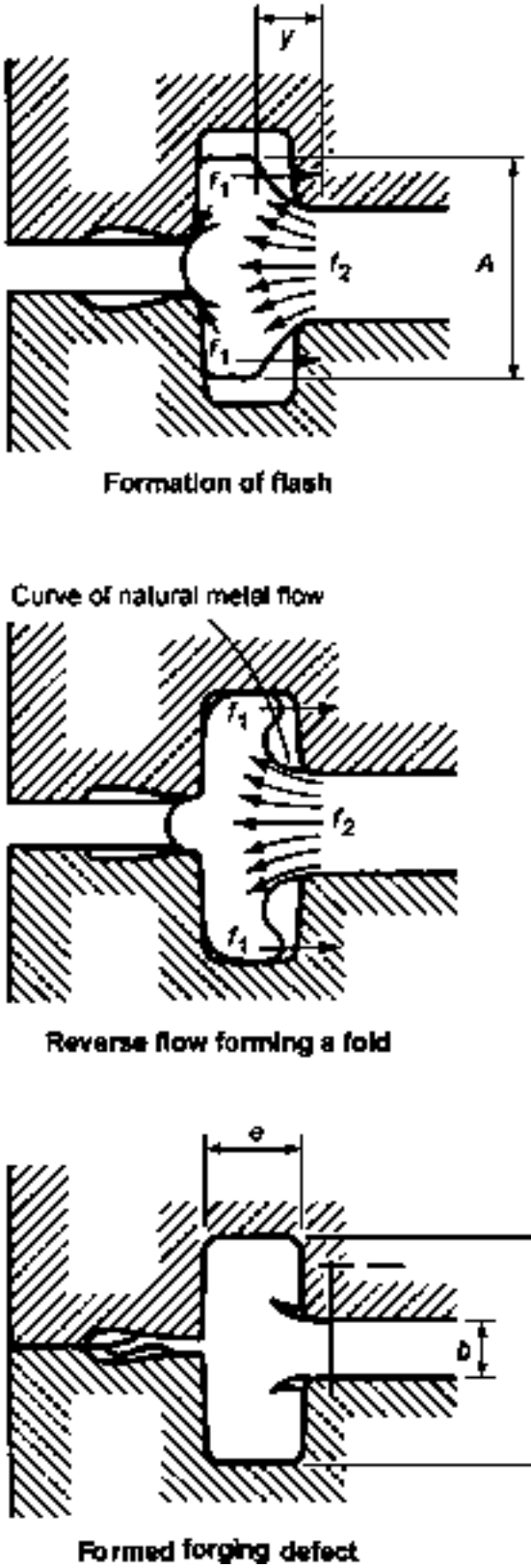


Fig. 21 Lap formation in a rib-web forging caused by improper radius in the preform die

Extrusion-Type Defects. The tail of an extrusion is unusable because of nonuniform flow through the extrusion die. This results in a center-to-surface-velocity gradient, with metal from the workpiece interior moving through the die at a slightly higher velocity than the outer material. The result shows up at the tail of the extrusion as a suck-in or pipe, and, for extrusions, the tail is simply cut off and discarded. Alternatively, a follower block of cheaper material may be added so that most of the defect falls in the cheaper material, and less length of the extruded workpiece is lost.

For forgings that involve forward or backward extrusion to fill a part section, the same situation can develop. Metal flow into a rib or hub can result in a suck-in defect, which, in a worst-case scenario, would show up as a fold on the face opposite to the rib, and a best case would be a depression on what otherwise should be a flat surface. One method of eliminating this type of defect is to position more material on the back face initially. Another method is to change the rib geometry (aspect ratio and/or angles). If neither of these changes can be accomplished, an extra forging step may be needed to limit the amount of extrusion that is done in any one step. Extrusion-type defects are formed when centrally located ribs formed by extrusion-type flow draw too much metal from the main body or web of the forging. A defect similar to a pipe cavity is thus formed (Fig. 22). Methods of minimizing the occurrence of these defects include increasing the thickness of the web or designing the forging with a small rib opposite the larger rib, as shown in Fig. 22.

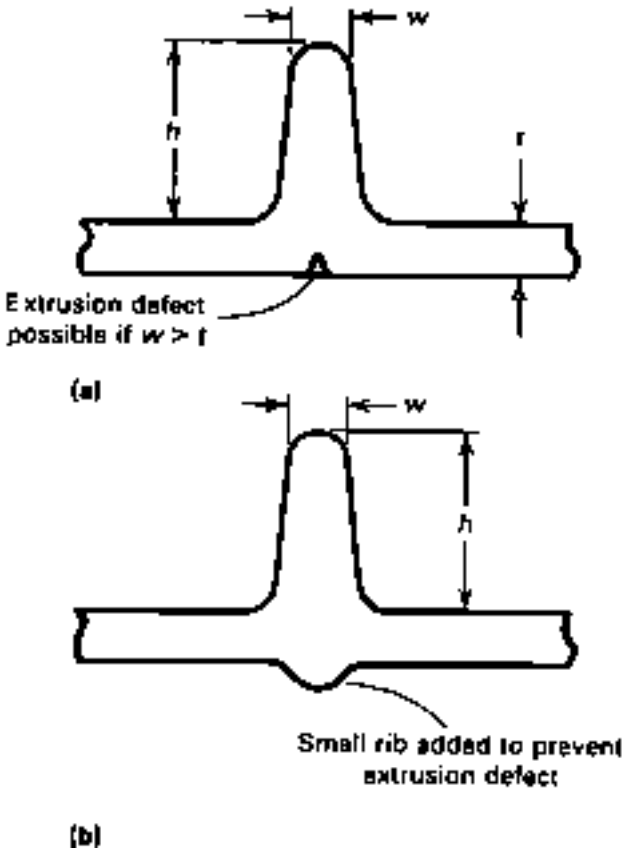


Fig. 22 Extrusion-type defect in (a) centrally located rib and (b) die-design modification used to avoid defect

Shear-Related or Flow-Through Defects. Shearing defects are also known as flow-through defects, because they result from excessive metal flow past a filled detail of the part. Flow-through defects form when metal is forced to flow past a recess after the recess has filled or when material in the recess has ceased to deform because of chilling. An example of this is shown in Fig. 23 for a trapped-die forging that has a rib on the top surface. The rib denoted by “2” is filled early in the forging sequence, and significant mass must flow past the rib in order to fill the inner hub, zone “4.” The result can be a complete shearing-off of the rib in the worst case, with a lesser case being the formation of a shear-type crack.

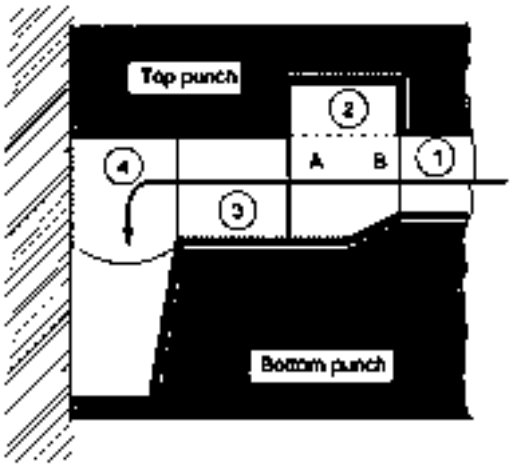


Fig. 23 Schematic of a flow-through crack at the base of a rib in a trapped-die forging. Excessive metal flow past region 2 causes a shear crack to form at A and propagate toward B.

Similar to laps in appearance, flow-through defects can be shallow, but they are indicative of an undesirable grain flow pattern or shear band that extends much deeper into the forging. An example is shown in Fig. 24. Flow-through defects can also occur when trapped lubricant forces metal to flow past an impression.

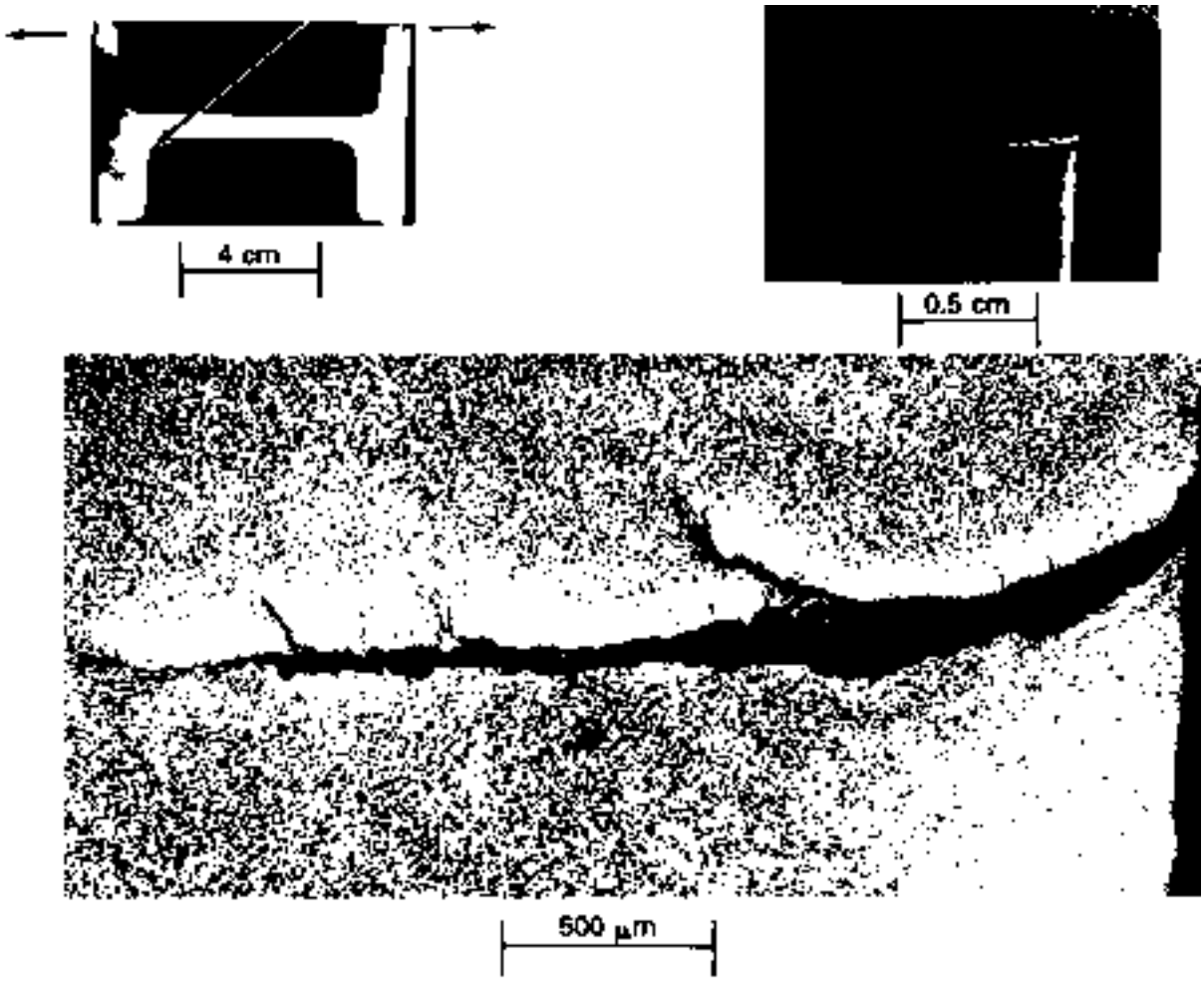


Fig. 24 Flow-through defect in Ti-6Al-4V rib-web structural part

Seams are crevices in the surface of the metal that have been closed, but not welded, by working the metal. Seams result from elongated trapped-gas pockets or from cracks. The surfaces are generally heavily oxidized and decarburized in steels. Depth varies widely, and surface areas sometimes may be welded together in spots. Seams may be continuous or intermittent, as indicated in Fig. 2(h). A micrograph of a typical seam is shown in Fig. 25. Oxides in a seam can be the result of exposure to elevated temperatures during processing.

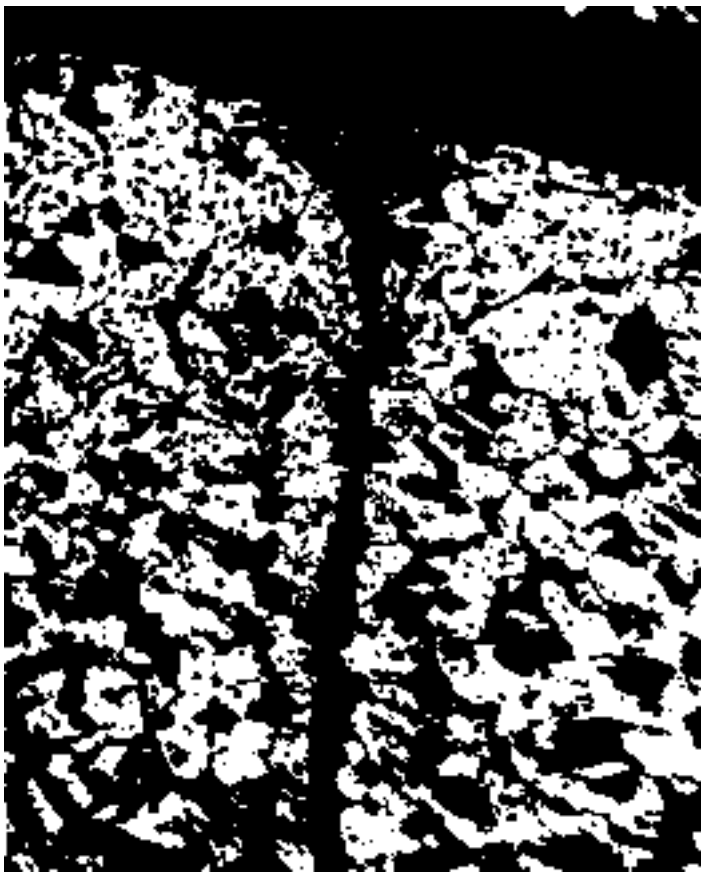


Fig. 25 Micrograph of a seam in a cross section of a 19 mm (0.75 in.) diameter medium-carbon steel bar showing oxide and decarburization in the seam. 350×

Seams seldom penetrate to the core of bar stock. The incomplete removal of seams from forging stock may cause additional cracking in hot forging and quench cracking during heat treatment. Seams are sometimes difficult to detect in an unused fastener, but they are readily apparent after a fastener has been subjected to installation and service stresses.

Seams have a large number of possible origins, some mechanical and some metallurgical. Seams can develop from cracks caused by working, or seams can develop from an imperfection in the ingot surface, such as a hole, that becomes oxidized and is prevented from healing during working. In this case, the hole simply stretches out during forging or rolling, producing a linear seam in the workpiece surface. Seams also result from trapped-gas pockets, cracks, a heavy cluster of nonmetallic inclusions, or a deep lap. Seams are normally closed tight enough that no actual opening can be visually detected without some nondestructive inspection techniques, such as magnetic-particle inspection (see *Nondestructive Evaluation and Quality Control,* Volume 17 of *ASM Handbook*)

Seams can be difficult to detect, because they may appear as scratches on the forging, or because a machining process may obliterate them. Seams may not become evident until the part has been subjected to installation and service stresses, or when the constraint exerted by the bulk of material is removed from the neighborhood of a seam. Figure 26 is an example of a seam detected by routine magnetic-particle inspection of a hot-rolled 4130 steel bar. No stringer-type inclusions were observed in the region of the flaw, but it did contain a substantial amount of oxide (Fig. 26b).

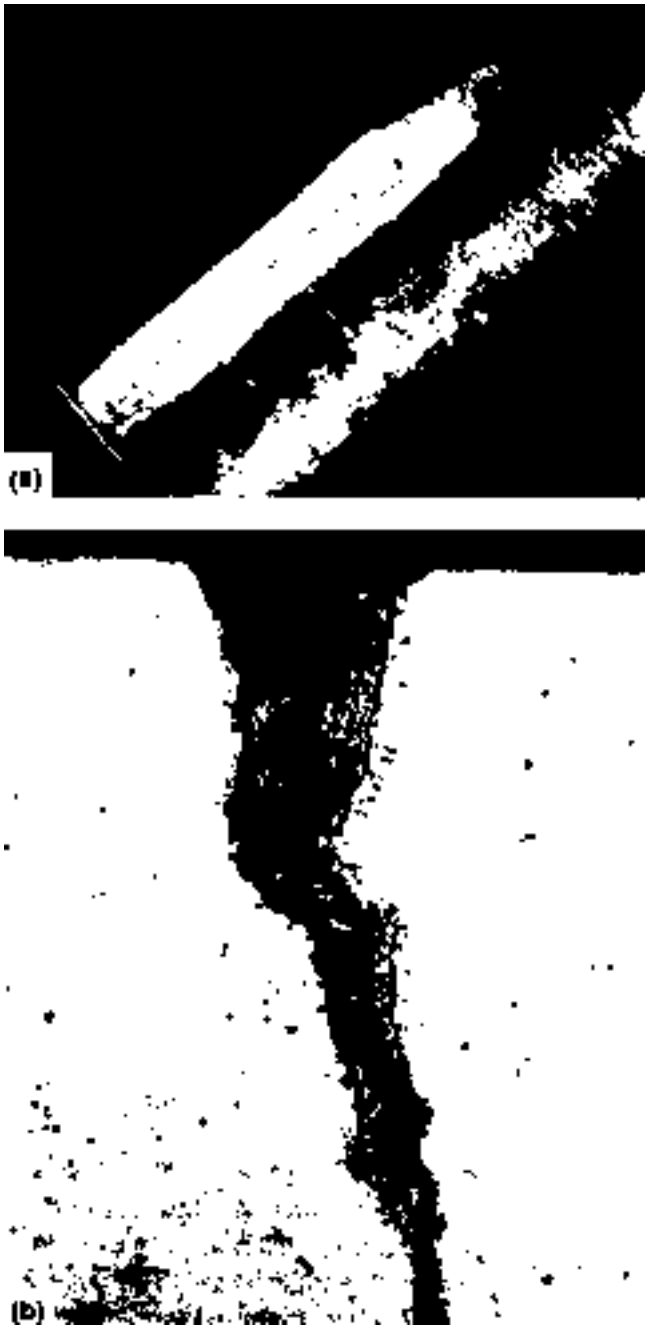


Fig. 26 Seam in rolled 4130 steel bar (a) Closeup of seam. Note the linear characteristics of this flaw. (b) Micrograph showing cross section of the bar. Seam is normal to the surface and filled with oxide. 30×

Central or Internal Bursts (Chevron Cracking). Chevrons are internal flaws named for their shape (Fig. 2k). A central burst, or chevron crack, is most commonly associated with extrusion and drawing operations, although it can be generated by forging and rolling processes as well. Central bursts are internal fractures caused by high hydrostatic tension. The severe stresses that build up internally cause transverse subsurface cracks. Some of the factors that can contribute to the formation of chevrons are incorrect die angles; either too great or too small a reduction of cross-sectional area; incomplete annealing of slug material; excessive work hardenability of the slug material; the presence of an excessive amount of seams and other imperfections in the slug materials; segregation in a steel slug that results in hard martensitic particles in the center of the slug, which act as barriers to material flow; and insufficient die lubrication.

Internal bursts in rolled and forged metals result from the use of equipment that has insufficient capacity to work the metal throughout its cross section. If the working force is not sufficient, the outer layers of the metal are deformed more than the inside metal, sometimes causing wholly internal, intergranular fissures that can act as initiation sites for further crack propagation during service loads. In forward cold extrusion, the occurrence

of central bursts or chevrons is nearly always restricted to isolated lots of material and usually to only a small percentage of the pieces extruded in any particular production run.

A change in deformation-zone geometry is usually sufficient to eliminate the problem. The conservative design approach is to ensure that no hydrostatic tension develops. Often, however, the part or tooling design cannot be changed sufficiently to eliminate hydrostatic tension. If the level of hydrostatic tension can be kept below a critical level, bursting can likely be avoided. This may be accomplished by a change in lubricant, die profile, temperature, deformation level, or process rate.

Internal bursts may also occur where the material is weak. For example, with ingot imperfections (such as pipe, porosity, segregation, or inclusions), tensile stresses can be sufficiently high to tear the material apart internally, particularly if the forging temperature is too high (Ref 12). Similarly, if the metal contains low-melting phases resulting from segregation, these phases may rupture during forging. Ingot pipe, unhealed center conditions, or voids associated with melt-related discontinuities may also induce center bursts if reduction rates are too severe or temperatures are incorrect during working. The conversion practice to bar or billet must impart sufficient homogenization or healing to produce a product with sound center conditions. An example of an unsound condition that did not heal is shown in Fig. 27.

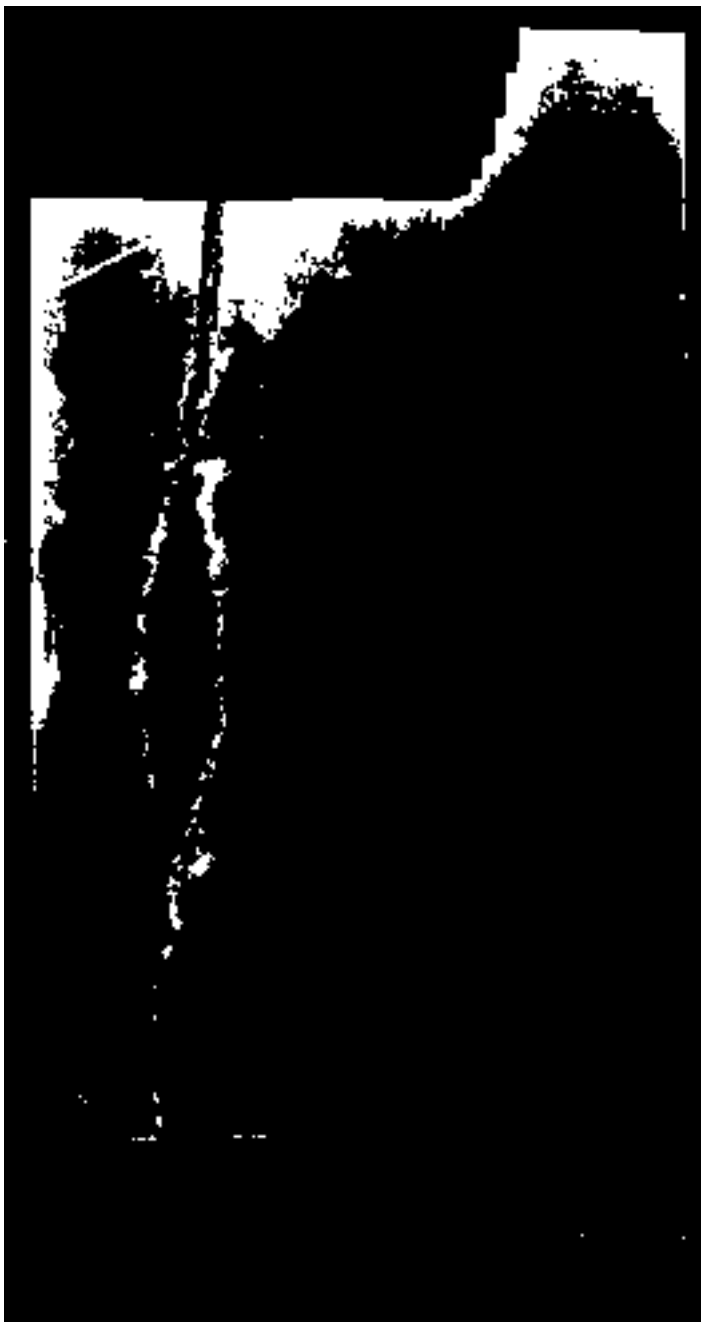


Fig. 27 Section through a heat-resistant alloy forging showing a central discontinuity that resulted from insufficient homogenization during conversion. Step machining was used to reveal the location of the rupture; original diameter is at right.

Macroetching and ultrasonic inspection methods are the most widely used for identifying regions of unsoundness. Bursts usually display a distinct pattern of cracks and do not show spongy areas, thus distinguishing them from pipes. Bursts are readily detected by macroetching. Figure 28 shows a large burst that occurred during the forging of an electroslag remelted ingot. The cause was traced to a weak solidification plane near the bottom of the ingot combined with higher forging temperatures than normal.

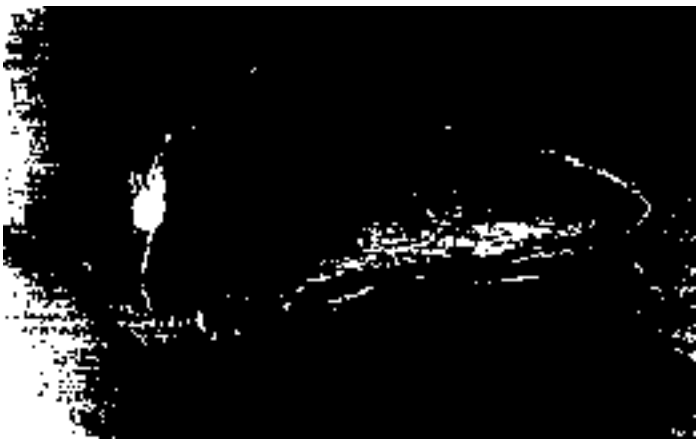


Fig. 28 Cross section of a forged bar showing a forging burst. The burst is located approximately at the centerline of the workpiece. Arrow indicates the direction of working.

Grain Flow and Anisotropy

Grain flow refers to the directional pattern of crystals and second-phase particles as they become aligned in the direction of greatest metal flow during the breakdown of cast ingots and the subsequent working of wrought forms. The grain flow consists of a combination of:

- The crystallographic reorientation of the grains during severe deformation
- The alignment of nonmetallic inclusions, voids, and chemical segregation

Together, these can produce the familiar fibering pattern of wrought products. Both types of preferred orientations can cause anisotropy in the mechanical and physical properties of a metal. The effects of anisotropy on mechanical properties vary among different metals and alloys. For example, a vacuum-melted steel of a given composition is generally less mechanically anisotropic than a conventionally killed, air-melted steel of the same composition. The occurrence and severity of fibering varies with such factors as composition, extent of chemical segregation, and the amount of work or reduction the workpiece receives. Furthermore, this condition is a consequence of processing and may not be adequately considered during the design of a particular product.

Anisotropy is most easily identified by differences in ductility, toughness, and fatigue resistance for longitudinal, transverse, and short-transverse directions. Grain flow also can be used effectively to produce a better product. If the grain flow can be aligned in the direction of maximum tension in the part during service, superior performance is achieved. The resistance to crack initiation and propagation also is highest in the direction normal to the grain flow, because the grain boundaries and elongated inclusions act to blunt advancing cracks. If the tension is normal to the fiber direction, the fiber enhances crack propagation.

Figure 29 shows grain-flow patterns for a channel section forged with different parting-line locations. Because the grain-flow pattern breaks out to the part surface at the parting line, any stress during service that is normal to the parting line acts perpendicular to the weakest planes. These planes are the easiest to fracture along, for example, the short-transverse equivalent position. Because a channel is usually stressed in a direction normal to

it (flexing the channel legs to open or close), the grain-flow pattern in Fig. 29(b) is least detrimental. The parting line of forging dies should be located so as to minimize disruption to the grain-flow lines.

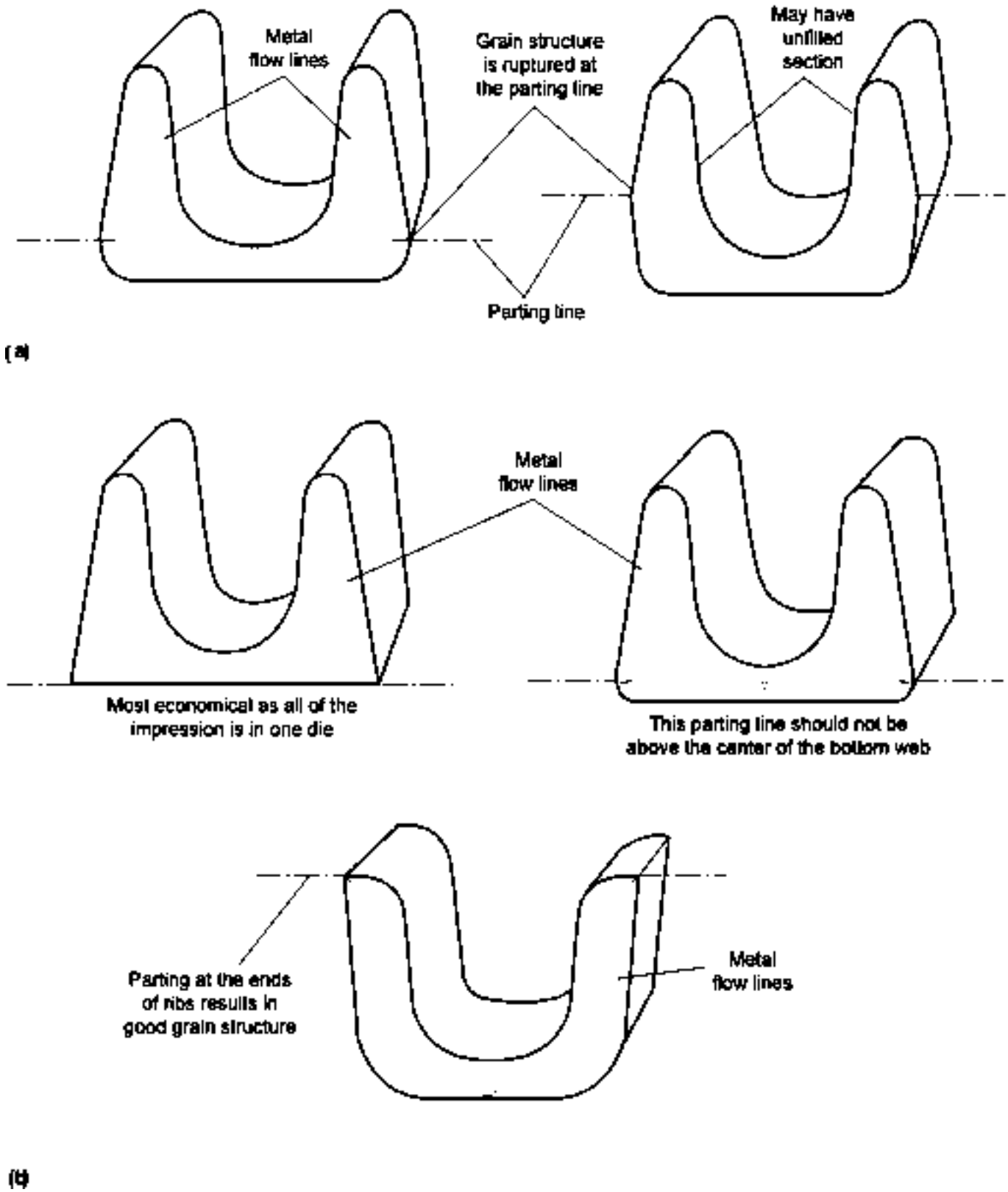


Fig. 29 Parting-line location and its influence on grain-flow pattern in a channel section forging. (a) Parting lines resulting in metal-flow patterns that cause forging defects. (b) Parting lines resulting in smooth flow lines at stressed sections

Although grain flow can be used to advantage, anisotropy of fracturing in wrought products is extremely widespread, occurring in both ductile and brittle materials and including some that are essentially pure metals. Probably the most dramatic evidence for the effects of mechanical fibering are those observed during fracture,

either in service or during laboratory testing or process deformation. When the stress direction is perpendicular to the lay of the stringer, the fracture direction that is parallel to the lay of the stringer results in a woody appearance to the fracture (e.g., see Fig. 21 in the "Atlas of Fractographs" in *Fractography*, Volume 12 of *ASM Handbook*.)

Thermal Effects and Heat Treatment

Improper heat treatment can result in failure to attain the desired microstructure in the metal and therefore the desired levels of mechanical or physical properties. Such deficiencies can cause failures in service.

Overheating. As alloys are heated above their recrystallization temperatures, grain growth occurs. As the temperature increases, so does grain growth, becoming quite rapid and resulting in large grains, often accompanied by many undesirable characteristics. The impairment that usually accompanies large grains is caused not only by the size of the grains but also by the more continuous films that are formed on larger grains by grain-boundary impurities, such as preferential precipitates and evolved gases. Finer grains, on the other hand, present a greater amount of total grain-boundary area over which the impurities may be distributed.

The detrimental effects of overheating depend on the temperature and the time of exposure as well as on the chemical composition of the alloy. For example, a short exposure of a high-speed tool steel to temperatures at approximately 1250 °C (2280 °F) is required to dissolve the carbides, but prolonged heating at that temperature causes grain growth and loss of mechanical properties. The damage caused by overheating is particularly significant in the high-carbon and medium-carbon steels, in which both strength and ductility are affected.

One of the most conspicuous indications that a metal has been overheated is the coarse-grain fracture surface that results. Usually, examination with a stereoscopic microscope shows the characteristic faceted surface of an intergranular fracture. Microscopically, the large grain size resulting from overheating is quite evident and can be measured and compared to a similar metal with a normal grain size. In addition to large grain size, fine oxide particles are often found dispersed throughout the grains, particularly near the surface. A Widmanstätten structure is often associated with coarse grains in an overheated steel forging with a controlled rate of cooling (neither extremely fast nor extremely slow).

Burning is a term applied when a metal is grossly overheated and permanent irreversible damage to the structure occurs as a result of intergranular penetration of oxidizing gas or incipient melting. The micrograph in Fig. 30(b) shows burning at the grain boundaries in a specimen of copper C11000 (electrolytic tough-pitch copper) heated to 1065 °C (1950 °F). Copper oxide migrated to grain boundaries, forming a continuous network that severely reduced strength and ductility.

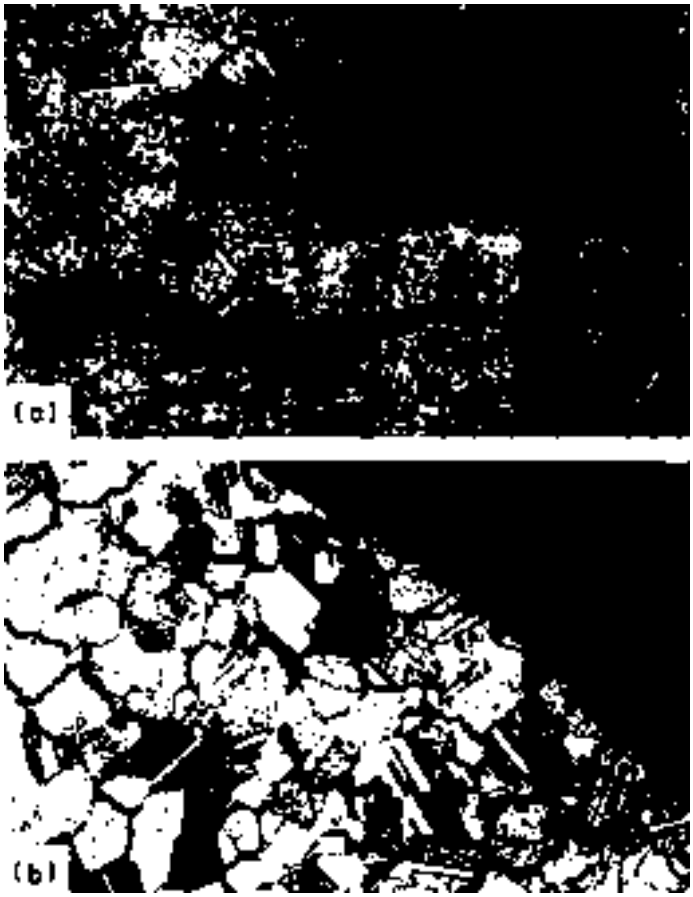


Fig. 30 Micrographs showing the effects of overheating and burning on microstructures of copper forgings. (a) Overheated copper C10200 forging showing oxides (black particles). The forging was heated to 1025 °C (1875 °F). (b) Burning (black outlines) at grain boundaries of a copper C11000 forging heated to 1065 °C (1950 °F)

In steels, burning may manifest itself with the formation of extremely large grains and incipient melting at the grain boundaries. Melting is particularly evident where segregation has occurred, with the marked formation of low-melting phases in the interdendritic regions and at the grain boundaries. Burning cannot be readily detected by visual examination. However, a metallographic examination of the structure shows the enlarged grains and the pronounced grain-boundary network.

The typical causes of burning include excessive furnace temperatures; welding and flame-cutting operations that are not properly controlled; and operations involving the direct heating of a workpiece surface, such as flame hardening, induction hardening, shrink fitting, and so on, that are inadequately controlled. Occasionally, burning occurs in adequately controlled furnaces simply because the flame is allowed to impinge on the metal surface, causing localized overheating. Another source of overheating, though less common, can be the conversion of mechanical energy into heat. If there are segregated areas near the center of the billet, and if the initial forging temperature is close to the melting point of the segregated regions, the additional heat supplied by the transformation of work into heat can cause localized burning during the forging operation. Burned material cannot be salvaged and should be scrapped, because the metallurgical changes that have occurred are irreversible.

Quench cracks are often the cause of failure in steel forgings. These cracks may be obvious and may prompt the removal of a component during production, or they may be obscured during manufacture and may be present in the shipped component. Quench cracks in steel result from stresses produced by the volume increase accompanying the austenite-to-martensite transformation. When a steel forging is quenched, untempered (hard and brittle) martensite is formed at the outer surfaces first.

As cooling continues, the austenite-transformation reaction progresses toward the center of the workpiece, accompanied by a volumetric expansion that can place the center portions of the part in tension. When the workpiece cannot adjust to the strains produced by this process, cracking generally occurs. Such cracking may take place during or after the quench, because the transformation product is hard and brittle before tempering.

This susceptibility to cracking is increased by the presence of stress raisers, such as sharp fillets, tool marks or other notches, massive inclusions, or voids. Quench cracks in forgings may also occur near the trim line of a closed-die forging, where some localized burning may have occurred. Other factors that affect quench cracking are the hardenability of the steel, the rate of cooling, and elapsed time between quenching and tempering. In general, an increase in any of these factors increases the likelihood of quench cracking.

Quench cracks have several recognizable features. Macroscopically, they are usually straight and extend from the surface (often initiating from a fillet or tool mark) toward the center of the component. The margins of the crack may have scale present if the part was tempered after quenching. The cracks are open to the surface and may be detected by magnetic-particle, ultrasonic, or eddy-current inspection. Microscopically, quench cracks are invariably intergranular and may be free of decarburization.

Carburization and Decarburization. Although surface treatment by carburization is commonly used to improve the wear resistance and overall strength characteristics of many forged steel components, particularly axles, the accidental alteration of the surface structure during heat treatment can produce disastrous results. Heating a metal surface that is contaminated with oil or carbon or heating in a carbon-rich atmosphere can produce a surface with a high carbon content. The mechanical properties of a carburized surface layer are different from those of the core, and cracking problems can arise, because a carburized surface was not considered in the original design. The opposite condition (decarburization) can occur by heating in an oxidizing atmosphere such as air. The result in this case is a surface layer with a lower carbon content than that of the core.

Reference cited in this section

12. V.N. Whittacker, *Nondestr. Test.*, Oct 1971, p 320

Workability

The term *workability* refers to the ability to deform and shape metals by bulk deformation without fracture. This ability to deform without fracture is highly variable, and the limits of plastic deformation for a given material are greatly affected by the prevailing state of stress, strain rate, and temperature in the metalworking process. Deformation is also variable among different alloys and conditions, and thus, the design or specification of metalworking operations is a basic factor during both the conceptual and detailed stages of design and materials selection.

Most metals and alloys can be worked over a wide range of temperatures, and hot working requires less force for plastic deformation. However, dimensional control of the workpiece often can be difficult in hot working operations, because the metal contracts nonuniformly during cooling. This particular problem is not encountered in the cold working temperature range; thus, cold working has advantages, even though workability is less than that of hot working. The term *cold forming* can have broad meaning, depending largely on the product forms to which it is applied. Cold forming is frequently defined as forming that occurs below the recrystallization temperature of the specific metal being deformed, and it usually begins at or near room temperature. Depending mainly on the severity of the forming, the temperature of the work metal may increase as much as several hundred degrees Fahrenheit during the deformation process.

Bulk working is commonly in the hot working regime, where metallic materials are more workable (Fig. 31) (Ref 13). In hot working (typically carried out at temperatures in excess of ~60% of the absolute melting point of the metal), individual grains are deformed; the deformed grains then immediately begin to recrystallize, that is, nucleate new stress-free grains that are roughly equiaxed in morphology. The hotter the metal, the more plastic it behaves, and the more easily it deforms. At excessively high temperatures, however, grain growth, incipient melting, phase transformation, and changes in composition can occur, which may cause degradation of the properties in the forging. At lower hot working temperatures, the metal is more difficult to work, yet the resultant grain size may be finer, and the product may have better mechanical properties.

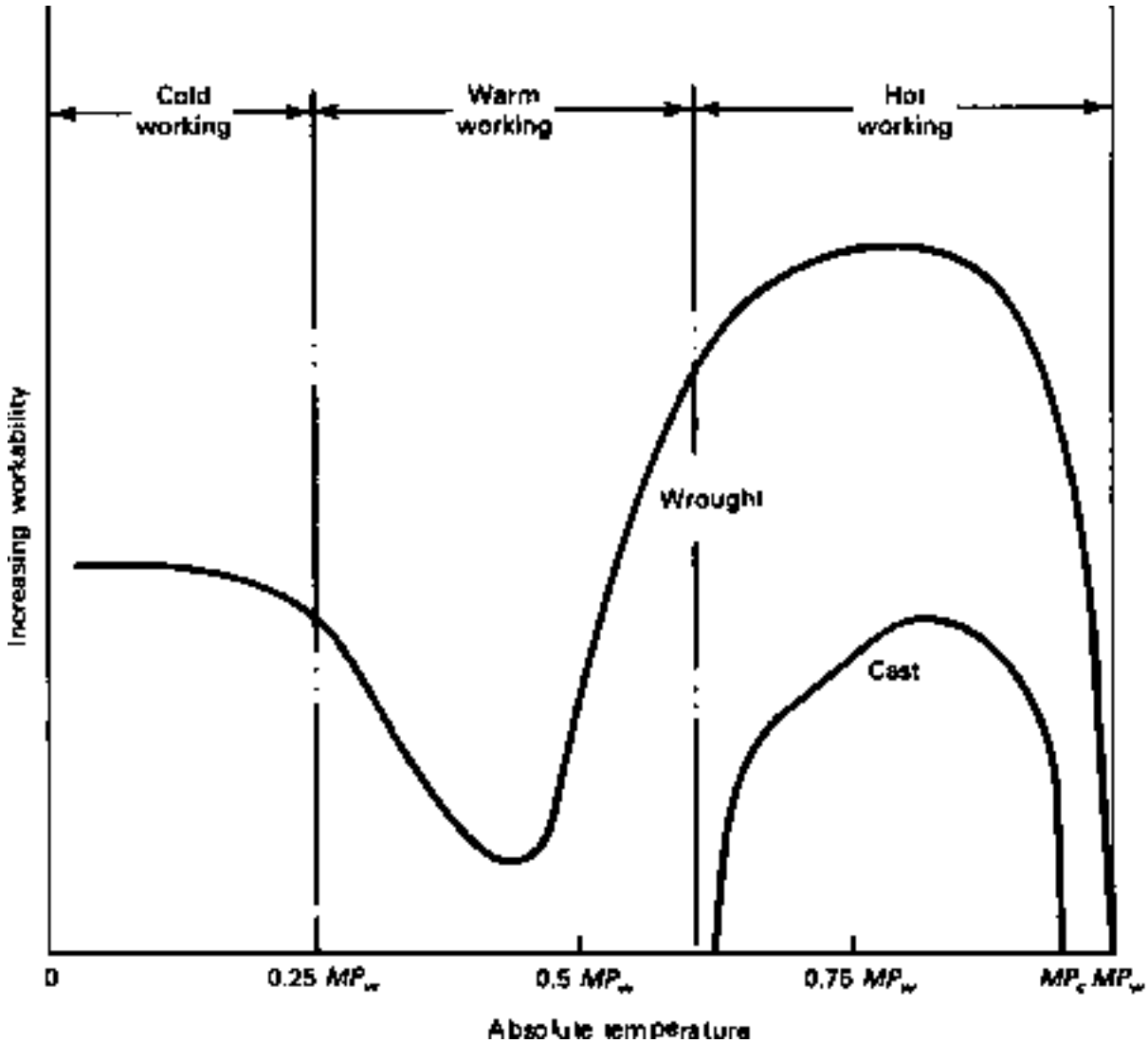


Fig. 31 Relative workability of coarse-grained cast materials and wrought-and-recrystallized metals as a function of temperature. The melting point (or solidus) is denoted as MP_c(cast) or MP_w(wrought).

Warm working is conducted in the range between 25 and 60% of the absolute melting temperature. Paradoxically, the workability of metals in the warm forging regime is considerably lower than that of cold forging (Fig. 31), due to reduced metal ductility in the warm working regime. For this reason, warm working is generally reserved for bulk working operations that require small deformations and for applications in which a decrease in the workpiece temperature offers an energy savings in terms of reduced heating costs. If working temperature is reduced still further below the recrystallization temperature (i.e., in the cold working regime), the deformed grains do not break up and form new grains but remain deformed and highly stressed. When the metal is in this condition, it can behave in a brittle manner, and cracking may occur.

Workability is a measure of the ability of a metal to endure deformation without cracking for bulk working processes such as forging, rolling, extrusion, and bending of thick sections. Typically, fracture in bulk deformation processing occurs as ductile fracture, rarely as brittle fracture. However, depending on temperature and strain rate, the details of the ductile fracture mechanism vary. At temperatures below approximately one-half the melting point of a given material (below the hot working region), ductile fracture usually occurs by dimple fracture. Cavitation or ductile rupture is the more dominant mode of ductile fracture when working temperatures are higher than one-half the melting point of a given material.

Ductility as measured by the tensile test is a good way to compare the inherent resistance of different materials or different microstructures for the same material. However, ductility alone is insufficient when judging fracture susceptibility of a deformation process, because fracture depends on the conditions of both the local strain and stress. Compressive stresses superimposed on tensile or shear stresses by the deformation process can have a significant influence on closing small cavities or limiting their growth and thus enhancing workability. Therefore, workability is not a property similar to tensile reduction of area at fracture. Workability depends not

only on material characteristics but also on process variables, such as strain, strain rate, temperature, and stress state.

Flow Localization. Workability problems can arise when metal deformation is localized to a narrow zone. This results in a region of different structures and properties that can be the site of failure in service. Localization of deformation can also be so severe that it leads to failure in the deformation process. In either mode, the presence of flow localization needs to be recognized.

Flow localization is commonly caused by the formation of a dead-metal zone between the workpiece and the tooling. This can arise from poor lubrication at the workpiece-tool interface. When the workpiece is constrained from sliding at the interface, it barrels, and the friction-hill pressure distribution is created over the interface. The inhomogeneity of deformation throughout the cross section leads to a dead zone at the tool interface and a region of intense shear deformation. A similar situation can arise when the processing tools are cooler than the workpiece; in this case, heat is extracted at the tools. Consequently, the flow stress of the metal near the interface is higher because of the lower temperature.

However, flow localization may occur during hot working in the absence of frictional or chilling effects. In this case, localization results from flow softening (negative strain hardening). Flow softening arises during hot working as a result of structural instabilities, such as adiabatic heating, generation of a softer texture during deformation, grain coarsening, or spheroidization. Flow softening has been correlated with materials properties (Ref 14) by the parameter:

$$a = -\frac{(g-1)}{m} \qquad \text{(Eq 1)}$$

for upset compression and:

$$a = \frac{-g}{m} \qquad \text{(Eq 2)}$$

for plane-strain compression, where $\gamma = (1/\sigma)d\sigma/d\varepsilon$ is the normalized flow-softening rate, and where the strain-rate sensitivity, m, is the differential relation of stress and strain rate:

$$m = d \log \sigma / d \log \varepsilon$$

In alpha-beta titanium alloys and other materials that exhibit a strong tendency toward flow localization, this phenomenon is likely to occur when the parameter α is greater than 5. Figure 32 (Ref 15) shows a crack that initiated in a shear band during the high-energy-rate forging of a complex austenitic stainless steel.

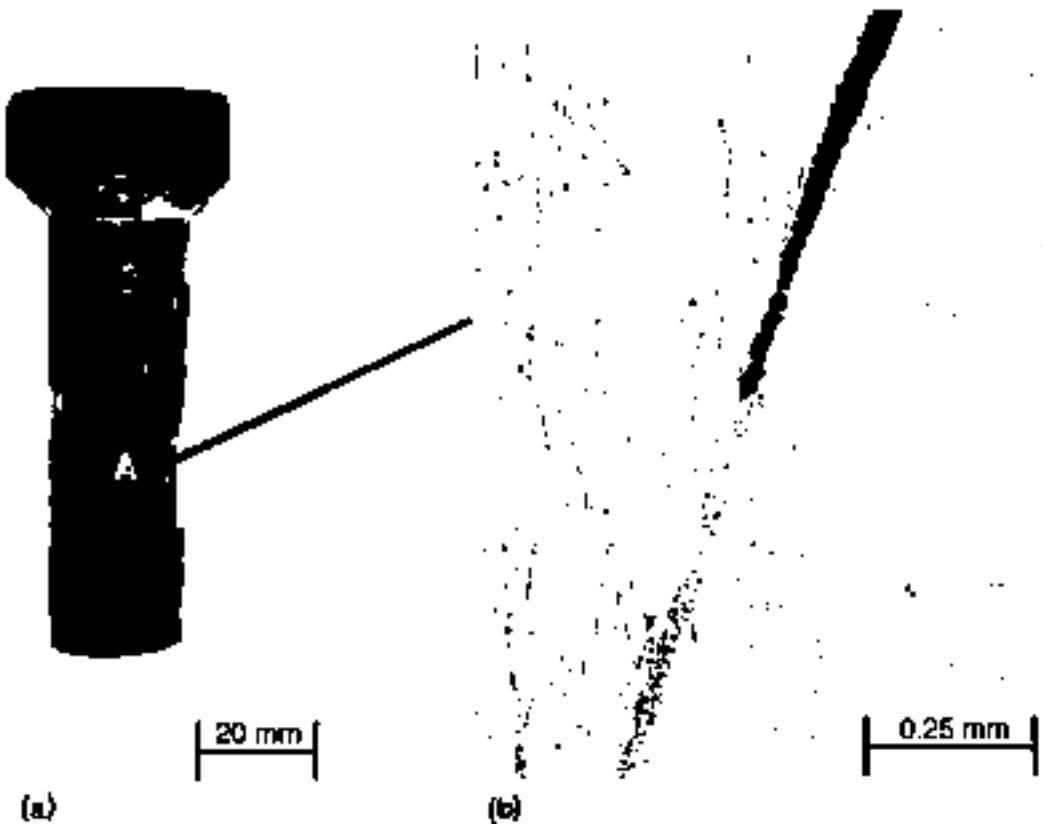


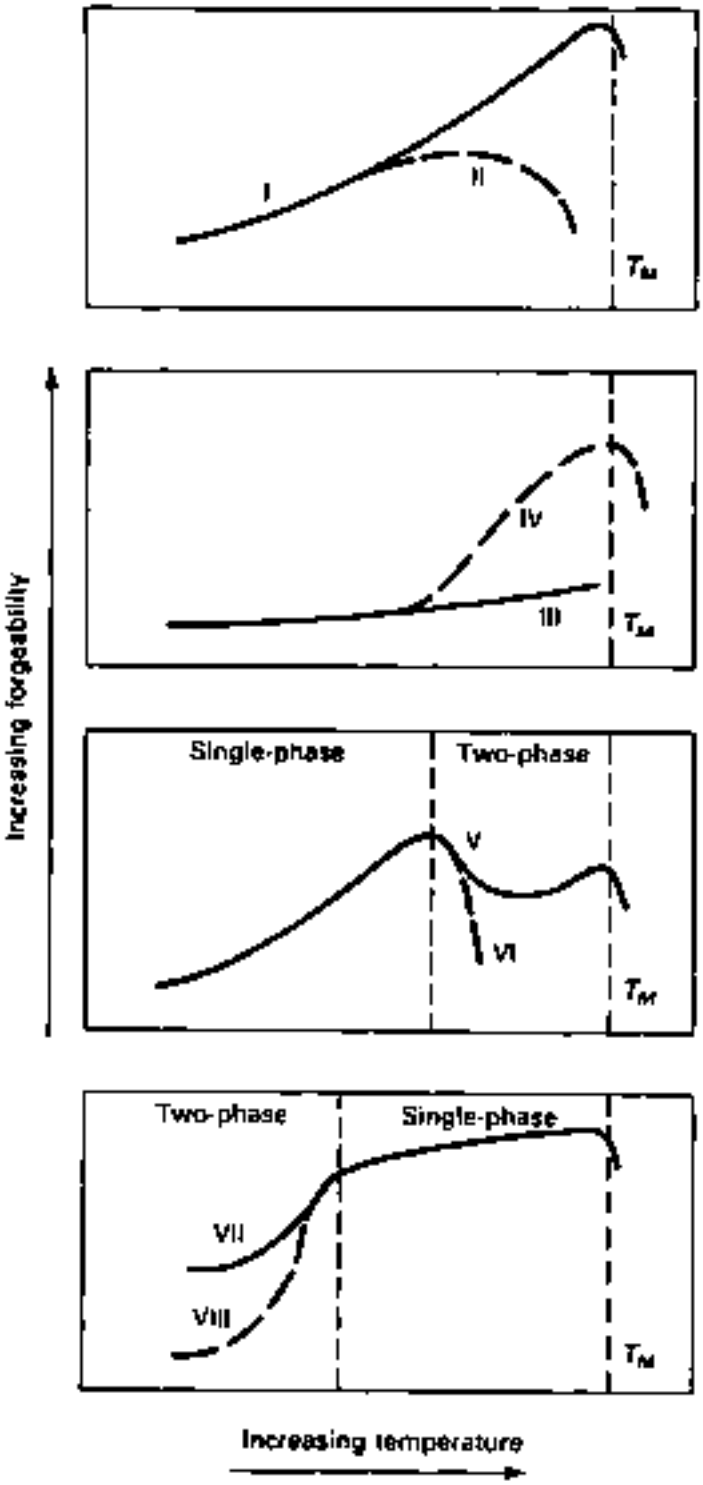
Fig. 32 Austenitic stainless steel high-energy-rate forged extrusion. Forging temperature: 815 °C (1500 °F); 65% reduction in area; $\varepsilon = 1.4 \times 10^3\ s^{-1}$. (a) View of extrusion showing spiral cracks. (b) Optical micrograph showing the microstructure at the tip of one of the cracks in the extrusion (area A). Note that the crack initiated in a macroscopic shear band that formed first at the lead end of the extrusion. Etchant: oxalic acid

Metallurgical Considerations. Workability problems depend greatly on grain size and grain structure. When the grain size is large relative to the overall size of the workpiece, as in conventionally cast ingot structures, workability is lower, because cracks may initiate and propagate easily along the grain boundaries. Moreover, with cast structures, impurities are frequently segregated to the center and top or to the surface of the ingot, creating regions of low workability. Because chemical elements are not distributed uniformly on either a micro- or a macroscopic scale, the temperature range over which an ingot structure can be worked is rather limited.

Typically, cast structures must be hot worked. The melting point of an alloy in the as-cast condition is usually lower than that of the same alloy in the fine-grain, recrystallized condition because of chemical inhomogeneities and the presence of low-melting-point compounds that frequently occur at grain boundaries. Deformation at temperatures too close to the melting point of these compounds may lead to grain-boundary cracking when the heat developed by deformation increases the workpiece temperature and produces local melting.

This fracture mode is called hot shortness. This type of fracture can be prevented by using a sufficiently low deformation rate that allows the heat developed by deformation to be dissipated by the tooling, by using lower working temperatures, or by subjecting the workpiece to a homogenization heat treatment prior to hot working. The relationship between the workability of cast and wrought structures and temperatures is shown in Fig. 31. The intermediate temperature region of low ductility shown in Fig. 31 is found in many metallurgical systems (Ref 16). This occurs at a temperature that is sufficiently high for grain-boundary sliding to initiate grain-boundary cracking but not so high that the cracks are sealed off from propagation by a dynamic recrystallization process.

The relationship between workability and temperature for various metallurgical systems is summarized in Fig. 33 (Ref 17). Generally, pure metals and single-phase alloys exhibit the best workability, except when grain growth occurs at high temperatures. Alloys that contain low-melting-point phases (such as γ′-strengthened nickel-base superalloys) tend to be difficult to deform and have a limited range of working temperature. In general, as the solute content of the alloy increases, the possibility of forming low-melting-point phases increases, while the temperature for precipitation of second phases increases. The net result is a decreased region for good forgeability (Fig. 34).



I. **Pure metals and single phase alloys**
Aluminum alloys
Tantalum alloys
Niobium alloys

II. **Pure metals and single-phase alloys exhibiting rapid grain growth**
Beryllium
Magnesium alloys
Tungsten alloys
All-beta titanium alloys

III. **Alloys containing elements that form insoluble compounds**
Resulfurized steel
Stainless steel containing selenium

IV. **Alloys containing elements that form soluble compounds**
Molybdenum alloys containing oxides
Stainless steel containing soluble carbides or nitrides

V. **Alloys forming ductile second phase on heating**
High-chromium stainless steels

VI. **Alloys forming low-melting second phase on heating**
Iron containing sulfur
Magnesium alloys containing zinc

VII. **Alloys forming ductile second phase on cooling**
Carbon and low-alloy steels
Alpha-beta and alpha-titanium alloys

VIII. **Alloys forming brittle second phase on cooling**
Superalloys
Precipitation-hardenable stainless steels

Fig. 33 Typical workability behavior exhibited by different alloy systems. T_M, absolute melting temperature. Source: Ref 17

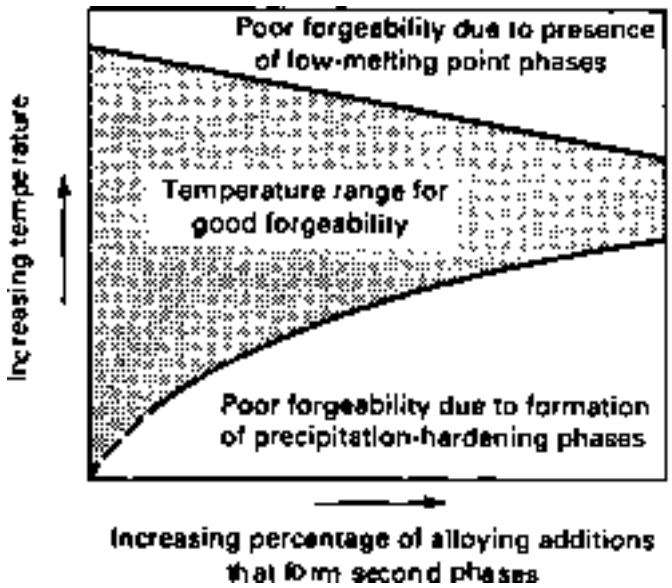


Fig. 34 Influence of solute content on melting and solution temperatures and therefore on forgeability

References cited in this section

13. S.L. Semiatin and J.J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization*, American Society for Metals, 1984, p 3

14. S.L. Semiatin and G.D. Lahoti, The Occurrence of Shear Bands in Isothermal Hot Forging, *Metall. Trans. A*, Vol 13, 1982, p 275–288

15. M.C. Mataya and G. Krauss, A Test to Evaluate Flow Localization during Forging, *J. Appl. Metalwork.*, Vol 2, 1981, p 28–37

16. F.N. Rhines and P.J. Wray, Investigation of the Intermediate Temperature Ductility Minimum in Metals, *Trans. ASM*, Vol 54, 1961, p 117

17. A.M. Sabroff, F.W. Boulger, and H.J. Henning, *Forging Materials and Practices*, Reinhold, 1968

Cracking in Bulk Working

The common failure modes occurring in deformation processing are summarized in Table 2. Three temperature regions are common: cold working, warm working, and hot working. The general types of fracture in bulk working are:

- Free-surface fracture
- Die-contact surface fracture
- Internal fracture

Table 2 Common failure modes in deformation processing

Temperature regime	Metallurgical defects in:	
	Cast grain structure	Wrought (recrystallized) grain structure
Cold working	(a)	Free surface fracture, dead metal zones (shear bands, shear cracks), centerbursts, galling
Warm working	(b)	Triple-point cracks/fractures, grain-boundary cavitation/fracture

Temperature regime	Metallurgical defects in:	
	Cast grain structure	**Wrought (recrystallized) grain structure**
Hot working	Hot shortness, centerbursts, triple-point cracks/fractures, grain-boundary cavitation/fracture, shear bands/fractures	Shear bands/fractures, triple-point cracks/fractures, grain-boundary cavitation/fracture, hot shortness

(a) Cold working of cast structures is typically performed only for very ductile metals (such as dental alloys) and usually involves many stages of working with intermediate recrystallization anneals.
(b) Warm working of cast structures is rare.
The most common types of failures in bulk working are free-surface fracture (at cold working temperatures) or triple-point cracking/cavitation (at hot working temperatures), but internal cracking (i.e., central bursts or chevron cracking, previously discussed) can be an extremely dangerous type of defect, because it cannot be seen visually. This section reviews the first two.
A vast array of mechanical tests and theoretical analyses have been developed for predicting failures of these and other types during forging, extrusion, rolling, and other bulk forming operations. These tests and analyses are summarized in the Section "Evaluation of Workability" of *Forming and Forging*, Volume 14 of *ASM Handbook.* Other common test techniques used to gage bulk workability include the uniaxial upset, flanged or tapered compression, notched-bar upset, and wedge tests.
The types of cracks that develop in cold forging by upsetting-type processes are discussed in Ref 18 and are shown in Fig. 35. Table 3 provides a detailed description of each type of crack.

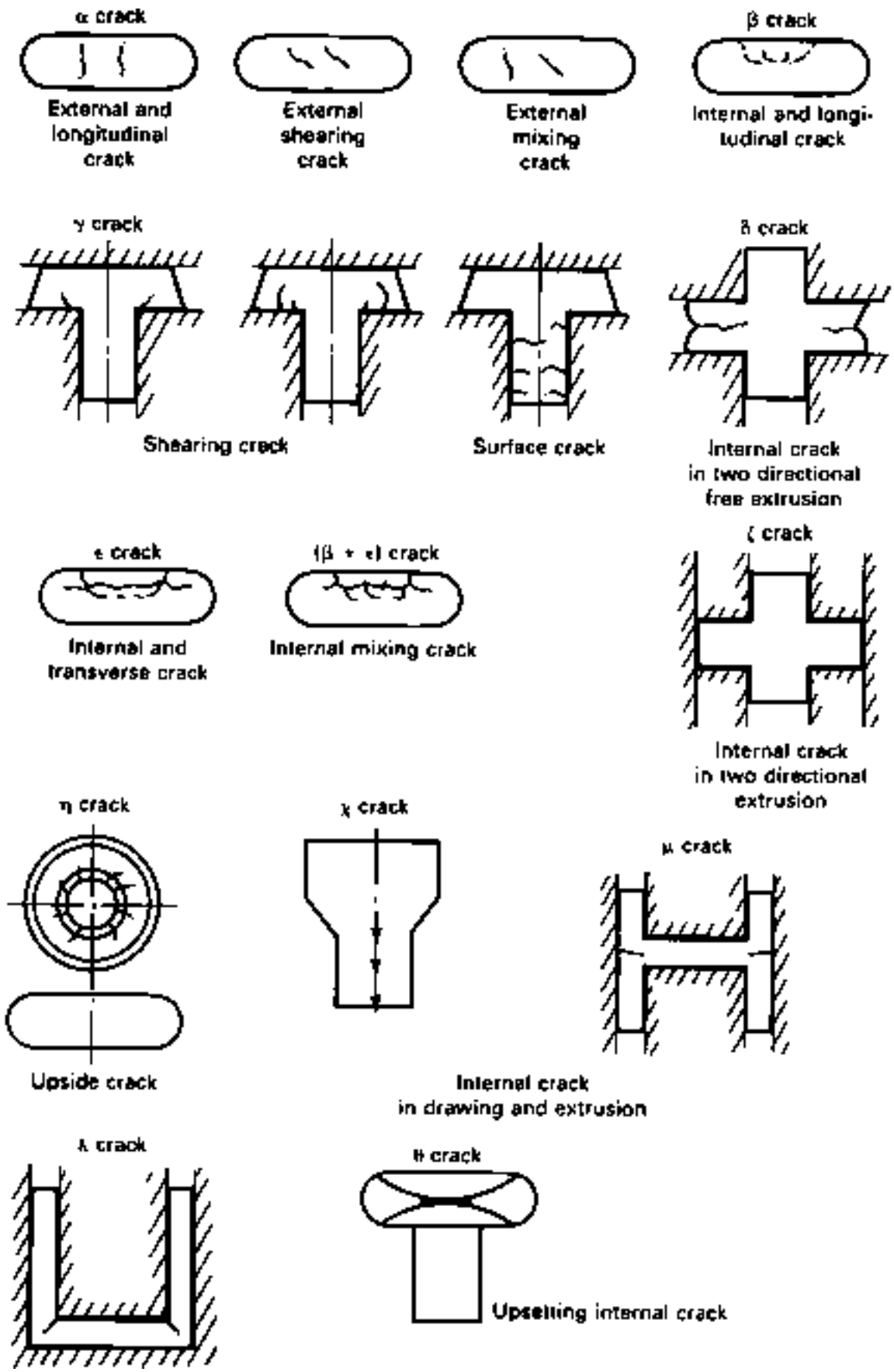
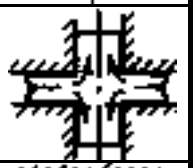
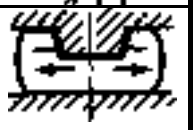


Fig. 35 Classification of cracks. Greek letters indicate type of cracking. See also Table 3. Source: Ref 18

Table 3 Characteristics of cracks and crack growth mechanism

Cracking type (Fig. 35)	Characteristics of cracking and estimation of crack growth mechanism	Crack
α	External cracking that appears at midheight of side surface of the specimen in the upsetting; two types of cracks, longitudinal and oblique, occur according to the degree of end constraint of specimen; they are all shear cracks	
β	Longitudinal cracking that occurs at the bottom of the concave part of the specimen (in upsetting by a circular truncated cone punch): it is a shear crack based on the section of the specimen and is caused by the expansion of material under the cone punch due to circumferential flow; prevention requires selection of a suitable punch shape	
γ	Shear cracking that appears at the corner of extruded material in free extrusion (coining); cracking occurs at the boundary between dead-metal and plastic zones; surface cracking is caused by cracks that occur at the corners; prevention requires selection of a suitable diameter for the die	
δ	Cracking that occurs at midheight of material in flange in two direction free extrusion without side constraint; the material is extruded forward and backward; therefore, cracking is caused by the depression of material in flange; prevention requires selection of a suitable diameter die	
ε	Cracking that occurs at midheight on inside surface of the concave part and is advanced in a circumferential direction in upsetting by the circular truncated cone punch; crack starts at the point where the material around the concave portion bends toward the inside; prevention requires selection of a suitable punch shape	
α + β	Cracking in which α and β cracks coexist	
ζ	Cracking that occurs at the center of material with the sides constrained and forward and backward extrusion; the cavity occurs at the center of material, as material is extruded forward and backward	
η	Cracking that appears to advance from upper and side surface of concave portion to the top of specimen (in upsetting by a circular truncated cone punch); in cross section, the cracks are distributed radially; in the case of a large tapered punch, cracks are caused by the expansion of upper part of specimen by the punch; prevention requires selection of a suitable punch shape	
κ	Cracking that occurs at the center of specimen when excessive reduction is imposed on the specimen in extrusion and drawing	
μ	Peripheral cracking that occurs at midheight on the side surface of the specimen in forward and backward piercing and 45° shearing crack with respect to longitudinal axis	
λ	Cracking that occurs at the bottom of the concave portion in piercing	
θ	Microscopic cracks that occur at the boundary between top and bottom dead metal and at the point of inflection of the metal flow in case of excessive upsetting of a bolt head in bolt forging: in practical use, the splitting off of the head is caused by these microscopic cracks	

Source: Ref 18

Free-Surface Cracking (Adapted from Ref 19). Free-surface fracture occurs at a free surface when there is no stress normal to the surface; the free surface expands and cracks due to the deformation process. For forging processes, this would include cracking on the exposed surfaces during upsetting and cracking on surfaces at the leading edge of localized extrusion during forging. For rolling processes, this may be edge cracking of rolled slabs, plates, or rings.

One of the most successful and useful design tools to come from bulk workability research is the workability diagram for free-surface fracture during the cold working of wrought and powder metals. An example of a workability diagram of this type is shown in Fig. 36 (Ref 20). The graph indicates the locus of free-surface normal strains (one tensile and one compressive) that cause fracture. The workability diagrams are used during process design by plotting calculated or estimated surface strain paths that are to be imposed during forming on the fracture locus diagram (Fig. 36). If the final strains lie above the locus, part failure is likely, and changes are necessary in preform design, lubrication, and/or material. The fracture locus concept has been used to prevent free-surface cracking in forging and to prevent edge cracking in rolling. With modifications, the fracture locus approach has also provided insight into such failure modes as center bursting in extrusion and forging and die-workpiece contact fractures in forging. These limits change with chemistry, grain size, temperature, second-phase content, and possibly with strain rate.

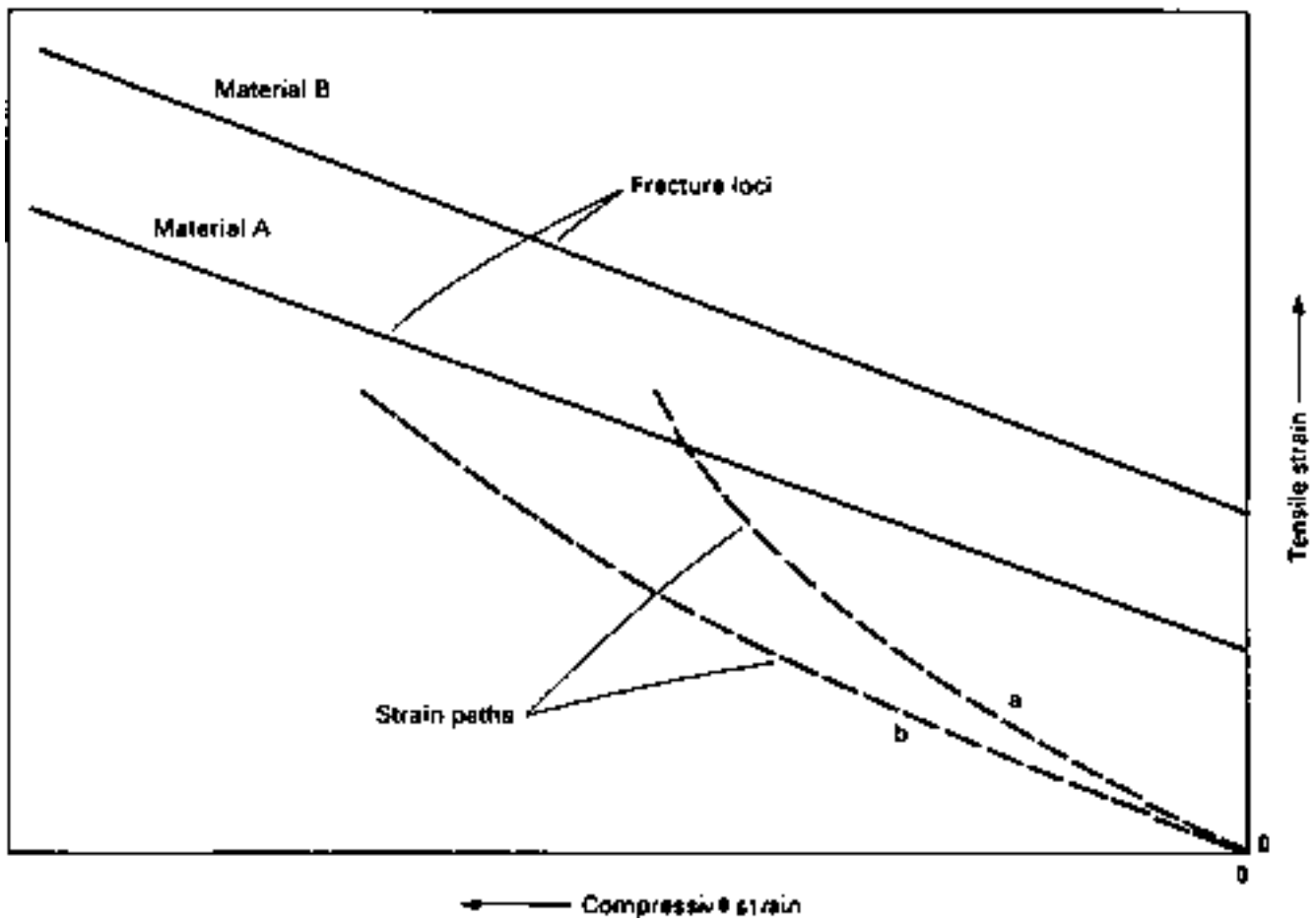


Fig. 36 Schematic workability diagrams for bulk forming processes. Strain path (a) would lead to failure for material A. Both strain paths (a and b) can be used for the successful forming of material B.

The concept of a working limit for free-surface fracture is important, because the workability of a metal may be characterized for a particular set of process conditions. In general, as the working temperature is increased, the location of the fracture line moves upward, indicating that higher deformation can be accommodated before fracture. However, higher temperatures are not always beneficial, as in the case of IN718 nickel-base superalloy, which has a temperature limit of approximately 1120 °C (2050 °F) for hot working. (Ref 21). Experimental observations have shown that the slope of the line increases with strain rate for some metals, most notably some brasses and austenitic stainless steels (Ref 22). The position of the line drops (lower workability) as the second-phase content increases, much like the tensile ductility decreases with second phase. The fracture line location is also sensitive to the microstructure. For example, a spheroidized structure for a high-carbon steel has a higher fracture line position than a pearlitic microstructure. More details and references on workability in terms of free-surface cracking are in Ref 19.

Die-Contact Surface Cracking (Ref 19). Cracking on surfaces in contact with a die is a common problem. Frequently, cracks occur during forging on surfaces that are in contact with the dies. One common location of

such defects is the vicinity of a die or punch corner. A combination of shear deformation with tension or low values of hydrostatic pressure in the vicinity of a die corner are responsible for surface cracking (Ref 23).

From the observation of a variety of such defects, it appears that a common characteristic is an abrupt change in frictional shear traction distribution in the region of the crack. High friction to retard metal flow in advance of the crack location is one method for preventing such defects. These cracks usually do not propagate deeply into the workpiece but instead result in unacceptable surface quality or unacceptable machining depths if that surface was to be finish machined. As the forging community has moved closer to net-shape forming, this type of defect has become an increasing problem. Causes of this problem include nonuniform lubrication, flow around die corners, and an improper starting workpiece surface.

References cited in this section

18. T. Okamoto, T. Fukuda, and H. Hagita, "Material Fracture in Cold Forging—Systematic Classification of Working Methods and Types of Cracking in Cold Forging," Sumitomo Search, No. 9, May 1973, p 46; *Source Book on Cold Forming,* American Society for Metals, 1975, p 216–226

19. B.L. Ferguson, Design for Deformation Processes, *Materials Selection and Design,* Vol 20, *ASM Handbook,* 1997, p 730–744

20. P.W. Lee and H.A. Kuhn, Cold Upset Testing, *Workability Testing Techniques,* G.E. Dieter, Ed., American Society for Metals, 1984, p 37

21. S.L. Semiatin, Workability in Forging, *Workability Testing Techniques,* G.E. Dieter, Ed., American Society for Metals, 1984, p 197–247

22. H.A. Kuhn and G.E. Dieter, Workability in Bulk Forming Processes, *Fracture 1977,* Vol 1, ICF4, (Waterloo, Canada), 1977, p 307–324

23. H.A. Kuhn and B.L. Ferguson, *Powder Forging,* Metal Powder Industries Federation, 1990

Sheet Forming

The major problems encountered in sheet metal forming are fracturing, buckling and wrinkling, shape distortion, loose metal, and undesirable surface textures. The occurrence of any one or a combination of these conditions can render the sheet metal part unusable. The effects of these problems are discussed subsequently.

Fracturing occurs when a sheet metal blank is subjected to stretching or shearing (drawing) forces that exceed the failure limits of the material for a given strain history, strain state, strain rate, and temperature. In stretching, the sheet initially thins uniformly, at least in a local area. Eventually, a point is reached at which deformation concentrates and causes a band of localized thinning known as a neck, which ultimately fractures. The formation of a neck is generally regarded as failure, because it produces a visible defect and a structural weakness. Most current formability tests are concerned with fracture occurring in stretching operations.

In shearing, fracture can take place without prior thinning. The most common examples of this type of fracture occur in slitting, blanking, and trimming. In these operations, sheets are sheared by knife edges that apply forces normal to the plane of the sheet. Shearing failures are sometimes produced in stamping operations by shearing forces in the plane of the sheet, but they are much less common than stretching failures.

Buckling and Wrinkling. In a typical stamping operation, the punch contacts the blank, stretches it, and starts to pull it through the blankholder ring. The edges of the blank are pulled into regions with progressively smaller perimeters. This produces compressive stresses in the circumferential direction. If these stresses reach a critical level characteristic of the material and its thickness, they cause slight undulations known as buckles. Buckles may develop into more pronounced undulations or waves, known as wrinkles, if the blankholder pressure is not sufficiently high.

This effect can also cause wrinkles in other locations, particularly in regions with abrupt changes in section and in regions where the metal is unsupported or contacted on one side only. In extreme cases, folds and double or

triple metal may develop. These may, in turn, lead to splitting in another location by preventing metal flow or by locking the metal out. Therefore, increasing the blankholder pressure often corrects a splitting problem.

Shape Distortion. On removal of the external forces, the internal elastic stresses relax. In some locations, they can relax completely, with only a very slight change in the dimensions of the part. However, in areas subjected to bending, through-thickness gradients in the elastic stresses occur; that is, the stresses on the outer surfaces are different from those on the inner surfaces.

If these stresses are not constrained or locked in by the geometry of the part, relaxation causes a change in the part shape known as shape distortion or springback. Springback can be compensated for in die design for a specific set of material properties but may still be a problem if there are large material property or process variations from blank to blank.

Loose metal occurs in undeformed regions and is undesirable, because it can be easily deflected. A phenomenon usually referred to as oil canning, in which a local area can be either concave or convex, can also be encountered. In stampings with two or more sharp bends of the same sign in roughly the same direction, such as a pair of feature lines, a tendency exists for the metal between them to be loose because of the difficulty involved in pulling metal across a sharp radius.

It is sometimes possible to avoid the problem by ensuring that the metal is not contacted by both lines at the same time; thus, some stretching can occur before the second line is contacted. There is a tendency for loose metal to occur toward the center of large, flat, or slightly curved parts. Increasing the restraining forces on the blank edges usually improves this condition.

Undesirable Surface Textures. Heavily deformed sheet metal, particularly if it is coarse grained, often develops a rough surface texture commonly known as “orange-peel.” This is usually unacceptable in parts that are visible in service. Another source of surface problems occurs in metals that have a pronounced yield point elongation, that is, materials that stretch several percent without an increase in load after yielding. In these metals, deformation at low strain levels is concentrated in irregular bands known as Lüders lines (or bands) or stretcher strains.

These defects disappear at moderate and high strain levels. However, almost all parts have some low-strain regions. These defects are unsightly and are not concealed by painting. Aged rimmed steels and some aluminum-magnesium alloys develop severe Lüders lines.

Cold Formed Parts

In general, parts produced by cold forming are susceptible to the same causes of failure as parts of the same metal composition and function produced by other forming processes. There are, however, some unique characteristics of cold formed parts that may influence their susceptibility to failure. Similar to parts formed by other metalworking methods, the prevention of failures in cold formed parts generally involves well-known preventive or corrective measures that may include changes in alloy composition, design, and/or process conditions (tooling, lubricant, heat treatment, etc.) or the elimination of surface discontinuities. However, various metals have their own specific peculiarities. For example, in the forming of certain stainless steels or iron-base heat-resisting alloys, severe forming encourages grain-boundary carbide precipitation if the metal is reheated to a temperature within the sensitizing-temperature range (400 to 870 °C, or 750 to 1600 °F). In the sensitized condition, the metal is extremely vulnerable to stress-corrosion cracking. Such a condition can be corrected only by the use of a recrystallizing anneal after forming, or it can be avoided by using a grade of stainless steel that is less susceptible to sensitization.

In some cases, limited heat is applied to the work metal before forming begins in order to increase its formability. Some metals—for example, certain magnesium alloys—can be difficult to form at room temperature. Preheating to temperatures as low as 120 °C (250 °F) makes these alloys much more formable. The practice of applying some heat to the material being formed without surpassing or closely approaching the recrystallization temperature is known as warm forming.

Regardless of the initial temperature, forming must be performed within the boundaries set by the material and tooling used. It is obviously possible to exceed the strength or elastic limits of a given material during a drastic deformation process. Care must therefore be taken to ensure that material specifications are accurately and realistically prepared in order to provide a product that fulfills the desired expectations. Designing a process that operates at the extreme upper limit of the capability of any material leads to an unacceptably high level of part failure because of the inevitable variations that occur from heat to heat or lot to lot.

Grain Deformation. Cold forming of any metal part deforms the grains; the amount of deformation depends on the severity of forming. Severe deformation of the grains in locations of sharp bends or where the metal has become stretched, as in deep-drawn parts, results in a condition that is more vulnerable to failure from such mechanisms as corrosion and fatigue. Susceptibility to corrosion is easily demonstrated by observing simple parts formed from low-carbon steel and exposed to the weather. It is often observed that the severely deformed areas rust before areas on the same part that have been subjected to little or no deformation. This condition prevails even when there are no signs of cracks in severely deformed areas.

Frequently, there are fine cracks in certain areas of formed parts where deformation was severe. Although such cracks are harmless in many applications, they help to accelerate any type of corrosion, and they also initiate fatigue cracks if the part is subjected to repeated stress.

Grain Size Variation in Thin-Section Product. Formability and grain size generally become more important as the metal being formed becomes thinner. As a rule, finer grain size improves formability. In some instances of coarse-grain metal, the locations where forming was the most severe may have only a very few highly deformed grains in a given cross section. It is important to select a fine-grain metal to accommodate the tendency of grains of varying orientation to deform in preferential directions, but this is not always sufficient; it may be necessary to follow forming with full annealing (recrystallization) or at least with a stress-relieving heat treatment.

Precautions are also needed in specifying a recrystallization-annealing treatment because of the possibility of producing very large grains, which may result in undesirable properties for some applications. Subcritical strain for recrystallization can be achieved in the transition region from a bend radius to a straight section in parts formed from sheet, strip, or wire; near the neutral axis in some bent structural shapes, such as I-beams or channels; and in some deep-drawn parts.

Precautions should also be taken regarding grain growth during other heat treating operations after forming. The following is an example of this for an aluminum connector tube that was solution treated after bending. Solution treatment induced excessive grain growth.

Other characteristics that are generally unique to cold-formed parts include certain types of tool marks, particularly indentations at inside bend radii, scoring that may occur during deep drawing or forming, and scoring on outer surfaces such as often occurs when forming in a press brake using V-dies.

Orange-peel results when a very coarse-grain metal is stressed beyond its yield strength, increasing vulnerability to corrosion and helping to initiate fatigue cracks if the part is subjected to repeated stress.

Corner Thinning causes concentration of stress flow lines at corners, which may help to cause failure by one or more mechanisms.

Work Hardening. Another unique characteristic of cold-formed parts involves local areas that have become severely work hardened, such as those areas that have been thinned by stretching or localized compression. Such areas are vulnerable to stress cracking.

Plating. Many formed parts are plated to protect them from corrosion. This often causes difficulties, because the open areas of the part where there was little or no deformation usually receive the thickest plating, whereas severely formed areas may be plating-starved—for example, the inside corners of a box-shaped part. Thus, some formed parts are more susceptible to various forms of corrosion simply because the more vulnerable areas were not sufficiently protected.

Spangles in Zinc-Coated Steel Sheet. In some cases, zinc-coated steels exhibit surface defects known as spangles. This phenomenon occurs only in hot-dipped products and is caused by the development of a coarse grain size in the galvanic coating, which makes the individual grains clearly visible. This problem can be corrected in the coating process. In addition to the previously mentioned occurrences, handling damage, dents caused by dirt or slivers in the die, and scoring or galling caused by a rough die surface or inadequate lubrication sometimes produce unacceptable surfaces.

References

1. J. Schey, Manufacturing Processes and Their Selection, *Materials Selection and Design,* Vol. 20, *ASM Handbook,* ASM International, 1997, p 687–704

2. L. Pease III and W. West, *Fundamentals of Powder Metallurgy,* Metal Powder Industries Federation, 2002

3. G.F. Vander Voort, *Metallography: Principles and Practice,* McGraw-Hill, 1984 (out of print), reprinted 1999 ASM International

4. R.K. Dayal, J.B. Gnanamoorthy, and P. Rodriguez, Failure of an I-Beam, *Handbook of Case Histories in Failure Analysis,* Vol 1, ASM International, 1992, p 378–381

5. E. Kauczor, *Prakt. Metallogr.* Vol 8, July 1971, p 443–446

6. G. Vander Voort, Embrittlement of Steels, *Properties and Selection: Iron, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, p 689–736

7. P.A. Thornton, *J. Mater. Sci.,* Vol 6, 1971, p 347

8. D.V. Edmonds and C.J. Beevers, *J. Mater. Sci.,* Vol 3, 1968, p 457

9. A.D. Wilson, Calcium Treatment of Plate Steels and Its Effect on Fatigue and Toughness Properties, *Proc. 11th Annual Offshore Technology Conference,* OTC 3465, May 1979, p 939–948

10. A.D. Wilson, Effect of Calcium Treatment on Inclusions in Constructional Steels, *Met. Prog.,* Vol 121 (No. 5), April 1982, p 41–46

11. M.D. Chaudhari, Failure Analysis of a Helicopter Main Rotor Bolt, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 392

12. V.N. Whittacker, *Nondestr. Test.,* Oct 1971, p 320

13. S.L. Semiatin and J.J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization,* American Society for Metals, 1984, p 3

14. S.L. Semiatin and G.D. Lahoti, The Occurrence of Shear Bands in Isothermal Hot Forging, *Metall. Trans. A,* Vol 13, 1982, p 275–288

15. M.C. Mataya and G. Krauss, A Test to Evaluate Flow Localization during Forging, *J. Appl. Metalwork.,* Vol 2, 1981, p 28–37

16. F.N. Rhines and P.J. Wray, Investigation of the Intermediate Temperature Ductility Minimum in Metals, *Trans. ASM,* Vol 54, 1961, p 117

17. A.M. Sabroff, F.W. Boulger, and H.J. Henning, *Forging Materials and Practices,* Reinhold, 1968

18. T. Okamoto, T. Fukuda, and H. Hagita, "Material Fracture in Cold Forging—Systematic Classification of Working Methods and Types of Cracking in Cold Forging," Sumitomo Search, No. 9, May 1973, p 46; *Source Book on Cold Forming,* American Society for Metals, 1975, p 216–226

19. B.L. Ferguson, Design for Deformation Processes, *Materials Selection and Design,* Vol 20, *ASM Handbook,* 1997, p 730–744

20. P.W. Lee and H.A. Kuhn, Cold Upset Testing, *Workability Testing Techniques,* G.E. Dieter, Ed., American Society for Metals, 1984, p 37

21. S.L. Semiatin, Workability in Forging, *Workability Testing Techniques,* G.E. Dieter, Ed., American Society for Metals, 1984, p 197–247

22. H.A. Kuhn and G.E. Dieter, Workability in Bulk Forming Processes, *Fracture 1977*, Vol 1, ICF4, (Waterloo, Canada), 1977, p 307–324

23. H.A. Kuhn and B.L. Ferguson, *Powder Forging*, Metal Powder Industries Federation, 1990

Failures Related to Casting

Introduction

FAILURES OF CASTINGS, like the failures of wrought materials, can occur from service conditions, improper design and/or materials selection, manufacturing deficiencies, or a combination of these. A design-related failure is one in which even a perfect casting (free from unexpected deficiencies and meeting specified requirements) does not perform to the requirements of the application. Correspondingly, a defect-related failure is one resulting from a condition that falls outside of the established acceptance criteria that the designer did not or could not anticipate. In many instances, both design and manufacture of the casting contribute to premature failure.

Clearly, both types of failure are preventable. On first inspection, it seems clear that the designer must accept responsibility for failures of the first type, and the founder the second. No such distinction should be drawn. The key to serviceable castings is communication between the designer and the foundry engineers. Foundrymen are an invaluable source of information on just what can be produced in cast form, both in terms of geometry and materials selection. This input should be sought early in the design process, to incorporate the suggestions where practical. This can also have a dramatic impact on the manufacturability of the final part, influencing the ultimate quality of the casting and thus reducing the likelihood of failure due to defective parts.

Included in the design process is the determination of specifications to which the foundry must work. Here, too, the foundry can be invaluable in providing the designer with input as to what materials should be used, what specifications are called out, what nondestructive testing techniques are applicable, what level of metallurgical testing should be performed, and, finally, what quality-assurance documentation is needed. A teamwork approach between the designer and foundry (and, where possible, the ultimate user) is the best (and lowest-cost) method of avoiding casting failures.

This article describes the general root causes of failure attributed to the casting process, casting material, and/or design. In failure analysis, it is fundamental to establish the objectives of the designer by reviewing drawings, specifications, and purchase orders. It is also important to determine what was manufactured, delivered, and installed, and to find if it agrees with the drawings. It is then necessary to ascertain what happened to the component before and at the time of failure (history, performance, environment, and so on). The evidence and investigation should lead to determination of the cause of the failure, in terms of faulty design or materials selection, faulty material or processing, and/or service conditions. To determine the true cause of failure, the investigator must fully consider the interplay among design, fabrication, material properties, environment, and service conditions. Once the failure is accurately classified, corrective action and recommendations can be made. Appropriate solutions can involve redesign, change in materials selection or processing, quality control, changes in maintenance or inspection schedules, protection against environment, or restrictions on service loads or service life.

Casting Discontinuities

Discontinuities are interruptions in the physical continuity of a casting, and the engineering community tends to make specific distinctions between the terms *discontinuity* (or *flaw*) and *defect*. By definition, a defect is a condition that must be removed or corrected. The word *defect*, therefore, should be carefully used, because it

implies that a casting is defective and requires corrective measures or rejection. An imperfection also may be a manufacturing defect, in the sense of being an unacceptable condition during an inspection regime, but not a functional defect. Critical engineering assessment of discontinuities is therefore necessary to define whether an imperfection is a defect or a harmless discontinuity that does not sacrifice the reliability or intended function of the casting.

Although the term *discontinuity* or *imperfection* is preferred when referring to deviations in less-than-perfect castings, the term *casting defect* is sometimes used for general imperfections as well. Unfavorable foundry practice can result in various casting imperfections that are detrimental in service and that may contribute to failure. However, many of the common causes of failures of castings occur from one or more aspects of design, materials selection, casting imperfections, faulty processing, improper assembly, or service conditions not initially anticipated. Some casting imperfections have no effect on the function or the service life of cast components, but will give an unsatisfactory appearance or will make further processing, such as machining, more costly. Many such imperfections can be easily corrected by shotblast cleaning or grinding.

Other imperfections or defects that can be more difficult to remove can be acceptable in some locations. Moreover, many castings (especially iron and steel castings) generally lend themselves to being weld repaired, with proper attention to the details of weld repair procedures. Imperfections and defects can be removed and a suitable weld deposit substituted to give a finished product fit for the intended service, although many failures of castings also result from improper repair or rework of casting defects (see the section “Welding” in this article).

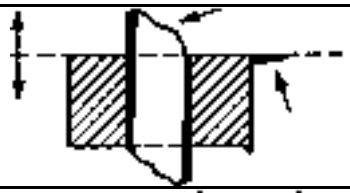
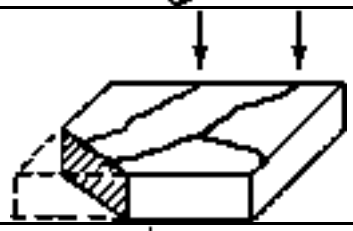
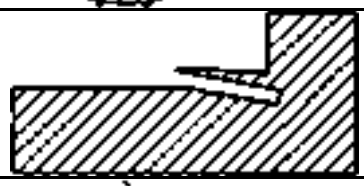
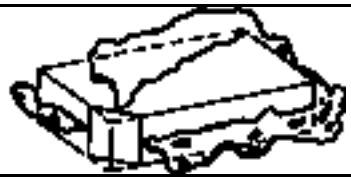
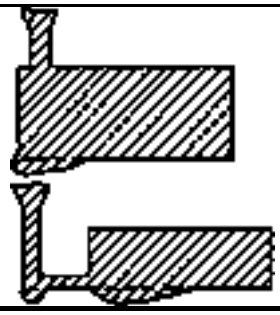
In summary, it is most critical that the casting designer specify the true design needs. Poor casting design can interfere with the ability of the foundry to use the best techniques to produce reliable castings. Defect-free castings can be produced at a price. The multitude of process variables, such as molding mediums, binder, gating and risering, melting and ladle practice, pouring technique, and heat treatment, that must be controlled to produce such sound castings requires a thorough understanding of the processes used and strict engineering supervision. The price of a casting often reflects the cost of such efforts. The attention to detail necessary to make good castings can reduce total cost of a manufactured part by significantly reducing machining, repair, and fit problems later in the assembly. Proper specification, as opposed to overspecification or underspecification, is the key to the successful application, that is, freedom from failure, of castings.

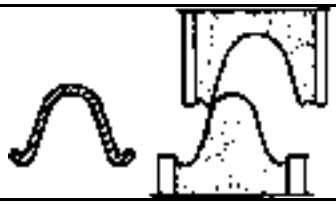
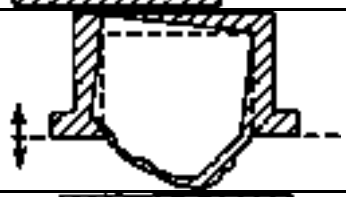
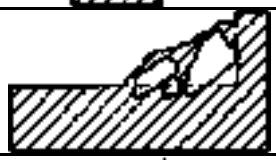
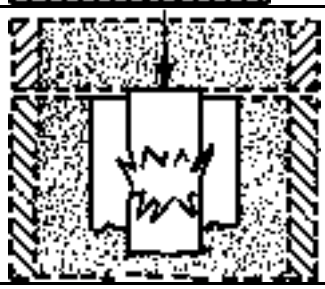
International Classification of Casting Defects. Foundrymen have traditionally used rather unique names, such as rattail, scab, buckle, and shut, to describe various casting imperfections. Some of these terms have been used to describe imperfections in cast iron and steels, as noted in the following section “Imperfections That Can Occur in Iron and Steel Castings.” However, because foundrymen sometimes use different nomenclature to describe imperfections or defects, the International Committee of Foundry Technical Associations (ICFTA) has standardized the nomenclature, starting with the identification of seven basic categories of casting imperfections (or defects, in the terminology used by the ICFTA classification scheme, Table 1):

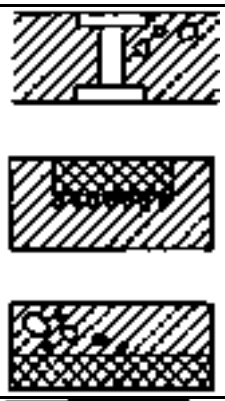
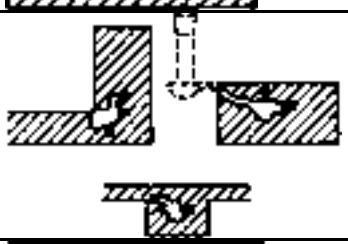
- Metallic projections
- Cavities
- Discontinuities
- Defects
- Incomplete casting
- Incorrect dimension
- Inclusions or structural anomalies

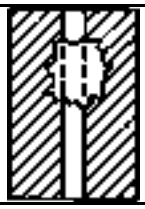
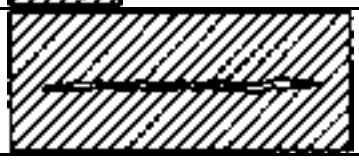
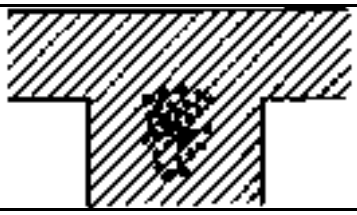
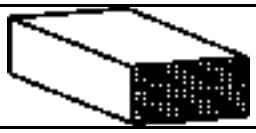
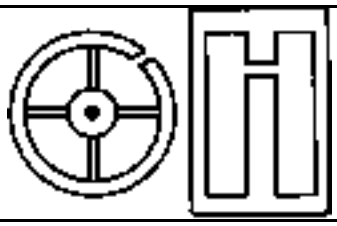
Table 1 International classification of common casting defects

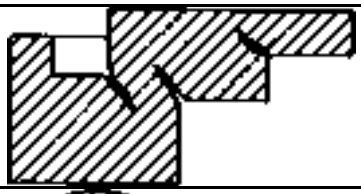
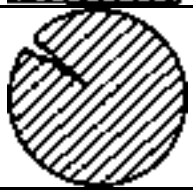
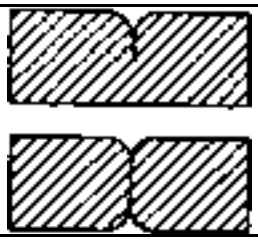
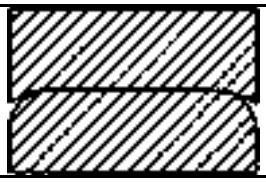
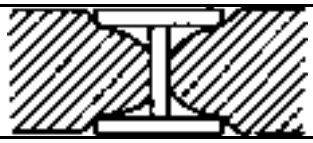
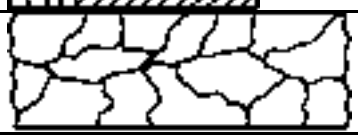
No.	**Description**	**Common name**	**Sketch**
Metallic Projections			
A 100:	**Metallic projections in the form of fins or flash**		
A	**Metallic projections in the form of fins (or flash) without change in principal casting dimensions**		

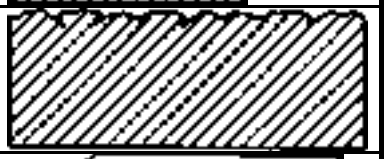
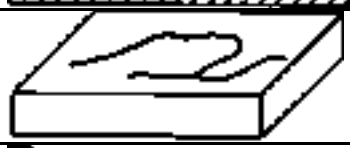
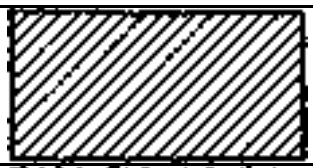
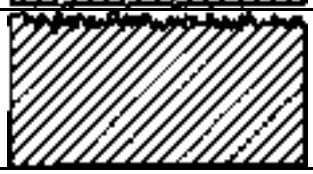
No.	Description	Common name	Sketch
Metallic Projections			
110:			
A 111	Thin fins (or flash) at the parting line or at core prints	Joint flash or fins	
A 112	Projections in the form of veins on the casting surface	Veining or finning	
A 113	Network of projections on the surface of die castings	Heat-checked die	
A 114[(a)]	Thin projection parallel to a casting surface, in re-entrant angles	Fillet scab	
A 115	Thin metallic projection located at a re-entrant angle and dividing the angle in two parts	Fillet vein	
A 120:	**Metallic projections in the form of fins with changes in principal casting dimensions**		
A 123[(a)]	Formation of fins in planes related to direction of mold assembly (precision casting with waste pattern); principal casting dimensions change	Cracked or broken mold	
A 200:	**Massive projections**		
A 210:	**Swells**		
A 212[(a)]	Excess metal in the vicinity of the gate or beneath the sprue	Erosion, cut, or wash	

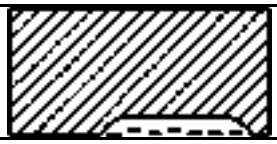
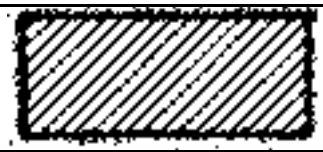
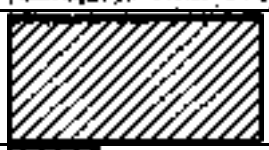
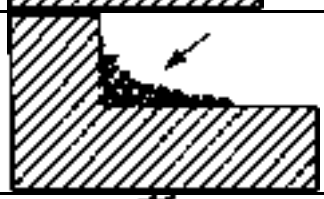
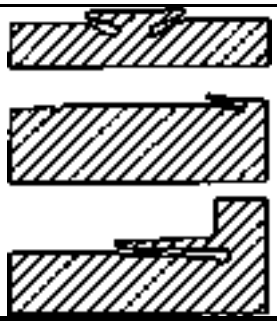
No.	Description	Common name	Sketch
Metallic Projections			
A 213[(a)]	Metal projections in the form of elongated areas in the direction of mold assembly	Crush	
A 220:	**Projections with rough surfaces**		
A 221[(a)]	Projections with rough surfaces on the cope surface of the casting	Mold drop or sticker	
A 222[(a)]	Projections with rough surfaces on the drag surface of the casting (massive projections)	Raised core or mold element cutoff	
A 223[(a)]	Projections with rough surfaces on the drag surface of the casting (in dispersed areas)	Raised sand	
A 224[(a)]	Projections with rough surfaces on other parts of the casting	Mold drop	
A 225[(a)]	Projections with rough surfaces over extensive areas of the casting	Corner scab	
A 226[(a)]	Projections with rough surfaces in an area formed by a core	Broken or crushed core	
Cavities			
B 100:	**Cavities with generally rounded, smooth walls perceptible to the naked eye (blowholes, pinholes)**		
B 110:	**Class B 100 cavities internal to the casting, not extending to the surface, discernible only by special methods, machining, or fracture of the casting**		
B 111[(a)]	Internal, rounded cavities, usually smooth-walled, of varied size, isolated or grouped irregularly in all areas of the casting	Blowholes, pinholes	

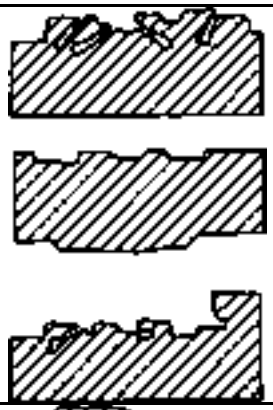
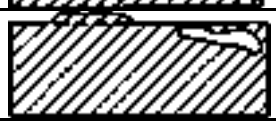
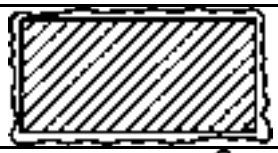
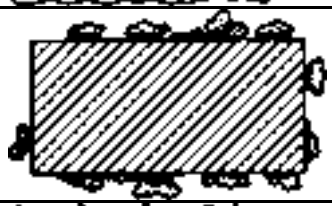
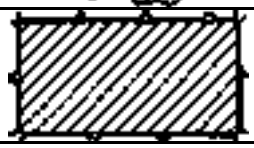
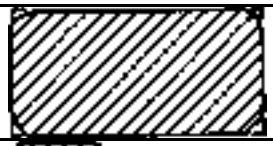
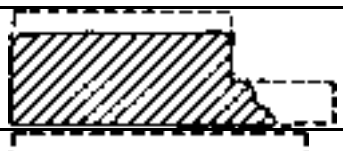
No.	Description	Common name	Sketch
Metallic Projections			
B 112[(a)]	As above, but limited to the vicinity of metallic pieces placed in the mold (chills, inserts, chaplets, etc.)	Blowholes, adjacent to inserts, chills, chaplets, etc.	
B 113[(a)]	Like B 111, but accompanied by slag inclusions (G 122)	Slag blowholes	
B 120:	**Class B 100 cavities located at or near the casting surface, largely exposed or at least connected with the exterior**		
B 121[(a)]	Exposed cavities of various sizes, isolated or grouped, usually at or near the surface, with shiny walls	Surface or subsurface blowholes	
B 122[(a)]	Exposed cavities, in re-entrant angles of the casting, often extending deeply within	Corner blowholes, draws	
B 123	Fine porosity (cavities) at the casting surface, appearing over more or less extended areas	Surface pinholes	
B 124[(a)]	Small, narrow cavities in the form of cracks, appearing on the faces or along edges, generally only after machining	Dispersed shrinkage	
B 200:	**Cavities with generally rough walls, shrinkage**		
B 210:	**Open cavity of Class B 200, sometimes penetrating deeply into the casting**		
B 211[(a)]	Open, funnel-shaped cavity; wall usually covered with dendrites	Open or external shrinkage	
B 212[(a)]	Open, sharp-edged cavity in fillets of thick castings or at gate locations	Corner or fillet shrinkage	

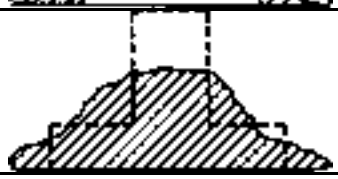
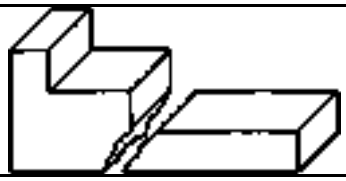
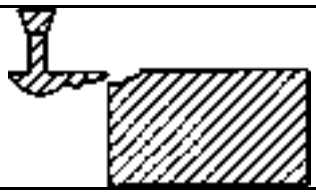
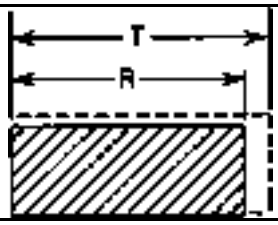
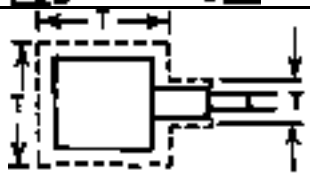
No.	Description	Common name	Sketch
Metallic Projections			
B 213(a)	Open cavity extending from a core	Core shrinkage	
B 220:	**Class B 200 cavity located completely internal to the casting**		
B 221[(a)]	Internal, irregularly shaped cavity; wall often dendritic	Internal or blind shrinkage	
B 222[(a)]	Internal cavity or porous area along central axis	Centerline or axial shrinkage	
B 300:	**Porous structures caused by numerous small cavities**		
B 310:	**Cavities according to B 300, scarcely perceptible to the naked eye**		
B 311[(a)]	Dispersed, spongy dendritic shrinkage within walls of casting; barely perceptible to the naked eye	Macro- or micro-shrinkage, shrinkage porosity, leakers	
Discontinuities			
C 100:	**Discontinuities, generally at intersections, caused by mechanical effects (rupture)**		
C 110:	**Normal cracking**		
C 111[(a)]	Normal fracture appearance, sometimes with adjacent indentation marks	Breakage (cold)	
C 120:	**Cracking with oxidation**		
C 121[(a)]	Fracture surface oxidized completely around edges	Hot cracking	
C 200:	**Discontinuities caused by internal tension and restraints to contraction (cracks and tears)**		
C 210:	**Cold cracking or tearing**		
C 211[(a)]	Discontinuities with squared edges in areas susceptible to tensile stresses during cooling; surface not oxidized	Cold tearing	
C 220:	**Hot cracking and tearing**		

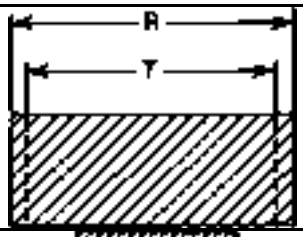
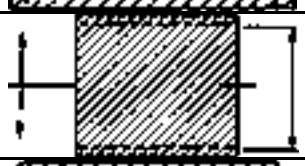
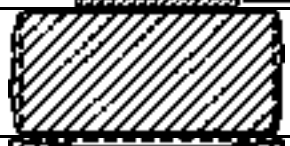
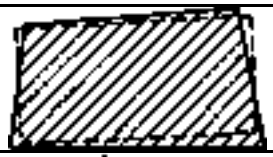
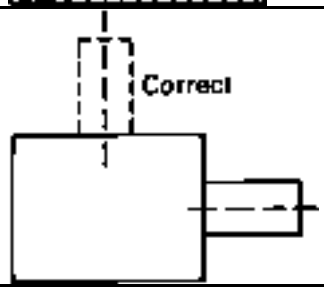
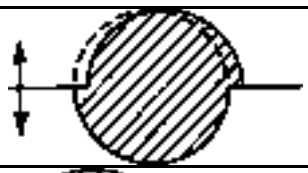
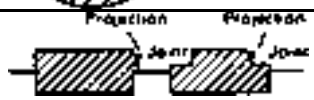
No.	Description	Common name	Sketch
Metallic Projections			
C 221[(a)]	Irregularly shaped discontinuities in areas susceptible to tension; oxidized fracture surface showing dendritic pattern	Hot tearing	
C 222[(a)]	Rupture after complete solidification, either during cooling or heat treatment	Quench cracking	
C 300:	**Discontinuities caused by lack of fusion (cold shuts); edges generally rounded, indicating poor contact between various metal streams during filling of the mold**		
C 310:	**Lack of complete fusion in the last portion of the casting to fill**		
C 311[(a)]	Complete or partial separation of casting wall, often in a vertical plane	Cold shut or cold lap	
C 320:	**Lack of fusion between two parts of casting**		
C 321[(a)]	Separation of the casting in a horizontal plane	Interrupted pour	
C 330:	**Lack of fusion around chaplets, internal chills, and inserts**		
C 331[(a)]	Local discontinuity in vicinity of metallic insert	Chaplet or insert cold shut, unfused chaplet	
C 400:	**Discontinuities caused by metallurgical defects**		
C 410:	**Separation along grain boundaries**		
C 411[(a)]	Separation along grain boundaries of primary crystallization	Conchoidal or “rock candy” fracture	
C 412[(a)]	Network of cracks over entire cross section	Intergranular corrosion	
Defective Surface			
D 100:	**Casting surface irregularities**		
D 110:	**Fold markings on the skin of the casting**		

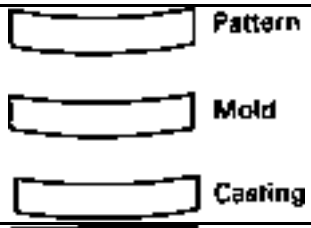
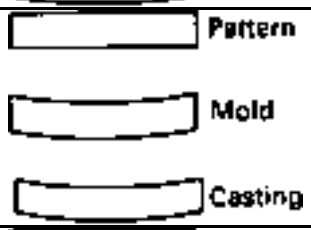
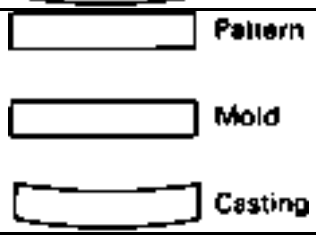
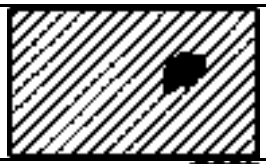
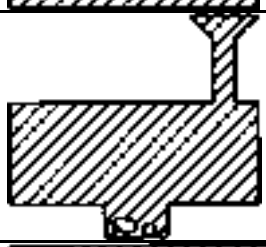
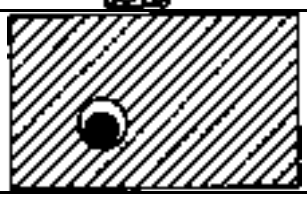
No.	Description	Common name	Sketch
Metallic Projections			
D 111	Fold markings over rather large areas of the casting	Surface folds, gas runs	
D 112	Surface shows a network of jagged folds or wrinkles (ductile iron)	Cope defect, elephant skin, laps	
D 113	Wavy fold markings without discontinuities; edges of folds at same level, casting surface is smooth	Seams or scars	
D 114	Casting surface markings showing direction of liquid metal flow (light alloys)	Flow marks	
D 120:	**Surface roughness**		
D 121	Depth of surface roughness is approximately that of the dimensions of the sand grains	Rough casting surface	
D 122	Depth of surface roughness is greater than that of the sand grain dimensions	Severe roughness, high pressure molding defect	
D 130:	**Grooves on the casting surface**		
D 131	Grooves of various lengths, often branched, with smooth bottoms and edges	Buckle	
D 132	Grooves up to 5.1 mm (0.2 in.) in depth, one edge forming a fold which more or less completely covers the groove	Rat tail	
D 133	Irregularly distributed depressions of various dimensions extending over the casting surface, usually along the path of metal flow (cast steel)	Flow marks, crow's feet	
D 134	Casting surface entirely pitted or pock-marked	Orange peel, metal mold reaction, alligator skin	
D 135	Grooves and roughness in the vicinity of re-entrant angles on die castings	Soldering, die erosion	
D 140:	**Depressions in the casting surface**		

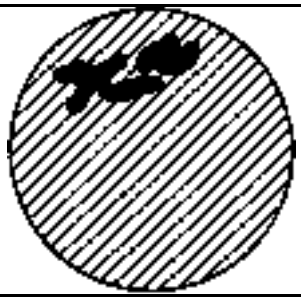
No.	Description	Common name	Sketch
Metallic Projections			
D 141	Casting surface depressions in the vicinity of a hot spot	Sink marks, draw or suck-in	
D 142	Small, superficial cavities in the form of droplets of shallow spots, generally gray-green in color	Slag inclusions	
D 200:	**Serious surface defects**		
D 210:	**Deep indentation of the casting surface**		
D 211	Deep indentation, often over large area of drag half of casting	Push-up, clamp-off	
D 220:	**Adherence of sand, more or less vitrified**		
D 221	Sand layer strongly adhering to the casting surface	Burn on	
D 222	Very adherent layer of partially fused sand	Burn in	
D 223	Conglomeration of strongly adhering sand and metal at the hottest points of the casting (re-entrant angles and cores)	Metal penetration	
D 224	Fragment of mold material embedded in casting surface	Dip coat spall, scab	
D 230:	**Plate-like metallic projections with rough surfaces, usually parallel to casting surface**		
D 231[(a)]	Plate-like metallic projections with rough surfaces parallel to casting surface; removable by burr or chisel	Scabs, expansion scabs	

No.	Description	Common name	Sketch
Metallic Projections			
D 232[(a)]	As above, but impossible to eliminate except by machining or grinding	Cope spall, boil scab, erosion scab	
D 233[(a)]	Flat, metallic projections on the casting where mold or core washes or dressings are used	Blacking scab, wash scab	
D 240:	**Oxides adhering after heat treatment (annealing, tempering, malleablizing) by decarburization**		
D 241	Adherence of oxide after annealing	Oxide scale	
D 242	Adherence of ore after malleablizing (white heart malleable	Adherent packing material	
D 243	Scaling after anneal	Scaling	
Incomplete Casting			
E 100:	**Missing portion of casting (no fracture)**		
E 110:	**Superficial variations from pattern shape**		
E 111	Casting is essentially complete except for more or less rounded edges and corners	Misrun	
E 112	Deformed edges or contours due to poor mold repair or careless application of wash coatings	Defective coating (tear-dropping) or poor mold repair	
E 120:	**Serious variations from pattern shape**		
E 121	Casting incomplete due to premature solidification	Misrun	
E 122	Casting incomplete due to insufficient metal poured	Poured short	

No.	Description	Common name	Sketch
Metallic Projections			
E 123	Casting incomplete due to loss of metal from mold after pouring	Runout	
E 124	Significant lack of material due to excessive shot-blasting	Excessive cleaning	
E 125	Casting partially melted or seriously deformed during annealing	Fusion or melting during heat treatment	
E 200:	**Missing portion of casting (with fracture)**		
E 210:	**Fractured casting**		
E 211	Casting broken, large piece missing; fractured surface not oxidized	Fractured casting	
E 220:	**Piece broken from casting**		
E 221	Fracture dimensions correspond to those of gates, vents, etc.	Broken casting (at gate, riser, or vent)	
E 230:	**Fractured casting with oxidized fracture**		
E 231	Fracture appearance indicates exposure to oxidation while hot	Early shakeout	
Incorrect Dimensions or Shape			
F 100:	**Incorrect dimensions; correct shape**		
F 110:	**All casting dimensions incorrect**		
F 111	All casting dimensions incorrect in the same proportions	Improper shrinkage allowance	
F 120:	**Certain casting dimensions incorrect**		
F 121	Distance too great between extended projections	Hindered contraction	
F 122	Certain dimensions inexact	Irregular contraction	

No.	Description	Common name	Sketch
Metallic Projections			
F 123	Dimensions too great in the direction of rapping of pattern	Excess rapping of pattern	
F 124	Dimensions too great in direction perpendicular to parting line	Mold expansion during baking	
F 125	Excessive metal thickness at irregular locations on casting exterior	Soft or insufficient ramming, mold-wall movement	
F 126	Thin casting walls over general area, especially on horizontal surfaces	Distorted casting	
F 200:	**Casting shape incorrect overall or in certain locations**		
F 210:	**Pattern incorrect**		
F 211	Casting does not conform to the drawing shape in some or many respects; same is true of pattern	Pattern error	
F 212	Casting shape is different from drawing in a particular area; pattern is correct	Pattern mounting error	
F 220:	**Shift or Mismatch**		
F 221	Casting appears to have been subjected to a shearling action in the plane of the parting line	Shift	
F 222	Variation in shape of an internal casting cavity along the parting line of the core	Shifted core	
F 223	Irregular projections on vertical surfaces, generally on one side only in the vicinity of the parting line	Ramoff, ramaway	
F 230:	**Deformations from correct shape**		

No.	Description	Common name	Sketch
Metallic Projections			
F 231	Deformation with respect to drawing proportional for casting, mold, and pattern	Deformed pattern	
F 232	Deformation with respect to drawing proportional for casting and mold; pattern conforms to drawing	Deformed mold, mold creep, springback	
F 233	Casting deformed with respect to drawing; pattern and mold conform to drawing	Casting distortion	
F 234	Casting deformed with respect to drawing after storage, annealing, machining	Warped casting	
Inclusions or Structural Anomalies			
G 100:	**Inclusions**		
G 110:	**Metallic inclusions**		
G 111[(a)]	Metallic inclusions whose appearance, chemical analysis or structural examination show to be caused by an element foreign to the alloy	Metallic inclusions	
G 112[(a)]	Metallic inclusions of the same chemical composition as the base metal; generally spherical and often coated with oxide	Cold shot	
G 113	Spherical metallic inclusions inside blowholes or other cavities or in surface depressions (see A 311). Composition approximates that of the alloy cast but nearer to that of a eutectic	Internal sweating, phosphide sweat	
G 120:	**Nonmetallic inclusions; slag, dross, flux**		
G 121[(a)]	Nonmetallic inclusions whose appearance or analysis shows they arise from melting slags, products of metal treatment or fluxes	Slag, dross or flux inclusions, ceroxides	

No.	Description	Common name	Sketch
Metallic Projections			
G 122[(a)]	Nonmetallic inclusions generally impregnated with gas and accompanied by blowholes (B 113)	Slag blowhole defect	
G 130:	**Nonmetallic inclusions; mold or core materials**		
G 131[(a)]	Sand inclusions, generally very close to the surface of the casting	Sand inclusions	
G 132[(a)]	Inclusions of mold blacking or dressing, generally very close to the casting surface	Blacking or refractory coating inclusions	
G 140:	**Nonmetallic inclusions; oxides and reaction products**		
G 141	Clearly defined, irregular black spots on the fractured surface of ductile cast iron	Black spots	
G 142[(a)]	Inclusions in the form of oxide skins, most often causing a localized seam	Oxide inclusion or skins, seams	
G 143[(a)]	Folded films of graphitic luster in the wall of the casting	Lustrous carbon films, or kish tracks	
G 144	Hard inclusions in permanent molded and die cast aluminum alloys	Hard spots	

(a) Defects that under some circumstances could contribute, either directly or indirectly, to casting failures. Adapted from *International Atlas of Casting Defects,* American Foundrymen's Society, Des Plaines, IL

In this scheme, the term *discontinuity* has the specific meaning of a planar separation of the metal, that is, a crack. Table 1 presents some of the common defects in each category. In general, defects that can serve as stress raisers or crack promoters are the most serious. These include preexisting cracks, internal voids, and nonmetallic inclusions.

Imperfections That Can Occur in Iron and Steel Castings. This section presents definitions of the imperfections and discontinuities that can occur in iron and steel castings and that may be the basis for rejection during inspection. Some of these imperfections, if undetected, can serve as initiation points of casting failures during service. The effects of these imperfections on casting performance depend largely on their severity and location.

buckle.

An indentation in a sand casting resulting from expansion of the sand.

cold shot.

A small globule of metal that has solidified prematurely and that is embedded in, but not entirely fused with, the surface of the casting.

cold shut.

A discontinuity on, or immediately beneath, the surface of a casting, caused by the meeting of two streams of liquid metal that failed to merge. A cold shut may have the appearance of a crack or *seam* with smooth, rounded edges.

gas porosity.

Voids caused by entrapped gas, such as air or steam, or by the expulsion of dissolved gases during solidification.

hot crack.

A crack or fracture caused by internal stresses that develop after solidification and during cooling from an elevated temperature (above 650 °C, or 1200 °F, for gray iron). A hot crack is less visible (less open) than a *hot tear* and usually exhibits less evidence of oxidation and decarburization.

hot tear.

A crack or fracture formed prior to completion of solidification because of hindered contraction. A hot tear frequently is open to the surface of the casting and thus, exposed to the atmosphere. This may result in oxidation, decarburization, or other metal-atmosphere reactions at the tear surface.

inclusions.

Nonmetallic materials in a solid metallic matrix. In castings, common inclusions include particles of refractory, slag, deoxidation products, or oxides of the casting metal.

mechanical (cold) crack.

A crack or fracture resulting from rough handling or from thermal shock, such as may occur at shakeout or during heat treatment.

metal penetration.

An imperfection on the surface of a casting caused by the penetration of molten metal into voids between refractory particles of the mold—usually the result of excessive pouring temperatures or overly coarse refractory.

misrun.

A casting not fully formed because of solidification of metal before the mold is filled.

off-dimension casting.

A casting of the wrong size or containing an error in shape as a result of a dimensional error in the pattern, incorrect design of the pattern and the mold equipment, mismatch between cope and drag, or shifting of a core during pouring.

rattail.

In a sand casting, a minor *buckle* occurring as a small irregular line on the surface.

sand hole.

A pit in the surface of a sand casting resulting from a deposit of loose sand on the surface of the mold.

scab.

An imperfection consisting of a flat volume of metal joined to a casting through a small area. Usually, a scab is set in a depression in the casting and is separated from the metal of the casting proper by a thin layer of sand.

seam.

An unfused fold or lap on a cast surface that appears as a crack. See also *cold shut.*

shrinkage porosity.

Voids resulting from solidification shrinkage. The growth of dendrites during the freezing process may isolate local regions, preventing complete feeding from the risers. Shrinkage cavities or depressions are the result of shrinkage porosity.

stress cracks.

Cracks that result from high residual stresses after the casting has cooled to below 650 °C (1200 °F). Stress cracks may form at room temperature several days after casting.

Effect of Casting Discontinuities (Adapted from Ref 1)

Because the designer must specify the quality requirements that ensure that the cast component will perform as desired, the effect of discontinuities on properties is important. As previously noted, over-specification causes needless expense and can be avoided by understanding the effect of discontinuities on casting performance and the effect of casting design on the tendency for discontinuities to form during the casting process. The existence of casting discontinuities does not, in and of itself, indicate that casting performance in service will be affected. Equally important are the size, location, and distribution of these discontinuities. Those discontinuities that are small and located near the center of the casting have little effect, while those located at or near the surface of the casting are usually damaging. Clustered discontinuities and those that occur in a regular array have a greater effect on properties than those that are isolated and randomly distributed.

The extent, cause, and type of imperfections or defects vary, to some extent, on the method of casting, depending on whether casting is done with the more traditional expendable (e.g., sand) mold or by die casting with a permanent mold. This section focuses on the general types of discontinuities or imperfections for traditional casting with sand molds. Permanent-mold methods are discussed in the next section. The important types of casting discontinuities discussed in this section are:

- Porosity
- Inclusions
- Oxide films
- Second phases
- Hot tears
- Metal penetration
- Surface defects

Each is discussed briefly in terms of cause, prevention, and effects on properties. In specifying acceptable levels of discontinuities, such as microporosity and inclusion sizes and distribution, the designer should determine the critical flaw size that deleteriously affects performance in service. This permits the foundry to design a casting practice that eliminates such discontinuities at minimum cost.

Porosity

Porosity is a common defect (imperfection) in castings and takes many forms. Pores may be connected to the surface, where they can be detected by dye-penetrant techniques, or they may be wholly internal, where they require radiographic techniques to discover. Macroporosity refers to pores that are large enough to see with the unaided eye on radiographic inspection, while microporosity refers to pores that are not visible without magnification. Microporosity is found not only in castings but also in heavy section forgings that have not been worked sufficiently to close it up.

Both macroporosity and microporosity are caused by the combined action of metal shrinkage and gas evolution during solidification. It has been shown (Ref 2, 3) that nucleation of pores is difficult in the absence of some sort of substrate, such as a nonmetallic inclusion, a grain refiner, or a second-phase particle. This is why numerous investigations have shown that clean castings, those castings that are free from inclusions, have fewer pores than castings that contain inclusions. Porosity and casting costs also are minimized in casting designs that emphasize progressive solidification toward a gate or riser, tapered walls, and the avoidance of hot spots.

Static properties are mostly unaffected by microporosity. Microporosity is found between dendrites and, similar to macroporosity, is caused by the inability of feed metal to reach the interdendritic areas of the casting where shrinkage is occurring and where gas is being evolved. However, because this type of porosity occurs late in solidification, particularly in long-range freezing (mushy-freezing) alloys, it is particularly difficult to eliminate. The most effective method is to increase the thermal gradient (often accomplished by increasing the solidification rate), which decreases the length of the mushy zone. This technique can be limited by alloy and mold thermal properties and by casting geometry, that is, the design of the casting.

The fatigue strength of cast alloys is only slightly reduced by the presence of shrinkage or gas porosity but is much reduced when the porosity extends to the surface, regardless of severity, because of the notch effect. If micropores are less than 0.2 mm (0.008 in.) in length, there is little or no effect on dynamic properties; fatigue properties of castings with pores that size or smaller are in the same range as those of castings where no micropores were found (Ref 4, 5, 6). The shape of the micropore is as important as its size, with elongated

pores having a greater effect than round pores (Ref 7). Areas where microporosity is expected can be predicted by solidification modeling, similar to the prediction of macroporosity.

When the shrinkage and the gas combine to form macroporosity, properties are deleteriously affected. Static properties are reduced at least by the portion of the cross-sectional area that is taken up with the pores (because there is no metal in the pores, there is no metal to support the load there, and the section acts as though its area was reduced). Because the pores also cause a stress concentration in the remaining material (Ref 8, 9), static properties can be reduced by more than the percentage of cross-sectional area that is caused by the macroporosity. Dynamic properties are also affected. A study of aluminum alloys showed that fatigue properties in some were reduced 11% when specimens having x-ray quality equivalent to ASTM E 155 level 4 were tested, and that they were reduced 17% when specimens having quality of ASTM E 155 level 8 were tested (Ref 10).

Example 1: Cracking in a Gray-Iron Cylinder Head Caused by Microporosity. A cracked cylinder head was removed from an engine after approximately 16,000 km (10,000 miles) of service. This cylinder head was submitted for failure analysis to determine the cause of cracking.

Investigation. The cylinder head was magnetic-particle inspected by both the head and coil methods. No indications were present on the cylinder head, except those shown in Fig. 1 and 2. The head was cracked on the rocker-arm pan rail next to the No. 3 intake port. The crack extends into the water jacket on the rocker-arm side of the head. Visual inspection did not reveal any indication of material deformation or mishandling. The crack was opened for examination, and no indication of internal discontinuities could be seen during optical examination. Wall thickness at the thinnest section of failure measured 5.3 mm (0.210 in.). The specification minimum is 4.6 mm (0.180 in.).

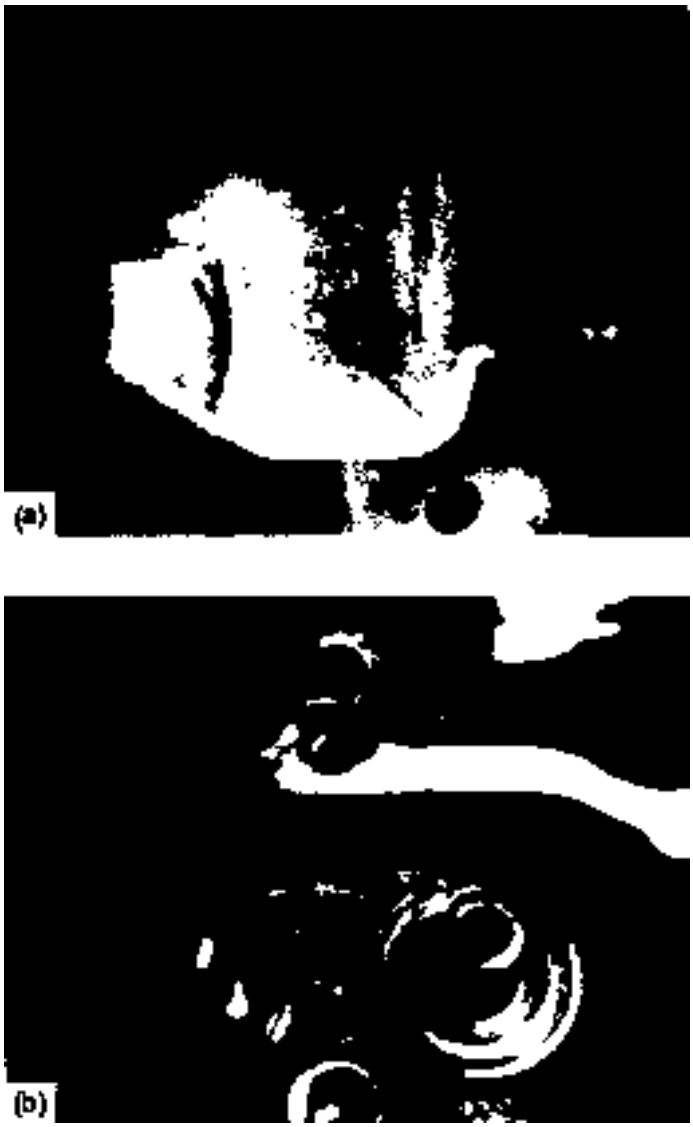


Fig. 1 Crack in a gray-iron cylinder head. (a) Crack on side of head next to manifold No. 3. (b) Another view of the same crack, which ends at the water jacket vent plug. Both 0.5×

Fig. 2 Additional views of the failed cylinder head shown in Fig. 1. (a) Fracture surface observed when the crack shown in Fig. 1 was opened. 0.5×. (b) Specimen taken from area adjacent to the vent plug showing microporosity. Etched with 2% nital. 100×. (c) Same as (b), but at a higher magnification. 400×

The hardness at the bottom deck of the cylinder head was 217 HB. The hardness at the pan rail adjacent to the crack was 217 HB. This hardness is within the specified range of 192 to 241 HB. The chemical composition of this cylinder head, as determined by spectrographic analysis, was within the specified range. Three specimens for metallographic examination were taken from the crack: one from the rocker-arm pan rail, one from the area next to the plug in the water jacket, and one from the area next to the No. 3 intake port. All sections were made perpendicular to the fracture plane.

Microporosity was found intersecting the crack in the sections taken from the water jacket next to the plug and the area next to the No. 3 intake port. Figures 2(b) and (c) show the microstructure of the specimen taken from near the plug. A wave of microporosity travels midway between the inner and outer surfaces of the casting. It varies in width from 3.8 mm (0.150 in.) at the fracture to 7.6 mm (0.30 in.) at the specimen section edge on the sample taken next to the plug. There was no indication of microporosity in the specimen taken from the rocker-arm rail.

The microstructures consist of predominantly types A and B graphite, flake size 4 to 6, in a matrix of medium-to-fine lamellar pearlite. The specimen taken near the No. 3 intake port has a higher-than-usual amount of free ferrite, approximately 5 to 10%.

Conclusion. The cracking is attributed to microporosity in the casting material. The high operational stresses of the engine acting on a material discontinuity, such as microporosity in a web member, can produce cracking.

Shrinkage porosity can result from use of an insufficient number of pouring gates, inadequate gate cross section, excessive pouring temperature, or inadequate feeding. Unless the casting is of simple design, the use of only one pouring gate often results in shrinkage near the gate. Shrinkage may be caused by too much metal passing over one area in the mold, which results in localized overheating of the mold wall and a slowdown in the solidification rate of the metal in that area of the casting. This type of shrinkage can usually be eliminated by use of two or more gates. Use of only one gate may be satisfactory when a lower pouring temperature can be used. A lower pouring temperature results in an improvement in the surface appearance of the casting.

Shrinkage is often aggravated by high pouring temperatures. If the casting does not run with a moderate pouring temperature, it may be because of back pressure due to inadequate venting, use of low-permeability molding material, or inadequate gating. Correction of one or more of these factors allows the pouring temperature to be lowered and alleviates the shrinkage difficulty.

In some instances, surface shrinkage cavities may be caused by inadequate feeding at the junction of a thin section and a thick section, where risering is insufficient to feed the shrinkage occurring in the thick section. The shape, location, and size of the risers govern the degree to which they feed the casting. Small errors in designing the risers often cause great differences in the results.

Gating can often affect the efficiency of a riser. A riser must contain hot metal, and the metal must remain liquid until after the casting has been adequately fed and has solidified. Where possible, it is generally good practice to gate through a riser into the casting, causing the riser to be the last portion of the mold to be filled. Risers that are located on the side of a casting opposite the gates are usually ineffective (or less effective), because they contain the coldest metal in the mold and are unable to feed the casting adequately.

Chills are an effective tool in controlling shrinkage. One function of a chill is to equalize the solidification rates of a thick section and an adjacent thin section, minimizing the likelihood of shrinkage. Chills also are useful in promoting directional solidification, so that a riser can act more effectively.

Depending on casting shape and size, variations in section thickness, and types of junctions between component members, dendritic growth during solidification may generate dendrite arms that isolate local internal regions, preventing complete feeding by the risers. Generally, the microscopic voids that may form between the dendrite arms do not significantly reduce density or adversely affect mechanical properties. However, even a small amount of shrinkage porosity formed *near the surface* can sometimes lead to failure. Consider, for example, a Y-shaped casting with a small included angle. If a small fillet is used between the arms of the casting, there is a thin, sharp edge of sand at the junction, which is heated to a high temperature by the contacting molten metal. This wedge of hot sand then acts as a thermal insulator, retarding solidification of the metal around it. As a result, a region is isolated from the risers by solid metal, and a shrinkage cavity forms just under the skin of the Y-junction. This cavity may remain undetected and become a site for crack initiation or corrosion penetration in service. Use of a larger-radius fillet could eliminate this problem.

Figure 3 illustrates shrinkage porosity at bolt-boss sites of an alloy ductile-iron cylinder head. This defect was detected by a leak test after completion of machining operations. If this porosity goes undetected, contamination of oil in the engine can result. Other examples are given subsequently.

Fig. 3 Shrinkage porosity at bolt-hole bosses in a ductile-iron cylinder head

Example 2: Tensile Fracture That Originated at Shrinkage Porosity in a Cast Low- Alloy Steel Connector. A sand-cast steel eye connector (Fig. 4a) used to link together two 54,430 kg (120,000 lb) capacity floating-bridge pontoons broke prematurely in service. The pontoons were coupled by upper and lower eye and clevis connectors that were pinned together.

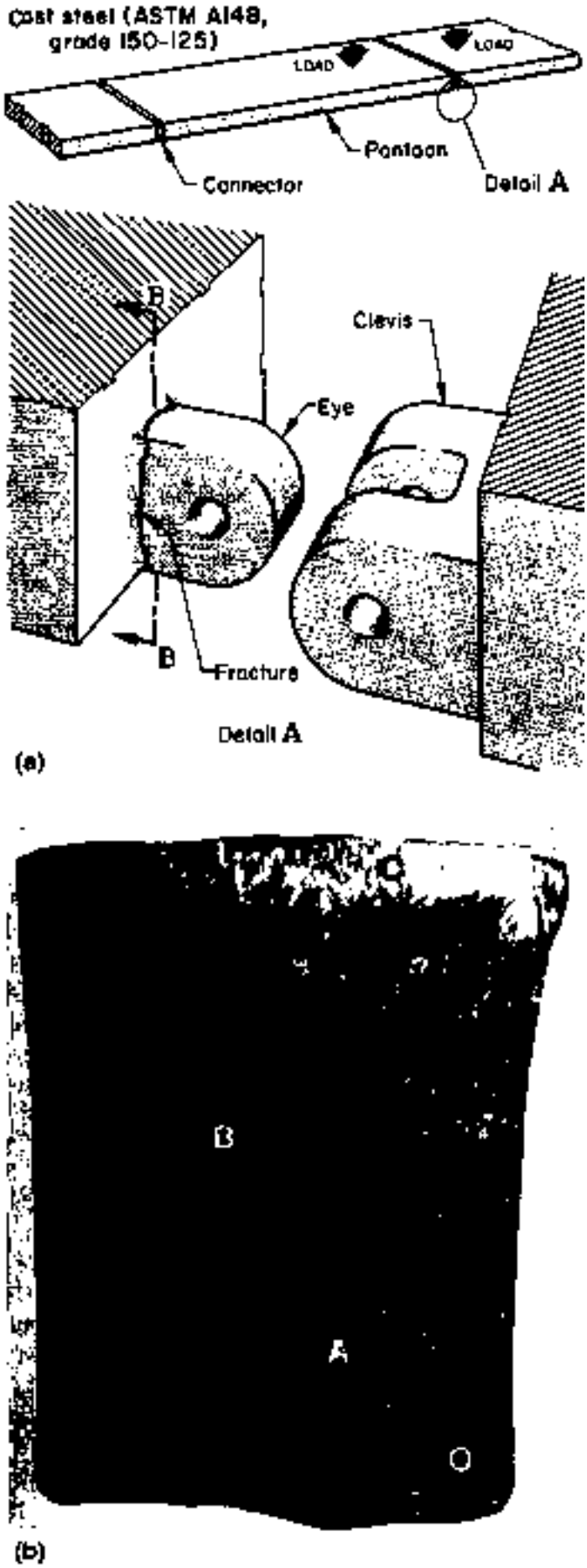


Fig. 4 Sand-cast low-alloy steel eye connector from a floating-bridge pontoon that broke under static tensile loading. (a) Schematic illustration of pontoon bridge and enlarged view of eye and clevis connectors showing location of fracture in eye connector. (b) A fracture surface of the eye connector.

Fracture origin is arrow O, area containing shrinkage porosity is at region A, area of sound metal is at region B, and the shear lip is at region C.

The eye connector was 8.9 × 12.7 cm ($3\frac{1}{2}$ × 5 in.) in cross section and was cast from low-alloy steel conforming to ASTM A 148, grade 150-125. Expected life of such connectors was 5000 or more load cycles.

Investigation. Examination of the fracture surface established that the crack had originated along the lower surface (arrow O, Fig. 4b), initially penetrating a region of shrinkage porosity (region A, Fig. 4b) that extended over the lower-right quarter of the fracture surface. Cracking then propagated in tension through sound metal (region B, Fig. 4b) and terminated in a shear lip at the top of the eye (region C, Fig. 4b). The stress state that produced the fracture approximated simple tension.

Conclusions. Fracture of the eye connector occurred by tensile overload because of shrinkage porosity in the lower surface of the casting. The design of the connector was adequate, because normal service life was more than 5000 load cycles.

Corrective Measures. Subsequent castings were inspected radiographically to ensure sound metal.

Example 3: Brittle Fracture of Cast Low- Alloy Steel Jaws Because of Shrinkage Porosity and Low Ductility of Case and Core. Eight pairs of specially designed sand-cast low-alloy steel jaws that had fractured were submitted for laboratory examination. The jaws were implemented to stretch the wire used in prestressed concrete beams. The wire was gripped at one end of the stretching machine between stationary semicylindrical jaws and at the other end between movable wedge-shaped jaws that were fitted in a block in a manner that forced them together as tension was applied to the wire. Figure 5(a) shows two pairs of movable jaws; Fig. 5(b), two pairs of stationary jaws.

Fig. 5 Wire-stretching jaws that broke because of shrinkage porosity and low ductility of case and core. The jaws, sand cast from low-alloy steel, were used to stretch wire for prestressed concrete beams. (a) Two pairs of movable jaws. 0.7×. (b) Two pairs of stationary jaws. 0.7×. (c) and (d) Fracture surfaces of two movable jaws. Approximately 2×

Investigation. It was found that all the fractures were macroscale brittle and exhibited very little evidence of deformation. Two of the jaws contained surface cracks (not distinguishable in Fig. 5a and b), but these cracks

did not appear to be related to the main fractures. Figures 5(c) and (d) show fracture surfaces of two individual movable jaws (from different pairs).

Little is known about the possible variations of stress that might have existed in this application. For example, the fit of the jaws in the block might have been imperfect, or there might have been too large an opening in the block, allowing the tips of the jaws to emerge without support. In such circumstances, the section of wire within the jaws might not have been in axial alignment with the wire under tension, which would have resulted in one jaw being subjected to a side thrust at the tip.

All surfaces of the jaws had been case carburized and heat treated to a surface hardness of 62 to 64 HRC and a core hardness of 36 to 39 HRC. Metallographic examination of a specimen prepared from a jaw established that the surface had been case carburized to a depth of 0.8 to 1.6 mm ($\frac{1}{32}$ to $\frac{1}{16}$ in.), the case structure was martensite containing small spheroidal carbides but no carbide network, the structure of the core consisted of martensite plus some ferrite, and the fracture appeared to be related to shrinkage porosity.

Chemical analysis of the core showed that the composition corresponded to 3310 steel except for low manganese, slightly high chromium, and lower silicon than would normally be expected in a cast steel.

The hardness range of the core conforms to data in end-quench hardenability curves for 3310 steel. It represents, in fact, essentially the maximum hardness values attainable and, as such, indicates that tempering was probably limited to about 150 °C (300 °F). In view of the hardenability of the core and assuming that the tapered slot in the block had a hardened surface, the low tempering temperature used may have contributed to the brittleness.

Conclusions. Failure of the jaws was primarily attributed to the presence of internal microshrinkage, although the brittleness of case and core probably contributed to failure.

Recommendations. The procedures used for casting the jaws should be revised to eliminate the internal shrinkage porosity. Tempering at a slightly higher temperature to reduce surface and core hardness would be desirable. Finally, the fit of the jaws in the block should be checked to ensure proper alignment and loading.

Example 4: Damaged Austenitic Cast Iron Impellers in a Rotary Pump Caused by Cracks Initiated at Zones of Localized Shrinkage Porosity (Ref 11). Two damaged impellers manufactured from austenitic cast iron came from a rotary pump used for pumping brine mixed with drifting sand. On impeller 1, pieces were broken out of the back wall in four places at the junction of the blades. Impeller 2 had cracks at the same location but did not yet break apart.

Investigation. Figure 6(a) shows the fractured edges following the shape of the blade. Numerous cavitation pits on the inner surface of the front wall are visible. Impeller 2, which did not exhibit such deep cavitation pits, was cracked in places along the line of the blades (Fig. 6b).

Fig. 6 Photographs of a damaged cast iron pump impeller. (a) Breaks in the back wall of impeller 1 and cavitation pits on the front wall. (b) Cracks following the shape of the blades on the back wall of impeller 2. Both 0.7×. Source: Ref 11

The microstructure shown in Fig. 7, illustrating lamellar graphite and carbides in an austenitic matrix, can be considered normal for the specified material (GGL NiCuCr 15 6 2). A specimen for metallographic examination was taken from one of the fracture edges. The material in this region was found to be in the final stages of disintegration (Fig. 8). A further section was taken through an as yet unbroken region, the essential

portion of which is shown at low magnification in Fig. 9(a). Cavitation pits can be seen on the internal surface of the front wall (arrow at top left), and casting pores (arrow at bottom right) are visible in the region corresponding to Fig. 8. It should be mentioned that even when the specimens were cut carefully with a water-cooled cutting wheel, the stem in the brittle porous zone crumbled away, and only on the third attempt was it possible to obtain a suitable whole specimen for examination. A section through a cracked region of impeller 2 revealed a porous zone in the corresponding position. Figure 9(b) shows a section of a cavitation region. The absence of corrosion products confirms that the surface has been damaged by cavitation.

Fig. 7 Micrograph of impeller specimen. This structure, which consists of lamellar carbides in an austenitic matrix, is considered normal for this material. Etched with V2A reagent. 500×. Source: Ref 11

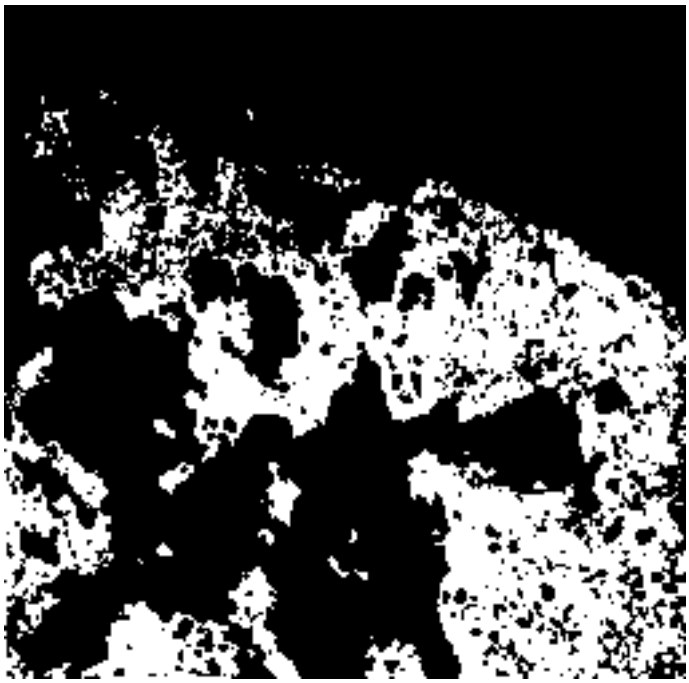


Fig. 8 Porous zone under a fracture edge of impeller 1. See also Fig. 9(a). 20×. Source: Ref 11

Fig. 9 Impeller specimen showing cavitation damage and porous zones. (a) Cavitation pits (top-left arrow) and porous zones (bottom-right arrow) in an unetched transverse section of impeller 1. 2.5×. (b) Microsection from a cavitation zone. Etched with V2A reagent. 100×. Source: Ref 11

Conclusions. The cause of impeller damage from cracking is attributed, in part, to the shrinkage porosity at the junction between the back wall and blades. Cavitation could have led to long-term damage even in a sound casting, although it did not appear to contribute to the fracture.

Recommendations. Casting practices should be implemented to avoid porosity, and an alloy austenitic cast iron (GGL NiCuCr 15 6 3) that has a higher chromium content and is more resistant to cavitation should be used.

Gas Porosity. Formation of cavities resulting from the presence of gas is influenced by many factors, including melting practice, pouring practice, and type of mold to be used. Three common causes of gas holes during solidification of castings are:

- Exceeding the solubility necessary to keep the gas in solution
- Reaction-type pinhole formed by segregation during solidification
- Reaction product between the metal and mold material

Gas solubility limits and gas porosity are some of the most serious problems in the casting of cast iron, aluminum, and copper. In determining if gas evolution can cause porosity, the thermodynamics for the reactions must be evaluated, as discussed in the article “Gases in Metals” in *Casting,* Volume 15 of *ASM Handbook*. The gases in cast iron that can cause porosity are hydrogen and nitrogen. Aluminum and copper are very susceptible to porosity from hydrogen. The most common method for removing hydrogen from aluminum and copper, which can also be used for removing hydrogen and nitrogen from cast iron, is inert gas flushing.

Hydrogen pinholes in cast iron are small, spherical, or pear-shaped cavities about 3 mm ($\frac{1}{8}$ in.) or less in diameter. A continuous graphite film is often found in these holes, although an iron oxide layer may be present if the castings are heat treated. Nonmetallic inclusions are not present in the voids, but some have been found to contain a small bead of metal. Pinholes are typically subsurface in nature and are therefore not visible until after heavy shotblasting or, more likely, after machining. Thin sections are more susceptible to this defect than thick sections, as are areas remote from the ingates.

A number of mold and metal factors have been associated with hydrogen pinhole formation. It is likely that these factors function in combination rather than individually. Aluminum levels in gray cast iron as low as 0.005% have been found to encourage dissociation of the water vapor arising from the mold, thus increasing the hydrogen content of the metal. Aluminum sources include ferroalloys used in iron treatment and scrap. Excessive moisture in the mold and lack of active carbonaceous mold additions also favor pinholing.

In the cast iron family, the susceptibility to form pinholes is related to the rate of solidification. Ductile iron, having the slowest solidification rate, is most susceptible, while white iron, having the most rapid rate, is least susceptible. By analyzing the pinhole shape and chemistry, it can be learned which gases are present. For example, a hydrogen-reaction pinhole has a graphite-free layer and a graphite film. A carbon monoxide (CO) reaction type has no graphite film and no graphite-free layer.

Pinholes resulting from insufficiently dried sand molds may lower the endurance limit of steel castings, or they may cause only poor surface appearance; however, if the number of pinholes is high, the casting may corrode in service. Internal gas cavities may reduce the effective cross section appreciably and may therefore impair load-carrying ability, depending on loading conditions but also stress concentration. In general, however, internal gas cavities have smooth surfaces that do not act as severe stress concentrators as a sharp, linear discontinuity.

Example 5: Retrieved Bone Screw Made From Co-Cr-Mo Alloy with Casting Defects. Portions of the threads of the screw shown in Fig. 10(a) had broken off, and other threads had cracked. The screw was made from a cast Co-Cr-Mo alloy. A longitudinal section through the screw revealed gas porosity, segregation of primary inclusions, and oxides (Fig. 10b, c). The screw threads broke because of mechanical weakness from gas holes, brittleness, and dissolution of oxides. With today's advanced casting techniques and testing methods, such implant defects can be avoided and eliminated.

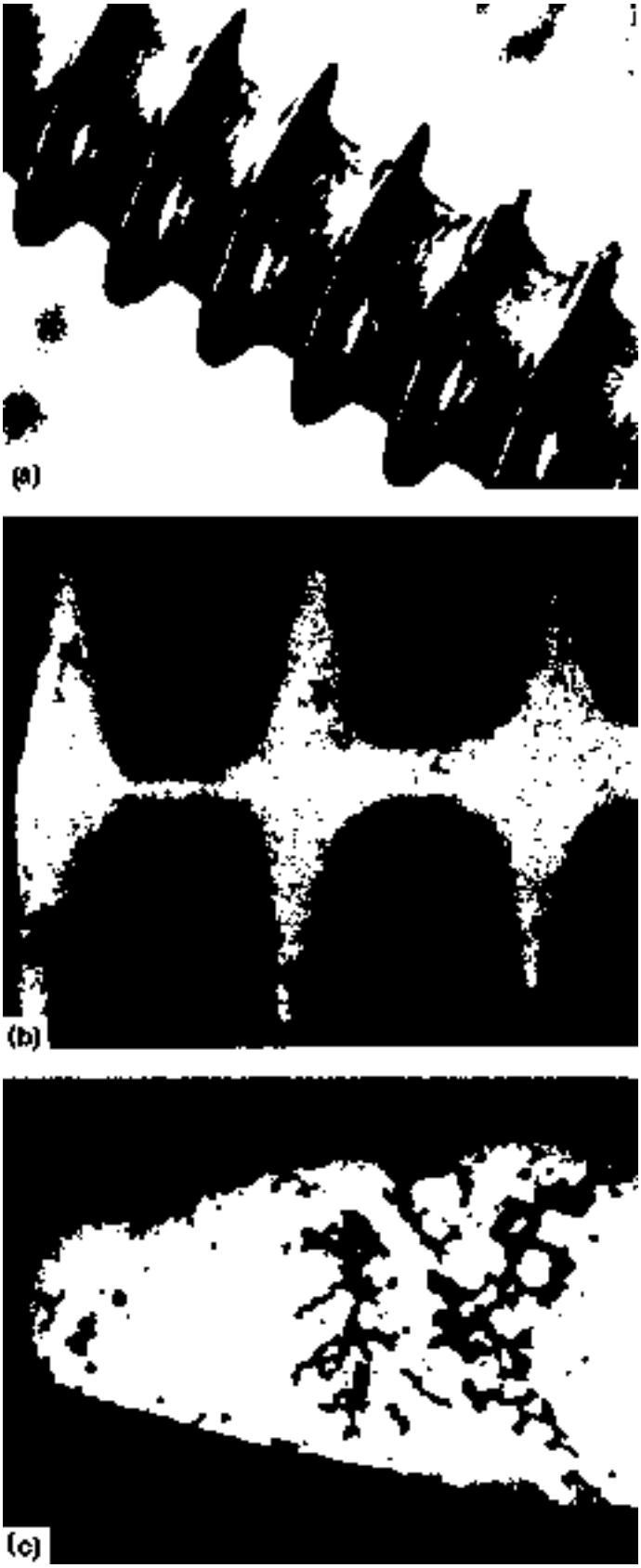


Fig. 10 Retrieved screw of cast Co-Cr-Mo alloy (type ASTM F75). (a) Defective screw threads from casting deficiencies. (b) Longitudinal section through threads showing porosity. 15×. (c) Enlarged thread of section shown in (b) with gas holes, segregation of primary phases, and dissolved oxides. 155×

Steam bubbles result from exposure of any liquid metal to moisture. Moisture in the air, in gases of combustion, in furnace and ladle linings, and in mold materials provides sources of steam bubbles. Whatever the source, the entrained water is released as steam bubbles on freezing of the molten metal, resulting in porosity in the castings that varies with the manner in which the water was entrained. If the water content in a molding sand is

too high or if water exists as condensate on the mold surface, porosity forms at the surface of the casting. If a slight amount of moisture is entrained in the melting furnace, it may be dissipated during cooling in the ladle, and the castings may be free of porosity. If a ladle lining is inadequately dried, the entire heat can be ruined by gross gas evolution during freezing in the molds, which causes all the risers to puff and overrun, creating large cavities throughout the casting.

Air bubbles as a cause of casting porosity are less detrimental than steam bubbles. Air bubbles usually arise when the conditions of pouring and the nature of the ingate passages combine to produce a turbulent flow that sucks air into the mold cavity. The amount of air in the cavity is sometimes low enough for all of it to be carried to the topmost riser, causing no porosity in the casting proper. In most instances, smoother pouring and a more streamlined ingate system will minimize air bubbles.

Inclusions

Inclusions are nonmetallic and sometimes intermetallic phases embedded in a metallic matrix. They are usually simple oxides, sulfides, nitrides, or their complexes in ferrous alloys and can include intermetallic phases in nonferrous alloys. There are essentially two classifications for inclusions:

- *Exogenous:* Those derived from external causes
- *Indigenous:* Those that are native, innate, or inherent in the molten metal treatment process

Examples of exogenous inclusions include slag, dross, entrapped mold materials, and refractories. In most cases, these inclusions would be macroscopic or visible to the naked eye at the casting surface (Fig. 11). When the casting is sectioned, they may also appear beneath the external casting surface, if they have had insufficient time to float out or settle due to density differences with respect to the molten metal. Exogenous inclusions are mostly oxides or mixtures of oxides and are primarily slag or dross particles, which are the oxides that result when the metal reacts with oxygen in the air during melting. These are removed from the melt by filtration before pouring.

Fig. 11 Ductile-iron crankshaft segment essentially free of exogenous inclusions (1, left) and with numerous exogenous inclusions (2, right). Low pouring temperature and poor mold filling practice were the cause of the inclusions in part 2.

Inclusions of refractory, slag, sand, oxides, and other compounds in the casting metal can affect the service life of the casting. Some of these, such as refractory and sand, are inadvertently introduced and can be minimized by proper maintenance of linings and by good housekeeping. The presence of slag can be controlled by careful pouring and the use of more effective skimming techniques. Oxides of the casting metal are formed largely during melting and can be avoided by thorough deoxidation. Some elements, for example, chromium in some steels, oxidize preferentially on the surface of the pouring stream. Skillful skimming is the only means of preventing this oxidized metal from entering the mold.

Indigenous inclusions include sulfides, nitrides, and oxides derived from the chemical reaction of the molten metal with the local environment. Such inclusions are usually small and require microscopic magnification for identification. They can be uniformly distributed on a macroscopic scale, but they are often in the last regions to solidity and therefore are typically interdendritic in the microstructure. Grain-boundary inclusions can be particularly damaging to mechanical properties. For example, grain-boundary sulfide films (type II sulfide inclusions) are detrimental to the ductility and toughness of cast steel and cause increased susceptibility to hot tearing. Other compounds, such as phosphides and sulfides, are detrimental to mechanical properties unless kept to a minimum. Globular manganese sulfides are beneficial as replacements of iron sulfides, which can form deleterious films at grain-boundary junctions and can result in incipient melting if the casting is heat treated.

There are cases in which indigenous inclusions such as oxides may react with mold materials to promote an exogenous inclusions variety that otherwise might not appear in the absence of the indigenous oxide. To circumvent the strict definition as indigenous or exogenous, a number of investigators prefer to catalog all inclusions as either macroinclusions of microinclusions. Although such definitions describe the anticipated effect on mechanical properties, there can be a loss in source identification of the inclusion.

Whatever definition is used, control of inclusions requires an understanding of their origin and the associated physical chemistry. For example, nonmetallic inclusions that form during solidification of cast steel depend on the oxygen and sulfur contents of the casting. Deoxidation decreases the amount of nonmetallic inclusions by eliminating the oxides, and it affects the shape of the sulfide inclusions. High levels of oxygen (>0.012% in the metal) form FeO, which decreases the solubility of manganese sulfides and causes the manganese sulfides to freeze early in solidification as globules. The use of silicon deoxidation alone normally causes the formation of globular (type I) sulfides.

Decreasing the oxygen to between 0.008 and 0.012% in the metal, through the use of aluminum, titanium, or zirconium, increases the solubility of the manganese sulfides so that they solidify last as grain-boundary films. These grain-boundary sulfide films are classified as type II inclusions. If the oxygen content is very low (<0.008% in the metal), then the deoxidizer forms complex sulfides that are crystalline and form early in solidification. These type III inclusions are less harmful than type II but are more harmful than type I.

Most inclusions found in steel castings arise from the oxidation of metal during the pouring operation (Ref 12). This is known as reoxidation and takes place when the turbulent flow of the metal in the gating system causes the metal to break up into small droplets, which then react with the oxygen in the air in the gating system or casting cavity to form oxides. Metal casters use computer analysis of gating systems to indicate when reoxidation can be expected in a gating system and to eliminate them. However, casting designs that require molten metal to “jet” through a section of the casting to fill other sections recreates these inclusions and should be avoided.

Although it is preferable to minimize undesirable inclusions in cast structures, for most applications there is an added expense in preparing castings that are essentially inclusion-free. Methods of reducing inclusions to very low levels include the use of special ceramic filters. Small inclusions that are located well within the center of the cross section of the casting have little effect, whereas larger inclusions and those located near the surface of the casting can be particularly detrimental to properties. Inclusions can also be a problem when machining surfaces, causing excessive tool wear and tool breakage. In addition, because inclusions and the surrounding metal have different coefficients of thermal expansion, thermally induced stresses may appear in the metal surrounding the inclusion during solidification (Ref 13). As a result, an inclusion might act as a stress raiser and

reduce properties. As in the case of microporosity, the size of the inclusion and its location determine its effect (Ref 14).

Figure 12 is an optical micrograph of a subsurface region with inclusions that played a major role in thread defects that occurred in ductile-iron tire-mold castings. In this case, the defect was found to be caused by stringers of magnesium-silicate inclusions, which are known to be hard and detrimental to tool life. This resulted in poor machinability and improper finish in a localized area of the tire-mold castings. Figure 12 illustrates optical micrographs of subsurface areas immediately beneath the defective casting surface. These micrographs show the presence of inclusion stringers adjacent to the surface (Fig. 12a) and the deterioration of the graphite nodule shape in the metal below (Fig. 12b, c). The long stringers of inclusions are very similar to film-type magnesium-silicate inclusions.

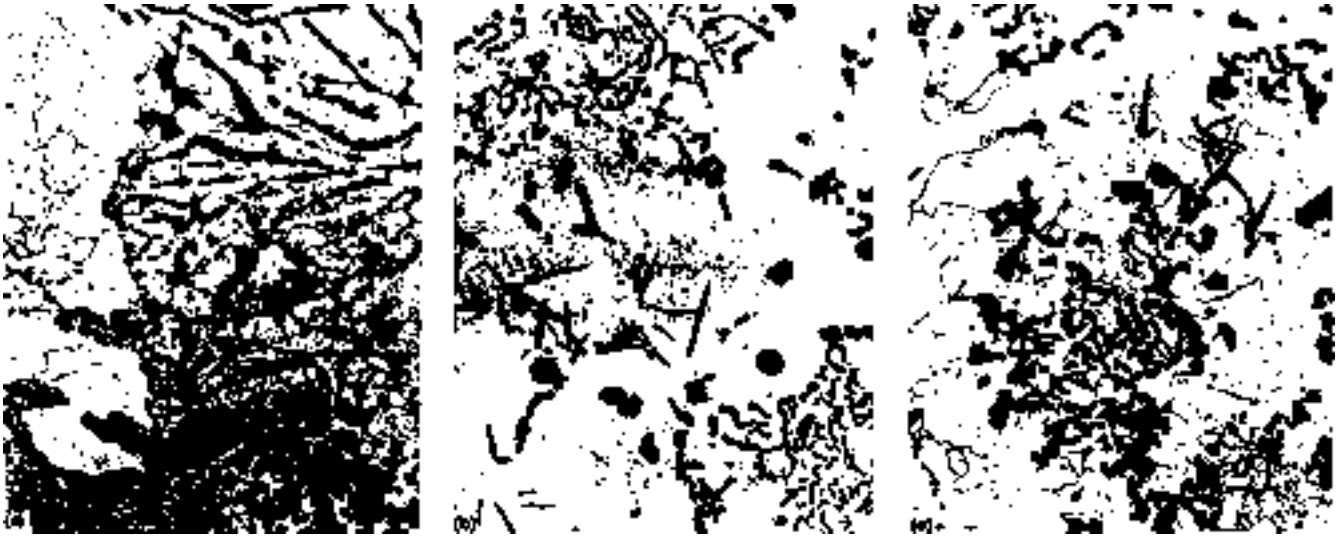


Fig. 12 Internal discontinuities in a tire-mold casting. (a) Stringer-type inclusions adjacent to the surface. As-polished. (b) Structure below the surface. Note the change in graphite shape. Unetched. (c) Ferrite matrix with degenerate vermicular graphite nodules. Etched with 2% nital. All 135×

Oxide Films

Oxide films are similar to inclusions and are related to the pouring characteristics (Ref 15, 16, 17). These form on the surface of the molten metal as it fills the mold. If this surface film is trapped within the casting instead of being carried into a riser, it is a linear discontinuity and an obvious site for crack initiation. It has been shown (Ref 18, 19) that elimination of oxide films, in addition to substantially improving static properties, results in a five-fold improvement of fatigue life in axial tension-tension tests.

Oxide films are of particular concern in nonferrous castings, although they also must be controlled in steel and stainless steel castings. (Because of the high carbon content of cast iron, oxide films do not form on that metal.) If the film folds over on itself as a result of turbulent flow or “waterfalling” (when molten metal falls to a lower level in the casting during mold filling), the effects are particularly damaging. Casting design influences how the metal fills the mold, and features of the design that require the metal to fall from one level to another while the mold is filling should be avoided so that waterfalls are eliminated. Oxide films are avoided by filling the casting from the bottom, in a controlled manner, by pumping the metal into the mold using pneumatic or electromagnetic pumps. If the casting is poorly designed, waterfalling results. An example is given in Fig. 13.

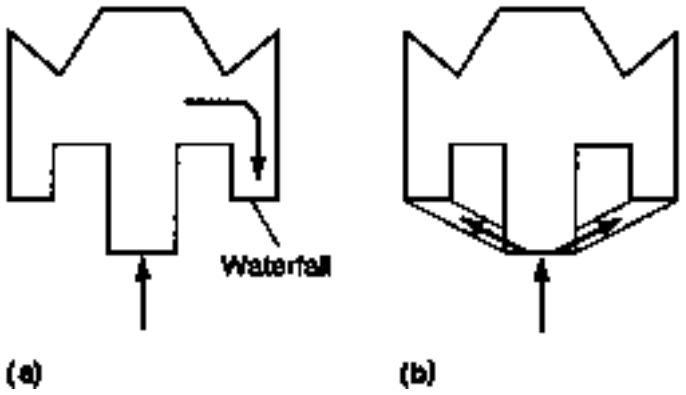


Fig. 13 Redesign of a casting to avoid waterfalling. (a) In this design, waterfalling results when casting is filled from the bottom. (b) Improved design provides a path for the metal to follow as it fills the mold.

Segregation

All metallic materials contain solute elements or impurities that can segregate during solidification. The variable distribution of chemical composition on the microscopic level in a microstructure, such as dendrites and grains, is referred to as microsegregation. Variation on the macroscopic level is called macrosegregation.

Microsegregation characterizes concentrations of elements in interdendritic regions that range in size from a few to several hundred microns. Microsegregation is present in the as-cast casting and may be removed by various degrees, depending on subsequent heat treatment. By contrast, macrosegregation is the gradient difference, measurable on a macroscale, in alloying elements from the surface to the center of an ingot or casting. Macrosegregation becomes more pronounced with increasing section size. Macrosegregation cannot be reduced by homogenization heat treatment.

Segregation generally deteriorates the physical and chemical properties of materials and should be kept to a minimum, as discussed further in the article "Microsegregation and Macrosegregation" in *Casting*, Volume 15 of *ASM Handbook*. An example of segregation is also given in Example 14 in this article. Macroetching is the most common method used to reveal the dendritic structure of an ingot or casting.

Second Phases

Second phases that form during solidification may also nucleate cracks, if they have the proper size and morphology. An example is aluminum-silicon alloys, where the silicon eutectic is present as large platelets, which nucleate cracks and provide a path for crack propagation (Ref 20, 21). The size of these platelets may be significantly reduced by modifying the alloy with additions of sodium or strontium. These modifications also change the platelets to a spheroid shape. However, such additions increase the size of micropores (Ref 22), and for this reason, many foundrymen rely on accelerated solidification of the casting to refine the silicon, because higher solidification rates refine the structure. Heavy sections are to be avoided if a fine structure is desired. Generally speaking, however, secondary phases in the structure of castings become important in limiting mechanical behavior of castings only in the absence of nonmetallic inclusions and microporosity (Ref 23, 24).

Hot Tears and Hot Cracking

Tearing or hot cracking can occur during solidification, when the greatest amount of shrinkage occurs and when the casting is least resistant to stresses imposed by the geometrical constraints of the mold. This type of cracking, sometimes termed hot shortness, is always intergranular and is a feature, to some extent, of all casting alloys.

As previously noted in the section "Imperfections That Can Occur in Iron and Steel Castings," a hot crack is defined as "a crack or fracture caused by internal stresses that develop after solidification and during cooling from an elevated temperature (above 650 °C, or 1200 °F, for gray iron)." A hot tear is defined as "a crack or fracture formed prior to completion of solidification because of hindered contraction." Both of these definitions apply equally well to a casting or a weldment.

A hot crack is less visible (less open) than a hot tear and usually exhibits less evidence of oxidation and decarburization. Because hot cracking is related to grain-boundary liquation, hot cracks are generally intergranular or interdendritic (similar to the appearance of the impeller fracture faces in Example 4 of this article). A hot tear frequently is open to the surface of the casting and thus exposed to the atmosphere. This can result in oxidation, decarburization, or other metal-atmosphere reactions at the tear surface. Hot tears form when casting sections are constrained by the mold from shrinking as they cool near the end of solidification. They are sometimes produced in external surfaces, such as in the fillet of a flange or a rib, where overall contraction of the casting is hindered by the mold.

Hot cracking problems can be minimized by proper selection of casting process, alloy selection, part and mold design, solidification control, and grain refinement. Hot strength (resistance to cracking at solidification temperature) is alloy-dependent. Casting alloys demonstrate varying degrees of resistance to cracking; therefore, the cracking sensitivity of alloys capable of meeting part requirements should be an important criterion in final alloy selection. Effective grain refinement to ensure a fine, equiaxed grain structure is especially important in hot-short alloys or in casting designs with features that aggravate cracking tendencies. Fine grain size reduces local stress concentrations by minimizing grain-boundary effects.

Improper mold design is a frequent cause of hot cracking. It is important to avoid abrupt changes in cross-sectional area and inadequately radiused or filleted angles or corners. These are design features that can sharply increase and concentrate stresses resulting from restrained contraction during and following solidification. The intelligent use of chills and/or antichills can also be beneficial in preventing hot cracks. An extensive discussion on how to prevent hot tears through casting design is provided in Ref 25.

Hot tears and hot cracks should normally be discovered during inspection. Hot tears are fairly large and are most often weld repaired. If not repaired, their effect is not readily predictable (Ref 26). While, generally, they are detrimental to casting properties, under some circumstances they do not affect them. Hot tears are caused by a combination of factors, including alloy type, metal cleanliness, and mold and core hardness. However, poor casting design is the primary cause. Castings should be designed so that solidifying sections are not subjected to tensile forces caused by shrinkage during solidification, because the solidifying alloy has little strength before it solidifies.

Example 6: Fatigue Fracture of a Sand-Cast Steel Axle Housing That Originated at a Hot Tear. A sand-cast medium-carbon steel (Fe-0.25C-1Mn-0.40Cr-0.35Mo) heavy-duty axle housing, which had been quenched and tempered to about 30 HRC, fractured after almost 5000 h of service.

Investigation. Figure 14 shows the fracture surface of the axle housing. Study of the casting revealed that the fracture had been initiated by a hot tear (region A, Fig. 14) about 10 cm (4 in.) long that had completely penetrated the 2.5 cm (1 in.) thick wall; this tear formed during solidification of the casting. The mass of a feeder-riser system located near the tear retarded cooling in this region, creating a hot spot.

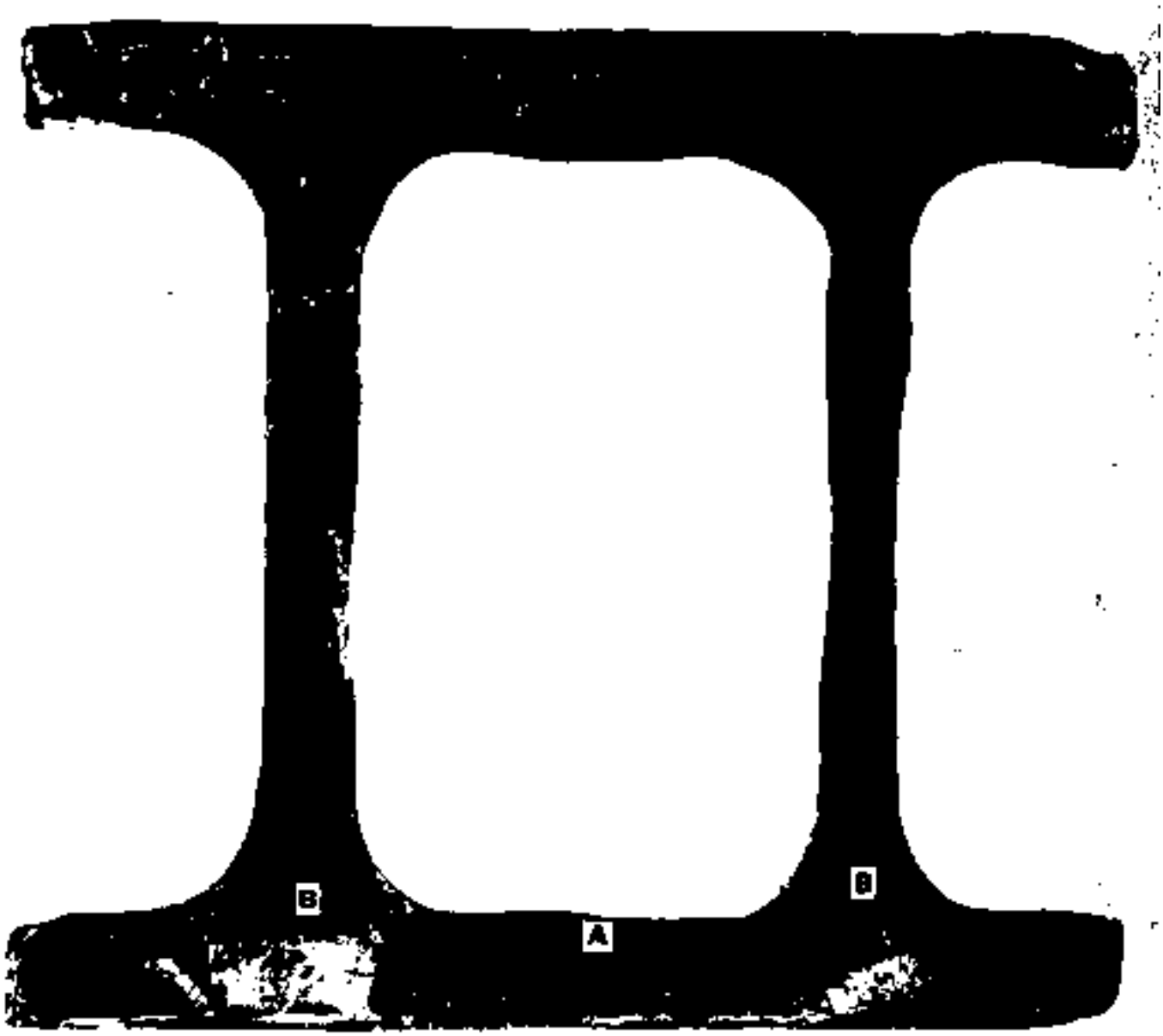


Fig. 14 Fracture surface of a sand-cast medium-carbon steel heavy-duty axle housing. Failure originated at a hot tear (region A), which propagated in fatigue (region B) until final fracture occurred by overload. 0.4×

Because this tear formed at a high temperature and was open to the surface, it was subject to oxidation, as indicated by the dark scale in region A in Fig. 14. Exposure to air also resulted in decarburization below the oxide scale and penetration of oxide along secondary cracks.

Both the size of the hot tear and the long service life of the housing indicated that the service stresses in the region of the tear had been relatively low. Stress concentrations at the tip of the tear, however, were sufficient to initiate the formation of two fatigue cracks (note the beach marks visible in regions B in Fig. 14). Ultimately, the effective thickness of the housing wall was reduced sufficiently for the applied load to cause final fracture. In this instance, the hot tear was located at an area of low stress, and the casting served a satisfactory life. However, had the local stress been high, even a very small discontinuity could have generated an early fracture.

Conclusions. Fracture of the axle housing originated at a hot tear that formed during solidification by hindered contraction and was enlarged in service by fatigue.

Corrective Measures. Changing the feeder location eliminated the hot spot and thus the occurrence of hot tearing.

Metal Penetration and Mold-Wall Deficiencies

Molten metal may penetrate the surface of the mold, forming a rough surface or, in extreme cases, actually becoming intimately mixed with the sand in the mold. Mold-wall deficiencies are common problems in gray-iron castings. They result in surface flaws such as scabs, rattails, cuts, washes, buckles, drops, and excessive metal penetration into space between sand grains. These flaws are generally detected by visual inspection.

In iron castings, metal penetration is normally the result of the combination of metallostatic head (the pressure exerted on the molten iron at the bottom of the mold by the weight of the metal on top of it) and the surface tension relationships between the liquid iron and molding materials (Ref 27). In cast iron, it is frequently also the result of the expansion of graphite at the end of solidification, forcing liquid metal into the mold if the casting is not properly designed with a tapered wall to promote directional solidification and to avoid hot spots.

Surface Condition

Several types of surface discontinuities can occur in castings, such as misruns, cold shots, scabs, buckles, rattails, pulldowns, pits, sand holes, and mechanical (cold) cracks. Hot tears and hot cracks also are sometimes produced on the external surface. Some of these imperfections contribute only to the poor appearance of a casting. A surface discontinuity is unlikely to decrease the service life of the casting, unless it is of critical size or is located at a portion of the casting surface that is subjected to major stress. The most damaging types of surface discontinuities are those that are linear in nature, those that generate notch-type stress raisers, and combinations of these. Among these are hot tears, cold cracks, and cold shuts (also surface porosity, inclusions, or shrinkage porosity extending to the surface).

Spikes are a form of surface shrinkage not normally visible to the naked eye, but they appear as a multitude of short, discontinuous surface cracks when subjected to fluorescent magnetic-particle inspection. Unlike true fractures, spikes do not propagate, but they are not acceptable where cyclic loading could result in fatigue failure. Spikes are usually seen as short indications approximately 1.6 mm ($\frac{1}{16}$ in.) long or less and never more than 75 μm (0.003 in.) deep. These defects do not have a preferred orientation but a random pattern that may or may not follow the direction of solidification.

Misruns, cold shots, scabs, buckles, pulldowns, pits, and sand holes are among those imperfections that contribute only to poor casting appearance and that rarely initiate failure. Misruns usually do not result in rejection of castings, unless good surface appearance is of primary importance. If a misrun is not severe, it should offer little difficulty in service, provided it is not in a highly stressed region of the casting.

Cold shots, which are the result of turbulence, occur when metal splatters during pouring, adheres to the mold wall, and fails to fuse with the metal that engulfs it as the mold is filled. Thus, cold shots are globules embedded in the casting and normally produce no ill effect except for appearance. Under certain conditions, however, the interface between the cold shot and the casting can serve as a corrosion cell that leads to pitting and the initiation of fatigue cracks.

Scabs, buckles, rattails, and pulldowns are surface discontinuities that result from local expansion of a sand mold. A scab is a thin, flat layer of metal formed behind a thin layer of sand and held in place by a thin ligament of metal. When the scab is removed (readily done by chipping), an indentation called a buckle is exposed. Rattails are minor buckles in the form of small irregular lines on the casting surface. A pulldown is a buckle in the cope portion of the casting. All these discontinuities affect casting appearance but usually have no influence on service life.

Pits and sand holes are dirt or sand inclusions in the surface of a casting and are generally removed in cleaning. The resulting cavities affect appearance but generally do not affect performance in service, unless they are larger and are in a critical area.

Mechanical (cold) cracks can be generated by rough treatment or by thermal shock in cooling, either at shakeout or during heat treatment. These cracks can be very difficult to detect without the aid of nondestructive testing. Magnetic-particle or liquid-penetrant inspection should reveal them clearly. Depending on the value and weldability of the casting, mechanical cracks should be removed and repair welded, unless they are large enough to warrant scrapping of the casting. Removal by grinding, chipping, or other methods must be complete, or the remaining portion of the crack will be a stress raiser beneath the weld. Magnetic-particle or liquid-penetrant inspection after grinding is effective in determining if a crack has been completely removed.

Surface Finish and Dimensional Tolerances. Ultimately, surface finish and dimensional tolerances are the important specification. Obviously, smoother surfaces are superior to rough surfaces in terms of dimensional

accuracy and surface integrity. Designers should be certain that fatigue data used in design calculations have been taken from as-cast surfaces rather than machined surfaces, because most surfaces on castings where stress concentrations might be expected are not machined. The influence of surface discontinuities on properties is reviewed in Ref 28, and 29.

The surface finish of casting is controlled ultimately by the molding. If the dimensional accuracy and surface finish requirements of the casting exceed the limits of the green sand molding process, other processes, such as shell molding, chemical bond molding, and permanent molding, can be used. Mold coatings are frequently applied to cores to enhance the casting surface finish and to reduce casting defects at the mold/metal interface. Ceramic molding is intended for the production of castings of high quality, not only in terms of their dimensional accuracy and surface finish but also in terms of soundness and freedom from nonmetallic inclusions.

Mold coatings are frequently applied to cores to enhance the casting surface finish and to reduce casting defects at the mold/metal interface. Coatings accomplish this by having a higher refractory value than the sand and/or by forming an impermeable barrier between the metal and the core. Most coatings are formulated with five major components:

- Refractory materials
- Carrier system
- Suspension system
- Binder system
- Chemical modifiers

Cold Shuts

Cold shuts, or laps, are seams that form when two streams of molten metal meet but do not completely merge. Cold shuts can occur in all types of castings and can be serious sources of stress concentration. In general, cold shuts are caused when molten metal is poured at too low a temperature or run over too long a distance. Isolated heavy sections completely surrounded by thin areas should be avoided. Thin, deep ribs also are likely to cause cold shuts and misruns.

Example 7: Fracture of a Cast Stainless Steel Lever Because of a Cold Shut. The lever shown in Fig. 15 was a component of the main fuel-control linkage of an aircraft engine. After a service life of less than 50 h, the lever fractured in flight. The lever had been machined from a casting made of AISI type 410 stainless steel, then surface hardened by nitriding.

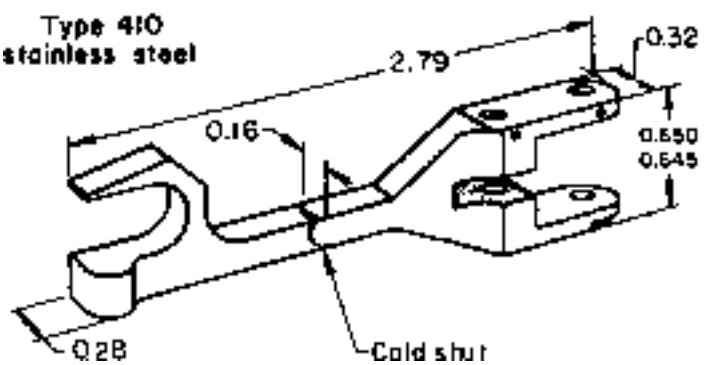


Fig. 15 Cast type 410 stainless steel fuel-control lever that fractured at a cold shut. Dimensions in inches

Investigation. Examination revealed that the lever had broken at a cold shut (Fig. 15) that extended through about 95% of the cross section of the lever arm. The remaining 5% of the cross section appeared to be sound metal.

Specifications for both the casting and the finished lever required radiographic inspection. This inspection method reveals internal voids, such as gas porosity or shrinkage porosity, but does not detect a crack or a cold shut, unless the plane of the imperfection is nearly parallel to the x-ray beam and unless there is a finite opening; a cold shut that has its faces in contact is rarely revealed by radiography.

Conclusions. The lever broke at a cold shut extending through approximately 95% of the cross section. The normally applied load constituted an overload of the remainder of the lever.

Corrective Measures. Magnetic-particle inspection was added to the inspection procedures for this cast lever. Because cold shuts extend to the surface of a casting, magnetic-particle inspection is a relatively reliable method of detecting them in ferromagnetic materials.

Example 8: Gray-Iron Paper-Drier Head Removed from Service. A paper-drier head manufactured from gray cast iron was removed from service as a result of nondestructive evaluation (NDE) detection of cracklike surface discontinuities. These imperfections were not removed by grinding. This component is subjected to internal steam pressure to provide heat energy for drying. Because the vessel is under pressure, a crack could present a safety hazard; therefore, it was removed from service for evaluation. The material specification of this cast iron was not known.

Investigation. Figure 16 shows close-up views of the surface casting defects. Sections of this casting at the area designated as 12 o'clock were removed and subjected to magnetic-particle inspection, which provided cracklike indications. These sections were further studied by chemical analysis, mechanical testing, and metallography.

(a)

(b)

(c)

Fig. 16 Close-up views of surface casting defects on a paper-drier head. (a) At the 12 o'clock position. (b) At the 9 o'clock position, with arrow indicating a surface defect. (c) At the 6 o'clock position. All approximately 0.2×

The chemical composition of a section near the 12 o'clock position was evaluated and was satisfactory for gray cast iron with a carbon equivalent of 4.56. However, the phosphorus content is significantly higher than the normal 0.02 to 0.07% range common for gray iron. The high phosphorus content can cause the formation of steadite, a ternary eutectic consisting of ferrite (containing some phosphorus in solution) and iron phosphide (Fe_3P). This undesirable, hard, brittle constituent degrades the strength of the material.

Tensile specimens prepared from this same region have a tensile strength of 101 MPa (14.7 ksi); the material had a hardness of 107 HB. These properties indicate a strength level lower than a low-strength gray iron (ASTM A 48, grade 20). This specification has a minimum expected tensile strength of 138 MPa (20 ksi). Another specification, ASTM A 159, grade G1800, has an expected hardness value of 187 HB, maximum. Both of these cast iron grades are expected to have tensile strength values exceeding 124 MPa (18 ksi). The low strength measured indicates that the casting might have substandard properties. However, the specifications for the material were not known and thus cannot be verified. It is also not correct to remove test sections from the manufactured component to determine adherence to a specification (i.e., grade/strength level). Adherence to specification is determined by testing separately poured test bars. Properties of the finished component depend on cooling rate/microstructure and therefore vary with section thickness.

Metallographic sections were made to include material at and beneath the crack. The graphite structure shown in Fig. 17(a) is essentially ASTM type A, size 4; ASTM type B graphite was also observed in the vicinity of the surface near the crack indications. The matrix microstructure was observed to be principally ferritic. The matrix also contained approximately 10% pearlite and some steadite (Fig. 17b,c). Ferritic grades of gray cast iron are not common, but they do exist, for example, ASTM A 159 G1800. The matrix explains the low strength and hardness. The presence of steadite reflects the higher phosphorus content.

Fig. 17 Microstructures of the gray-iron drier head shown in Fig. 16. (a) Typical graphite distribution in the casting—type A, size 4 graphite. As-polished. 54×. (b) The matrix microstructure consists of ferrite with approximately 10% pearlite and 15% steadite. Etched with nital. 54×. (c) Higher magnification view of (b). Etched with nital. 215×

At the crack-indication site revealed by NDE, a demarcation line was observed progressing from the surface of the casting to the interior. Figure 18 shows this demarcation line for the condition and revealed that the structure of the metal on one side of the demarcation line was different than that observed on the other side

(Fig. 19). This condition generally results when a cold shut forms during the casting process. A cold shut reflects two different solidification fronts in the mold, which explains the reason for the structure variances observed on either side of the demarcation line. Although not a crack, the demarcation can serve as a site for crack initiation. In this instance, however, the demarcation line had not propagated beyond the point at which the structure was different. The demarcation line terminates at approximately the same area where the difference in structure terminates (Fig. 19). This condition further indicates that the demarcation had not propagated beyond the original condition that occurred during the casting operation.

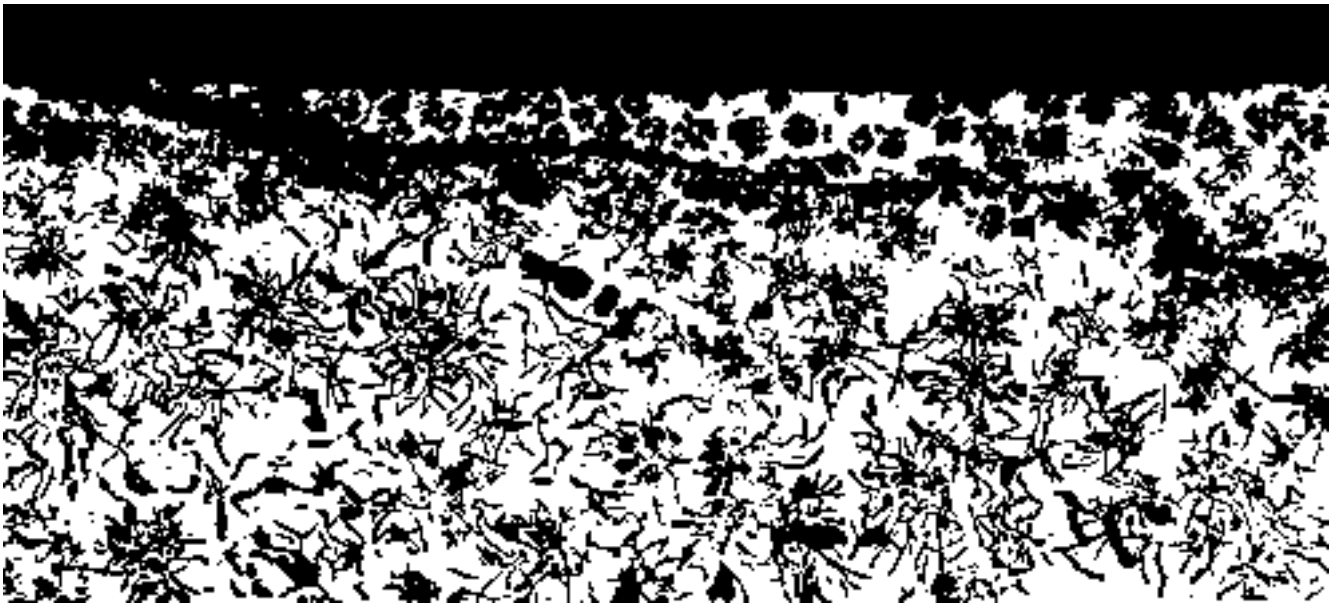


Fig. 18 Demarcation line at the crack, showing the differences in graphite type and distribution on either side of the crack. Etched with nital. 33×

Fig. 19 Higher magnification view of Fig. 18 showing the two different solidification structures. Etched with nital. 230×

Conclusions. The NDE indications were the consequence of a cold-shut condition in the casting. This conclusion is based on the presence of two significantly different microstructures representing two separate solidification fronts that met but did not merge. The cold shut serves as a stress-concentration site and is therefore a potential source of crack initiation. The combination of low material strength and a casting defect is a potential source of unexpected fracture during service, because the component is under pressure from steam.

Recommendations. Other dryer heads exhibiting similar discontinuities and/or material quality should be removed from service.

Example 9: Fracture of Cast Steel Equalizer Beams. Two sand-cast low-alloy steel equalizer beams designed to distribute the load to the axles of a highway truck broke after an unreported length of service. Normal service life would have been about 805,000 km (500,000 miles) of truck operation. The cast equalizer beams were made according to ASTM A 148, grade 105–85, which specifies that all castings be heat treated (either fully annealed, normalized, normalized and tempered, or quenched and tempered, usually at the option of the foundry).

Investigation. Figure 20(a) shows a fracture surface of one of the beams. Internal shrinkage porosity and evidence of a cold shut were found near the tip of one flange (detail A, Fig. 20). Decarburization of the fracture surface indicates that the initial crack (region B, detail A, Fig. 20) was formed before heat treatment—most likely as a mechanical crack resulting from rough handling during shakeout. The crack originated in a region of internal shrinkage porosity and entrapped oxides.

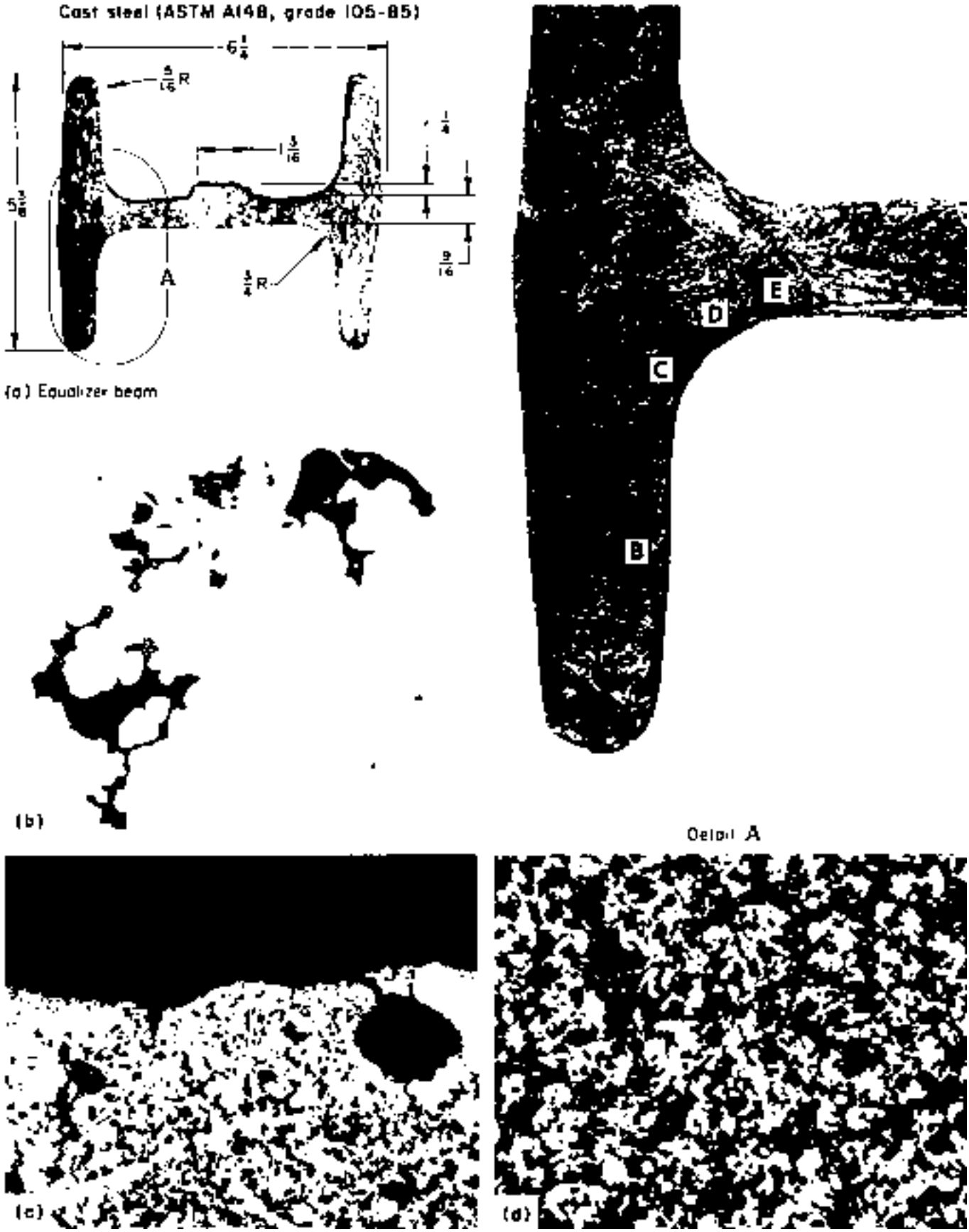


Fig. 20 Highway-truck equalizer beam, sand cast from low-alloy steel, that fractured because of mechanical cracking. (a) Fracture surface; detail A shows increments (regions B, C, D, and E) in which crack propagation occurred sequentially. Dimensions given in inches. (b) Micrograph of an unetched specimen showing internal shrinkage porosity and entrapped oxides. 65×. (c) Micrograph of a specimen etched in 1% nital showing internal tensile crack and cold shut adjacent to fracture surface; note faint

evidence of dendritic segregation at lower left. 65×. (d) Micrograph showing nonhomogenized microstructure consisting of transformation products and ferrite. 65×

The appearance of the fracture surface suggests that propagation of the crack occurred in a sequence of four large increments (regions B, C, D, and E, detail A, Fig. 20). The detectable difference in darkening of these fracture segments indicates that they existed in turn for sufficient lengths of times to undergo different cumulative amounts of surface oxidation. This behavior indicates infrequent but high loads. The fracture segment indicated as region E in detail A in Fig. 20 shows a band of fatigue beach marks that nearly penetrated the fillet between the upper portion of the left flange and the web of the beam. Final fracture occurred at this stage, because the remaining section of the beam was unable to sustain the applied load or as the result of a high load, such as those that caused earlier crack advances.

Examination of the fracture surface of the second beam showed that only two increments of crack propagation occurred before final fracture and that there was no fatigue cracking. Failure in the second beam occurred after far less crack penetration than in the companion beam, indicating that a high load, rather than a uniform load, caused final fracture.

Both beams exhibited evidence of internal shrinkage porosity and entrapped oxides; a typical condition is shown in Fig. 20(b). Areas of decarburization were found below the fracture surfaces of both beams. Internal cracking and cold shuts, such as those shown in Fig. 20(c), were evident adjacent to the fracture surfaces.

The microstructure of the beams consisted of transformation products and ferrite (Fig. 20d). The microconstituents exhibited some evidence of dendritic segregation.

Chemical analysis of the two beams, which were cast from separate heats of steel, showed close agreement in composition. A tensile specimen and a Charpy V-notch test bar were machined from one flange of each beam. The properties of the two beams were in close agreement. Average values were tensile strength, 712 MPa (103.25 ksi); yield strength (0.2% offset), 381 MPa (55.25 ksi); elongation in 5 cm (2 in.), 20%; reduction of area, 48%; and room-temperature impact strength, 32.5 J (24 ft · lbf). These properties indicated that the steel was too soft for the application—probably due to improper heat treatment.

Conclusions. Fracture of the equalizer beams resulted from growth of mechanical cracks in one flange that were formed before the castings were heat treated. The pattern of crack growth suggested a service condition involving many relatively low applied stresses and occasional high loads.

Recommendations. Three changes in processing were recommended: better gating and risering in the foundry to achieve sounder castings, better shakeout practice to avoid mechanical damage and better inspection to detect imperfections, and normalizing and tempering to achieve better mechanical properties.

References cited in this section

1. T. Piwonka, Design for Casting, *Materials, Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 723

2. E.L. Rooy, Hydrogen: The One-Third Solution, *AFS Trans.,* Vol 101, 1993, p 961

3. N. Roy, A.M. Samuel, and F.H. Samuel, Porosity Formation in Al-9 Wt Pct Mg—3 Wt Pct Cu Alloy Systems: Metallographic Observations, *Met. Mater. Trans.,* Vol 27A, Feb 1996, p 415

4. B. Skallerud, T. Iveland, and G. Härkegård, Fatigue Life Assessment of Aluminum Alloys with Casting Defects, *Eng. Fract. Mech.,* Vol 44 (No. 6), 1993, p 857

5. M.J. Couper, A.E. Neeson, and J.R. Griffiths, Casting Defects and the Fatigue Behaviour of an Aluminium Casting Alloy, *Fatigue Fract. Eng. Mater. Struct.,* Vol 13 (No. 3), 1990, p 213

6. J.C. Ting and F.V. Lawrence, Jr., Modeling the Long-Life Fatigue Behavior of a Cast Aluminum Alloy, *Fatigue Fract. Eng. Mater. Struct.,* Vol 16 (No. 6), 1993, p 631

7. J.T. Berry, Linking Solidification Conditions and Mechanical Behavior in Al Castings—A Quarter Century of Evolving Evidence, *AFS Trans.,* Vol 103, 1995, p 837

8. M.K. Surappa, E. Blank, and J.C. Jaquet, Effect of Macro-Porosity on the Strength and Ductility of Cast Al-7Si-0.3Mg Alloy, *Scr. Metall.,* Vol 20, 1986, p 1281

9. C.H. Cáceres, On the Effect of Macroporosity on the Tensile Properties of the Al-7%Si-0.4%Mg Casting Alloy, *Scr. Metall.,* 1994, submitted for publication

10. C.M. Sonsino and K. Dietrich, Einflub der Porosität auf das Schwingfestigkeitverhalten von Aluminium-Gubwerkstoffen, Giessereiforschung, Vol 43 (No. 3 and 4), 1992, p 119–140

11. P.M. Unterweiser, Ed., *Case Histories in Failure Analysis,* American Society for Metals, 1979

12. C.E. Bates and C. Wanstall, Clean Steel Castings, "Metalcasting Competitiveness Research," final report, DOE/ID/13163-1 (DE95016652), Department of Energy, Aug 1994, p 51

13. I.P. Volchok, Non-Metallic Inclusions and the Failure of Ferritic-Pearlitic Cast Steel, *Cast Met.,* Vol 6 (No. 3), 1993, p 162

14. P. Heuler, C. Berger, and J. Motz, Fatigue Behaviour of Steel Casting Containing Near-Surface Defects, *Fatigue Fract. Eng. Mater. Struct.,* Vol 16 (No. 1), 1992, p 115

15. J. Campbell, J. Runyoro, and S.M.A. Boutorabi, Critical Gate Velocities for Film-Forming Alloys, *AFS Trans.,* Vol 100, 1992, p 225

16. N.R. Green and J. Campbell, Proc. Spring 1993 Meeting 4–7 May, 1993 (Strasbourg, France), European Division Materials Research Society

17. N.R. Green and J. Campbell, The Influence of Oxide Film Filling Defects on the Strength of Al-7Si-Mg Alloy Castings, *AFS Trans.,* Vol 102, 1994, p 341

18. C. Nyahumwa, N.R. Green, and J. Campbell, "The Effect of Oxide Film Filling Defects on the Fatigue Life Distributions of Al-7Si-Mg Alloy Castings," presented at the International Symposium on Solidification Science and Processing, Dec 1995 (Honolulu, HI), Japan Institute of Metals and TMS

19. J. Campbell, The Mechanical Strength of Non-Ferrous Castings, *Proc. 61st World Foundry Cong.* (Beijing), 1995, p 104; available from the American Foundrymen's Society, Des Plaines, IL

20. F.T. Lee, J.F. Major, and F.H. Samuel, Effect of Silicon Particles on the Fatigue Crack Growth Characteristics of Al-12 Wt Pct-0.35 Wt Pct Mg-(0 to 0.02) Wt Pct Sr Casting Alloys, *Met. Mater. Trans.,* Vol 26A (No. 6), June 1995, p 1553

21. J.F. Major, F.T. Lee, and F.H. Samuel, "Fatigue Crack Growth and Fracture Behavior of Al-12 wt% Si-0.35 wt% Mg (0–0.02) wt% Sr Casting Alloys," Paper 96-027, *AFS Trans.,* Vol 104, 1996

22. D. Argo and J.E. Gruzleski, Porosity in Modified Aluminum Alloy Castings, *AFS Trans.,* Vol 96, 1988, p 65

23. T.L. Reinhart, Fatigue and Fracture Properties of Aluminum Alloy Castings, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 813

24. T.L. Reinhart, "The Influence of Microstructure on the Fatigue and Fracture Properties of Aluminum Alloy Castings," presented at Aeromat '96, June 1996 (Dayton, OH), ASM International

25. A.L. Kearney and J. Raffin, *Heat Tear Control Handbook for Aluminum Foundrymen and Casting Designers,* American Foundrymen's Society, 1987

26. M.J. Couper, A.E. Neeson, and J.R. Griffiths, Casting Defects and the Fatigue Behaviour of an Aluminium Casting Alloy, *Fatigue Fract. Eng. Mater. Struct.,* Vol 13 (No. 3), 1990, p 213

27. D.M. Stefanescu et al., "Cast Iron Penetration in Sand Molds, Part I: Physics of Penetration Defects and Penetration Model," Paper 96-206, *AFS Trans.,* Vol 104, 1996

28. R.L. Naro and J.F. Wallace, Effect of Mold-Steel Interface Reactions on Casting Surface and Properties, *AFS Trans.,* Vol 75, 1967, p 741

29. R.L. Naro and J.F. Wallace, Effect of Mold-Steel Interface Reactions on Casting Surfaces, *AFS Trans.,* Vol 100, 1992, p 797

Casting Design

Good casting design and communication with the foundry are essential in the ability to use the best techniques to produce reliable and economical castings. Good design practice also requires attention to stress concentration from surface details, depending on the materials. Many cast materials are less notch sensitive in fatigue loading than comparable wrought alloys. However, notch sensitivity of castings depends on the material. For example, very little allowance need be made for a reduction in fatigue strength caused by notches or abrupt changes of section in gray cast iron, because the graphite flakes in gray iron act as internal notches. In contrast, the notch sensitivity of ductile iron may be more important. In one case, for example, the minimum fillet in the crankpin of a ductile-iron crankshaft was set at 1.65 mm (0.065 in.) thick, in contrast to 1.14 mm (0.045 in.) for the pearlitic malleable iron that it replaced.

Figure 31 shows a fracture of an anchor link that rides on a post with a spherical surface. No bevel or radius was called for, leaving an extremely sharp edge. A fatigue crack grew from the top sharp edge (note beach marks in Fig. 31). The casting had good toughness and strength (1448 MPa, or 210 ksi, tensile strength), but this fatigue failure was the result of poor design. A generous radius was recommended to avoid similar failures.

Fig. 31 Fatigue failure of an anchor link. The arrow indicates the sharp edge that acted as the stress concentrator that led to fatigue crack growth.

Process Design and Properties (Adapted from Ref 1). In terms of the casting process, design details of castings depend on the particular combination of material and casting process. In designing a casting, the designer should use properties from test bars that have solidified at the actual cooling rate in critical sections of the casting. The properties of a cast component depend on microstructure, which is determined by cooling rate and therefore also section size. This cooling rate can be determined by instrumenting a casting and measuring the cooling rate in various sections or by simulating its solidification using a commercial solidification simulation program.

Section size, of course, is an important factor in design. However, cooling rate depends not only on section size but also on pouring temperature, mold material, gating system, the presence or absence of chills in the mold, mold coatings, and insulation. Thus, within limits (some very broad), the properties can be controlled by the metal caster to produce those which the designer desires.

Typically, the design phase may initially consider the “section size” effect, a factor by which the local casting properties are adjusted based on local dimensions. Bulky sections cool slower, while thin sections solidify faster than the thick sections. Certain other geometric features also influence solidification rate. For instance, concave sections, or reentrant angles, solidify more slowly than fins or protrusions, again affecting the resultant local structure and the properties. In other words, property variation within a casting is caused by local differences in cooling rate that can be influenced by several factors in addition to section size.

The overarching principle of good casting design is that casting sections should freeze progressively, allowing the risers to supply liquid metal to feed shrinkage that occurs during solidification. Designing for progressive solidification requires tapering walls, so that they freeze from one end to the other, as shown in Fig. 32 (Ref 31), and avoiding situations where two heavy sections are separated by a thin section. Junctions also concentrate heat, leading to areas in the casting where heat is retained. Concave corners concentrate heat, while convex corners lose heat faster. There are a number of excellent summaries of principles of good casting design (Ref 31, 32, 33), and a comprehensive booklet was published on design of premium-quality aluminum alloy castings (Ref 34). Product and process design of castings is also described in more detail in *Casting* Volume 15 of *ASM Handbook.*

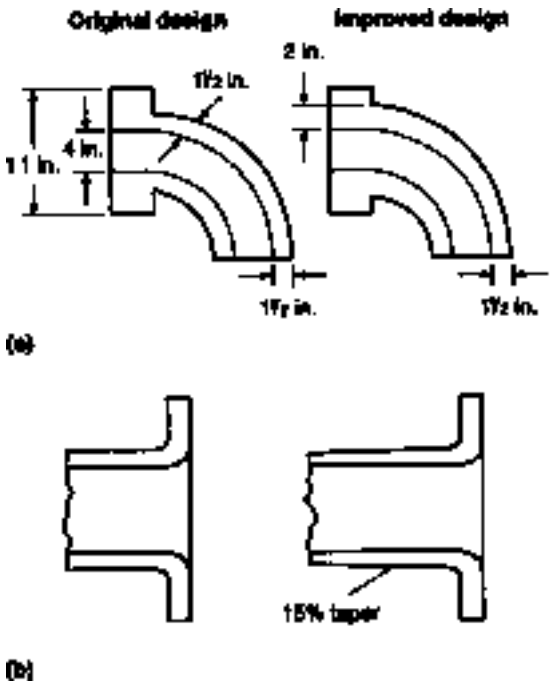


Fig. 32 Redesign of castings to provide progressive solidification through the use of tapered walls. (a) Elbow design. (b) Valve fitting design. Source: Ref 31

Stress Concentration from Residual Stresses. The presence of significant residual internal stresses in a casting can occur from local plastic deformation imposed by thermal gradients created during processing. These gradients may be caused by local cooling when the casting is very hot (in the mold or during shakeout, or in quenching from heat treatment) or by local heating, such as in repair or assembly welding. The effect is plastic stretching (or compression) of the weakest (hottest) portion of the casting. By the time the temperature has equalized, the stretched or compressed region is too long or too short to match the size of the other portion of the casting. The result may be a sizable residual internal stress that may be compressive or tensile. The behavior in service then depends on whether these residual internal stresses are added to or subtracted from the applied service loads.

Predicting residual stresses in casting designs is a relatively recent development, because finite-element simulation is now sufficiently advanced to enable the study of the effect of various process and design parameters on the residual stresses in castings. A current reference on this topic is (Ref 35). In general, the shape of a casting can contribute to residual stresses if it entails sections of markedly different thicknesses,

which can undergo different cooling rates in the mold. This effect can be minimized if the thinner sections are well insulated, so that they cool at a lower rate. The use of chills to accelerate cooling of heavy sections can also be helpful. Stress relief can diminish internal stresses, if the cast alloy is amenable to such treatment.

Example 11: Cracking in Gray-Iron Cylinder Blocks Caused by Casting Stresses (Ref 11). During the operation of tractors with cantilevered bodies, the lateral wall of the cylinder blocks cracked repeatedly. Three of the blocks were examined.

Investigation. Figures 33(a) and Figures 33(b) exhibit one of the failures, as seen from the outside and inside of the cylinder block. The crack was always located at the approximate center of the sidewall, and it ran in a horizontal direction. The sectional view (Fig. 34a) shows more clearly that the thin case wall is reinforced at this location by a thicker rib. The crack was located in the thin-wall portion next to the transition to the rib, as shown in the larger scale view in Fig. 34(c). At this location, high residual stresses might be expected, due to the earlier solidification in the thin wall compared with the heavier secondary sections during cooling of the casting. The stresses reduced wall thickness at this point by stretching the iron while at a high temperature. The arrow in Fig. 34(a) indicates this reduced section.

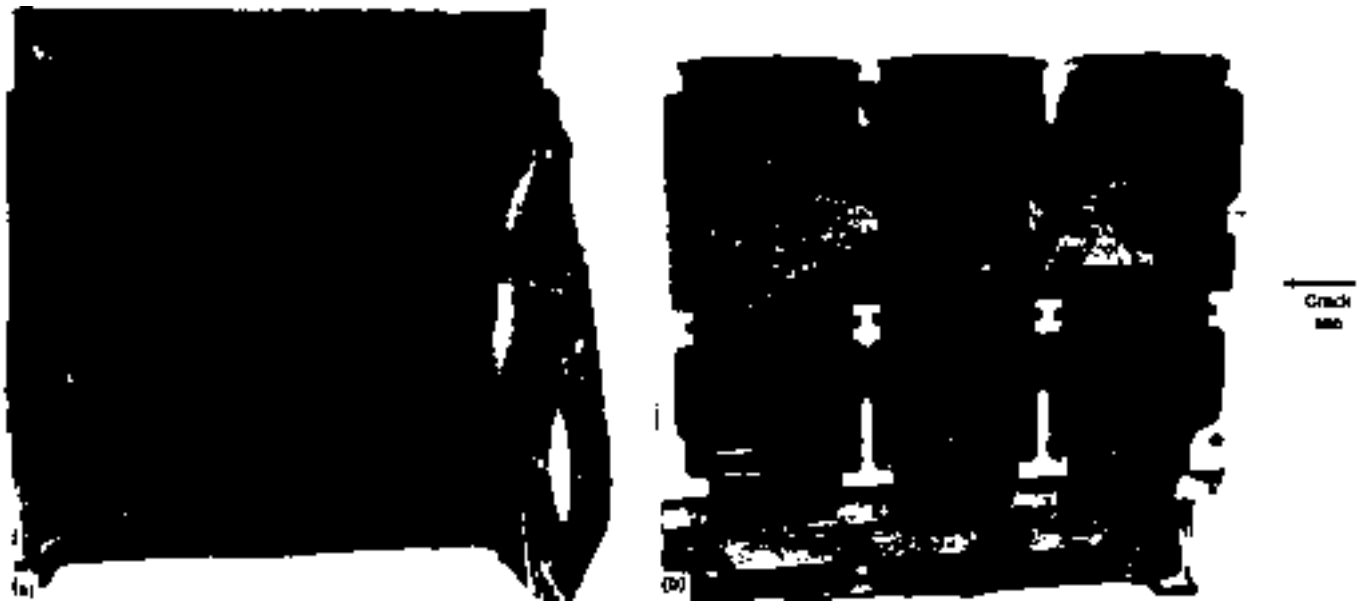


Fig. 33 Gray-iron cylinder block that cracked due to casting stresses. (a) External view. (b) Internal view showing crack site (arrow). 0.25×. Source: Ref 11

Fig. 34 Sections through gray-iron cylinder block. (a) Original design. 0.25×. (b) Improved design. 0.25×. (c) Enlarged section from (a). Actual size. Source: Ref 11

Chemical analysis of the three failed blocks gave the following compositions:

Block	Composition, %					
	C	Mn	P	S	Si	Fe
1	3.30	0.66	0.104	0.054	2.48	bal
2	3.29	…	0.100	…	2.44	bal
3	3.30	…	0.168	…	2.28	bal

Note: Manganese and sulfur were not determined for blocks 2 and 3.

From these data, the carbon-saturation values were determined. (Methods of determining carbon saturation are discussed later in this article in the section “Cast Irons.”) The three values were less than 1 (0.96, 0.96, and 0.95), demonstrating that the cast iron was hypoeutectic. The composition was acceptable for the specified gray iron.

Metallographic examination was performed on sections from these castings to determine if cracking was promoted by material defects. The microstructure of the thick-wall portion consisted of uniformly distributed, medium flake size, type A graphite in a matrix of pearlite with little ferrite (Fig. 35a). The thin-wall part had a structure of type D graphite, with dendrites of austenite forming and later transforming to a pearlitic-ferritic structure; the interdendritic residual liquid formed the type D graphite with a ferritic matrix (Fig. 35b). The structure is formed by supercooling of the molten iron before solidification. In this case, it is promoted by fast cooling of the thin wall.

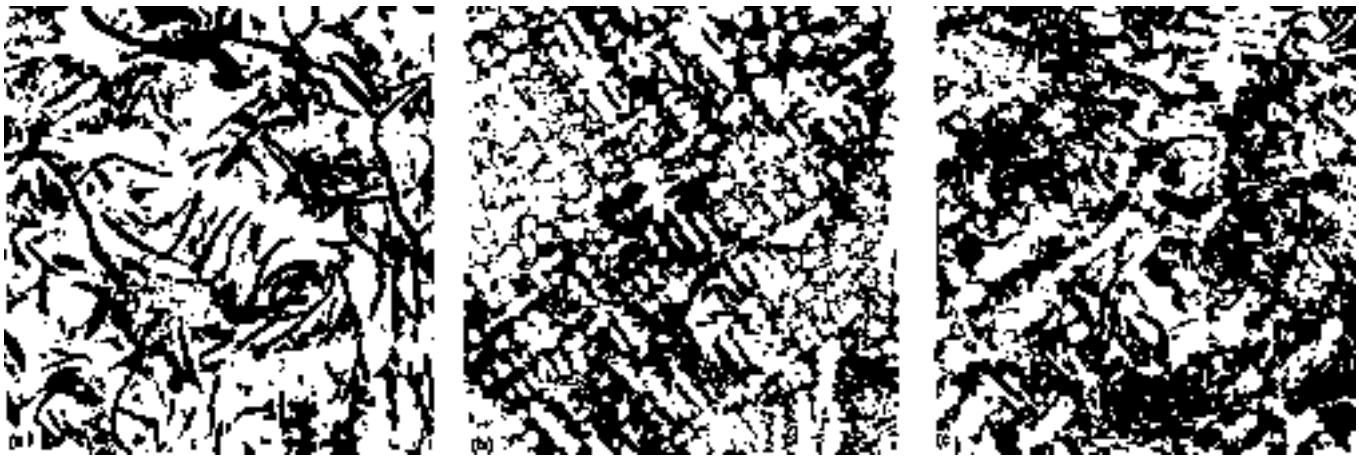


Fig. 35 Micrographs of the gray-iron cylinder block. (a) Section from the thick-wall part. (b) Section from the thin-wall part. (c) Thin-wall part of reinforced case. All etched in picral. 100×. Source: Ref 11

Corrective Measure. The material for the stuffing box, ASTM A 536 ductile iron, was changed from grade 60-45-10 to grade 86-60-03. This eliminated subsequent failures, except those that occurred in extremely corrosive applications.

Conclusions. Fracture in the cylinder blocks can be attributed primarily to casting stresses, which could be alleviated by better filleting of the transition cross section.

Recommendations. The case wall at the thin-wall section should be reinforced. Shrinkage stresses and supercooling could be alleviated or avoided by prolonged cooling of the casting inside the mold. Based on these recommendations, the design was changed with slower cooling (Fig. 34b). Figure 35(c) shows the microstructure of the reinforced wall with the revised design. This is a clear improvement over the previous design (Fig. 35b).

Example 12: Cracked Gray-Iron Crankcases Caused by Casting Stresses (Ref 11). The front wall of a gray cast iron crankcase cracked at the transition from the comparatively minor wall thickness to the thick bosses for the drilling of the bolt holes (Fig. 36). The dark coloration of a part of the fracture plane indicated that the crack originated in the thin-wall part. Four failed crankcases were examined to determine the cause of failure.

Fig. 36 Failed gray-iron crankcase. (a) Fragments of crankcase. (b) Fracture surface of crankcase. 1×. Source: Ref 11

Investigation. Figure 36(b) shows the abrupt change in section thickness. These adjoining thick- and thin-wall sections solidify and shrink at different rates. Metallographic examination indicates a differential structure at the thick- and thin-wall sections (Fig. 37). The thick-wall section of the first sample (Fig. 37a) has a more favorable graphite size and shape, while the thin-wall section (Fig. 37b) has a rosette-type graphite. Crankcase 2 also had the differential graphite condition at thick- and thin-wall sections, which can be observed by comparing Fig. 38(a) (thick-wall) with Fig. 38(b) (thin-wall).

Fig. 37 Microstructures of failed crankcase shown in Fig. 36. (a) Normal flake graphite from a thick-wall section. (b) Type B rosette graphite from a thin-wall section. Both etched with picral. 200×. Source: Ref 11

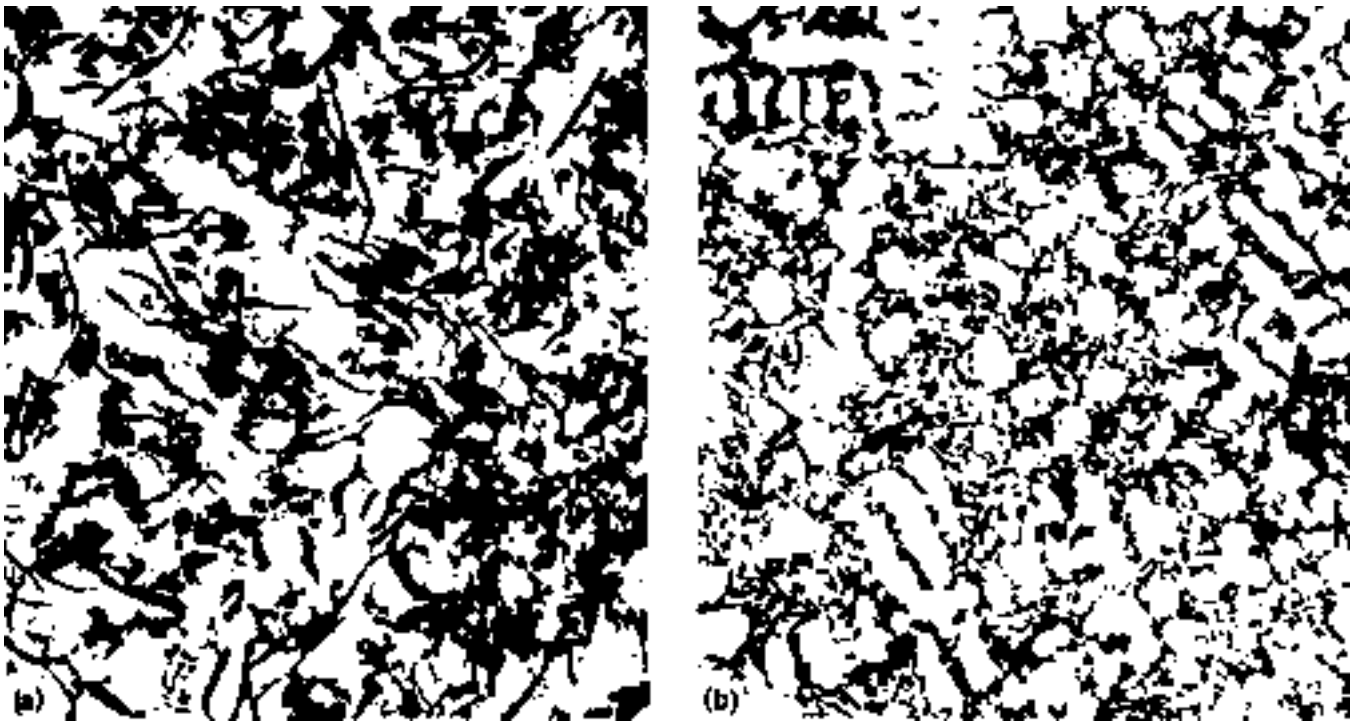


Fig. 38 Microstructures of crankcase 2. (a) Section through the thick-wall part. (b) Section through the thin-wall part with crack. Both etched with picral. 100×

Crankcase 3 had fractures at the frontal end wall and at the thinnest point between two bore holes. The fracture occurred at the point at which the crankcase is only 5 mm (0.2 in.) thick. The structure of this thin-wall part of crankcase 2 consisted of pearlitic dendrites and a connecting enveloping network of ferrite and fine type D graphite. Crankcase 4 had cracked in the borehole of the frontal face. One crack was so narrow that it was discernible only by magnetic-particle inspection. The structure of this casting was permeated with clusters of types B and D graphite surrounded by ferrite near the surface as well as in the interior (Fig. 39).

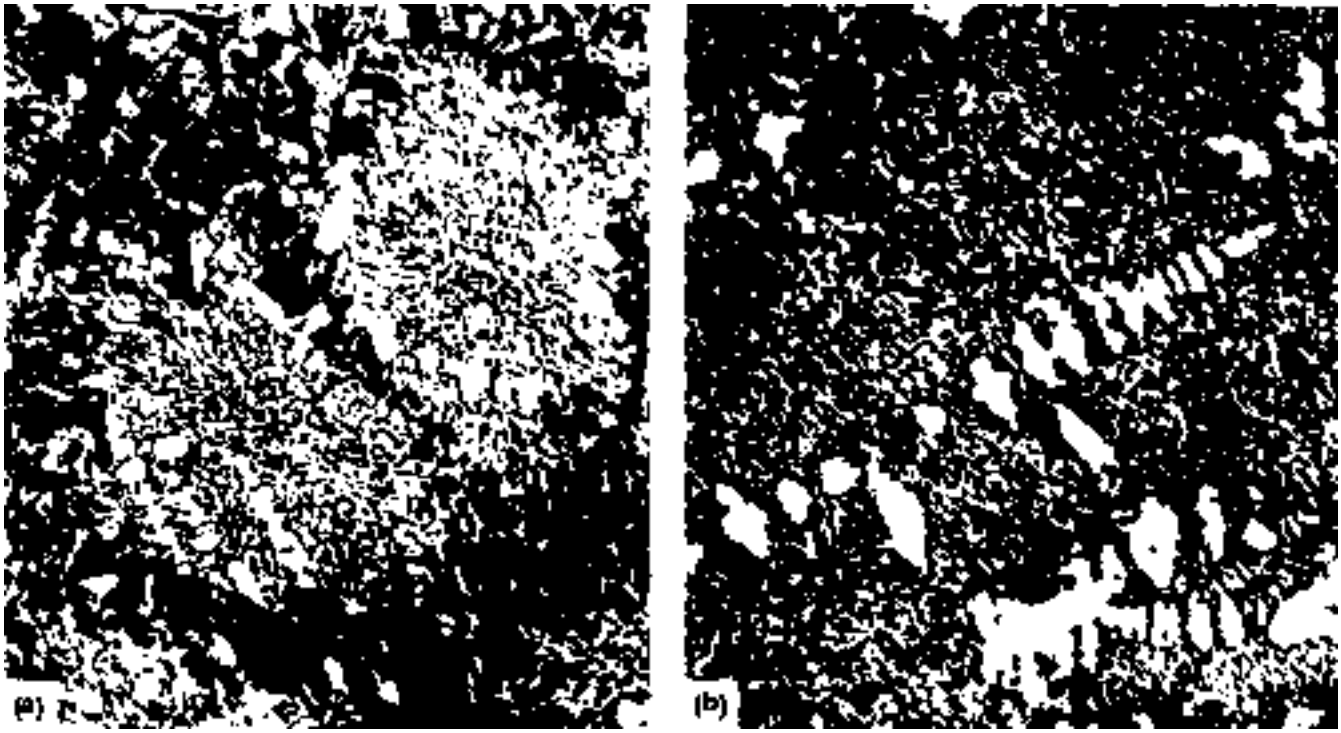


Fig. 39 Microstructures of crankcase 4. (a) Specimen taken from near the surface. (b) The internal structure. Both etched with picral. 100×. Source: Ref 11

Conclusions. In the first three cases, excessive thermal stresses created by the casting process (rapid cooling) as well as pour geometry were responsible for the crack formation. In the fourth case, the cause of the fracture is considered to be thermal stresses at the bore hole. Crack propagation led to the conclusions that shrinkage stresses were the cause of the cracks, because the adjoining thick- and thin-wall cast parts solidify and shrink at different rates. The faster a piece cools, the more damage is caused by differential shrinkage, that is, the higher the stresses at the cross-sectional transition. The amount of stress built up at the transition is also proportional to the abruptness of the change in section thickness. These stresses may add to crack formation during cooling, or they may remain in the piece as residual stresses, which are later superimposed on operating stresses and thus favor the fracture at this particular point.

Example 13: Fatigue Fracture of a Stuffing Box That Originated at the Inner End of a Lubrication Hole. The stuffing box shown in Fig. 40(a) was sand cast from ASTM A 536, grade 60-45-10, ductile iron and began leaking water after two weeks of service. The machine was operating at 326 rpm, with discharge water pressure of 21.4 MPa (3100 psi). The stuffing box was removed from the machine after a crack was discovered in the sidewall.

Fig. 40 Stuffing box sand cast from ASTM A 536, grade 60-45-10, ductile iron. (a) Configuration and dimensions (given in inches). (b) Micrograph showing the structure consisting of graphite nodules in a ferritic matrix with remnants of a pearlite network. Etched with nital. 100×

Investigation. The vertical portion of the crack passed through a tapped 8 mm (0.3 in.) diameter lubrication-fitting hole that was centered on the parting line of the casting (Fig. 40a.) The direction of the crack changed abruptly to nearly horizontal at an internal shoulder approximately 100 mm (4 in.) below the top of the externally threaded end (view A-A,Fig. 40a). A second nearly horizontal crack (faintly visible in view A-A) at the bottom of the vertical crack extended to the right and slanted slightly downward from the parting line for at least 19 mm (0.75 in.). This second crack was very tightly closed.

The surface of the fracture was closely examined, and indications of fatigue beach marks were found. As is normal for fatigue of cast iron, these indications were faint. The beach marks may originally have been slightly more distinct, but if so, they were partly eroded by leakage of water through the crack at high velocity.

A metallographic specimen cut from the casting displayed a structure consisting of well-rounded graphite nodules in a ferrite matrix that also contained remnants of a network of pearlite (Fig. 40b). Evidently, the heat treatment of the stuffing box had not been fully effective in eliminating the pearlite. The mechanical properties, however, were acceptable: 527 MPa (76.5 ksi) tensile strength, 393 MPa (57 ksi) yield strength, and 12% elongation in 50 mm (2 in.) at 179 HB.

Conclusion: Examination of the fracture surface suggested that the crack had been initiated at the inner edge of the lubrication hole and had propagated toward both the threaded and flange ends of the casting. An appreciable residual-stress concentration must have been present and caused propagation of the crack, because the operating pressure of 21.4 MPa (3100 psi), combined with the inside radius of 41 mm ($1\frac{5}{8}$ in.) and the section thickness at the roots of the external threads of 16 mm ($\frac{5}{8}$ in.), indicated a hoop stress of only about 69 MPa (10 ksi). The residual stress might have been caused when a fitting was tightly screwed into the lubrication hole, and it might have been concentrated by notches at the inner end of the hole that were created when the drill broke through the sidewall to the stuffing box.

References cited in this section

1. T. Piwonka, Design for Casting, *Materials, Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 723

11. P.M. Unterweiser, Ed., *Case Histories in Failure Analysis,* American Society for Metals, 1979

31. Manufacturing Design Considerations, *Steel Castings Handbook,* 6th ed., M. Blair and T.L. Stevens, Ed., Steel Founders' Society of America and ASM International, 1995

32. Manufacturing Considerations in Design, *Steel Castings Handbook,* 5th ed., P.F. Wieser, Ed., Steel Founders' Society of America, 1980, p 5–6

33. P.R. Beeley and R.F. Smart, Ed., *Investment Casting,* The Institute of Materials, 1995, p 334

34. S.P. Thomas, Ed., *Design and Procurement of High-Strength Structural Aluminum Castings,* American Foundrymen's Society, 1995

35. R.W Lewis, K.N Seetharamu, and A.Y. Hassan, Residual Stress Formation during Casting, *Handbook of Residual Stress and Deformation of Steel,* ASM International, 2002, p 361–371

Cast Irons

The type of material and manufacturing method is developed during the process of design and materials selection, and this, of course, requires several steps in the review of the process/operating conditions and evaluation of materials (as briefly outlined in the article “Materials Selection for Failure Prevention” in this Volume). However, each material has unique processing aspects and quality-control factors that influence the variability in chemical composition, material condition, and properties.

Improper chemical composition and mechanical properties, although not ordinarily considered to be casting “imperfections,” may result in unsatisfactory performance in service. For example, segregation in castings can result in improper microstructure and deviations from specified hardness or mechanical-property requirements. Other microstructural features that influence the performance of iron castings in service include:

- Graphite of unfavorable shape, size, or distribution
- Embrittling intergranular networks (carbides)
- Surface-structure gradients, such as accidental carburization or decarburization
- Unsatisfactory microstructures resulting from improper materials selection, improper compositions for section size (in gray and ductile irons), or incorrect heat treatment

One problem unique to cast iron is pull-out of graphite particles, which impart a pock-marked finish. Pull-out is a function of the graphite flake size and force applied on the surface by the tool. The presence of carbides in microstructure can also have a deleterious effect on surface finish.

Cast Iron Composition and Microstructure

Carbon Equivalent and Carbon Saturation. Carbon equivalent (CE) is a simplified method of evaluating the effect of composition. Several equations are used; Eq 1 is the most common:

$$CE = TC + \frac{\%Si + \%P}{3} \qquad \text{(Eq 1)}$$

where TC is total carbon content. This value is important, because it can be compared with the eutectic composition (4.3% C) to indicate whether the cast iron will behave as a hypoeutectic or hypereutectic alloy during solidification.

Carbon saturation, SC, can be calculated as:

$$SC = \frac{TC}{4.3 - \left(\frac{\%Si + \%P}{3}\right)} \qquad \text{(Eq 2)}$$

This value is important, because it is a measure of the tendency toward carbide decomposition, and it can be used to estimate the tensile strength for a given composition and section size of cast iron. Hypereutectic irons tend to form graphite, while hypoeutectic irons tend to form carbides during solidification (Ref 36, 37). Eutectic constituents in gray irons can be revealed by etching to exhibit cell boundaries.

Graphite. The size, shape, orientation, and distribution of graphite in castings of gray iron, ductile iron, and malleable iron determine the mechanical properties of the castings. For any given gray-iron composition, the rate of cooling from the solidification temperature to below approximately 650 °C (1200 °F) determines the ratio of combined-carbon to graphitic-carbon phases; this ratio controls the hardness and strength of the iron. Therefore, the effect of section size is considerably greater for gray-iron castings than it is for castings made of more homogeneous ferrous metals, for which the rate of cooling does not affect the form of the carbon content on a macroscopic scale.

Ductile cast iron, previously known as nodular iron or spheroidal-graphite cast iron, is cast iron in which the graphite grows as spheres, rather than as flakes of any of the forms characteristic of gray iron, from additives introduced in the molten iron before casting. Malleable iron is a type of cast iron that has most of its carbon in the form of irregularly shaped graphite. Malleable iron, similar to ductile iron, possesses considerable ductility and toughness, because of its combination of nodular graphite and a low-carbon ferrite matrix. Consequently, malleable iron and ductile iron are suitable for some of the same applications requiring good ductility and toughness.

Malleable iron is produced by first casting the iron as a white iron and then heat treating the white cast iron to convert the iron carbide into the irregularly shaped nodules of graphite. This form of graphite in malleable iron is called temper carbon, because it is formed in the solid state during heat treatment. Because solidification of white iron throughout a section is essential in the production of malleable iron, ductile iron has a clear advantage when the section is too thick to permit solidification as white iron. Malleable iron castings are produced in section thicknesses ranging from about 1.5 to 100 mm ($\frac{1}{16}$ to 4 in.) and in weights from less than 0.03 to 180 kg ($\frac{1}{16}$ to 400 lb) or more. Malleable iron in the United States is principally produced as ferritic malleable, while some pearlitic malleable is produced in Europe. Higher strength and lower ductility are achieved for pearlitic malleable than for ferritic matrix.

Ductile iron also has clear advantages over malleable iron when low solidification shrinkage is needed. Malleable iron cannot tolerate porosity, because voids migrate to the surface of hot spots, such as fillets, and appear as cracks. Also, because ductile iron does not require heat treatment to produce graphite nodules (as does malleable iron to produce temper-carbon nodules), it can compete with malleable iron even though it requires a treatment and inoculation process. The mold yield with ductile iron is normally higher than with malleable iron. Strength and elongation of ferritic ductile cast iron also tend to be higher than ferritic malleable iron (because irregular nodules do not have the same extent of benefit as graphite spherulites in ductile iron). In other applications, however, malleable iron has a distinct advantage over ductile iron. Malleable iron is preferred in the following applications:

- Thin-section casting
- Parts that are to be pierced, coined, or cold formed
- Parts requiring maximum machinability
- Parts that must retain good impact resistance at low temperatures
- Parts requiring wear resistance (martensitic malleable iron only)

Intergranular Networks. Grain-boundary networks of either proeutectoid ferrite (in hypoeutectic alloys) or iron carbide (in hypereutectic alloys) are sometimes encountered in as-cast iron castings that have been cooled through the austenite temperature range at a slow rate. The carbide network can also result in in situ graphitization of the carbide during malleableizing heat treatment, resulting in a grain-boundary network of graphite. Such networks impart low ductility and low toughness to castings and constitute paths of weakness because of the brittleness of the carbides. The presence of such networks is indicated by low impact energy in a Charpy V-notch test and can be detected by metallographic examination. A normalizing treatment can eliminate networks, provided that austenitizing has been done at the proper temperature and time. The casting can then be tempered to the desired hardness or, if higher strength is desired, can be quenched from just above the critical temperature and tempered appropriately.

Inverse chill is a condition in gray iron, ductile iron, or malleable iron in which the interior is chilled or white, while the outer sections are mottled or gray. This condition is observable on fracture casting specimens and generally occurs in lighter casting sections. Some of the causes of inverse chill are sulfur content not balanced by manganese; too low of a degree of nucleation for the section, arising from high melting and pouring temperatures; and the presence of tellurium, chromium, and other severe carbide stabilizers (Ref 38, 39).

Excess Ferrite. A generally acceptable microstructure occasionally may not perform satisfactorily because of some particular aspect. Figure 41 shows a micrograph of a section through the tip of a gray-iron cam lobe that sustained rapid lubricated wear because of an excessive amount of ferrite. The microstructure contained a dendritic network of pearlite and islands of free ferrite in a field consisting of type D graphite in a ferrite matrix. The large amount of free ferrite present was detrimental to wear resistance, although the cam surface was well lubricated. A lower silicon content and a slightly higher chromium content would reduce the amount of free ferrite and suppress the formation of type D graphite in the gray iron.

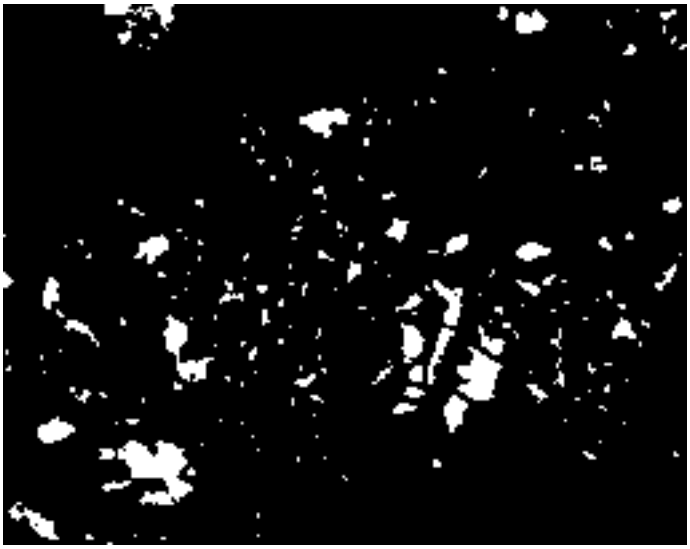


Fig. 41 Section through the worn tip of a gray-iron cam lobe. The white islands are free ferrite in a field of type D graphite. The matrix is ferritic. Etched with 2% nital. 235×

Carburization and Decarburization. Meticulous control of the furnace atmosphere is needed to prevent unintentional carburization and decarburization. Such control is not difficult with properly instrumented heat treating equipment.

In shell-mold castings of gray iron, the high thermal conductivity of the mold produces an initial chill of the iron surface. Subsequent annealing of the surface by the heat of the casting results in a skin that is almost 100% ferrite; such a skin has poor wear properties and should be removed by machining for applications requiring wear resistance.

Gray Cast Iron

The typical microstructure of gray iron is a matrix of pearlite with graphite flakes dispersed throughout. Too-rapid cooling (relative to carbon and silicon contents) produces mottled iron, which consists of a pearlite matrix with both primary cementite (iron carbide) and graphite. Very slow cooling of irons that contain larger percentages of silicon and carbon is likely to produce considerable ferrite as well as pearlite throughout the matrix, together with coarse graphite flakes. Under some conditions, very fine graphite (type D) can be formed near casting surfaces or in thin sections. Frequently, such graphite is surrounded by a ferrite matrix; therefore, soft areas result. Type D graphite is usually associated with a ferrite matrix. This structure occurs due to a lack of nucleation sites, a large undercooling, and usually at thin casting sections. Dimensional configuration resembles the shape of sea coral.

Gray-iron castings are susceptible to most of the imperfections generally associated with castings, with additional problems resulting from the relatively high pouring temperatures. These additional problems result in a higher incidence of gas entrapment, inclusions, poor metal structure, interrupted metal walls, and mold-wall deficiencies. Macroetching of cast iron is used to reveal such phenomena as chill formation and segregation.

Carbon flotation, sometimes called kish graphite, is another type of internal discontinuity in cast irons. This form of segregation appears at and near the upper surfaces of a casting, always in the cope portion. This is caused by the carbon equivalent being too high. Carbon flotation is free graphite separated from molten iron during solidification. This gross segregation occurs in spheroidal graphite metal (ductile iron). Carbon flotation is remedied by increasing the pouring temperature, by lowering the carbon equivalent value by closer chemistry control, and by increasing the cooling rate (Ref 39, 40).

Example 14: Premature Engine Failure Due to Improper Surface Finish Caused by Carbon Flotation. Premature failures were occurring in an engine within the first 1600 km (1000 miles) of operation. The failures involved the nodular cast iron crankshaft and its main-bearing inserts. These failures were indicated by internal noise, operation at low pressure, and total seizing. Concurrent with the incidence of engine field failures was a manufacturing problem: the inability to maintain a similar microfinish on the cope and drag sides of a cast

main-bearing journal. The failure analysis and investigation of the manufacturing problem showed that the cause was related to this manufacturing problem.

Investigation. During the in-plant investigation, it was found that the surface finish of the cope side was rougher than that of the drag side. The surface roughnesses of the 50 sample crankshaft main bearings checked ranged from 0.4 to 0.8 mm (15 to 30 mils) on the cope side and from 0.25 to 0.5 mm (10 to 20 mils) on the drag side. The requirement was 0.5 mm (20 mils) maximum.

Surface micrographs were taken for comparison purposes at the cope and drag sides of the main-bearing journals to identify the cause of the nonuniform surface finish. Figure 42 illustrates the finishes on the cope and drag sides. The cope side had areas where the surface was broken open, and it showed some raised areas. The drag side indicated fewer raised areas and almost none of the broken-open surface areas.

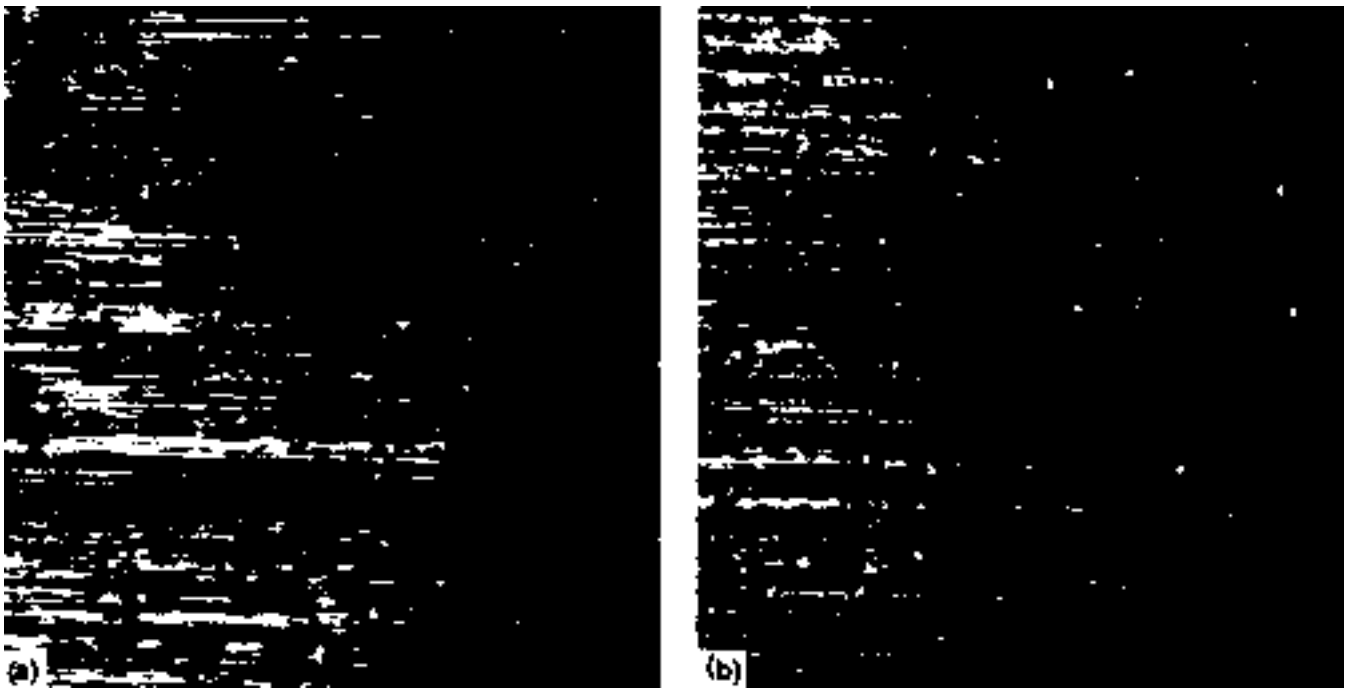


Fig. 42 Micrographs of a failed crankshaft main-bearing journal. (a) Cope side showing the surface broken open with raised areas. (b) Drag side, with fewer raised areas and almost no broken-open surface. Both 50×

Longitudinal metallographic sections were made of the cope and drag portions of the main-bearing journal. Figure 43 shows that the cope portion had larger and more numerous graphite nodules and ferrite envelopes than the drag portion. At the higher magnification shown in Fig. 44, the areas of the cope side that had the appearance of a broken-open surface in Fig. 42(a) were confirmed. The broken-open areas are called ferrite caps. Due to the pressure and heat of polishing during manufacture, the larger graphite nodules break open, and the ferrite envelopes (caps) become actual burrs. These burrs rise 4 to 13 μm above the normal surface (Fig. 44).

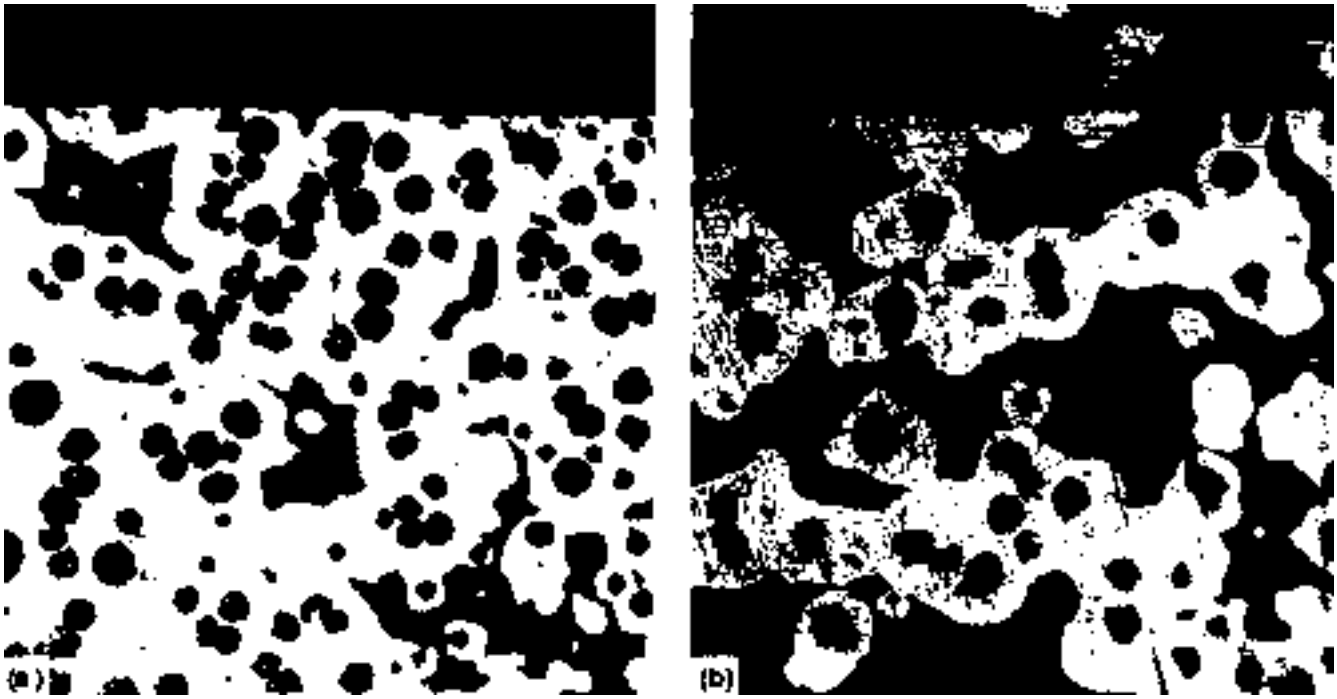


Fig. 43 Longitudinal sections from the main-bearing journal. (a) From the cope side. (b) From the drag side. Note the more numerous graphite nodules in (a). Both etched with nital. 60×

Fig. 44 Micrograph from the cope side of the main-bearing journal. Ferrite caps, which partially cover the graphite nodules and cause the broken-open surface condition, are shown. The burrs rise above the surface from 4 to 13 µm. Etched with nital. 220×

The height of these protrusions above the normal surfaces explains the variance in surface finish of cope versus drag areas. This evidence confirms the cause of crankshaft and crankshaft-bearing failures. The ferrite caps (burrs) can shear the lubricating oil film. Also, the burrs can actually penetrate the bearing-insert material. In either case, the net result is failure of the crankshaft and crankshaft-bearing insert interface due to heat generation or physical damage.

Conclusions. The root cause of the failure is carbon flotation. Chemical analysis of the crankshafts involved in the failures showed a higher-than-normal carbon content and/or carbon equivalent. Therefore, during the casting process, the carbon or graphite would tend to rise in the mold to the cope side, resulting in larger and more numerous graphite nodules in that area and in the presence of more ferrite and less pearlite on the cope side. This layer could be 2.5 mm (0.1 in.) below the surface. The larger nodules break open, causing the ferrite caps or burrs. They then become the mechanism of failure by breaking down the oil film and eroding the beating material. A by-product is heat, which assists the failure.

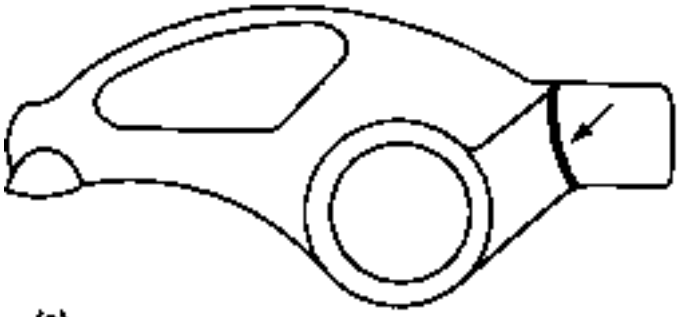
Recommendations. Steps were taken to establish closer control of chemical composition and foundry casting practices to alleviate the carbon-flotation form of segregation. Additionally, some nonmetallurgical practices in journal-finishing techniques were modified to ensure optimal surface finish.

Malleable Irons

Blowholes and spikes are defects often found in malleable iron castings. Shrinkage or open structure in the gated area is a defect often found in malleable iron castings that may be overlooked by visual inspection, although it is readily detected by either liquid penetrant or magnetic-particle inspection.

Example 15: Fatigue Failure of Malleable Iron Diesel-Engine Rocker Levers Originating at Spiking Defects. Several diesel-engine rocker levers failed at low hours in overspeed, overfuel, highly loaded developmental engine tests. The sand-cast rocker levers were malleable iron similar to ASTM A 602, grade M7002. Identical rocker levers had performed acceptably in normal engine tests.

Investigation. The rocker levers were failing through the radius of an adjusting screw arm (Fig. 45a). The typical fracture face exhibited two distinct modes of crack propagation (Fig. 45b). The upper portion, extending approximately 6.4 mm (0.25 in.) from the top surface, appeared very dark and rough, which would indicate overload at final fracture. The majority of the fracture face was very light gray and smooth in appearance. The fanlike fracture surface would be characteristic of the brittle, transgranular propagation of a fatigue fracture.



(a)

(b)

Fig. 45 Failed malleable iron rocker lever. (a) Illustration showing failure location (arrow). (b) Typical appearance of fracture face with dark overload fracture at top, light-gray fatigue fracture, and black casting defect (arrow). 1.5×

The fatigue fracture originated at a small, black-appearing casting defect (arrow in Fig. 45b), approximately 1 mm (0.04 in.) maximum depth by 1 mm (0.04 in.) maximum width, adjacent to the radius on the underside of the adjusting screw arm, near the casting parting line. The black appearance was a result of heavy oxidation along interdendritic voids connected to the surface by dendrite ribs. A cross section through a similar defect in another rocker lever revealed interdendritic-appearing voids surrounded by heavy oxidation and partial decarburization of the adjacent tempered martensitic matrix structure (Fig. 46 and 47).

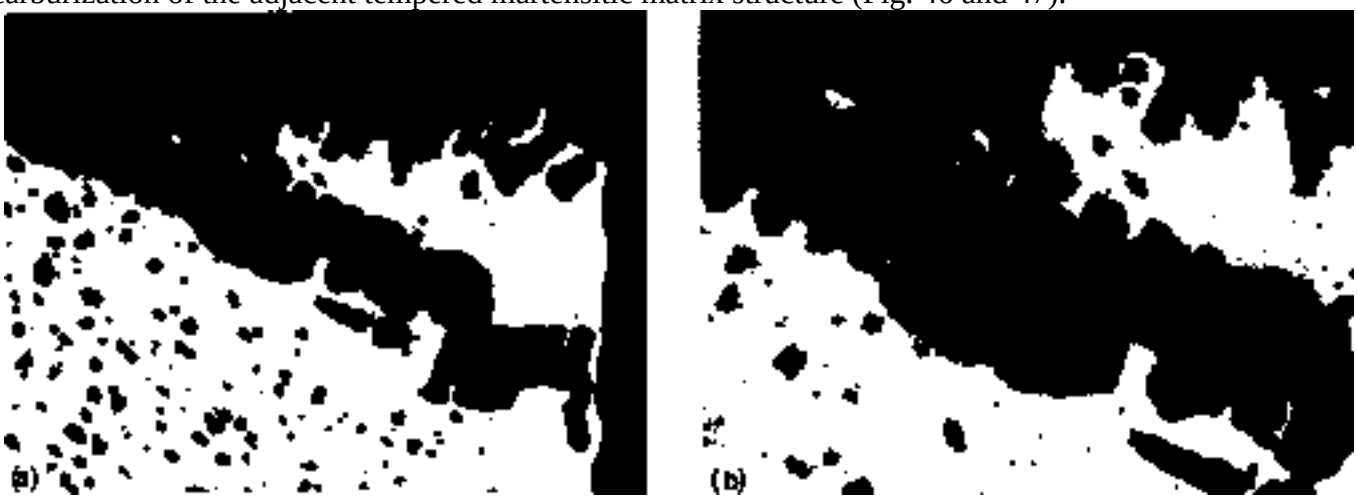


Fig. 46 Section through a spiking defect in rocker lever similar to that shown in Fig. 45. (a) 30×. (b) 60×

Fig. 47 Microstructure at and near the spiking defect shown in Fig. 46. (a) Heavy oxidation within an interdendritic void. As-polished. (b) Partial decarburization of the tempered martensitic matrix adjacent to a spiking defect. Etched with nital. Both 300×

Efforts to sort defective rocker levers by means of magnetic-particle inspection and radiography proved ineffective. Because fatigue testing is costly and time-consuming, a method was devised to evaluate the effectiveness of foundry efforts to correct casting defects. The test method involved sectioning the rocker levers to reduce the cross section in the region of the defects and mechanically overloading to initiate fracture in the general location of the defects. A large percentage of old and new castings were found to contain the dendritic-appearing defects. The defects located, both with and without oxidation, ranged in size to a maximum of 2.5 mm (0.10 in.) deep by 1 mm (0.04 in.) wide on the fractured surfaces.

Conclusions. The rocker levers failed in fatigue, with casting defects acting as stress raisers to initiate failures in highly loaded engine tests. The casting defects, known as spiking, were attributable to a solidification condition by which large austenite dendrites develop that can create an impossible feeding condition and thus

the interdendritic voids. Limiting the size of austenite dendrites by increased nucleation should eliminate the spiking defects.

Corrective Measures. Because initial foundry efforts to eliminate the spiking condition were not totally effective, shot peening of the levers was introduced as an interim measure to reduce the possibility of failure. Redesign to increase the cross-sectional area of the levers has since eliminated failures through the adjusting screw arm, as verified by fatigue testing.

Ductile Iron

Effect of Composition. The properties of ductile iron depend first on composition. Composition should be uniform within each casting and among all castings poured from the same melt. Many elements influence casting properties, but those of greatest importance are the elements that influence the matrix structure or the shape and distribution of graphite nodules.

Carbon influences the fluidity of the molten iron and the shrinkage characteristics of the cast metal. Excess carbon not in solution but in suspension reduces fluidity. The volume of graphite is 3.5 times the volume of iron. As ductile iron solidifies, the carbon in solution precipitates out as graphite and causes an expansion of the iron, which can offset the shrinkage of the iron as it cools from liquid to solid. The amount of carbon needed to offset shrinkage and porosity is indicated in the following formula:

$$\%C + \frac{1}{7}\%Si \geq 3.9\%$$

Carbon contents greater than this amount begin to decrease fatigue strength and impact strength before the effect is noticed on tensile strength. The size and the number of graphite nodules formed during solidification are influenced by the amount of carbon, the number of graphite nuclei, and the choice of inoculation practice. Normal graphite-containing ductile iron has 10% less weight than steel of the same section size.

In ductile iron, all properties relating to strength and ductility decrease as the proportion of nonnodular graphite increases, and those relating to failure, such as tensile strength and fatigue strength, are more affected by small amounts of such graphite than properties not involving failure, such as proof strength. This is illustrated in Fig. 48 (Ref 41), which shows the typical effects of graphite form on tensile and offset yield (proof) strength. The form of nonnodular graphite is important, because thin flakes of graphite with sharp edges have a more adverse effect on crack initiation.

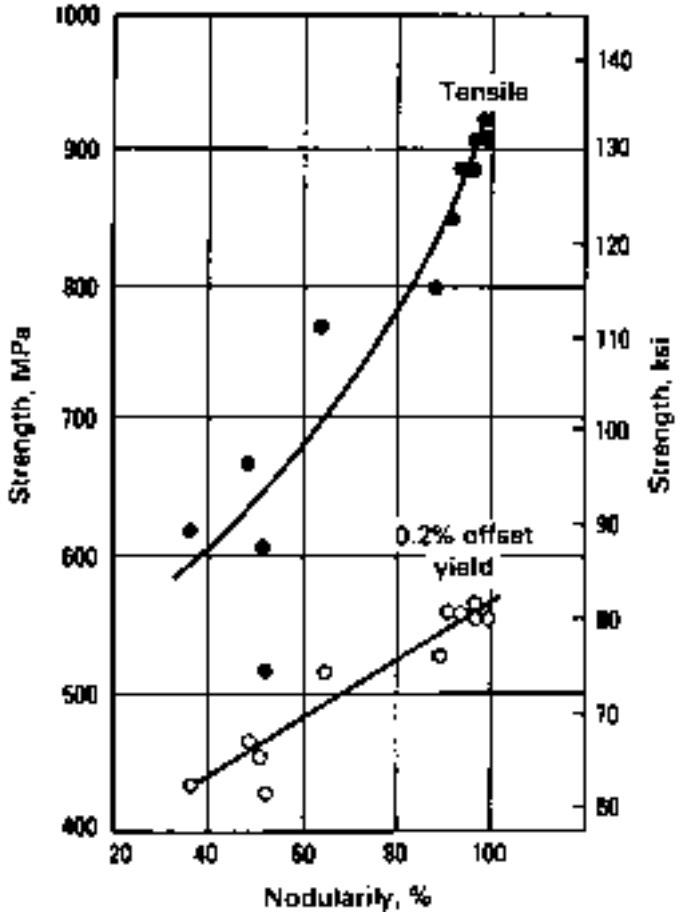


Fig. 48 Tensile and yield strength of ductile iron versus visually assessed nodularity. Source: Ref 41

Example 16: Fracture of Ductile-Iron Pistons for a Gun Recoil Mechanism as Affected by Type of Graphite. The gun mount used for two types of self-propelled artillery (175 mm guns and 205 mm howitzers) involved a recoil cylinder and a sand-cast ductile-iron piston containing orifices through which oil was forced in order to absorb the recoil energy of the gun. A rod attached to the gun tube engaged a thread inside the small end of the piston. The piston was stressed in tension by oil pressure on the flange at the large end, opposing the direction of motion of the rod. Several pistons that had cracked or fractured in service were submitted for analysis.

Investigation. Figures 49(a) and (b) show two views of a piston, illustrating the orifices specially designed to control oil flow and the pattern of cracks that developed. In Fig. 49(c), which shows a composite of the fractures found in several pistons, the piston surface has been developed as a flat-plan view. Orifices A and B correspond to those in Fig. 49(a) and (b), respectively; orifices A′ and B′ are 180° from orifices A and B. The vertical fracture (arrow) shown in Fig. 49(a) is unusual in origin and direction of propagation; most of the fractures originated at the upper right-hand corners of the large orifices (A and A′, Fig. 49c) and propagated in an approximately horizontal direction. Some of these nearly horizontal cracks intersected the small orifices (B and B′, Fig. 49c); others propagated around the piston to the opposite large orifice, joining it at the left side of the narrow opening at top left.

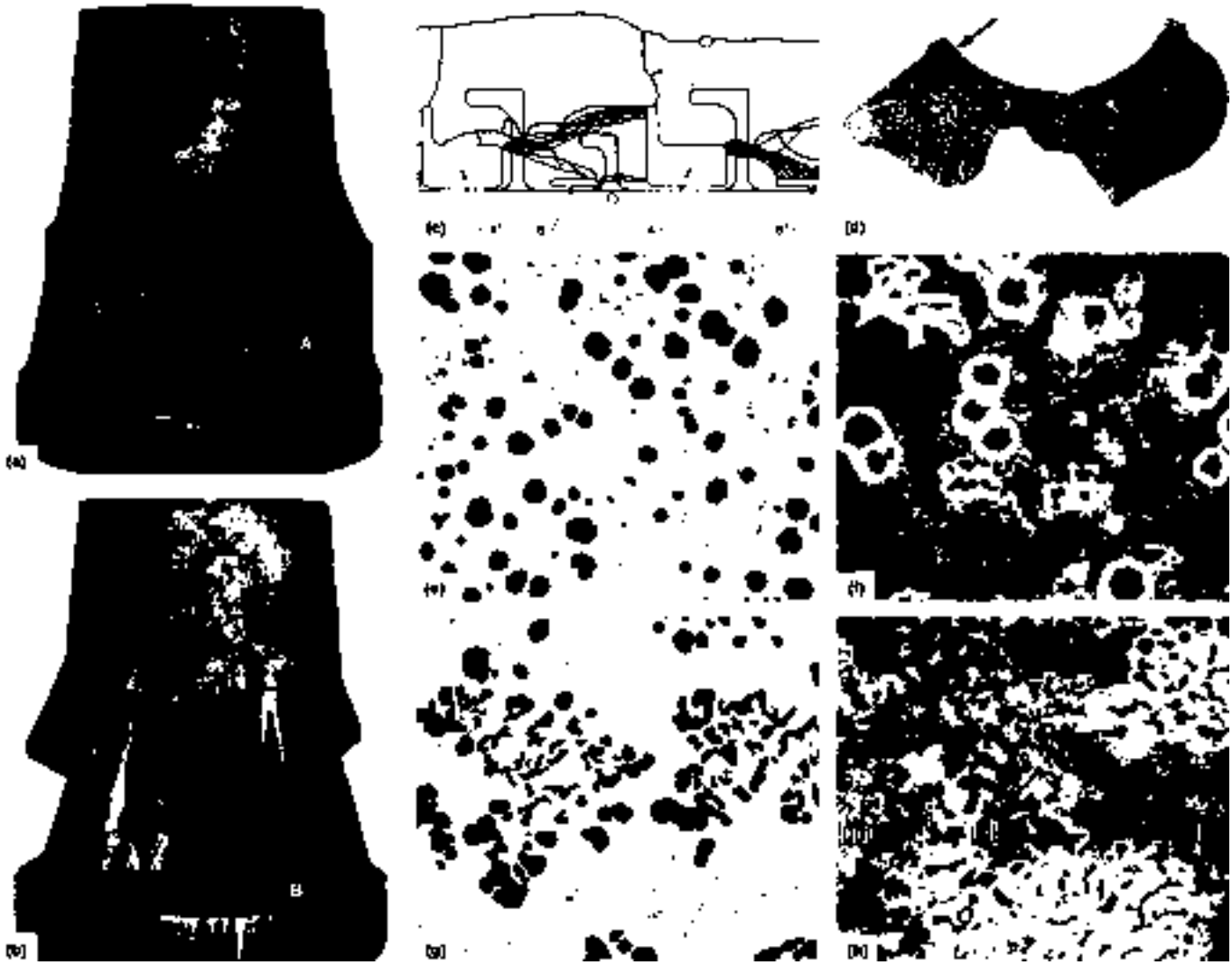


Fig. 49 Piston for a gun-recoil mechanism, sand cast from ductile iron conforming to MIL-I-11466, grade D7003, that fractured in fatigue because of vermicularity of graphite. (a) and (b) Two different views of the piston showing fractures; A and B indicate orifices (see text). Approximately 0.35×. (c) Flat-plan view showing composite pattern of fractures in several pistons. (d) Fracture surface showing mottled structure caused by vermicular graphite; arrow points to a fatigue zone. Approximately 0.25×. (e) to (h) Micrographs showing graphite with nodularity of rank 2, at 50× (e); rank 8, at approximately 60× (f); rank 13, at 50× (g); rank 16, at approximately 65× (h). Specimens in (e) and (g) were not etched; specimens in (f) and (h) were etched with 2% nital.

Figure 49(d) shows the surface of one of the horizontal fractures, which has a blotchy, mottled appearance. The pattern was created by the graphite present; there was relatively little of the massive, hypereutectic carbide that is, by definition, present in iron with a mottled structure. The arrow in Fig. 49(d) indicates a location that exhibits the characteristics of a fatigue zone. The uppermost fracture shown in Fig. 49(a) is met by the more or less vertical break at virtually a right angle. This signifies that the upper fracture took place earlier.

Because some of the pistons exhibited regions with an appearance suggesting fatigue (Fig. 49d), SEM fractographs were prepared to attempt to locate fatigue striations and to estimate the number of cycles to failure. These fractographs showed configurations similar in appearance to fatigue striations, but it was concluded that they were pearlite lamellae, because the spacings agreed with pearlite dimensions observed in optical micrographs. Scanning electron microscope views of the area of final fracture revealed dimples in some regions and features resembling cleavage facets in others. However, what appeared to be facets may have been interfaces of pearlite lamellae oriented nearly parallel to the direction of crack propagation.

The castings were made according to the specifications of MIL-I-11466, grade D7003, which requires minimum properties of 690 MPa (100 ksi) tensile strength, 483 MPa (70 ksi) yield strength, and 3% elongation in 50 mm (2 in.); these properties are usually acquired by heat treatment (normalizing for these castings). The

specification stipulated that the graphite be substantially nodular and required a metallographic test for each lot of castings. There was no requirement concerning the matrix. Of 17 pistons that were sectioned for testing (2 or more test specimens per piston), 11 averaged 814 MPa (118 ksi) tensile strength, but the remaining 6 averaged only 535 MPa (77.6 ksi); 13 met the yield strength requirement, averaging 552 MPa (80 ksi), while 4 did not, averaging 399 MPa (58 ksi); in elongation, 9 pistons were satisfactory, averaging 4.7%, but 8 averaged only 2.1%. However, depending on how the 690 MPa (100 ksi) specification is defined (i.e., from separately poured test bars or specimens cuts from casting), those variances may be expected.

Examination of the microstructures of these 17 pistons, along with those of additional pistons, revealed that 77% of those that failed did not contain the required nodular graphite. The range in nodularity was sufficient for establishment of a microstructure rating system in which quality of graphite was ranked from 1 (nodular) to 16 (vermicular with essentially no nodules). Structures containing graphite ranked 2, 8, 13, and 16 are shown in Fig. 49(e) through (h), respectively. Figure 49(h), which shows the structure of the piston with the mottled surface in Fig. 49(d), exhibits massive carbides but not enough to form a mottled structure in the usual sense of the term. For this application, graphite ranked from 1 to 7 was considered acceptable, graphite ranked from 8 to 12 borderline, and graphite ranked from 13 to 16 unacceptable.

The degree of nodularity of the graphite was correlated with the results obtained in ultrasonic testing in order to appraise the quality of pistons that had already been installed. It was found that vermicular graphite damped the back-reflection signal to a significantly greater degree than nodular graphite did. Comparison of the results obtained in the ultrasonic tests and in the metallographic examinations yielded a correlation coefficient of 0.84—an indication of a high degree of correlation.

Smooth-bar rotating beam tests were performed on specimens prepared from pistons specifically selected for a comparison of fatigue test behavior and graphite structure. Fatigue tests in simple tensile loading were conducted on new, complete pistons containing nodular graphite. Four tests at a stress corresponding to the gun recoil load produced fatigue fractures at 20,000 to 35,000 cycles. All these fractures, unlike most of those observed in service, occurred at the base of the threads in the top of the piston, which is the region of the upper crack shown in Fig. 49(a) and (b). The section through the threads had been calculated during design as the location of maximum stress under simple tension. Because nearly all service fatigue cracks did not occur at the threads but initiated from the oil orifices, cracking can likely be attributed to slight misalignment that caused bending as well as tension and to the force of the oil on the surface immediately above and below the large orifices, increasing the stresses at the corners of these orifices. These tests showed that the life of the pistons was about 20,000 recoil loadings instead of the indefinite service life that had been anticipated.

Conclusions. Fracture of the pistons occurred under high strain-rate axial tension modified by superimposed bending loads from oil pressure and/or normal alignment tolerance. Most of the service fractures occurred in pistons containing vermicular graphite instead of the specified nodular graphite. Brittleness caused by massive carbides might have shortened the service lives of a few pistons.

Corrective Measures. Ultrasonic testing was adopted for inspection of pistons already in the field to identify and reject those containing vermicular graphite. Metallographic control standards were applied to a production of new pistons to ensure structures containing substantially nodular graphite and no massive carbides.

References cited in this section

36. H.T. Angus, *Cast Iron: Physical and Engineering Properties,* 2nd ed., Butterworths, 1976

37. C.F. Walton, *Gray and Ductile Iron Castings Handbook,* Iron Founder's Society, Des Plaines, IL, 1971

38. S.I. Karsay, *Ductile Iron-Production Practices,* American Foundrymen's Society, Des Plaines, IL, 1975, p 167

39. G.W.S. Anselman et al., *Analysis of Casting Defects,* 2nd ed., American Foundrymen's Society, Des Plaines, IL, 1966

40. F. Hudson et al., *Atlas of Defects in Castings,* Institute of British Foundrymen, London, 1961

41. A.G. Fuller, Evaluation of the Graphite Form in Pearlitic Ductile Iron by Ultrasonic and Sonic Testing and the Effect of Graphite Form on Mechanical Properties, *Trans. AFS*, Vol 85, 1977, p 509–526

Cast Steels

Carbon steel parts are normally supplied in the normalized or normalized-and-tempered condition but can be quenched and tempered or annealed. Some are also used in the as-cast condition. Heat treating of carbon steel should be selected to ensure adequate toughness and weldability. Carbon steel castings, if they fail due to improper heat treatment, most often fail in a service that requires high impact toughness, particularly at low temperatures.

Low-Alloy Steel Castings

Low-alloy steel castings are susceptible to heat treating problems. In addition to toughness and ductility, strength is also determined by the heat treat cycles, and these can be varied over a wide range. The best combinations of mechanical properties are developed in the quenched-and-tempered condition, with the tempering conditions (temperature, time at temperature, and quenching temperature) controlling the hardness, ductility, strength, and toughness.

Many failures in low-alloy steel castings are due to lack of strength or a combination of factors leading to fatigue or overload fractures. A combination of cyclic loading (thermal fatigue), an aggressive service environment (steam), and internal defects resulted in gradual crack propagation in Example 17. Low hardness can also lead to premature wear of wear-resistant low-alloy castings. Low strength can be caused by several heat treating errors. A slack quench caused by lack of good tank circulation, over-loaded heat treat baskets, or early (hot) removal from the quench will lead to low strength. This is due to the lower-hardness bainitic microstructures that are formed instead of the harder martensitic ones. Low strength can occur with a good quench when it is followed by tempering at too high a temperature. An extended hold time in the heat treat furnace at the tempering temperature may produce the same result.

Example 17: Failure of a Thick-Wall Casing for a Steam Turbine by Cracking. When a crack developed in a cast steel steam-turbine casing, a small section of the casing was removed by torch cutting to examine the crack completely and to determine its origin. The crack was discovered during normal overhaul of the turbine. The chemical composition and processing of the casting were in accordance with ASTM A 356, grade 6 (1.25%Cr-0.5%Mo). The mechanical properties of the casting were yield strength, 352 MPa (51 ksi); tensile strength, 490 MPa (71 ksi); elongation, 29.5%; and reduction of area, 70.5%, all of which exceeded the requirements of the specification. Charpy V-notch impact energy values for specimens taken from the casting were 14.6 J (10.75 ft · lbf) at 25 °C (75 °F), 88 J (65 ft · lbf) at 0 °C (32 °F), and 5.4 J (4.0 ft · lbf) at -18 °C (0 °F). When the fracture surface was examined visually, an internal-porosity defect was observed adjoining the tapped hole. A second, much larger cavity was also detected (upper right portion, Fig. 50a).

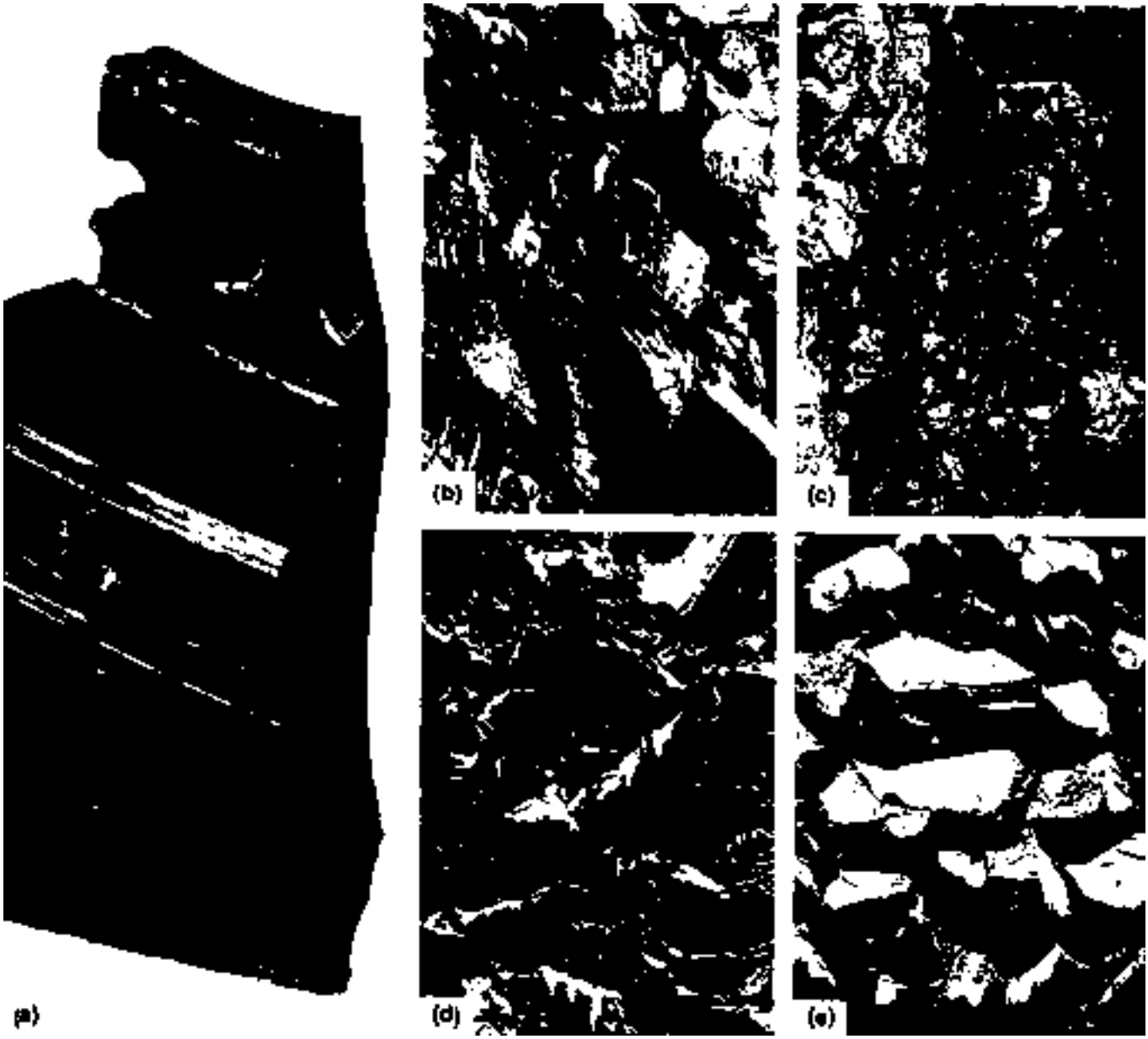


Fig. 50 ASTM A 356, grade 6 (1.25%Cr-0.5%Mo), cast steel turbine casing that failed by cracking. (a) Segment removed from the casing, showing the fracture surface at right. A large porosity defect can be seen at the upper right corner, near the broken-open tapped hole. (b) to (e) Transmission electron microscopy fractographs of four locations on the fracture surface. 7500×. Intergranular modes of fracture are shown in (b) and (c); a transgranular mode, typical of corrosion fatigue, is shown in (d). The fractograph in (e) was taken in the region of crack arrest and indicates an intergranular mode of fracture.

Metallographic Examination. An examination of the microstructure revealed a concentration of precipitates on slip planes within grain boundaries. Fatigue loading increases the concentration of point and line defects, and these defects assist diffusion. Therefore, precipitation, which is a diffusion-controlled process, is more advanced in the slip bands of a material subjected to fatigue. In this case, the fatigue was thermal fatigue caused by repeated temperature changes in the turbine over a period of about ten years. These temperature changes were more frequent than usual, because the turbine was on standby service for several years.

In regions near the main fracture surface, small subsidiary cracks—both transgranular and intergranular—were noted. At surfaces contacted by steam, there was evidence of stress-corrosion cracking (SCC) and intergranular attack.

Electron Fractography. Replica impressions were taken at eight locations on the main fracture surface for examination in an electron microscope. Depending on location, there was evidence of intergranular and transgranular cracking. For example, Fig. 50(b) and (c) indicate an intergranular crack propagation; however,

Fig. 50(d), taken farther down the fracture surface, indicates a transgranular cracking. Figure 50(e), which was taken from the region of crack arrest, indicates a completely intergranular mode.

Conclusions. Failure occurred through a zone of structural weakness that was caused by internal casting defects and a tapped hole. The combination of cyclic loading (thermal fatigue), an aggressive service environment (steam), and internal defects resulted in gradual crack propagation, which was, at times, intergranular—with or without corrosive attack—and, at other times, was transgranular.

Temper Embrittlement. Low-alloy steel castings can suffer from temper embrittlement when they are tempered in the region of 260 to 540 °C (500 to 1000 °F). Slow cooling through this region from a higher temperature can also lead to the problem. It is advisable to specify a water quench from the tempering temperature whenever practical to avoid this. The low toughness caused by temper embrittlement can lead to brittle fractures. More highly alloyed steels, such as the 4300 series and the HY steels, tend to be more susceptible to the problem.

Temper embrittlement, similar to any type of embrittlement that occurs from heat treating, is not detected by the common acceptance criteria of hardness testing. Annealed steels also differ from normalized-and-tempered steels in toughness but not in hardness. Data in Table 4 illustrate this. Two heats of 0.22C-0.12Mo steel were treated and tested. One was annealed at 925 °C (1700 °F) and furnace cooled. The other was normalized at 925 °C (1700 °F), air cooled, then tempered at 670 °C (1240 °F). While the tensile strengths of approximately 483 MPa (70 ksi) and the hardnesses of approximately 140 HB were the same, the impact results were quite different. The annealed material exhibited only 41 to 46 J (30 to 34 ft · lbf) Charpy V-notch impact toughness, while the normalized-and-tempered heat had 83 to 87 J (61 to 64 ft · lbf) impact toughness. Toughness testing is a good method of ensuring correct heat treatment.

Table 4 Composition and typical mechanical properties of 12 heats of closely similar cast carbon-molybdenum steel[(a)]

Item	Composition, %						Tensile strength		Yield strength		Elongation, %	Reduction of area, %	Hardness, HB	Charpy V-notch, J (f · lbf) at 21°C (70 °F)
	C	Mn	Si	Ni	Cr	Mo	Mpa	ksi	MPa	ksi				
1[(b)]	0.24	0.54	0.41	0.70	0.11	0.15	507	73.5	276	40.0	28.0	46.0	143	60, 54, 46, 53 (44, 40, 34, 39)
2[(b)]	0.26	0.59	0.45	0.09	0.14	0.14	528	76.6	293	42.5	27.0	46.3	149	46, 49, 46, 47 (34, 36, 34, 35)
3[(c)]	0.23	0.55	0.43	0.18	0.10	0.11	486	70.5	286	41.5	27.5	43.5	140	57, 54 (42, 40)
4[(b)]	0.23	0.55	0.43	0.18	0.10	0.11	467	67.8	255	37.0	28.0	39.5	138	41, 38 (30, 28)
5[(c)]	0.21	0.56	0.50	0.16	0.10	0.14	470	68.2	296	43.0	35.0	53.3	140	64, 76 (47, 56)
6[(b)]	0.21	0.56	0.50	0.16	0.10	0.14	452	65.5	248	36.0	32.0	48.5	138	41, 38 (30, 28)
7[(c)]	0.22	0.52	0.50	0.15	0.10	0.12	496	72.0	300	43.5	33.0	49.0	141	87, 83 (64, 61)
8[(b)]	0.22	0.52	0.50	0.15	0.10	0.12	474	68.7	256	37.2	27.5	40.0	138	46, 41 (34, 30)
9[(c)]	0.21	0.57	0.34	0.14	0.16	0.13	476	69.0	292	42.3	33.0	53.0	138	79, 62 (58, 46)
10[(b)]	0.21	0.57	0.34	0.14	0.16	0.13	462	67.0	256	37.2	35.0	46.3	136	47, 43 (35, 32)
11[(b)]	0.20	0.56	0.42	0.07	0.11	0.07	455	66.0	259	37.5	28.0	54.1	…	54 (40)
12[(b)]	0.20	0.58	0.49	0.06	0.05	0.06	483	70.0	276	40.0	33.0	55.2	…	56, 54 (41, 40)

(a) Equivalent to ASTM A27, grade 65–35; properties from a 8 by 13 by 18 cm (3 by 5 by 7 in.) keel block; basic electric steel, deoxidized with 1.1 kg (2.5 lb) of aluminum per ton of molten metal.
(b) Annealed at 925 °C (1700 °F) for 12 h. furnace cooled. Microstructure, 30% fine pearlite, 70% primary ferrite.
(c) Normalized 925 °C (1700 °F), 8 h. air cooled, tempered 670 °C (1240 °F) 2 h. air cooled. Structure, 25% fine pearlite, 75% primary ferrite

Temper embrittlement does not occur in plain carbon steels, only in alloy steels. The degree of embrittlement varies with alloy steel composition, and certain impurity elements must be present if temper embrittlement is to occur. The embrittling impurities are, in decreasing order of influence on a weight percent basis, antimony, phosphorus, tin, and arsenic. Of these elements, phosphorus is most commonly present in alloy steels, and it has captured the most attention in research studies. Manganese and silicon also increase the susceptibility to embrittlement. Little or no temper embrittlement occurs if manganese is kept below ~0.5 wt%. Although alloy steels are ferritic in the tempered condition, fracture below the ductile-brittle transition temperature occurs along prior-austenite grain boundaries, where both alloying elements and impurity elements are concentrated.

Section Size. The effect of section size on mechanical properties must be understood by those responsible for the design and specification of castings, but it is particularly important when considering low-alloy steels. Mechanical properties of a part also are determined by the quench rate during the hardening cycle and the hardenability of the specified alloy. It is standard practice in the foundry industry to use separately cast keel blocks, for example, ASTM A 370 dimensions, to determine the mechanical properties of a heat of steel. Where a historical relationship exists, tying test bar properties to field service, this approach is a sound one.

There is always a danger, however, that the assumption is made that the mechanical properties obtained from a 3.8 cm (1.5 in.) thick test bar are equivalent to those found in a casting with a much greater thickness. In the vast majority of the cases, this is not true. When no experience exists in correlating heat treatment/alloy/section size/service/test bar properties, it is prudent to request mechanical properties from test bars roughly equivalent in section thickness to the critical sections of the castings. Alternatively, this relationship should be established through field trials under controlled conditions or by destructively testing actual castings. In many cases, the foundry may have established this information.

Failures Due to Composition. While improper chemical composition would not appear to be a defect, many problems can be traced back to composition. Errors in chemical composition can occur in three ways. The first involves selecting an inappropriate nominal (as per specification) composition for a part and the service conditions it will encounter. Off-analysis is the second way. It involves making an error in manufacturing, leading to the part or casting being produced to improper chemistry. Fortunately, with foundries reasonably equipped to produce castings, off-analysis heats are rarely a major problem. Still, they do occur. Probably the most severe problems arise with deviations in carbon content (low or high), but difficulties can occur with other elements, such as nickel, chromium, or manganese. These are generally less severe. Also, standard specifications permit a composition *range* or, in some instances, a maximum value for a given element. Slightly exceeding a nominal limit for a given element seldom results in root cause failure. Exceeding the limit of maximum value element—for example, phosphorus—can cause significantly more trouble.

Finally, trace elements, such as arsenic, tin, antimony, aluminum, nitrogen, and hydrogen, which are not a part of most specifications, can lead to unexpected failures. Unspecified elements and trace elements of importance include the gases nitrogen, oxygen, and hydrogen and the residuals, such as sulfur, phosphorus, aluminum, antimony, arsenic, and tin. Copper routinely shows up more often in analysis and may be a concern. It is well documented that high nitrogen contents in low-alloy steels can be very detrimental in combination with aluminum (which is used for deoxidation) through the formation of aluminum nitride along grain boundaries. The intergranular fracture that results is the so-called rock candy structure, which can cause catastrophic failures under severe conditions of embrittlement if not detected. A high-resolution SEM image shows the relics of the rodlike precipitate responsible for the embrittlement (Fig. 51). Scanning electron microscopy analysis for this mode of embrittlement is one way to differentiate this mode of failure from temper embrittlement or grain-boundary ferrite. In the case shown, a Ni-Cr-Mo low-alloy steel was embrittled by very high levels of aluminum (0.23%) combined with good low-nitrogen (36 ppm) content. Nitrogen also can result in strain-age embrittlement.

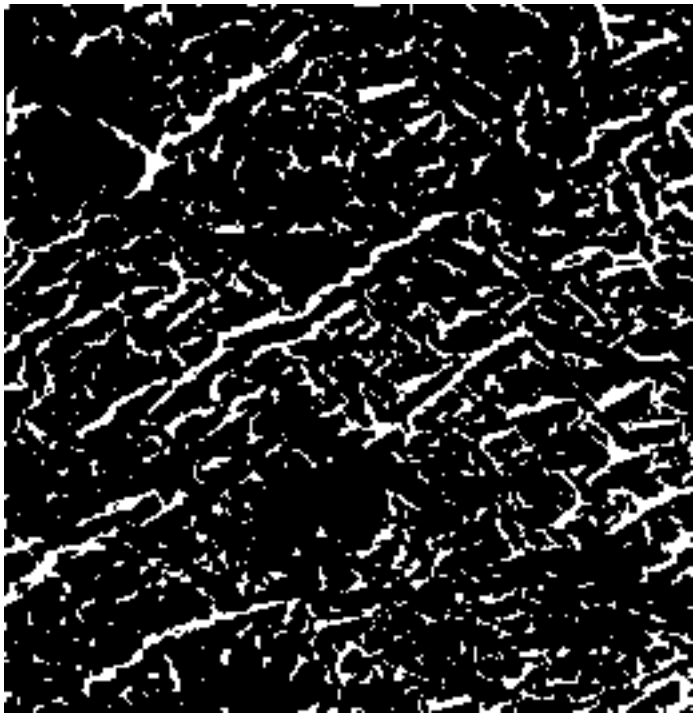


Fig. 51 Scanning electron micrograph illustrating the characteristic rodlike artifacts associated with aluminum nitride embrittlement. This characteristic appearance is confirmation of aluminum nitride embrittlement as opposed to ferrite films or temper embrittlement, which also lead to intergranular failure.

As in wrought products, hydrogen can be a cause of failure in low-alloy steel castings. Hydrogen-related defects can vary from open flakes to fisheyes that appear on fractured surfaces. This can be extremely detrimental in heavy section castings and particularly in those heat treated to high strength (yield strengths above approximately 690 MPa, or 100 ksi). Oxygen, although not such a potentially dangerous residual, can be cause for dirty steel and low mechanical properties when care is not taken in the manufacturing process to keep oxygen under control.

The effect of high sulfur contents on mechanical properties is well known and should be of decreasing importance as a defect cause. Relatively little is known of the effect of arsenic, tin, and antimony on castings, except that it is generally agreed that low levels, that is, 0.005% or lower, are important to minimize failures due to toughness degradation in heat treated high-strength castings.

High-Temperature Corrosion. Many castings are produced for high-temperature service. The control of composition inherent in castings, coupled with the ability to increase the carbon contents of standard high-temperature alloys beyond the wrought steel levels, has led to a significant number of specialty cast grades. The techniques used for analysis and the general types of cast steel failure at high-temperature modes are similar to wrought alloys, but high-temperature corrosion of low-alloy cast steels is briefly described for general oxidation, sulfidation, metallic oxide attack, and carburization. Many alloys are developed specifically for elevated temperature, and it is beyond the scope of this article to develop the reasoning for each alloy and its appropriate applications. The main intent is to examine the types of failures caused by high-temperature environments.

General Oxidation. The rate of general oxidation is one of the primary concerns of extended high-temperature service. The first endeavor in examining a high-temperature failure should be to determine what portion of the original section is left unaffected by the general oxidation. If the oxidation rate is found to be excessive, the alloy should be replaced with an alloy of greater oxidation resistance.

Temperature limits for low-alloy steels depend on various exposure and application factors. For example, short-time service may be limited to 510 °C (950 °F), while long-time service may be limited to 400 °C (750 °F). Many things begin to happen in low-alloy steels when the temperature reaches 400 °C (750 °F), such as significant loss in modulus, rapid oxidation, no scaling resistance, graphitization of cementite and pearlite, and so on. At temperatures above 540 to 650 °C (1000 to 1200 °F), general oxidation increases to rates that become unacceptable. A fluffy oxide or scale occurs and sloughs off, leading to decreased service life. A whole series

of alloys have been developed to cover a wide range of elevated-temperature applications. Most of these alloys contain chromium or chromium and nickel plus other elements, depending on the application requirements. Chromium stabilizes the carbides so that no in situ graphitization occurs and also provides scaling resistance. There also are nickel- and cobalt-base alloys used in critical high-temperature applications.

Sulfidation and Preferential Oxidation. Not all oxidation or corrosion is uniform, because oxidation and corrosion can also result in preferential attack and premature failure. Sulfidation is one form of attack at the grain boundaries. Sulfidation attack is a preferential deterioration at the grain boundaries due to a combination of sulfur with nickel in high-nickel alloys. This type of attack can be found in the stack areas of systems burning fuel oils or coal high in sulfur.

This type of preferential attack may be confirmed by electron microprobe analysis of the products found in the attacked grain boundary. One must be careful to differentiate between this type of attack enhanced by sulfur in the environment and preferential oxidation of the grain boundaries. Because most high-temperature alloys contain higher carbon than can be dissolved in the structure, these alloys tend to form eutectic carbides in grain boundaries and can show preferential oxidation of this area even in the absence of sulfur (Fig. 52).

Fig. 52 Example of preferential oxidation of the grain boundaries in a cast high-temperature alloy steel

Preferential oxidation can also be interdendritic in nature and is primarily due to the chemistry gradients across the dendrites (Fig. 53). While sulfidation attack can be intergranular and very rapid in high-nickel alloys, sulfidation also takes on rapid scaling or exfoliation characteristics in low alloys and some other alloys. This is attributed to the modification of the protective film produced on the alloy. In most cases, sulfidation resistance is improved by increasing chromium content, and aluminum and silicon have also been shown to be beneficial. Where hydrogen sulfide is the primary environmental factor, it has been found that intermediate amounts of chromium (5 to 12%) are actually deleterious and increase the rate of sulfidation. Chromium contents greater than 17% have been found to be adequate in hydrogen sulfide atmospheres, and protective layers of aluminum have been found to protect chromium steels totally from hydrogen sulfide attack.

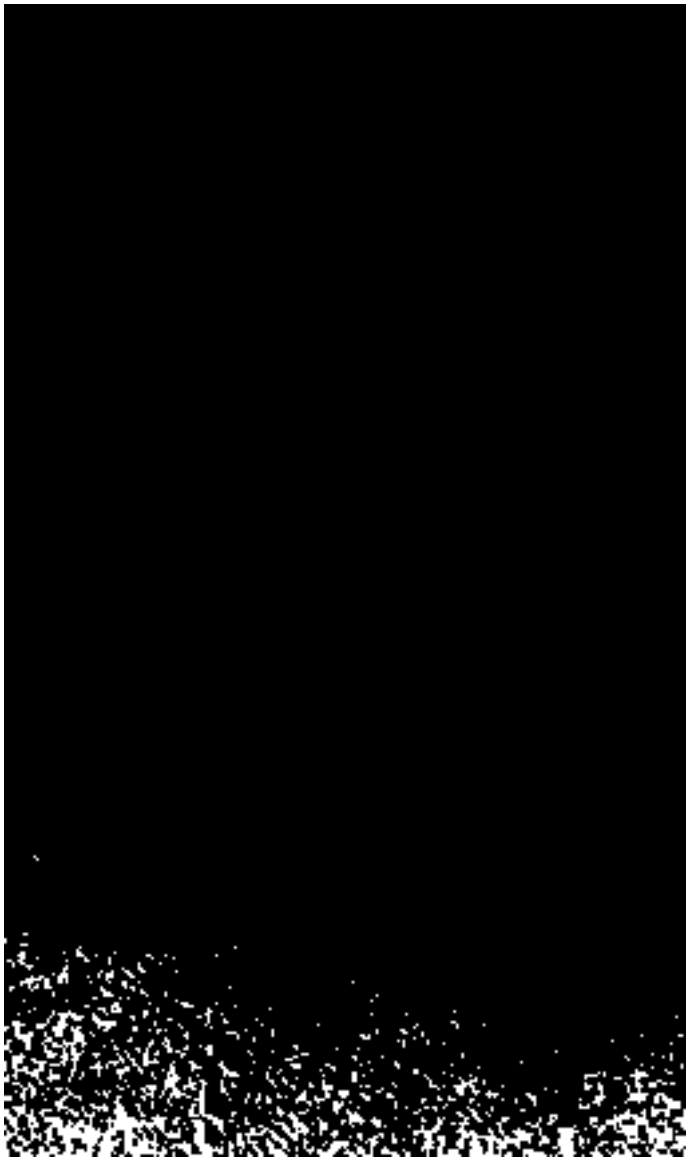


Fig. 53 Interdendritic preferential oxidation

Metal-Oxide Attack. Vanadium pentoxide, compounds containing sodium, welding fluxes, and some other reactive compounds can create a fluxing action of the normal protective oxide on the metal. These compounds tend to drop the melting temperature of the protective film and to allow the transport of oxygen or sulfur to the active metal surface. In this fashion, the contaminating compound acts as a catalyst, promoting the accelerated attack in the localized area involved. It should be noted that these fluxes can continue to cause damage even after the source has been removed, because they remain as a part of the oxide system. This type of problem is not easily handled, because there are no known metallurgical solutions. Most efforts to solve this problem require complete removal of all oxides from the affected surfaces, keeping these metal surfaces below the fusion temperature of the combined oxides, and require addition of additives to raise the fusion temperature, in the case of vanadium oxide in residual fuels.

A similar phenomenon, known as catastrophic oxidation, is encountered in those high-temperature alloys containing significant quantities of molybdenum and tungsten. These elements are widely used for enhancing creep-resistance properties in high-temperature alloys; however, they form oxides, such as molybdenum trioxide, that are gaseous at relatively low temperatures. The vaporization temperature of molybdenum appears to be 800 °C (1475 °F). Again, the molybdenum trioxide tends to change the character of what would otherwise be protective oxide films. Under the proper conditions of stagnation and temperature, alloys containing these metals produce thick, fluffy oxide coatings at a very rapid rate, resulting in both general and pitting-type metal loss. In general, this can be prevented by total protection from oxide production or by the removal of gaseous molybdenum trioxide by entrainment in a moving gas stream, including air.

Carburization is another mode of surface deterioration. At elevated temperatures, carbon in contact with the surface metal can react to form chromium carbides and, to a lesser degree, iron carbides, resulting in carbide enrichment of the surface layer of metal. This enrichment can go quite deep into the metal, with resulting increases in creep strength and decreases in metal ductility and toughness. Surface oxidation is not a primary concern under carburizing-type conditions.

A reducing atmosphere that produces carburization can remove significant amounts of chromium, allowing such phenomena as sulfidation to become significant. Also, the depletion of chromium can allow normal corrosion to occur during shutdown periods; during these cooling-down periods, liquids can wet and therefore corrode these surfaces. While the corrosion properties of these carburized surfaces are important, a loss of ductility in heavily carburized parts has been the cause of many failed parts. Increasing chromium content has been shown to slow the diffusion rate of carbon into the metal. Increasing nickel content has also been found to be beneficial. There are several alloys designed for carburizing service.

Austenitic Steels

Care is essential in heat treating more highly alloyed steel castings, especially those used for elevated-temperature service. Low-carbon austenitic stainless steels have some usefulness at elevated temperatures due to their oxidation resistance, but most of the alloys designed for elevated-temperature service have higher carbon contents than the typical 304 type of austenitic stainless steel and are thus susceptible to precipitation of grain-boundary carbides and embrittlement (see “Precipitation Embrittlement at Elevated Temperatures” in this section). An improperly heat treated stainless steel, such as Alloy Casting Institute (ACI) CF-8, corrodes faster in service due to carbide precipitate in grain boundaries, leaving other areas depleted of the chromium that gives the steel its corrosion resistance. This sensitization is frequently the cause of intergranular corrosion (see Example in this article). Too low a temperature or slow cooling from the solution-annealing temperature can produce the same detrimental effect.

Austenitic manganese (Hadfield) steels can suffer from a similar problem microstructurally, but in this case, poor wear resistance and decreased ductility are the results. Austenitic manganese steel exhibits its best properties when it gets a good water quench from 1095 °C (2000 °F). Otherwise, carbides precipitate heavily at the grain boundaries or, in severe cases, within the austenite, reducing the work-hardening capability of the matrix and lowering overall toughness and ductility. Reheating manganese steel can give a similar effect, as shown in the following example.

Example 18: Brittle Fracture of a Cast Austenitic Manganese Steel Chain Link. The chain link shown in Fig. 54(a), a part of a mechanism for transferring hot or cold steel blooms into and out of a reheating furnace, broke after approximately four months of service. The link was cast from 2% Cr austenitic manganese steel and was subjected to repeated heating to temperatures of 455 to 595 °C (850 to 1100 °F).

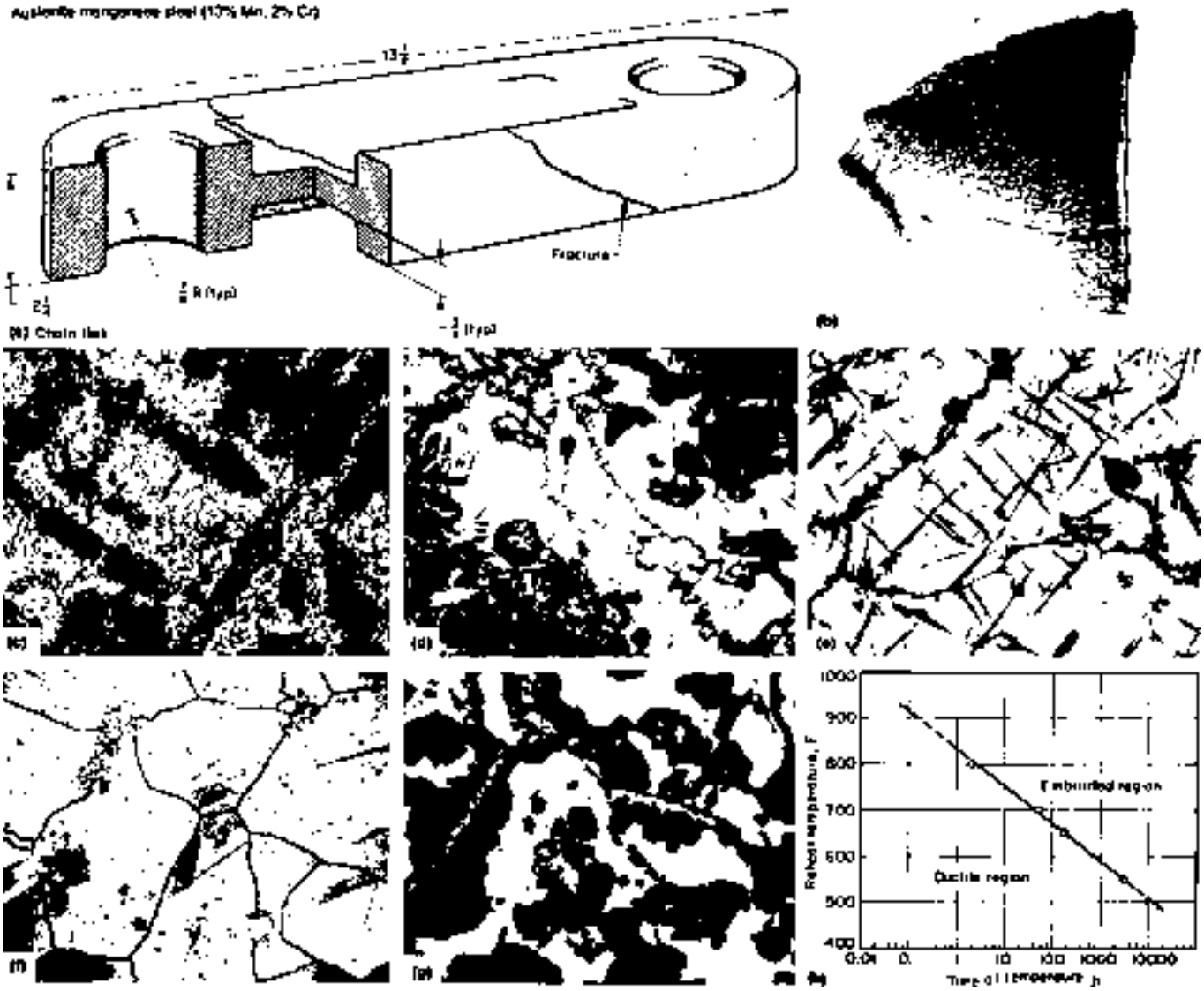


Fig. 54 Reheating-furnace chain link, sand cast from austenitic manganese steel, that failed by brittle fracture, because material was not stable at operating temperatures. (a) Chain link showing location of fracture. Dimensions given in inches. (b) Macrograph of a nital-etched specimen from an as-received chain link. 1.85×. (c) and (d) Micrographs of a nital-etched specimen from an as-received chain link. 100 and 600×, respectively. (e) Normal microstructure of as-cast standard austenitic manganese steel. 100×. (f) Micrograph of a nital-etched specimen that had been austenitized 20 min at 1095 °C (2000 °F) and air cooled. 315×. (g) Micrograph of the specimen in (f) after annealing 68 h at 480 °C (900 °F). 1000×. (h) Chart showing time at temperature needed to embrittle austenitic manganese steel after heat treating for 2 h at 1095 °C (2000 °F) and water quenching

Investigation. Chemical analysis revealed that the link had been cast from 13% Mn austenitic manganese steel to which 1.5 to 2.0% Cr was added to improve wear resistance by forming chromium carbides. Examination with a hand magnet indicated that the entire link was magnetically responsive. When properly heat treated, austenitic manganese steel is a tough, nonmagnetic alloy; however, composition changes that occur at the surface during heating may produce a magnetic skin.

Further examination of the magnetic properties of the link was performed with the aid of a Tinsley thickness gage, which is normally used for measuring the thickness of nonmagnetic coatings on ferritic steel surfaces. With this instrument, a completely nonmagnetic substance is given a rating of 10, and a magnetic material, such as plain carbon steel, is assigned a rating of 0. Specimens cut from the link were tested for magnetism after various heat treatments with the following results:

Condition[(a)]	Tinsley reading	Degree of magnetism
Web (as-received)	4.5	Partial
Flange (as-received)	4.65	Partial

Condition[a]	Tinsley reading	Degree of magnetism
1095 °C (2000 °F), WQ	10.0	Nonmagnetic
1095 °C (2000 °F), AC	10.0	Nonmagnetic
1095 °C (2000 °F), WQ; 4 h at 455 °C (850 °F)	9.0–9.6	Trace at edges
1095 °C (2000 °F), FC	8.0	Partial
1095 °C (2000 °F), FC; 68 h at 480 °C (900 °F)	4.8	Partial
1095 °C (2000 °F), AC; 68 h at 480 °C (900 °F)	4.75	Partial

(a) WQ, water quenched; AC, air cooled; FC, furnace cooled

Micrographs of specimens from the as-received link (Fig. 54c and d) display a segregated structure of complex carbides (black needlelike phase), $(FeMn)_3C$ carbides (white-on-white area), and pearlite (gray areas) in an austenite matrix, which accounts for the macrostructure shown in Fig. 54(b). Figure 54(b) is a macrograph of a polished and nital-etched specimen from the link as received at the laboratory, showing dendritic segregation and microshrinkage pores (dark spots at upper left). The segregation follows the grain boundaries, but the appearance is modified by dark oxide patches left by the nital etch.

The effects of the repeated heating of the link on the carbides can be observed by comparing the micrographs shown in Fig. 54(c) and (d) with that shown in Fig. 54(e) for a normal microstructure of as-cast standard austenitic manganese steel (1.15C-12.8Mn-0.50Si). Figure 54(e) displays a pearlite and/or martensite matrix, with carbide precipitation at grain boundaries and along crystallographic planes.

Figure 54(f) shows the microstructure of a specimen from the chain link that was austenitized for 20 min at 1095 °C (2000 °F) and air cooled. In this small specimen, the cooling rate from the austenitizing temperature was sufficient to retain the austenitic structure. The globular carbides did not dissolve at this temperature. The effect of reheating this structure to 480 °C (900 °F) for 68 h is illustrated in the micrograph shown in Fig. 54(g), which depicts carbides and austenite decomposed to pearlite in an austenite matrix. Many patches of fine pearlite formed both along the grain boundaries and within the grains. Note the similarity of the microstructure in Fig. 54(g) to that in Fig. 54(d). Brinell hardness tests performed on the link adjacent to the fracture surface yielded values of 415 HB for the web section and 363 HB for the flange; many fine cracks were found in the hardness indentations that attested to a brittle condition. These hardness values were in distinct contrast to the normal level of 180 to 200 HB expected in as-quenched austenitic manganese steel.

One question arising from the data on magnetism and from the results of metallographic analysis was whether the broken link had been properly heat treated before installation in service and magnetized as the result of service temperatures, or whether it had inadvertently been shipped from the foundry as-cast and thus partly magnetic. This question was resolved by the pronounced dendritic segregation observed in the microstructure shown in Fig. 54c. The presence of the dendritic pattern indicates that the link was not heat treated after casting, because austenitizing followed by rapid cooling would have homogenized the microstructure and largely eliminated the dendritic pattern.

Although quantitative data are lacking concerning the mechanical-property changes that result from tempering of austenitic manganese steel, metallographic data do exist that predict when embrittlement may be expected during reheating at various temperatures. Figure 54(h) shows a graph of the time at temperature needed to embrittle austenitic manganese steel after an initial heat treatment of 2 h at 1095 °C (2000 °F), followed by water quenching. The data used to plot this curve are based on the first evidences of transformation products visible under the microscope.

Conclusions. The chain link failed in a brittle manner, because the austenitic manganese steel from which it was cast became embrittled after being reheated in the temperature range of 455 to 595 °C (850 to 1100 °F) for prolonged periods of time. The alloy was not suitable for this application, because of its metallurgical instability under service conditions.

The chain links had not been solution annealed after casting. This was not as significant a factor in the failure as reheating above 425 °C (800 °F).

Precipitation Embrittlement at Elevated Temperatures. Most austenitic steels for elevated-temperature service have higher carbon contents than the typical 304 type, and they often contain such carbide-stabilizing elements as titanium or niobium to intentionally precipitate carbides for increased creep resistance. Other elements, such as molybdenum and tungsten, are often added to increase the creep resistance, and the matrix microstructure is usually austenitic for greater strength and stability.

The amount of carbides, the size, and the distribution are dependent on the operating temperature and time. However, there is a general tendency for a semicontinuous carbide film to be present in the grain boundaries (Fig. 55). This carbide film also tends to influence the overall toughness of these materials and the general fracture mode both at elevated temperatures and at ambient temperature (Fig. 56). Weld repair of these alloys can become difficult, especially if the surface material has become contaminated, and they often require a high-temperature heat treatment to dissolve carbides (1095 °C, or 2000 °F, or higher), with relatively fast cooling to improve the ductility to allow for acceptable welding characteristics.

Fig. 55 Cast heat-resistant alloy HH, type II, showing the effects of long-term exposure to temperatures between 705 and 925 °C (1300 and 1700 °F)

Fig. 56 Cast heat-resistant alloy HH, type II, showing cracking through intergranular carbides

Intermetallics, such as σ phase, are important sources of failures in high-temperature materials. This phase is a particular combination of iron and chromium that produces a hard, brittle second phase in the temperature range of 480 to 955 °C (900 to 1750 °F). Chromium contents above 20%, especially with the addition of significant amounts of silicon, molybdenum, or tungsten in an iron-base material, tend to allow the formation of this phase. It tends to precipitate after long exposures at the indicated temperature range, and, most of the time, it can be eliminated by reasonable cooling rates from a heat treatment at or above 995 °C (1850 °F).

Sigma phase tends to precipitate in those regions of high chromium content, such as the chromium carbides in the grain boundaries and (where present) within ferrite pools. A relatively small quantity of σ phase, when it is nearly continuous in a grain boundary, can lead to very early failure of high-temperature parts. It is important to understand that σ phase cannot undergo any significant plastic deformation; instead, it fractures even at relatively low strain levels. This is true at elevated temperatures but even more so at ambient temperatures. Where toughness and ductility are an important part of the system design, σ phase cannot be tolerated (Fig. 57). However, there are applications where σ phase has been present and has not played a significant role regarding failure.

Fig. 57 Sigma (σ) phase in cast heat-resistant alloy HH, type II. Intermetallic phases, such as σ, can greatly reduce the ductility of many high-temperature alloys in service at temperatures from 480 to 955 °C (900 to 1750 °F).

Corrosion-Resistant Castings

The fundamentals of corrosion problems are the same for both cast and wrought components with a few exceptions, although casting defects can play a major role in corrosion-related failures. Porosity can easily act as an initiation site for localized corrosion or as a stress concentrator leading to SCC. The same is true for other defects, such as hot tears, cold laps, oxide laps, and inclusions. Although wrought materials have their own weaknesses, they are free of many of the defects that may be associated with a casting.

It is significant to note that castings also have some striking advantages over comparable alloys in wrought form. Cast 18-8-type stainless steels—for example, ACI CF-8 and CFR-8M—contain small amounts of ferrite in cast form (typically 5 to 15%), while their nearest wrought equivalents (AISI type 304 and 316, respectively) do not. This yields a material with enhanced resistance to chloride SCC due to the crack-arrest (blunting) properties of this duplex microstructure. Inclusions in cast alloys tend to be spherical; these may be less harmful in terms of stress concentration than the planar or elongated inclusions characteristic of wrought materials. Finally, foundries offer relatively small heat sizes, which allow considerable freedom to modify the composition of an alloy to meet a specific application.

A survey from the 1980s (Ref 42) documents responses from surveyed foundries, pump and valve manufacturers, and others on the causes of premature failure in castings. The following causes were noted:

- Intergranular corrosion accounts for about 50% of the premature service failures of chromium-bearing corrosion-resistant castings.

- Solidification-type discontinuities (hot tears, porosity, and shrinkage) account for about 30 to 40% of the premature service failures of corrosion-resistant castings.
- Some premature service failures have been associated with repair welds.
- Hot tears and cracks in CN-7M, CD-4MCu, and some of the nickel-base alloys are particularly troublesome to the producing foundries and may be responsible for some premature service failures. Although information is available on the cause of hot tearing, improvements in the technology are probably required to produce acceptable castings more consistently.
- Crevice corrosion and SCC were seldom cited as the cause of premature service failures.

Intergranular corrosion is one of the most common causes of failure in cast corrosion-resistant components. In most cases, its cause is some step in its thermal history that leaves it susceptible. This could include missing the essential solution-annealing heat treatment or the heat associated with an un-heat-treated weld. Both welding and improper heat treatment can give rise to sensitization. This involves chromium-rich carbide precipitation at grain boundaries, accompanied by depletion of chromium in the adjacent region. The low-chromium area is then preferentially attacked by the corrosive solution. Intermetallic second phases, such as σ phase, can also give rise to preferential grain-boundary corrosion.

Example 19: Corrosion of a Neck Fitting. A neck fitting exhibiting extreme corrosion with large, deeply pitted areas was examined. It had been in service in a sulfite digester at 140 °C (285 °F) and 689 kPa (100 psi). The liquor was calcium bisulfite, with 9% total weight of SO_2 at a pH of 1.5 to 1.8. Chloride content was reported to be low. The composition was a cast equivalent of AISI type 317 (18Cr-10Ni-3.2Mo). A sample was taken from the deeply attacked area (Fig. 58). The corrosion was occurring along grain boundaries. Micrographs from these regions showed σ phase present in the ferrite (Fig. 58b) and also carbides (Fig. 58c). A 10 N KOH electrolytic etch was used to bring out the σ phase, while Murakami's reagent showed the carbides at the grain boundaries.

(a)

(b)

(c)

Fig. 58 Intergranular corrosion. (a) Sample from a cast stainless steel neck fitting. (b) Region adjacent to the intergranular corrosion revealing extensive σ-phase precipitation at grain boundaries; electrolytic etching using 10 N KOH. (c) Same area as (b) after repolishing and etching with Murakami's reagent to reveal substantial grain-boundary carbide precipitation

Conclusions. It appears that this casting never received a proper solution anneal. Solution treating is essential for CF-8M castings to avoid both carbide and intermetallic precipitates. Corrosion-screening tests in accordance with ASTM A 262 may be needed to ensure adequate corrosion resistance.

Reference cited in this section

42. D.B. Roach and F.H. Beck, Performance and Reliability of Corrosion-Resistant Alloy Castings, *MTI Manual No. 5*, The Materials Technology Institute at The Chemical Process Industries, Inc., Columbus, OH, 1981

Cast Aluminum Alloys

Aluminum alloy castings are affected by discontinuities such as gas porosity, microinterdendritic shrinkage, surface oxide concentrations, hot tears, and dross inclusions such as ferrous castings. For example, discontinuities provide sites for early crack initiation and thus shorter fatigue life than for wrought forms of aluminum. Casting imperfections can also lead to brittle fracture, as illustrated in Example 20 at the end of this section. However, aluminum cast alloys have specific concerns, such as gas porosity from hydrogen and other compositional factors, as briefly described subsequently.

Hydrogen Porosity. Two types or forms of hydrogen porosity may occur in cast aluminum. Of greater importance is interdendritic porosity, which is encountered when hydrogen contents are sufficiently high that hydrogen rejected at the solidification front results in solution pressures above atmospheric. Secondary (micron-sized) porosity occurs when dissolved hydrogen contents are low, and void formation is characteristically subcritical. During a high-temperature heat treatment, macroscopic interdendritic porosity may become rounded.

Finely distributed hydrogen porosity may not always be undesirable. Hydrogen precipitation may alter the form and distribution of shrinkage porosity in poorly fed parts or part sections. Shrinkage is generally more harmful to casting properties. In isolated cases, hydrogen may actually be intentionally introduced and controlled in specific concentrations compatible with the application requirements of the casting in order to promote superficial soundness. Nevertheless, hydrogen porosity adversely affects mechanical properties in a manner that varies with the alloy.

It is often assumed that hydrogen may be desirable or tolerable in pressure-tight applications. The assumption is that hydrogen porosity is always present in the cast structure as integrally enclosed, rounded voids. In fact, hydrogen porosity may occur as rounded or elongated voids and, in the presence of shrinkage, may decrease rather than increase resistance to pressure leakage.

Gravity Casting. Regardless of the types of melting and holding furnaces and the particular gravity casting process used, there is great concern for reducing or eliminating dissolved hydrogen and entrained oxides. These procedures are less frequently employed for pressure die casting, in which concerns are focused on the dominant process-related causes of casting unsoundness, namely, entrapped gas and pouring injection-associated inclusions.

Sensitivity to melt quality varies with the casting process and part design and necessitates special consideration of relevant criteria for each application. In general, the melt is processed to achieve hydrogen reductions and the removal of oxides to meet specific casting requirements. Modification and grain-refiner additions are made as appropriate to the given alloy and end product.

Die Casting. Different melt preparation practices are employed in die casting operations, because process-related conditions are more dominant in the control of product quality than those controlled by melt treatment. For this reason, degassing for the removal of hydrogen, grain refinement, and modification or silicon refinement in the case of hypereutectic silicon alloys are often intentionally neglected. The movement toward

higher-integrity die castings has brought into focus the importance of the same melt quality parameters established and used in the gravity casting of aluminum alloys.

In high-production die casting operations, the consumption of internal and external scrap is of primary importance in reducing base metal costs for the predominantly secondary alloy compositions that are consumed. Scrap crushing, shredding, and pretreatment of various types precede melting, often in efficient induction systems. Oxides entrained in the melt as a result of this sequence of operations are dealt with through the use of salts and/or reactive gas fluxing. Melt treatment is typically confined to this and to rudimentary fluxing in holding furnaces to remove gross oxide and to facilitate the maintenance of minimum furnace cleanliness.

A concern in die casting is the formation of complex intermetallics that are insoluble at melt-holding temperatures and/or precipitate under holding conditions or during transfer to and injection from the hot or cold chamber. These intermetallics (sludge) affect furnaces, transfer systems, and, by inclusion, the quality of the castings produced. Die casters are familiar with composition limits that prevent sludge formation. A common rule is that iron content plus two times manganese content plus three times chromium content should not exceed the sum of 1.7%. This limit is arbitrary and inexact, it is often assigned values from 1.5 through 1.9%, and it is subject to the specific composition and actual minimum process temperature.

The corrosion resistance of aluminum casting alloys in service depends on such factors as alloy composition and the environment to which the alloy is exposed. Corrosion can be minimized by designing the application to include protective measures and by selecting the alloy that economically meets all design requirements, including those pertaining to corrosion. In many applications, the corrosion resistance of various casting alloys is similar enough that alloy selection can be based on other factors. Only in severe environments, such as marine atmospheres or heavily polluted urban industrial atmospheres, do differences in the general corrosion resistances of various aluminum casting alloys become apparent.

Example 20: Premature Torquing Failures of Cast A356 Aluminum Actuators Due to Hot Tear (Ref 43). Two investment-cast A356 aluminum alloy actuators used for handles on passenger doors of commercial aircraft fractured during torquing at less than the design load. The actuator castings were being torque tested to a torque of 365 N · m (3240 lbf · in.) during static quality-control testing but failed prematurely at 180 and 200 N · m (1600 and 1800 lbf · in.), respectively. The cracks in both castings were at similar locations, through a machined side hole (Fig. 59). Fractography revealed that the cracks originated in hot tear locations in the castings.

Fig. 59 Crack (arrows) in casting that developed during torque testing. ~1.1×

Investigation. The actuator castings were manufactured in accordance with Aerospace Material Specification 4218, class 2, grade C and were not 100% radiographically inspected. The chemical composition requirements

for A356 aluminum alloy were met (nominal composition, Al-7.0Si-0.35Mg). Visual examination showed that cracking had occurred through a machined side hole.

The cracks were opened to expose the fracture surfaces. The crack origin areas were slightly dull in appearance compared with the remainder of the fracture surface, and they were present on the inner-diameter surfaces of the large-diameter hole along the length of the actuator casting. The exposed fracture surfaces were examined in a SEM, and both samples revealed similar fractographic features. The dull crack origin areas revealed an as-solidified free surface with a wavy appearance and some ligaments at the subgrain boundaries that fractured during the torquing process (Fig. 60, 61, 62). These features indicate the presence of a hot tear. Microporosity was also observed at random locations on the fracture surface (Fig. 60b), which is typical in this type of aluminum casting. The fracture features away from the hot tear locations revealed a typical macroscale brittle overload fracture (Fig. 63).

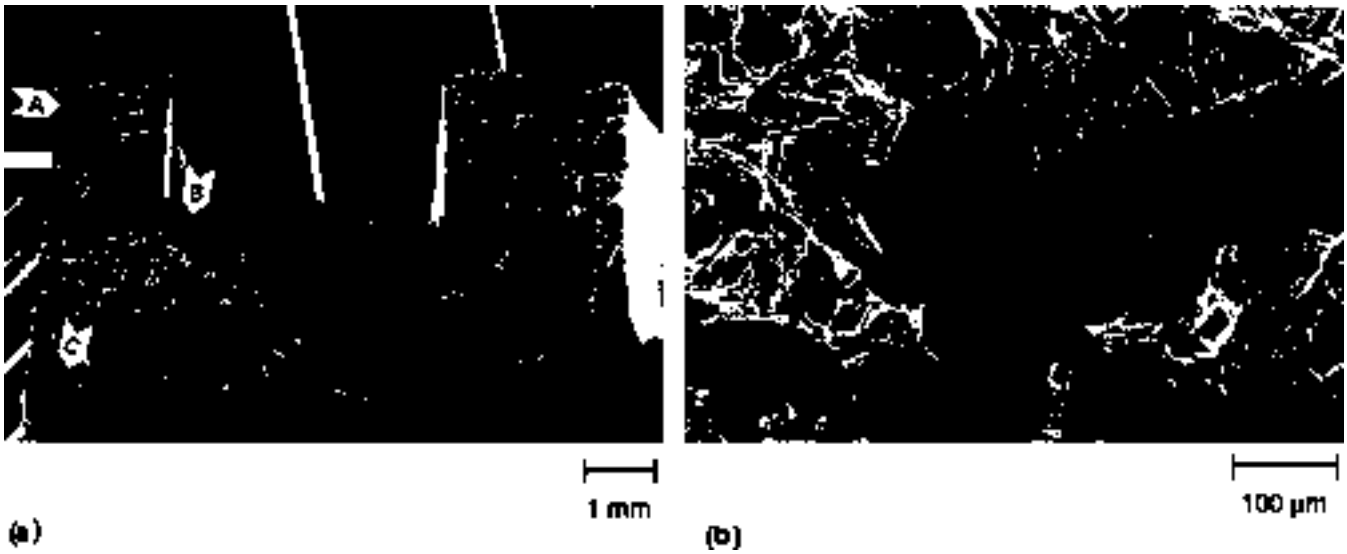


Fig. 60 SEM micrograph showing (a) overall fracture surface in crack initiation area and (b) detail at area C in (a), showing shrinkage porosity

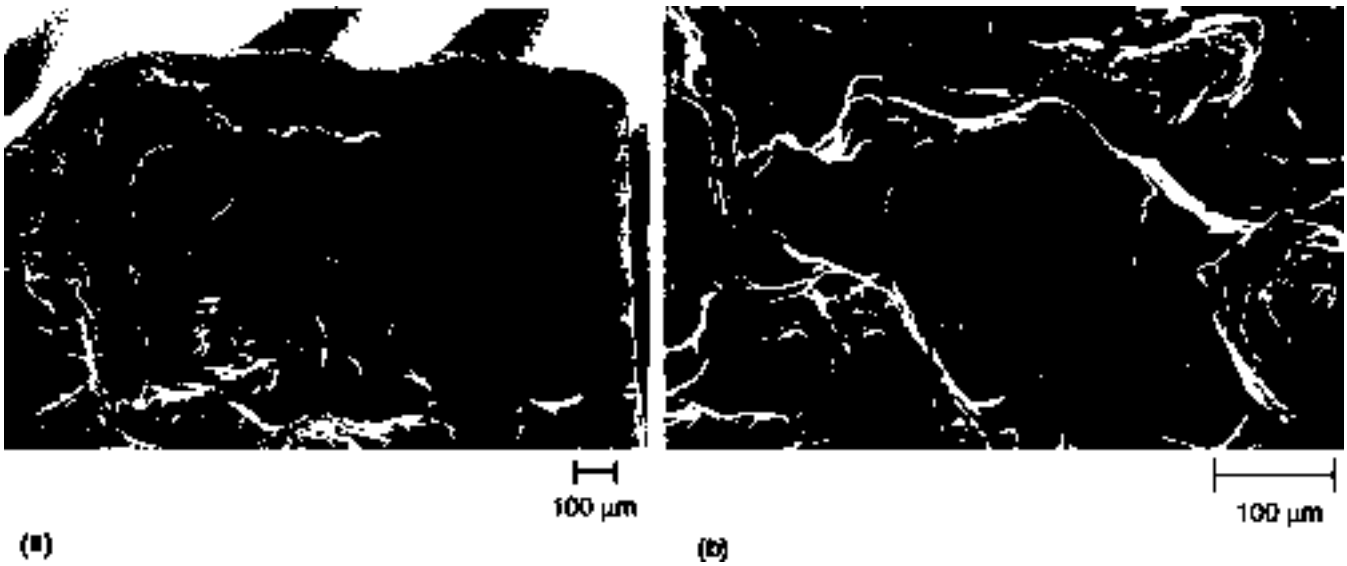


Fig. 61 SEM of hot tear in region A of Fig. 60

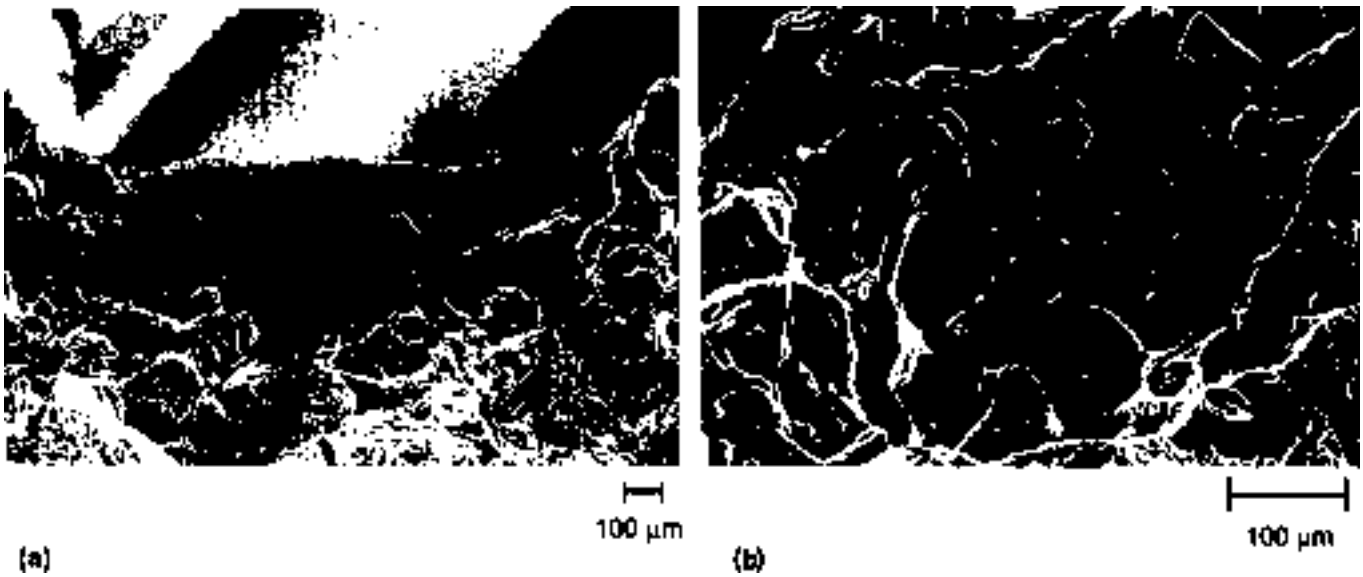


Fig. 62 SEM of hot tear in region B of Fig. 60

Fig. 63 SEM micrograph from overload region showing typical surface of brittle fracture

An energy-dispersive x-ray (EDX) microprobe module attached to the SEM microscope analyzed the elemental chemical composition in the hot tear regions and in the brittle overload areas. The EDX microprobe spectra (Fig. 64) obtained in a hot tear area and in an overload fracture area revealed the presence of aluminum and silicon, the principal constituents of A356 aluminum alloy. However, the hot tear area (Fig. 64a) exhibited a much higher silicon-to-aluminum ratio compared with the overload fracture area (Fig. 64b). The liquid in the interdendritic and grain-boundary regions becomes silicon enriched during solidification, and hot tears form in these regions by the stresses generated by solidification shrinkage. The overload fracture also follows the silicon-rich low-toughness phase and exhibits a higher silicon-to-aluminum ratio than expected from the base alloy composition.

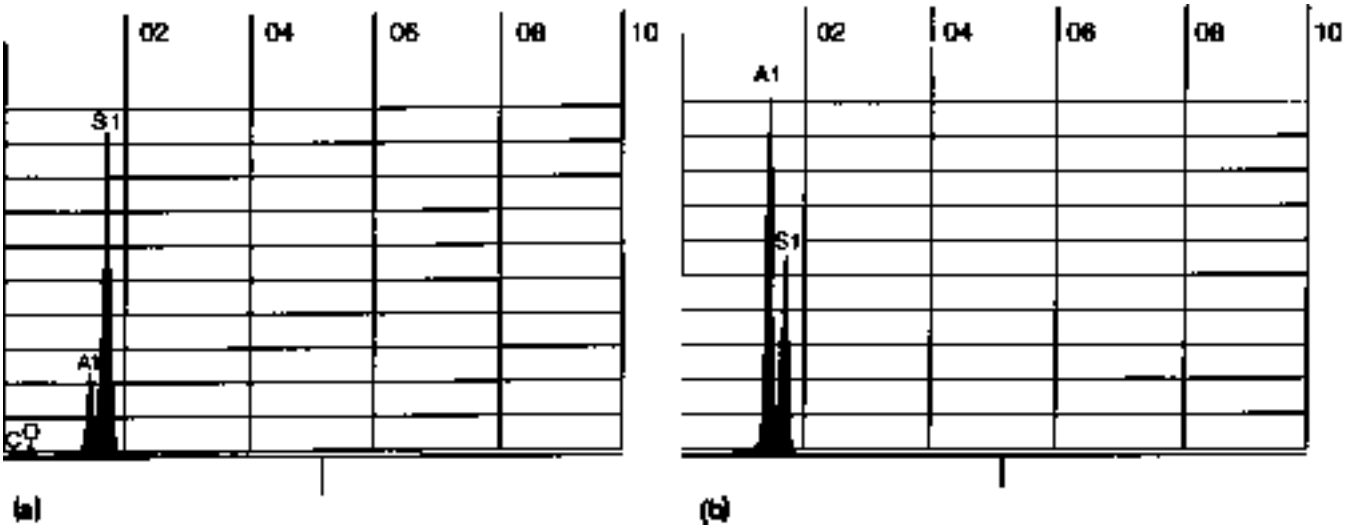


Fig. 64 Energy-dispersive x-ray spectra (windowless detector, 20kV electron beam) for (a) hot tear zone and (b) overload fracture zone

Metallographic examination of a section obtained through the origin area adjacent to the hot tear revealed no microstructural differences compared with the rest of the casting. The metallurgical microstructure contained a network of silicon particles that formed in the interdendritic aluminum-silicon eutectic, typical of A356-T61 castings. Normal amounts of shrinkage pores were also present. In summary, no microstructural anomalies were found that could be related to the observed failure.

Conclusion. The strength of the casting was compromised by the presence of hot tear casting defects that occurred during the solidification process. In both failed samples, the hot tears were present near the inner-diameter surface of the large-diameter longitudinal hole. The presence of a machined side hole (through which the crack had propagated) rendered this area mechanically weaker, causing the overload fracture to initiate because of the already present hot tear defects.

Example 21: Failure of a Sprocket Drive Wheel in a Tracked All-Terrain Vehicle (Ref 44). A sprocket drive wheel in a tracked all-terrain vehicle failed. The part was sand cast in alloy LM6M of British Standard (BS) 1490 (alloy A413.0). The wheel, approximately 380 mm (15 in.) in diameter, showed extensive cracking around each of the six bolt holes and evidence of considerable deformation in this region of the wheel.

Sections were taken from the cracked region and prepared for metallographic examination, and the structures were typical of modified aluminum-silicon alloys. Some variations in the fineness of the silicon eutectic were apparent, presumably as a result of various chilling rates.

The structure also contained a relatively high concentration of a thin, platelike constituent (appearing as needles in the cross sections shown in Fig. 65). Attempts were made to establish its composition but were not conclusive. The usual platelike constituent found in these alloys is known as β(Al-Fe-Si), and the amount present is proportional to the iron content. In this case, the etching reaction of the needles did not correspond to that of β(Al-Fe-Si), and it was presumed that other minor impurities, such as copper and manganese, were also present. In fact, many of the plates had a duplex structure, as shown at higher magnification in Fig. 66, which is from a section through a crack. Figure 66 also shows the strong influence that the plates had on the path of the fracture. Most of the fracture surfaces shown are either through a plate or at the interface between a plate and the aluminum matrix. Also, several incipient cracks are associated with plates.

Fig. 65 Section showing comparatively coarse eutectic structure and (dark) platelike constituents. Etched in 20% H_2SO_4 at 70 °C (160 °F) at 134×

Fig. 66 Section through fracture showing influence of platelike constituent. Etched in 20% H_2SO_4 at 70 °C (160 °F) at ~235×

Drillings were taken from the wheel and submitted for chemical analysis, and the composition of the wheel was within specification limits. The iron content, although below the specified maximum, was higher than usual for this class of material and may have had a deleterious effect on the ductility of the alloy. Four tensile test pieces, 75 mm (3 in.) long and 6.4 mm ($\frac{1}{4}$ in.) in diameter, were taken from the flange between the bolt holes, and the results of tensile tests are shown in Table 5, together with the minimum specified properties for separately cast test bars of alloy LM6M.

Table 5 Tensile test results for alloy LM6M (BS 1490) from failed sprocket drive wheel

Specimen	Ultimate tensile strength		Yield strength				Elongation (4*D*), %
			0.2%		0.1%		
	MPa	ksi	MPa	ksi	MPa	ksi	
A	154.4	22.4	85.5	12.4	84.1	12.2	1.0
B	172.4	25.0	93.8	13.6	80.7	11.7	2.0
C	156.5	22.7	91.7	13.3	82.0	11.9	2.0
D	155.8	22.6	86.9	12.6	75.8	11.0	2.0

Specimen	Ultimate tensile strength		Yield strength				Elongation (4*D*), %
			0.2%		0.1%		
	MPa	ksi	MPa	ksi	MPa	ksi	
Average	160.0	23.2	88.3	12.8	80.7	11.7	2.0
BS 1490:1963 (LM6M) minimum specification	162.0	23.5	…	…	53.8	7.8[(a)]	5.0

(a) Minimum not specified: for information only

There is no provision in BS 1490 for testing bars from castings, but, in North America, it is common for the minimum specifications in the casting to be 75% of the ultimate strength and 25% of the elongation of the corresponding separately cast test bars. Using that criterion, the properties of the sprocket wheel would be considered satisfactory in alloy LM6M. Tensile testing showed that the casting was as strong as could be expected for this alloy. However, somewhat better ductility might have resulted if the iron content had been lower and, consequently, less of the embrittling, platelike constituent had been present. However, the extensive cracking and evidence of substantial deformation in the region around each of the six bolt holes indicated that failure of the part resulted from gross overload. More extensive deformation before fracture would have been of little or no value.

Conclusion and Recommendations. Most probable cause is that overload resulted in considerable deformation and cracking and thus ultimate failure, although iron contents may have contributed to crack susceptibility (Ref 44). However, several other elements have low solubility and result in acicular phases in the microstructure. In addition, when the β(Al-Fe-Si) phase was not indicated, the specimens were not repolished and etched with different etchants, as part of standard phase-identification procedure for aluminum alloys. Also, the possibility of energy-dispersive spectroscopy or wavelength-dispersive spectroscopy analysis in the SEM should have been considered. It is not clear that a cleaner alloy, based on the incomplete analysis, is a conclusive recommendation.

References cited in this section

43. S. Ensha and A. Kumar, Premature Torquing Failures of Cast A356 Aluminum Actuators, *Handbook of Failure Analysis Case Histories,* Vol 1, ASM International, 1992 p 47–50

44. W.A. Pollard, Failure of a Sprocket Drive Wheel in a Tracked All-Terrain Vehicle, *Handbook of Failure Analysis Case Histories,* Vol 2, ASM International, 1993 p 58–60

Welding

In addition to casting defects, failures can occur from various secondary or postcasting operations such as machining, grinding, heat treatment, and surface treatment. Machining of castings, for example, can introduce stress raisers; tool marks, notches, and sharp edges such as those around drill holes are among the most common stress raisers produced during machining. After machining, these cast parts are hardened by heat treatment, then ground to close tolerances. Cracks can result from grinding burns caused by localized frictional heat during grinding. The effects of machining and grinding, however, also apply to wrought products and are not peculiar to castings.

Welding is an integral part of casting production. Foundries often perform modest amounts of welding on castings (always prior to heat treatment) to correct surface defects. This action returns the casting to the “drawing-specified” condition and is thus called “rework” (as opposed to “weld repair”). Weld repair differs from rework in that postweld heat treatment is not done; it is sometimes done for maintenance to render damaged castings usable. Of course, mechanical properties can be greatly affected as a result. For example, in alloy A357-T6, yield strength may be reduced by 50% or more by repair welding. Consequently, weld repairs are performed only in cases where service stresses are low.

Welding rework of castings is common for iron, steels, and nonferrous alloy such as aluminum. Virtually all of the fusion-welding processes are used for welding steel castings. However, casting failures resulting from welding defects are not uncommon. Such welding defects as porosity, trapped slag, undercut, cracks, and lack

of fusion often act as fracture-initiation sites because of their stress-concentration effects. Methods of combating these types of failure include welding with qualified procedures and welders as well as various forms of nondestructive examination.

Postweld heat treatment should include a proper stress-relief heat treatment to relieve welding stresses and improve resistance to cracking. A common error in postweld heat treatment of some alloy steels is to air cool from the postweld heat treatment temperature. When this is done, stresses are lowered, but the material may also suffer temper embrittlement. Because temper embrittlement greatly increases susceptibility to hydrogen-assisted cracking, the net result of air cooling from the postweld heat treatment temperature can be increased risk of failure.

Example 22: Hot Cracking of a Pump Impeller from a Nuclear Plant (Ref 45). Liquid penetrant inspection of an ASTM A 296 grade CA-15 residual heat-removal pump impeller from a nuclear plant revealed a cracklike indication approximately 125 mm (5 in.) long near the upper wear ring (Fig. 67). The direction of this indication approximated the outer contour of the wear ring. There was no evidence of crack indications on the wear rings. The crack and three hardness specimens were then removed from the impeller.

Fig. 67 Close-up view of penetrant-enhanced crack in impeller

The cracking was predominantly interdendritic and was associated almost entirely with areas of weld repair. The observation of cracks in the repair area, but not on the exterior surfaces of a weld repair, led to the conclusion that failure was caused by hot cracking related to original weld repairs performed on the impeller casting. The failure might have been avoided if dye-penetrant inspection had been performed after repair welding.

Example 23: Failure Analysis of a Cast- Steel Crosshead (Ref 46). An ASTM A 48 grade 105-85 cast-steel crosshead of an industrial compressor, which represented a new design, was in service for less than two years. It was returned by the end user when a crack was detected during routine maintenance. The manufacturing sequence that the foundry used before shipping the castings consisted of casting, normalizing, preheating for weld repairing of areas of shrinkage porosity, normalizing, quenching and tempering, drawing to the required Brinell hardness, and nondestructive testing (dry magnetic-particle inspection), and machining.

Investigation. The fracture surface from one side of the failed crosshead was sawed off to obtain specimens for optical microstructural examination. Figure 68 clearly shows the dark region where the crack initiated and propagated radially away from this region. Figure 69 shows an enlarged fractograph from the other side of the crosshead. Beach marks appear to radiate from the weld-repair region. The cracks initiated at multiple sites and finally formed a single front. The twisted region at the left end of the specimen appears to be where mating surfaces were separated to retrieve the samples for this investigation.

Fig. 68 Fracture surface of failed crosshead showing slag inclusions (dark) and fatigue beach marks

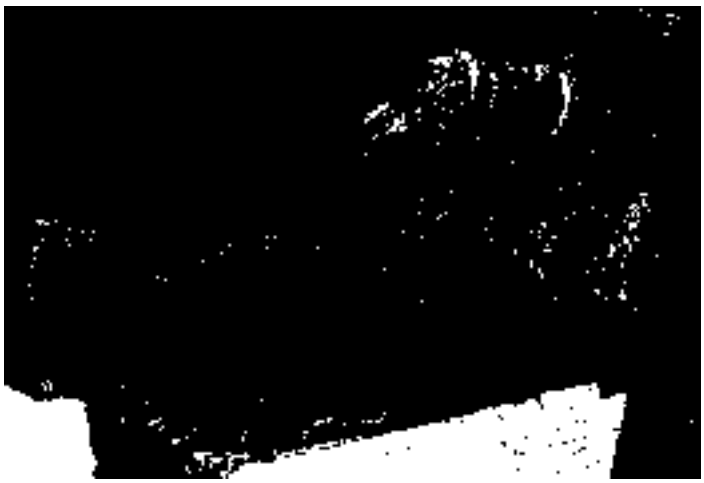


Fig. 69 Close-up view of fracture surface at the other web. Note beach marks.

The fracture surface was examined by SEM. Figure 70 is a SEM fractograph showing a crack-initiation site and a slag inclusion (dark area). Standard energy-dispersive spectroscopy showed one of the slag inclusions to contain a high titanium content (Fig. 71), which was suspected to have come from the welding flux.

Fig. 70 Close-up view of the crack-initiation area in Fig. 69, showing slag inclusions (dark spots). ~10×

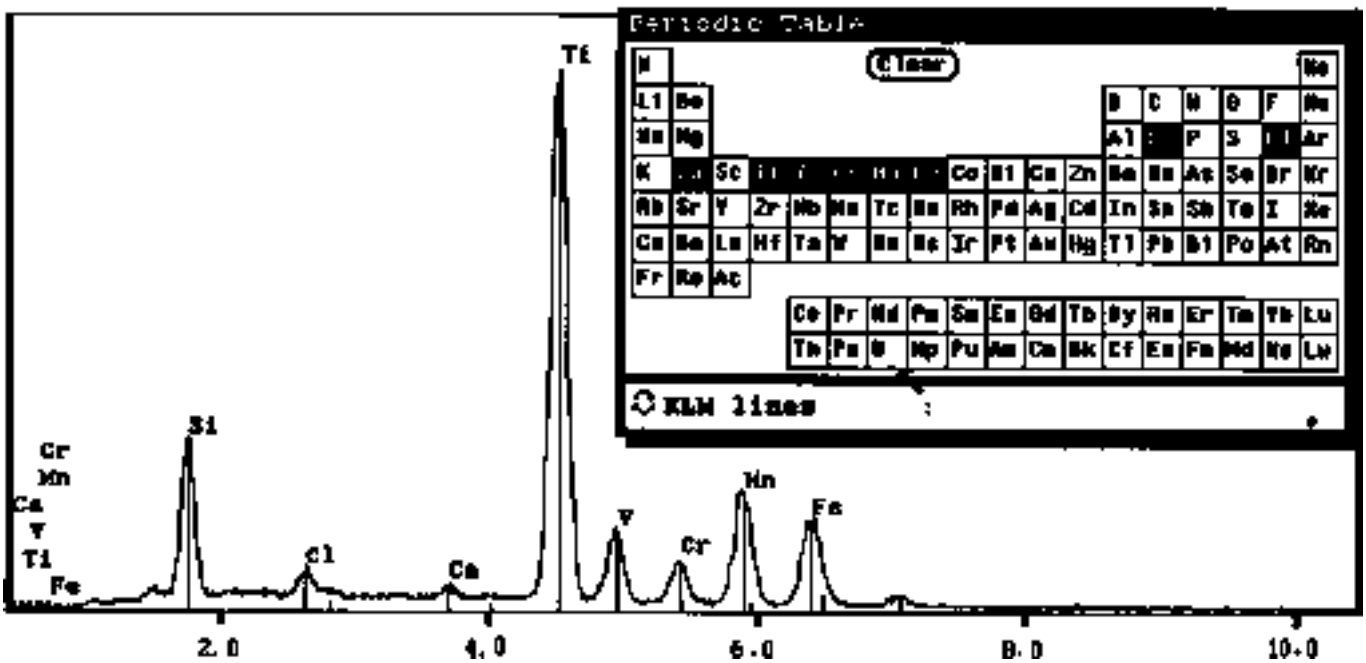


Fig. 71 Energy-dispersive spectroscopy spectrum indicating presence of slag

Conclusions. On the basis of the observations reported and available background information, it was concluded that failure began with the initiation of cracks at slag inclusions and sharp fillets in weld-repair areas in the casting. The weld-repair procedures were unsatisfactory. The cracks propagated by fatigue, as determined by the observation of beach marks. The casting quality was judged unacceptable because of the presence of excessive shrinkage porosity and slag inclusions.

Recommendations. It was recommended that crosshead castings be properly inspected before machining. Revision of foundry practice to improve the weld-repair technique and to reduce or eliminate porosity was also recommended.

References cited in this section

45. C. Czajkowski, Hot Cracking of a Pump Impeller from a Nuclear Plant, *Handbook of Case Histories in Failure Analysis,* Vol 1, ASM International, 1992 p 273–277

46. M. Chaudhari, Failure Analysis of Cast Steel Crosshead, *Handbook of Case Histories in Failure Analysis,* Vol 1, ASM International, 1992 p 291

Acknowledgments

The following contributors provided case histories or other useful information: Charles O. Cambridge, Navistar International, Inc.; George M. Goodrich, Taussig Associates, Inc.; Renita G. King, Cummins Engine Company; Dilip K. Subramanyam, Abex Corporation; John F. Wallace, Case Western Reserve University; Charles F. Walton, professional engineer; Peter F. Wieser, Case Western Reserve University; Frank J. Worzala, University of Wisconsin; Carl J. Czajkowski, Department of Nuclear Energy, Brookhaven National Laboratory, Upton, New York; and Mohan D. Chaundhari, Columbus Metallurgical Services, Inc., Columbus, Ohio.

References

1. T. Piwonka, Design for Casting, *Materials, Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 723

2. E.L. Rooy, Hydrogen: The One-Third Solution, *AFS Trans.,* Vol 101, 1993, p 961

3. N. Roy, A.M. Samuel, and F.H. Samuel, Porosity Formation in Al-9 Wt Pct Mg—3 Wt Pct Cu Alloy Systems: Metallographic Observations, *Met. Mater. Trans.,* Vol 27A, Feb 1996, p 415

4. B. Skallerud, T. Iveland, and G. Härkegård, Fatigue Life Assessment of Aluminum Alloys with Casting Defects, *Eng. Fract. Mech.,* Vol 44 (No. 6), 1993, p 857

5. M.J. Couper, A.E. Neeson, and J.R. Griffiths, Casting Defects and the Fatigue Behaviour of an Aluminium Casting Alloy, *Fatigue Fract. Eng. Mater. Struct.,* Vol 13 (No. 3), 1990, p 213

6. J.C. Ting and F.V. Lawrence, Jr., Modeling the Long-Life Fatigue Behavior of a Cast Aluminum Alloy, *Fatigue Fract. Eng. Mater. Struct.,* Vol 16 (No. 6), 1993, p 631

7. J.T. Berry, Linking Solidification Conditions and Mechanical Behavior in Al Castings—A Quarter Century of Evolving Evidence, *AFS Trans.,* Vol 103, 1995, p 837

8. M.K. Surappa, E. Blank, and J.C. Jaquet, Effect of Macro-Porosity on the Strength and Ductility of Cast Al-7Si-0.3Mg Alloy, *Scr. Metall.,* Vol 20, 1986, p 1281

9. C.H. Cáceres, On the Effect of Macroporosity on the Tensile Properties of the Al-7%Si-0.4%Mg Casting Alloy, *Scr. Metall.,* 1994, submitted for publication

10. C.M. Sonsino and K. Dietrich, Einflub der Porosität auf das Schwingfestigkeitverhalten von Aluminium-Gubwerkstoffen, Giessereiforschung, Vol 43 (No. 3 and 4), 1992, p 119–140

11. P.M. Unterweiser, Ed., *Case Histories in Failure Analysis,* American Society for Metals, 1979

12. C.E. Bates and C. Wanstall, Clean Steel Castings, “Metalcasting Competitiveness Research,” final report, DOE/ID/13163-1 (DE95016652), Department of Energy, Aug 1994, p 51

13. I.P. Volchok, Non-Metallic Inclusions and the Failure of Ferritic-Pearlitic Cast Steel, *Cast Met.,* Vol 6 (No. 3), 1993, p 162

14. P. Heuler, C. Berger, and J. Motz, Fatigue Behaviour of Steel Casting Containing Near-Surface Defects, *Fatigue Fract. Eng. Mater. Struct.,* Vol 16 (No. 1), 1992, p 115

15. J. Campbell, J. Runyoro, and S.M.A. Boutorabi, Critical Gate Velocities for Film-Forming Alloys, *AFS Trans.,* Vol 100, 1992, p 225

16. N.R. Green and J. Campbell, Proc. Spring 1993 Meeting 4–7 May, 1993 (Strasbourg, France), European Division Materials Research Society

17. N.R. Green and J. Campbell, The Influence of Oxide Film Filling Defects on the Strength of Al-7Si-Mg Alloy Castings, *AFS Trans.,* Vol 102, 1994, p 341

18. C. Nyahumwa, N.R. Green, and J. Campbell, "The Effect of Oxide Film Filling Defects on the Fatigue Life Distributions of Al-7Si-Mg Alloy Castings," presented at the International Symposium on Solidification Science and Processing, Dec 1995 (Honolulu, HI), Japan Institute of Metals and TMS

19. J. Campbell, The Mechanical Strength of Non-Ferrous Castings, *Proc. 61st World Foundry Cong.* (Beijing), 1995, p 104; available from the American Foundrymen's Society, Des Plaines, IL

20. F.T. Lee, J.F. Major, and F.H. Samuel, Effect of Silicon Particles on the Fatigue Crack Growth Characteristics of Al-12 Wt Pct-0.35 Wt Pct Mg-(0 to 0.02) Wt Pct Sr Casting Alloys, *Met. Mater. Trans.,* Vol 26A (No. 6), June 1995, p 1553

21. J.F. Major, F.T. Lee, and F.H. Samuel, "Fatigue Crack Growth and Fracture Behavior of Al-12 wt% Si-0.35 wt% Mg (0–0.02) wt% Sr Casting Alloys," Paper 96-027, *AFS Trans.,* Vol 104, 1996

22. D. Argo and J.E. Gruzleski, Porosity in Modified Aluminum Alloy Castings, *AFS Trans.,* Vol 96, 1988, p 65

23. T.L. Reinhart, Fatigue and Fracture Properties of Aluminum Alloy Castings, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 813

24. T.L. Reinhart, "The Influence of Microstructure on the Fatigue and Fracture Properties of Aluminum Alloy Castings," presented at Aeromat '96, June 1996 (Dayton, OH), ASM International

25. A.L. Kearney and J. Raffin, *Heat Tear Control Handbook for Aluminum Foundrymen and Casting Designers,* American Foundrymen's Society, 1987

26. M.J. Couper, A.E. Neeson, and J.R. Griffiths, Casting Defects and the Fatigue Behaviour of an Aluminium Casting Alloy, *Fatigue Fract. Eng. Mater. Struct.,* Vol 13 (No. 3), 1990, p 213

27. D.M. Stefanescu et al., "Cast Iron Penetration in Sand Molds, Part I: Physics of Penetration Defects and Penetration Model," Paper 96-206, *AFS Trans.,* Vol 104, 1996

28. R.L. Naro and J.F. Wallace, Effect of Mold-Steel Interface Reactions on Casting Surface and Properties, *AFS Trans.,* Vol 75, 1967, p 741

29. R.L. Naro and J.F. Wallace, Effect of Mold-Steel Interface Reactions on Casting Surfaces, *AFS Trans.,* Vol 100, 1992, p 797

30. J.A. Schey, *Introduction to Manufacturing Processes,* 2nd ed., McGraw-Hill, 1987

31. Manufacturing Design Considerations, *Steel Castings Handbook,* 6th ed., M. Blair and T.L. Stevens, Ed., Steel Founders' Society of America and ASM International, 1995

32. Manufacturing Considerations in Design, *Steel Castings Handbook,* 5th ed., P.F. Wieser, Ed., Steel Founders' Society of America, 1980, p 5–6

33. P.R. Beeley and R.F. Smart, Ed., *Investment Casting,* The Institute of Materials, 1995, p 334

34. S.P. Thomas, Ed., *Design and Procurement of High-Strength Structural Aluminum Castings,* American Foundrymen's Society, 1995

35. R.W Lewis, K.N Seetharamu, and A.Y. Hassan, Residual Stress Formation during Casting, *Handbook of Residual Stress and Deformation of Steel,* ASM International, 2002, p 361–371

36. H.T. Angus, *Cast Iron: Physical and Engineering Properties,* 2nd ed., Butterworths, 1976

37. C.F. Walton, *Gray and Ductile Iron Castings Handbook,* Iron Founder's Society, Des Plaines, IL, 1971

38. S.I. Karsay, *Ductile Iron-Production Practices,* American Foundrymen's Society, Des Plaines, IL, 1975, p 167

39. G.W.S. Anselman et al., *Analysis of Casting Defects,* 2nd ed., American Foundrymen's Society, Des Plaines, IL, 1966

40. F. Hudson et al., *Atlas of Defects in Castings,* Institute of British Foundrymen, London, 1961

41. A.G. Fuller, Evaluation of the Graphite Form in Pearlitic Ductile Iron by Ultrasonic and Sonic Testing and the Effect of Graphite Form on Mechanical Properties, *Trans. AFS,* Vol 85, 1977, p 509–526

42. D.B. Roach and F.H. Beck, Performance and Reliability of Corrosion-Resistant Alloy Castings, *MTI Manual No. 5,* The Materials Technology Institute at The Chemical Process Industries, Inc., Columbus, OH, 1981

43. S. Ensha and A. Kumar, Premature Torquing Failures of Cast A356 Aluminum Actuators, *Handbook of Failure Analysis Case Histories,* Vol 1, ASM International, 1992 p 47–50

44. W.A. Pollard, Failure of a Sprocket Drive Wheel in a Tracked All-Terrain Vehicle, *Handbook of Failure Analysis Case Histories,* Vol 2, ASM International, 1993 p 58–60

45. C. Czajkowski, Hot Cracking of a Pump Impeller from a Nuclear Plant, *Handbook of Case Histories in Failure Analysis,* Vol 1, ASM International, 1992 p 273–277

46. M. Chaudhari, Failure Analysis of Cast Steel Crosshead, *Handbook of Case Histories in Failure Analysis,* Vol 1, ASM International, 1992 p 291

Failures Related to Welding

Introduction

WELDMENT FAILURES may be divided into two classes: those rejected after inspection and mechanical testing, and failures in service that may arise from fracture, wear, corrosion, or deformation (distortion). Causes for rejection during inspection may be either features visible on the weldment surface or subsurface indications that are found by nondestructive testing methods. Surface features that are causes for rejection include:

- Excessive mismatch at the weld joint
- Excessive bead convexity and bead reinforcement
- Excessive bead concavity, underfill, and undersized welds
- Sharp undercut and overlap at the weld toe
- Cracks—hot or cold, longitudinal or transverse, crater and at weld toe
- Gas porosity
- Arc strike
- Spatter
- Backing piece left on: failure to remove material placed at the root of a weld joint to support molten weld metal

Subsurface features that are causes for rejection include:

- Underbead cracks
- Gas porosity
- Inclusions—slag, oxides, or tungsten metal
- Incomplete fusion
- Inadequate penetration
- Solidification cracks, liquation cracks

Failure to meet strength, ductility, or toughness requirements is another cause for rejection of weldments. Details of test methods for welds are provided in the standards of the American Welding Society (AWS), particularly American National Standards Institute (ANSI)/AWS B4.0, "Standard Methods for Mechanical Testing of Welds." A general treatment of mechanical testing of welds is also given in Ref 1.

Evaluation and nondestructive inspection of welds is also described in *Welding, Brazing, and Soldering,* Volume 6 of *ASM Handbook.*

Service failures of welds, similar to those of any other structural component, depend on the operating environment and the nature of the applied load and may include failures from:

- Overload failure by brittle cracking mechanisms
- Overload failure by ductile cracking mechanisms
- Plastic collapse
- Buckling
- Fatigue
- Corrosion fatigue
- Corrosion
- Stress-corrosion cracking
- Hydrogen-induced cracking

If elevated temperatures were involved, then creep deformation and stress rupture should be added to this list.

The first four modes of failure occur under static load. Ductile fracture, plastic collapse, and buckling are often preceded by significant plastic deformation, which may provide some warning or indication before final fracture. In contrast, brittle crack propagation is preceded by negligible macroscopic deformation and can lead to rapid (unstable) crack growth and catastrophic failure without warning. Failure of welded structures by brittle fracture is rare, but it can be dramatic, often leading to serious consequences. Perhaps the most dramatic and well-known example is the brittle failure of the World War II Liberty ships, which split in half while still at dockside (Fig. 1, Ref 2). The fractures originated at weld discontinuities and propagated through carbon-steel plate, which had coarse pearlite formed by furnace cooling. The existence of a coarse pearlitic structure was a key condition for the brittle failures experienced by the welded Liberty ships of World War II. In this case, cracking was minimized by the use of normalized plate, which has fine-grain pearlite that is much more tolerant of crack initiation at weld discontinuities.

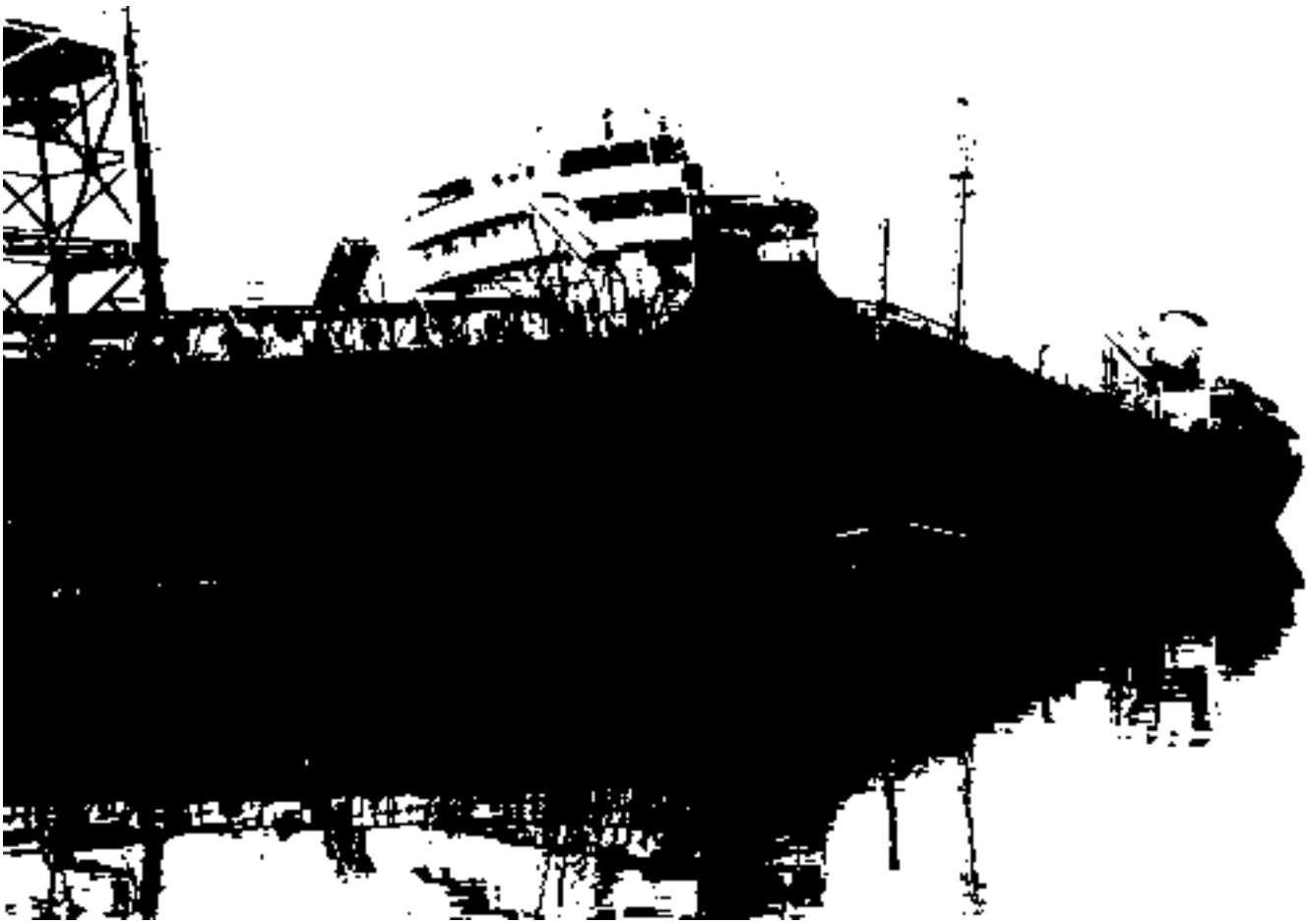


Fig. 1 Example of brittle fracture of welded ship structure. S.S. *Schenectady*, which fractured at its outfitting dock, was one of 19 Liberty ships that experienced brittle cracking of the welded structure. Source: Ref 2

Fatigue of welds is an important concern, because the joint itself is a discontinuity and a potential site for initiation of fatigue cracks. Joint design is one of the most critical factors of fatigue behavior, and this article devotes attention to the effects of joint design on weldment integrity. In terms of corrosion, it is also not unusual to find that, although the wrought form of a metal or alloy is resistant to corrosion in a particular environment, the welded counterpart is not. Conversely, there are also many instances in which the weld exhibits corrosion resistance superior to that of the unwelded base metal. There also are times when the weld behaves in an erratic manner, displaying both resistance and susceptibility to corrosive attack. Corrosion of weldments is discussed in more detail in *Corrosion,* Volume 13 of *ASM Handbook.*
There are many reasons why in-service failures can occur, including:

- Lack of knowledge of service loads and cycles
- Lack of knowledge of the operating environment
- Improper specification of the design parameters
- Improper use of the design method and lack of consideration of key failure modes
- Use of incorrect material properties for design
- Improper selection of materials and welding procedures
- Lack of inspection during fabrication and variability in fabrication practice
- Operation of equipment/component beyond design specification

Poor workmanship and improper selection of welding procedures and filler-metal composition account for numerous arc-weld failures. Other reasons for weldment failure include:

- Inappropriate joint design
- Improper weld size

- Unfavorable heat input
- Improper preweld and postweld heating
- Improper fit-up
- Incorrect parent (base) material composition
- Incorrect filler composition
- Embrittlement during welding
- Unfavorable cooling rate in the weld metal or heat-affected zone (HAZ)
- High residual stresses
- Environmental conditions not contemplated in the design of the weld. These comprise accidental overload, continual loads higher than intended, fatigue, abnormal temperatures, and marine or other corrosive atmospheres.

This article briefly reviews the general factors of weldment failures, which may arise from rejection after inspection or failure to pass mechanical testing as well as loss of function in service. The general factors to consider in the analysis of a service failure are briefly reviewed first. However, causes for rejection during inspection are just as important as a failure in service, and the principles of failure analysis certainly apply to the process failures in the fabrication of welded structures. Therefore, a major portion of this article focuses on the general discontinuities observed in welds, how some imperfections may be tolerable, and how other imperfections may be root-cause defects in service failures. The intent is to focus on the common origins of discontinuities in typical welds and some of the discontinuities that can serve as failure origins in welds made by a particular process.

References cited in this section

1. *W. Moyer,* Mechanical Testing of Welded Joints, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* 2000, p 845–852

2. E.R. Barker, *Brittle Behavior of Engineering Structures,* John Wiley & Sons, 1957

Weld Discontinuities and Service Conditions

Discontinuities are interruptions in the physical continuity of a part and or structure, and a weld is an inherent discontinuity. Critical engineering assessment of weld discontinuities is therefore necessary to define whether it is a harmless imperfection that does not sacrifice weldment reliability or if it is a defect (which, by definition here, is a condition that must be removed or corrected). A discontinuity is not necessarily a defect until it either affects the fitness-for-service of a weld or is defined as such by an acceptance/rejection criteria related to an examination. As such, the engineering community now tends to make specific distinctions between the terms *discontinuity* and *defect* (or flaw). Neither construction materials nor engineered structures are free from imperfections, and welds and weld repairs are not exceptions. All welded structures contain imperfections, to some level of examination.

Since the early 1980s, a number of fracture-mechanics-based assessment procedures have been developed that enable the significance of weld discontinuities to be assessed on a “fitness-for-service” basis. The adoption of fitness-for-service concepts in several codes has resulted in the development of more-rational flaw-acceptance criteria. Fitness-for-service is a concept of weld evaluation that seeks a balance between quality, reliability, and economy of welding procedure. Fitness-for-service is not a constant. It varies, depending on the service requirements of a particular welded structure as well as on the properties of the material involved. A number of different fitness-for-service assessment methodologies for calculating allowable or critical flaw sizes are currently in use, as described in the section “Fitness-for-Service Codes” in this article.

The discontinuities found in welds vary considerably in their importance as failure origins. Their location (surface or subsurface) and geometry must be considered in evaluating the significance of weld discontinuities. The spatial location and shape of welding imperfections (Fig. 2) can be classified into three broad categories: planar imperfections, volumetric imperfections, and geometrical imperfections. These spatial categories, as

discussed subsequently, all have distinct mechanical characteristics in terms of weld reliability and performance.

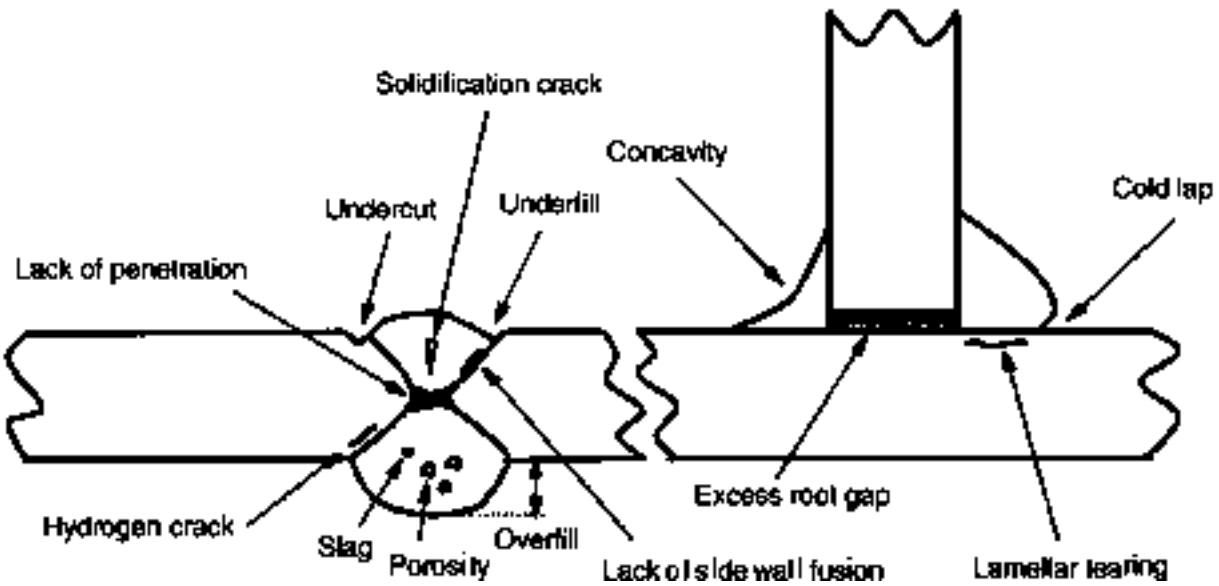


Fig. 2 Schematic of defects and discontinuities in welded joints

General Types of Weld Discontinuities

Weld discontinuities (or imperfections) can be caused by various design, processing, or metallurgical factors. Joint design, edge preparation, fit-up, cleanness of base metal and filler metal, shielding, and welding technique all affect weld quality and must be carefully controlled to prevent porosity, cracks, fissures, undercuts, incomplete fusion, and other weld imperfections. The sources of weld imperfections (or defects) are also discussed in this article, as they relate to service failures or rejection during inspection.

This section briefly discusses only the general characteristics and types of weld discontinuities. The roles of cooling rate (base metal thickness and joint configuration), composition of the base metal (carbon equivalent), level of hydrogen in the welding process, pre-heat and post-heating, as well as the heat input of the process are not addressed in detail.

Planar imperfections are sharp cracklike features that can substantially reduce the fatigue strength of a welded joint or cause initiation of brittle fractures. Examples include hydrogen cracks, lamellar tears, lack of fusion, reheat cracks, solidification cracks, and weld-toe intrusions. The last group are intrusions at the weld-toe region that act as crack starters for fatigue. They are present on a microscopic level, and, on average, they are typically 0.1 mm (0.004 in.) in depth and can be as much as 0.4 mm (0.016 in.) deep for steels (Ref 3). These features are the primary reason why fatigue in welds is a dominant failure mechanisms.

Volumetric imperfections include porosity and slag inclusions. Because these types of imperfections tend to be nearly spherical in form, their notch effect is minor, and they usually have little or no influence on the fatigue behavior. However, they do reduce the load-bearing area of the weld and hence reduce the static strength of the joint.

Geometrical imperfections include misalignment, overfill, stop/starts, undercut, and weld ripples. Geometric imperfections have the effect of locally elevating the stress over and above the nominal stress due to stress concentration from the joint geometry. Methods for calculating the stress concentration factor due to misalignment are readily available (Ref 4). However, unintentional misalignment that occurs during fabrication cannot be allowed for during design, and this can only lead to early failures. Guidance is also available on acceptable levels for undercut, overfill, and location of stop/starts (Ref 5). Weld ripples become stress concentrators for fatigue only when the weld is loaded axially.

Weld Cracks. Cracks are the most detrimental of all weld discontinuities because of their sharp extremities. They have a high probability of initiating brittle fracture. Cracks generally should not be tolerated in welded constructions and are considered defects, especially during manufacture. However, if a crack is detected in the welded product during service, the fitness-for-purpose concept based on fracture mechanics principles still can be used to determine the most appropriate course of action.

Cracks caused by welding essentially always form within the weld zone. In terms of location, surface cracks are most harmful, partly because they are exposed to environmental effects. This exposure amplifies the adverse influence of cracks on the structure, particularly in low-toughness materials. The formation of weld cracks is discussed in more detail in the section “Metallurgical Discontinuities of Welds” in this article.

Cracks can occur in a wide variety of shapes and types and can be located in numerous positions in and around a welded joint (Fig. 3). Cracks associated with welding can be categorized according to whether they originate in the weld itself or in the base metal by either hot or cold cracking. Hot cracking occurs during cooling near the solidus temperature and is characterized by interdendritic or intergranular cracking as the result of hot shortness or localized planar shrinkage. Cold cracking is due to transgranular separations produced by stresses exceeding the strength of the material.

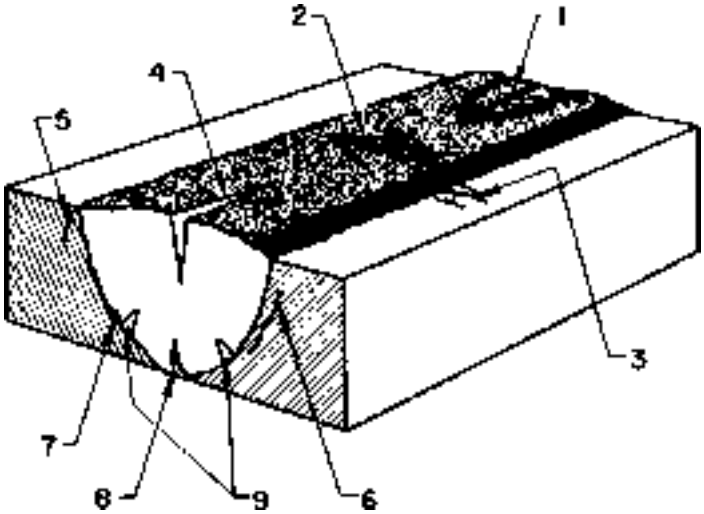


Fig. 3 Identification of cracks according to location in weld and base metal. 1, crater crack in weld metal; 2, transverse crack in weld metal; 3, transverse crack in HAZ; 4, longitudinal crack in weld metal; 5, toe crack in base metal; 6, underbead crack in base metal; 7, fusion-line crack; 8, root crack in weld metal; 9, hat cracks in weld metal

Four types commonly occur in the weld metal: transverse, longitudinal, crater, and hat cracks. Base metal cracks can be divided into seven categories: transverse cracks, underbead cracks, toe cracks, root cracks, lamellar tearing, delaminations, and fusion-line cracks.

Transverse cracks in weld metal (No. 2, Fig. 3) are formed when the predominant contraction stresses are in the direction of the weld axis. Transverse cracks lie in a plane normal to the axis of the weld and are usually open to the surface. They usually extend across the entire face of the weld and sometimes propagate into the base metal.

Transverse cracks in the HAZ (No. 3, Fig. 3) occur on the surface in or near the HAZ. They are the result of the high residual stresses induced by thermal cycling during welding. High hardness, excessive restraint, and the presence of hydrogen promote their formation in ferritic steels. Such cold cracks can propagate into the weld or beyond the HAZ into the base metal as far as is needed to relieve the residual stresses.

Underbead cracks (No. 6, Fig. 3) are similar to transverse cracks in that they form in the HAZ but are not open to the surface. In ferritic steels, they also are caused by high hardness, excessive restraint, and the presence of hydrogen. Their orientation follows the contour of the HAZ.

Longitudinal cracks (No. 4, Fig. 3) may exist in three forms, depending on their positions in the weld. Check cracks are open to the surface and extend only partway through the weld. Root cracks extend from the root to some point within the weld. Full centerline cracks may extend from the root to the face of the weld metal.

Check cracks are caused either by high contraction stresses in the final passes applied to a weld joint due to cold cracking, or by a hot-cracking mechanism.

Root cracks are the most common form of longitudinal weld metal crack, because of the relatively small size of the root pass. If such cracks are not removed, they can propagate through the weld as subsequent passes are applied. This is the usual mechanism by which full centerline cracks are formed, and they can be formed by either a hot or cold cracking mechanism.

Centerline cracks may occur at either high or low temperatures. At low temperatures, cracking is generally the result of poor fit-up, overly rigid fit-up, or a small ratio of weld metal to base metal. At high temperatures, the cracking is generally the result of a large depth-to-width ratio of the weld bead.

All three types of longitudinal cracks are usually oriented perpendicular to the weld face and run along the plane that bisects the welded joint. Seldom are they found at the edge of the joint face.

Crater cracks (No. 1, Fig. 3) are related to centerline cracks. As the name implies, crater cracks occur in the weld crater formed at the end of a welding pass. Crater cracks, formed in the residual puddle after the arc has been broken, are caused by failure to fill the crater before breaking the arc. When this happens, the outer edges of the crater cool rapidly, producing stresses sufficient to crack the interior of the crater. This type of crack may be oriented longitudinally or transversely or may occur as a number of intersecting cracks forming the shape of a star. Longitudinal crater cracks can propagate along the axis of the weld to form a centerline crack. In addition, such cracks may propagate upward through the weld if they are not removed before subsequent passes are applied.

Crater cracks are potential sources of failure because, although frequently small, they usually exist at the end of the weld where stress concentration is greatest. Their occurrence may be minimized by pausing before breaking the arc, by breaking and restarting the arc several times to feed the shrinkage pipe, or by terminating welding on runoff tabs that are cut away later. If crater cracks are found, they should be chipped out and the area rewelded, because it is very difficult to melt out a crater crack.

Hat cracks (No. 9, Fig. 3) derive their name from the shape of the weld cross section with which they are usually associated. This type of weld flares out near the weld face, resembling an inverted top hat. Hat cracks are the result of excessive voltage or welding speed. The cracks are located about halfway up through the weld and extend into the weld metal from the fusion line of the joint. These cracks are formed by a hot cracking mechanism.

Toe and root cracks (No. 5 and 8, Fig. 3) can occur at the notches present at notch locations in the weld when high residual stresses are present. Both toe and root cracks propagate through the brittle HAZ before they are arrested in more ductile regions of the base metal. Characteristically, they are oriented almost perpendicular to the base metal surface and run parallel to the weld axis. These cracks are typically formed by a cold cracking mechanism.

Fusion-line cracks (No. 7, Fig. 3) can be classified as either weld metal cracks or base metal cracks, because they occur along the fusion line between the two. There are no limitations as to where along the fusion line these cracks can occur or how far around the weld they can extend, and they can form by either a hot or cold cracking mechanism.

Occurrence of weld metal cracking in steel is likely if combinations of the following conditions are present:

- Incompatible composition and microstructure (high hardenability or carbon equivalent)
- High restraint
- High hydrogen content in the weld zone
- High tensile stress (residual or applied)
- High weld travel speed
- Absence of preheat or postheat
- Low interpass temperature
- Poor weld profile
- Poor electrode manipulation
- Thin weld-bead cross section (particularly in the first pass)
- Improper fit-up
- Low weld current
- Small electrode diameter
- Improper polarity
- Improper joint geometry (included angle or root opening too small)
- Improper backgouging (narrow or shallow)
- Unbalanced heat input

To prevent cracking in multiple-pass groove or fillet welds:

- Increase bead size of the first pass by using lower travel speed, a short arc, and an uphill (~5°) welding mode.
- Change weld profile to flat or convex.
- Control interpass temperature and weld while the base metal is hot. (Also postheat, which is more effective than preheat)
- Use filler metal and/or slag compositions that promote good wettability with the base metal.

To prevent cracking in craters:

- Fill crater sufficiently before breaking the arc.
- Employ a backstepping technique (terminate each weld on the crater of the previous weld).

To prevent cracking under restraint conditions:

- Eliminate or decrease the degree of restraint.
- Leave a gap between plates to permit shrinkage.
- Weld in the direction of no restraint.
- Peen the passes while hot. Caution: avoid peening first and final passes, due to the danger of either causing or covering cracks or of interfering with inspection.

Analysis of In-Service Failures

The initial task of the failure analyst is to seek out and compile as complete a history as possible of the failed weldment and its preparation. The success in arriving at a correct determination of the cause of failure may be greatly influenced by the amount of background information obtained as early as possible. There are no set rules as to the sequence in which the details are sought, but it is important to secure promptly all oral reports of the failure while the event is still fresh in the minds of the observers. The following is a suggested checklist of information that will be useful later in the analysis:

- Determine when, where, and how the failure occurred. Interview all operators involved. How was the part treated after failure? Was it protected? How was the fracture handled? Did the failure involve any fire, which could have altered the microstructure of the weld or of the base metal?
- Establish the service history—loads, atmospheric exposure, service temperature, and length of service. Was an accident involved? Have there been other similar failures? Is the service history consistent with design criteria? Were operating parameters exceeded? Obtain inspection records if they exist.
- Obtain drawings of the weldment design, calculations of service stresses, and estimation of service life. What were the specified and actual base metal and filler metal? Obtain, if possible, the actual chemical composition, heat treatment and mechanical properties of the base metal, and the actual chemical composition of the filler metal. Gather as much information as possible on the function of the welded part that failed.
- Ascertain the cleaning and fit-up procedures specified and those actually used. Obtain details of the welding procedures, specified and used, and any repairs made. Was preweld or postweld heating applied?
- Establish how the weldment was finished and what tests were performed. How long was the weldment stored, and under what conditions? When and how was it shipped for installation? Was it promptly and properly installed? Request copies of inspection procedures and inspection reports.

After the background information relating to the failed weldment has been secured, the analyst is ready to undertake the study of the failed part itself.
Examination Procedures. The study should begin with visual examination or detailed assessment of the weldment and fracture-surface features, accompanied by preparation of sketches and photographs (both general and closeups) to provide complete documentation of the work performed and the results. This should be followed by careful examination under a low-power (5 to 30×) stereomicroscope. These procedures should be followed for all failures, regardless of whether the failure is caused by fracture, wear, corrosion, or deformation.

The analyst should determine whether the failure was located in the weld, the HAZ, or the base metal; whether the weld met specifications; and whether those specifications were appropriate for the service conditions encountered. The subsequent steps used for identifying the exact cause of the failure are greatly influenced by the outcome of these early observations. One of the most important aspects of this early examination is the determination of origin location(s) that lead to failure.

If the failure is not a fracture, it may be necessary to follow low-power microscopic examination with examination at higher magnification. If so, this should be undertaken before chemical analysis of the weldment surface alters the surface characteristics. Examination of the weldment surface at high magnification should reveal whether there has been selective attack of the part by corrosion, erosion, or other failure mechanisms and may provide clues as to its nature. If the failure is a fracture, much can be learned from the location of fracture origins, the mechanisms and directions of crack propagation, and the types of loading that were involved, by replica or direct examination in a scanning electron microscope.

Once the macroscopic and microscopic examinations are completed, analysis of the material and of the weldment surface may include chemical analysis, mechanical testing, metallography, or examination of the topology and/or chemical constituents of the fracture surface. These methods are discussed in more detail in the Section "Tools and Techniques in Failure Analysis" in this Volume.

Nondestructive Testing. Certain nondestructive tests must be used with caution to avoid altering evidence or creating confusing evidence. For example, liquid penetrant tests are very effective in revealing fine cracks, but if stress-corrosion cracking (SCC) is a potential cause for evaluation, the use of a penetrant might provide false clues, for example, by introducing the penetrant or a corrodent into the crack. Application of the penetrant should be delayed until after completion of the pertinent chemical analyses and metallographic examination.

Magnetic-particle testing can also create problems. In one instance, welds were being made successfully on a steel containing more than 0.35% C, but because the normal borderline weldability of this material made it susceptible to cracking, a magnetic-particle test was done to check for cracks. No cracks were found, but a few weeks later, the weldments were returned, because cracks had initiated at minute arc strikes that were caused by the prods used to make electrical contact for the magnetic-particle test. Due to their small size, the arc strikes cooled rapidly and created a hard, untempered martensitic structure that was susceptible to cracking. Magnetic-particle tests conducted with a magnetic field induced by a yoke that does not pass current through the part avoids this difficulty.

With proper precautions, NDE can provide important information in a failure analysis. Radiography, liquid penetrant inspection, and magnetic-particle examination can give very significant information both at and near the failure as well as on other representative welded joints. Ultrasonic inspection is also valuable not only for thickness measurements but also for analysis of volumetric discontinuities. More detailed information on the NDE of welds is provided in *Nondestructive Evaluation and Quality Control*, Volume 17 of *ASM Handbook.*

Metallographic Sectioning. When study of the fracture surface is complete, sectioning may be performed. The weld profile microstructure of each zone of the weldment can be revealed and compared to that expected for the welding procedure reported and the composition of base and filler metals. Sectioning also permits appraisals of inclusion distributions and shapes, disclosure of whether the fracture and any secondary cracks were transgranular or intergranular, and observation as to whether the weld fusion was incomplete or the penetration was inadequate. Such a microscopic examination of a distorted but unfractured weldment would also be appropriate as a means of discovering if the distortion was caused by loading at an excessively high temperature. Microhardness traverses of the weld zones could yield valuable corroborative data as to whether or not the microstructure had acceptable properties. Additional information on sectioning of welded joints can be found in the article "Weldments: Metallographic Techniques and Microstructures" in *Metallography and Microstructures,* Volume 9 of *ASM Handbook,* formerly 9th edition *Metals Handbook.*

Analysis of Information. The final stage in analysis of a weldment failure consists of comparing and integrating all of the information that has been gathered. It is particularly important to avoid arriving at a decision concerning the cause of failure before this final stage. This is because nearly all welds contain discontinuities of one type or another that, at initial investigation, might be considered to be a contributing factor to the failure. Therefore, it is important for the analyst to keep in mind that many welds containing discontinuities may have given satisfactory performance based on fitness-for-service assessments.

Fractography can be used to focus on the origin or origins of fracture, the mechanisms of crack propagation, and the paths that the propagating crack followed. The fracture origin should reveal whether a stress concentrator might have contributed to failure. The metallographic specimens should disclose the existence of

any improper microstructures, such as decarburized layers, carburized skin, alloy segregation, high-strength martensite in the HAZ, and inadequate weld penetration, as well as excessive oxides or porosity within the weld zone. Any one of these items might prove to be the critical contributing factor in a particular instance, but generally, the two most important items of information are the location of the fracture origin and the fine details of the fracture surface. From these two items it may be deduced how the fracture began and how it propagated; whether it was generated by a single overload or by repetitive stresses; whether the loading was in tension, torsion, bending, shear, or some combination; whether or not environmental effects or corrosion played a role; whether fracture was ductile by microvoid coalescence, brittle cleavage, or by integranular separation. However, all of the other data are necessary to ensure that a less common or less obvious causes of failure are not missed.

The outcome of the failure analysis should provide a root cause for the failure. Usually a recommendation that avoids recurrence is also provided. Such recommendations necessarily differ according to the identified root cause(s). The procedure may consist of using a better weld contour, a more favorable filler metal, a more appropriate heat input (a different cooling rate for the weld metal and the HAZ), a more accurate fit-up, a higher-quality base metal, or different service or maintenance conditions.

Fitness-for-Service Codes

A number of different fitness-for-service assessment methodologies for calculating allowable or critical flaw sizes are currently in use throughout the world. A comprehensive review can be found in Ref 6. Common methodologies in use today include:

- British Standard (BS) 7608:1993 (Ref 7)
- BS PD 6493:1991 (Ref 8)
- BS 7910:1999 (Ref 9)
- Central Electricity Generating Board (CEGB) R6 (Ref 10)
- Electric Power Research Institute (EPRI) GE J and crack tip opening displacement (CTOD) estimation scheme (Ref 11)
- Deformation plasticity failure assessment diagram (Ref 12)
- Welding Engineering Society (WES) 2805 (Ref 13)
- International Institute of Welding (IIW) approach (Ref 14)
- American Petroleum Institute (API) 579 approach (Ref 15)
- American Society of Mechanical Engineers (ASME) IX approach (American Society of Mechanical Engineers, *Boiler & Pressure Vessel Code,* Section IX)

Of these methods, BS PD 6493:1991 (Ref 8) has been widely used, and a brief description of it is provided subsequently. Additional information on fitness-for-purpose evaluation is given in the article “Failure Assessment Diagrams” in this Volume.

BS PD 6493:1991 includes three different fracture assessment routes to enable structures to be assessed at a level of complexity appropriate to the problem under consideration. The level 1 approach is based on the CTOD design curve method, which forms the basis of the elastic-plastic fracture assessment procedure. This approach has a deliberate safety factor built into the calculations, and it results in computation of “tolerable” crack sizes.

In BS PD 6493:1991, the level 1 approach is presented as a method for performing preliminary assessments. The level 2 approach is based on a plane-stress plastic collapse modified strip yield model, and there is no inherent safety factor. This method is the preferred assessment level for the majority of applications. The level 3 method is the most sophisticated assessment level in the proposed revisions and would normally be used only for the assessment of high-work-hardening materials when the level 2 approach has proved too restrictive or when a tearing instability analysis is required. The fracture assessment model adopted in the level 3 approach requires extensive material characterization data, including details of the stress-strain behavior of the material in which the defect is located. This presents a challenge for the HAZ. All three levels can be undertaken in terms of the stress-intensity factor, K (or K derived from fracture resistance, J), or the CTOD fracture mechanics parameters. The level 2 approach is based on procedures similar to CEGB R6 Rev. 2, and the level 3 approach is based on CEGB R6 Rev. 3. These codes have been extensively used in the electric utilities industry in the United Kingdom.

In all three levels of fracture assessment, the resistance of a structure to failure is determined by using a failure assessment diagram (FAD). The *y*-axis of the FAD indicates the resistance of the structure to brittle fracture, while the *x*-axis assesses its resistance to plastic collapse. The failure assessment curve interpolates between these two limiting failure modes.

The level 2 assessment procedure, which is recommended as the primary method, has a FAD that is defined by:

$$k_r, \sqrt{d_r} = S_r \left[\frac{8}{p^2} \ln \sec\left(\frac{p}{2}\right) S_r \right] \qquad \text{(Eq 1)}$$

where:

- K_r is the ratio of the applied stress-intensity factor to the fracture toughness, K.
- δ_r is the ratio of the applied CTOD to the material CTOD.
- S_r is the stress ratio that is taken as the ratio between the net-section stress created by the applied loads and the flow strength of the material.
- ln is the natural logarithm.

For a given flaw, both K_r or $\sqrt{d_r}$ and S_r are determined using appropriate solutions, and these are plotted as shown in the FADs (Fig. 4). If the point falls within the envelope of the failure locus, the flaw is considered safe. If the point falls outside the locus, the flaw is unacceptable.

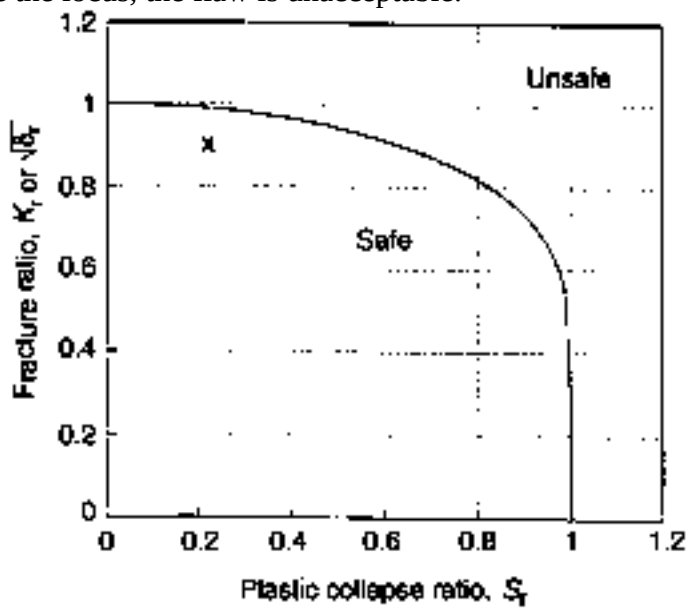


Fig. 4 Failure assessment diagram concept for assessing cracked components for brittle fracture and plastic collapse

The FAD to assess the integrity of a structure assumes that the size of the flaw is known. In order to compute limiting flaw size, for example, for inspection purposes, a point that lies on the failure assessment line and corresponds to the applied loading conditions has to be determined. This is achieved using iterative techniques by plotting a series of assessment points as a function of flaw size. The point that intersects the failure line gives the limiting size of a flaw. By performing these calculations for a series of crack aspect ratios, tolerable flaw size plots can be generated. The plots can be used either to set inspection levels or to allow known flaws to be assessed. In the same way, an iterative process has to be used to generate the critical fracture toughness for some hypothetical (or assumed) flaw size and assumed loading conditions. This computed value can then be directly used for materials selection.

Another approach that is sometimes used to ensure that catastrophic failure does not occur is leak-before-break. This design philosophy has been developed primarily for applications where the crack extends in a stable manner by such mechanisms as stable tearing or fatigue crack growth and is employed in the ASME Section VIII Boiler and Pressure Vessel Code. This approach can be invoked only if the crack shape at breakthrough can be precisely predicted. In most cases, it is generally assumed that the defect length at breakthrough is 2 to 3

times the plate thickness. However, there have been instances where defects up to 10 times the plate thickness have been found during inspection of petrochemical plants. Leak-before-break becomes less likely as the defect length increases, so it should be invoked with great caution.

Brittle Fracture

The occurrence of unstable fractures initiating at weld discontinuities is governed primarily by material toughness, level of tensile stress, rate of loading, and severity of stress intensity for a discontinuity type. The problem of macroscale brittle fracture of steel structures can be greatly reduced by selecting steels that have sufficient toughness at the lowest anticipated service temperatures. Unfortunately, the thermal process of welding may accentuate some undesirable characteristics, such as reducing the toughness of steel welds, and raise its ductile-brittle transition temperature in the HAZ.

Welding also introduces residual stresses that can approach yield point levels. The largest residual stress usually occurs in regions near the weld parallel to the weld axis (longitudinal stress), while the amount of transverse stress is not as high (unless lateral movement of the joint is severely restrained). However, transverse residual stresses can be of significance, because the orientation of many common planar discontinuities (incomplete penetration, incomplete fusion) is normal to this stress. Residual stresses are therefore important, because they can have a significant effect on the initiation of fractures from planar discontinuities.

The deleterious effect of residual stress on the performance of welded constructions can be significant, even when the external load is rather low, if the weldment is of low toughness, subject to adverse environmental degradation (e.g., SCC or high-temperature damage), or if significant discontinuities are present. In one case of a catastrophic brittle fracture of a pressure vessel (Ref 16), the cause of failure was felt to be cracking that developed during the installation of a new shroud ring. From reconstruction of the brittle fracture and subsequent metallographic examination of the fracture origin, the indications revealed lamellar tears that originated at nonmetallic inclusions on planes parallel to the plate surface, caused by weld shrinkage stresses. The cracks caused by the root weld pass were not by themselves large enough to initiate fracture, but subsequent thermomechanical cycling during the remainder of the welding process caused dynamic strain aging and localized embrittlement at the crack tips. The recommendation from this analysis was to perform stress-relief heat treatments to reduce residual-stress levels after welding non-pressure-retaining components to certain pressure shells. This treatment should reduce residual stresses and lessen the embrittlement effects of strain aging.

Preweld and postweld heating are used for relief of internal stress and elimination of hydrogen. However, temperatures to which parts are heated should be selected with care; carbon steels and many alloy steels may undergo a decrease in toughness from embrittlement phenomena when heated in certain critical temperature ranges, such as the blue brittleness temperature around 260 °C (500 °F).

In many cases of brittle fracture of welded steel structures, the recommendation may be the use of tougher steel. It may also involve a re-design of the welded components to eliminate sharp corners. The classic case of the Liberty ships is an example of this, as previously mentioned at the beginning of this article. Another example subsequently follows.

Example 1: Low-Temperature Brittle Fracture in a Steel Tank Car Because of Weld Imperfections. A railway tank car developed a fracture in the region of the sill and shell attachment during operation at -34 °C (-30 °F). On either side of the sill-support member, cracking initiated at the weld between a 6.4 mm ($\frac{1}{4}$ in.) thick frontal cover plate and a 1.6 mm ($\frac{5}{8}$ in.) thick side support plate. The crack then propagated in a brittle manner upward through the side plate, through the welds attaching the side plate to the 25 mm (1 in.) thick shell plate, and continued for several millimeters in the shell plate before terminating. The fracture surfaces displayed chevron marks pointing back to the weld between the frontal cover and side support plates at the point of fracture origin.

Investigation. Examination of the weld at the fracture origin revealed imperfections, including small root cracks, unfused regions, and small hard spots on the surfaces of the weld and adjacent plates. Metallographic specimen and hardness values (up to 45 HRC) showed that the hard spots had been rapidly cooled, but why they formed was not ascertained. The presence of these hard spots indicated that the assembly had not been adequately stress relieved after structural modifications that involved welding in the region of the sill and shell attachment.

The shell plate met the chemical-composition requirements of ASTM A212, grade B, steel. Other plates involved were not positively identified but were generally classified as semikilled carbon steels.

The toughness of the shell and side support plates were measured and were found to be inadequate for the service conditions in the presence of weld imperfections. Charpy V-notch testing gave a 20 J (15 ft · lbf) transition temperature of -7 to 5 °C (20 to 40 °F) for the shell plate and -1 to 5 °C (30 to 40 °F) for the side support plates. Drop-weight tests on the shell plate gave a nil-ductility temperature of 5 to 10 °C (40 to 50 °F).

Conclusions. The fracture was initiated by weld imperfections and propagated in a brittle manner as a result of service stresses acting on plate having low toughness at the low service temperatures encountered.

Recommendations. On the basis of fracture toughness studies, it was suggested that the specifications for the steel plates be modified to include a toughness requirement: the Charpy V-notch requirement of 20 J (15 ft · lbf) at -46 °C (-50 °F), which is specified in ASTM A 300 for ASTM A212, grade B, steel, should be adequate for this application. In addition, improved welding and inspection practices were recommended to reduce the incidence of weld imperfections. It should be noted that ASTM A212, grade B, steel has been superseded by ASTM A516, grade 70, steel, which has a low-temperature requirement of 20 J (15 ft · lbf) at -34 °C (-30 °F), as specified in ASTM A 20/A 20M.

References cited in this section

3. J. Signes, R.G. Baker, J.D. Harrison, and F.M. Burdekin, Factors Affecting the Fatigue Strength of Welded High Strength Steels, *Br. Weld. J.*, Vol 14 (No. 3), 1967, p 108–116

4. S.J. Maddox, “Fitness-for-Purpose Assessment of Misalignment in Transverse Butt Welds Subject to Fatigue Loading,” Document XIII-1180–85, International Institute of Welding, 1985

5. S.J. Maddox, Recent Advances in the Fatigue Assessment of Weld Imperfections, *Weld. J.*, July 1993, p 42–51

6. J.R. Gordon, “A Review of Currently Available Fracture Analysis Methods and Their Applicability to Welded Structures,” paper presented at Sixth Annual North American Welding Research Conference, Edison Welding Institute and American Welding Society, Oct 1990

7. “Code of Practice for Fatigue Design and Assessment of Steel Structures,” BS 7608, British Standards Institution, 1993

8. “Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures,” BS PD 6493:1991, British Standards Institution, Aug 1991 (superseded by BS 7910)

9. “Guide on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures,” BS 7910, British Standards Institute, 1999

10. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, “Assessment of the Integrity of Structures Containing Defects,” R/H/R6-Rev. 3, Central Electricity Generating Board (United Kingdom), 1987

11. V. Kumar, M.D. German, and C.F. Shih, “An Engineering Approach for Elastic-Plastic Fracture Analysis,” Technical Report NP 1931, Electric Power Research Institute, July 1981

12. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, STP 803, Vol II, ASTM, 1983, p 206

13. “The Method of Assessment for Defects in Fusion Welded Joints with Respect to Brittle Fracture,” WES 2805, Japan Welding Engineering Society

14. “IIW Guidance on Assessment of the Fitness for Purpose of Welded Structures,” Draft for Development, IIW/IIS-SST-1157-90, International Institute of Welding, 1990

15. “Recommended Practice for Fitness-for-Service of Refinery Equipment,” API 579, American Petroleum Institute

16. A.R. Cuiffreda and R.D. Merrick, Brittle Fracture of a Pressure Vessel: Study Results and Recommendations, *Proceedings of the American Petroleum Institute,* Vol 62, 1983

Joint Design (Adapted from Ref 17)

The term *joint design* refers to the product-related considerations of a joint based on structural design requirements. An ideal joint is one that effectively transmits forces among the joint members and throughout the assembly, meets all structural design requirements, and can still be produced. Problems with design or structural details include choice of the wrong type of weld joint for a given application, introduction of areas of stress concentrations, or undesirable changes in cross section.
The first consideration, of course, is the ability of the joint to transfer load, usually within a substantial safety margin. Joint type is selected on the basis of product configuration and loading. The types of joints (Fig. 5) are:

- *Butt joints* between two members lying approximately in the same plane where continuity of section is desired. Although they may be welded with no preparation, butt joints are often grooved, especially with thicker and heavier plate. Generally, when the metal thickness exceeds approximately 5 mm ($\frac{3}{16}$ in.), the butt joint requires grooving to prevent a lack of fusion from weakening the weld. Welding from both sides may also be employed.
- *T-joints,* which are frequently used in fabricated forms. T-joints are well suited for no-preparation fillet welds, but preparation (such as a bevel) may be necessary in welding thick materials. Fillet-welded T-joints should not be used for fusion welding with ordinary gas or arc welding of containers, because corrosive solids may be trapped.
- *Corner joints* with members located at approximately right angles to each other in various ways (Fig. 6). The corner-to-corner joint (Fig. 6a) is difficult to assemble, and a small electrode with low welding current must be used to prevent the first welding pass from melting through. Simple corner with fillet welds (Fig. 6b and c) are easy to assemble, do not melt through easily, and require half the amount of weld metal of the corner-to-corner joint of Fig. 6(a). With thick plates, a partial-penetration groove weld with beveling (Fig. 6d) is often used in a corner joint. For a deeper weld, a J-groove preparation may be chosen instead of a bevel (Fig. 6e).
- *Lap joints* with two overlapping members. Lap joints do not require edge preparation and are necessary for use in seam and spot welding of sheet metals.
- *Edge joints* between the edges of two or more parallel or nearly parallel members. Edge joints may not require the addition of filler metal for welding and are frequently used to joint thin turned-out edges, such as those formed where the sides of a container meet.

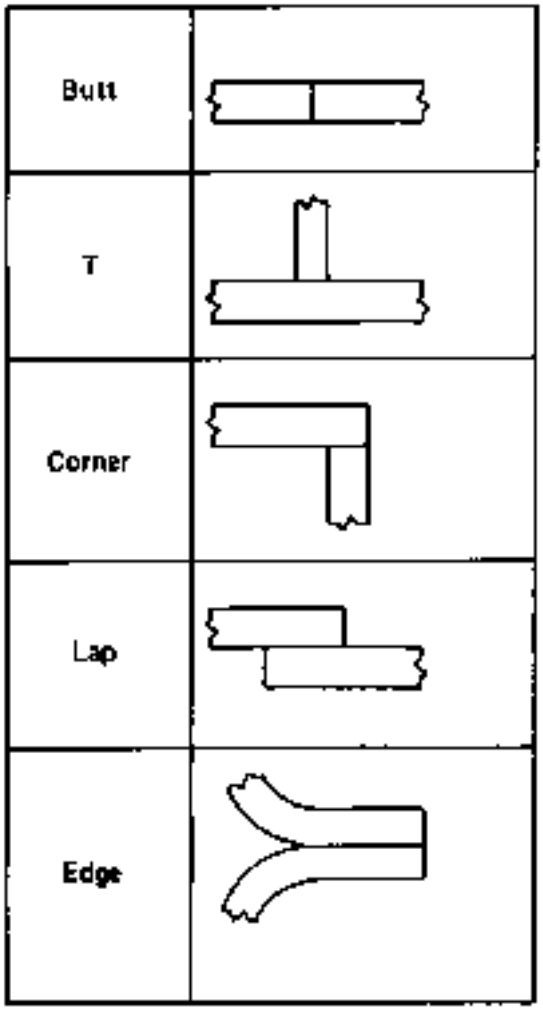


Fig. 5 Types of joints

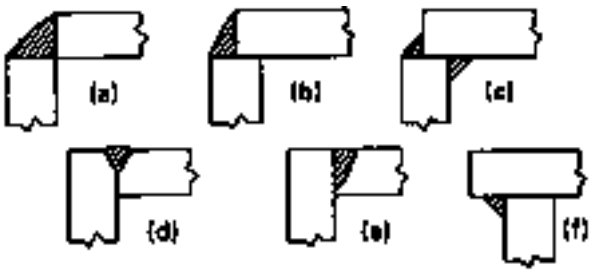


Fig. 6 Corner joints. (a) Corner-to-corner joint. (b) Simple corner joint with single fillet weld. (c) Simple corner joint with double fillet welds. (d) V-groove weld. (e) J-groove weld. (f) Inconspicuous fillet weld in a corner

The design or selection of appropriate joint type is determined primarily from the type of service loading. For example, butt joints are preferred over tee, corner, lap, or edge joints in components subjected to fatigue loading. The specific joint design aspects, such as the size, length, and relative orientation of the joint, are based on a formal stress analysis. Location of welds and their start and finish point are also an important part of joint design. Location should be chosen such that they are not in regions subjected to high stress. The starting point of the weld, which may act as a more pronounced stress raiser, may also be a factor. In one case, for example, failures caused by bending were occurring in the circumferential welds of two tubular members. The failures were prevented when the weld was started on the neutral axis for the intended load condition.

Joint design also includes many factors for the economical production of welds. Joints must be placed in locations that allow operators to readily make the joints. Weld joint designs also employ bevel angles and root openings to enhance accessibility to the welding torch (or electrode) and provide adequate weld penetration. Joint and weld types are also designed to reduce the required amount of filler metal to avoid unnecessary

expense. For example, single-bevel groove or double-bevel groove welds (Fig. 7) are used to provide adequate accessibility and reduce the amount of weld metal required for completion of the joint. Double-groove weld joints reduce the amount of weld filler metal required for single-groove preparations by about half. The decreased welding of double-groove weld also reduces distortion and facilitates alternating weld passes on each side of the joint, which further reduces distortion. Double bevel joints are typically used in thicker plates.

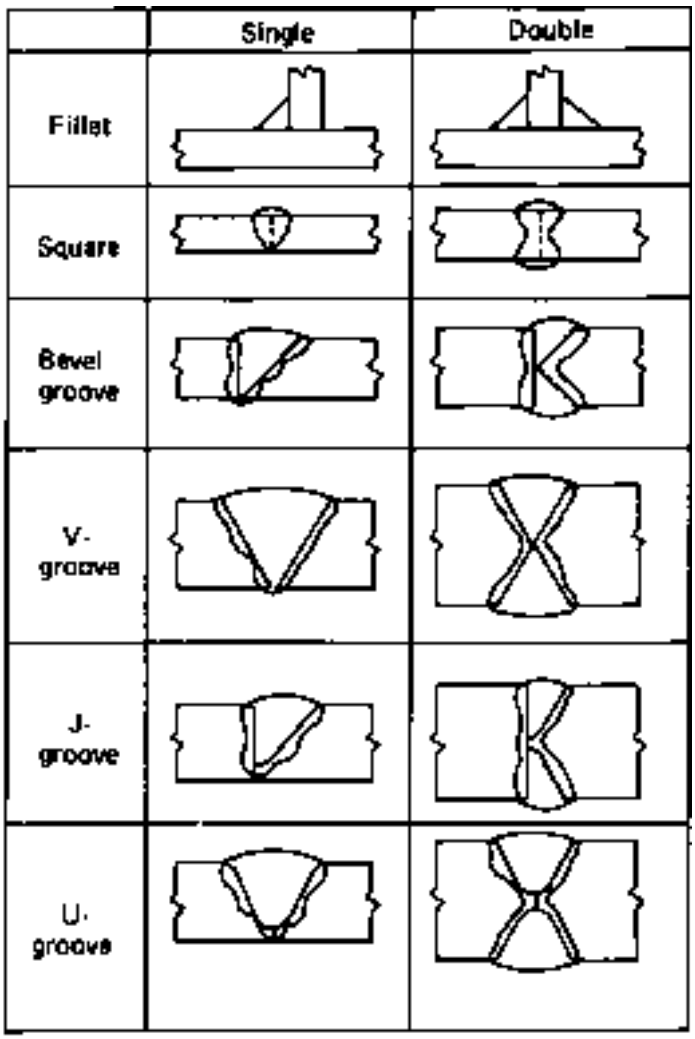


Fig. 7 Types of welds

The types of welds are illustrated in Fig. 7. Cost is the major consideration in the choice between fillet or groove welds. Although simple fillet-welded joints are the easiest to make, they may require excessive weld filler metal for larger sizes. For example, the fillet welds in Fig. 8(a) are easy to apply and require no special plate preparation. The welds also can be made with large-diameter electrodes using higher welding current. However, the amount of weld metal increases in relation to the square of the leg size (ω). In contrast, double-bevel groove welds (Fig. 8b) have about one-half the weld area of the fillet welds (Fig. 8a) but require extra preparation and the use of smaller-diameter electrodes with lower welding currents to place the initial pass without melt-through. As plate thickness increases, this initial low-deposition region becomes a less important factor, and the higher cost factor (preparation and operation expenses) decreases in significance.

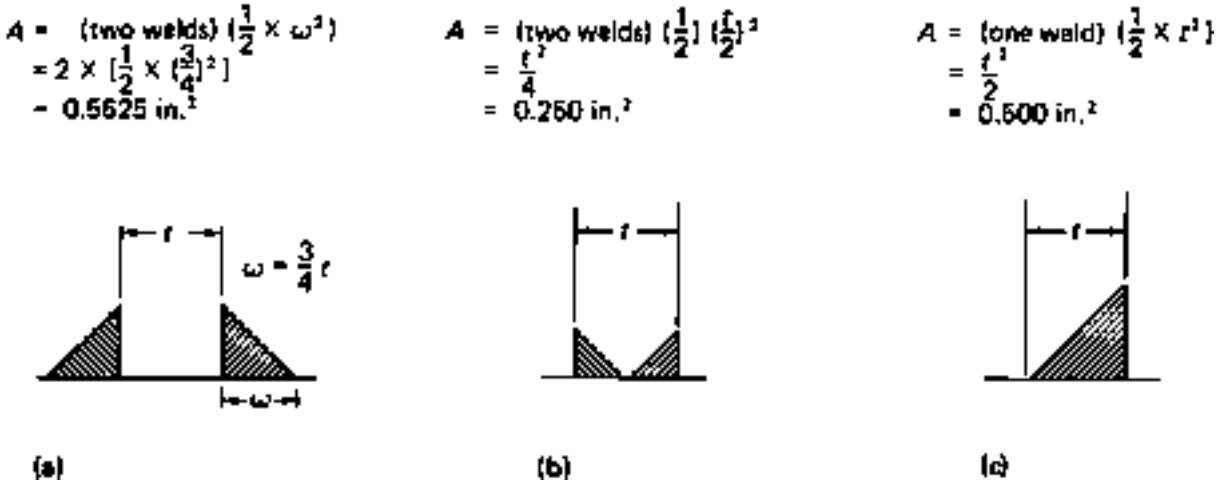


Fig. 8 Comparison of fillet and bevel-groove welds. ω is the leg size of the fillet weld in inches = $\frac{3}{4}$ in.; *A* is the cross-sectional area of the weld in square inches = $\frac{1}{2}\omega^2$; *t* is the plate thickness in inches = 1.0 in. (a) Fillet welds. (b) Double-bevel groove weld. (c) Single-bevel groove weld

The single-bevel groove weld in Fig. 8(c) requires about the same amount of weld metal as the fillet welds in Fig. 8(a). Thus, there is no apparent economic advantage. From a design standpoint, however, the single-bevel groove weld offers a more direct transfer of force through the joint, which provides better service under fatigue loading. Although the full-strength fillet welds (Fig. 8a) would be sufficient, many welding codes have lower allowable stress limits for fillet welds and may require a leg size equal to the plate thickness due to their poor performance in fatigue. In those cases, the cost of the fillet-welded joint may exceed the cost of the single-bevel groove in thicker plates.

The size of the weld is an important design factor. It includes both the width as well as the length of the weld. Weld size of a fillet weld should always be designed with reference to the size of the thinner member. The joint cannot be made any stronger by matching the weld size to the thicker member. The throat (*T*) of a fillet weld (Fig. 9) is a better index of strength of a weld than leg size (ω), because the allowable stress is applied along the throat. However, the leg size of a fillet weld influences the size of the throat. When sizing fillet welds, the direction of loading should be considered. Extensive test data have shown a one-third increase in weld strength under transverse versus parallel loading. Therefore, weld size may be reduced under transverse loading conditions.

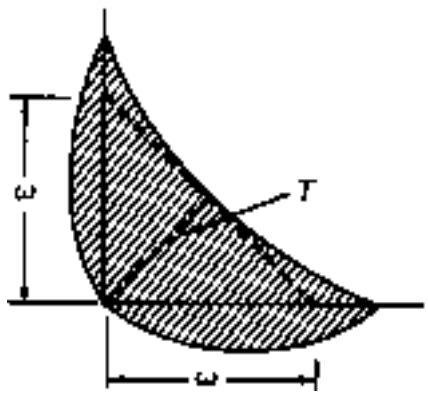


Fig. 9 Leg size (ω) and throat (*T*) of a fillet weld. The throat is the shortest distance between the root of the joint and the face of the fillet weld.

Combinations of partial-penetration groove welds and fillet welds are used for many joints. When full-strength fillet welds are not required in the design, savings can often be achieved by using partial-penetration groove welds. Similar to the use of minimum throat for fillet or partial-penetration groove welds, the minimum throat is used for designing a partial-penetration combination groove weld. Figure 10 shows, the correct minimum

throat for a combination weld. Figure 11 illustrates the faulty calculations that result when the incorrect throat is used to determine the allowable unit force on a combination weld.

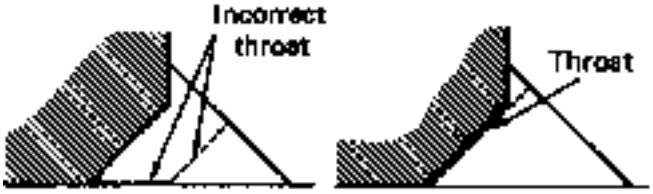


Fig. 10 Determination of minimum throat

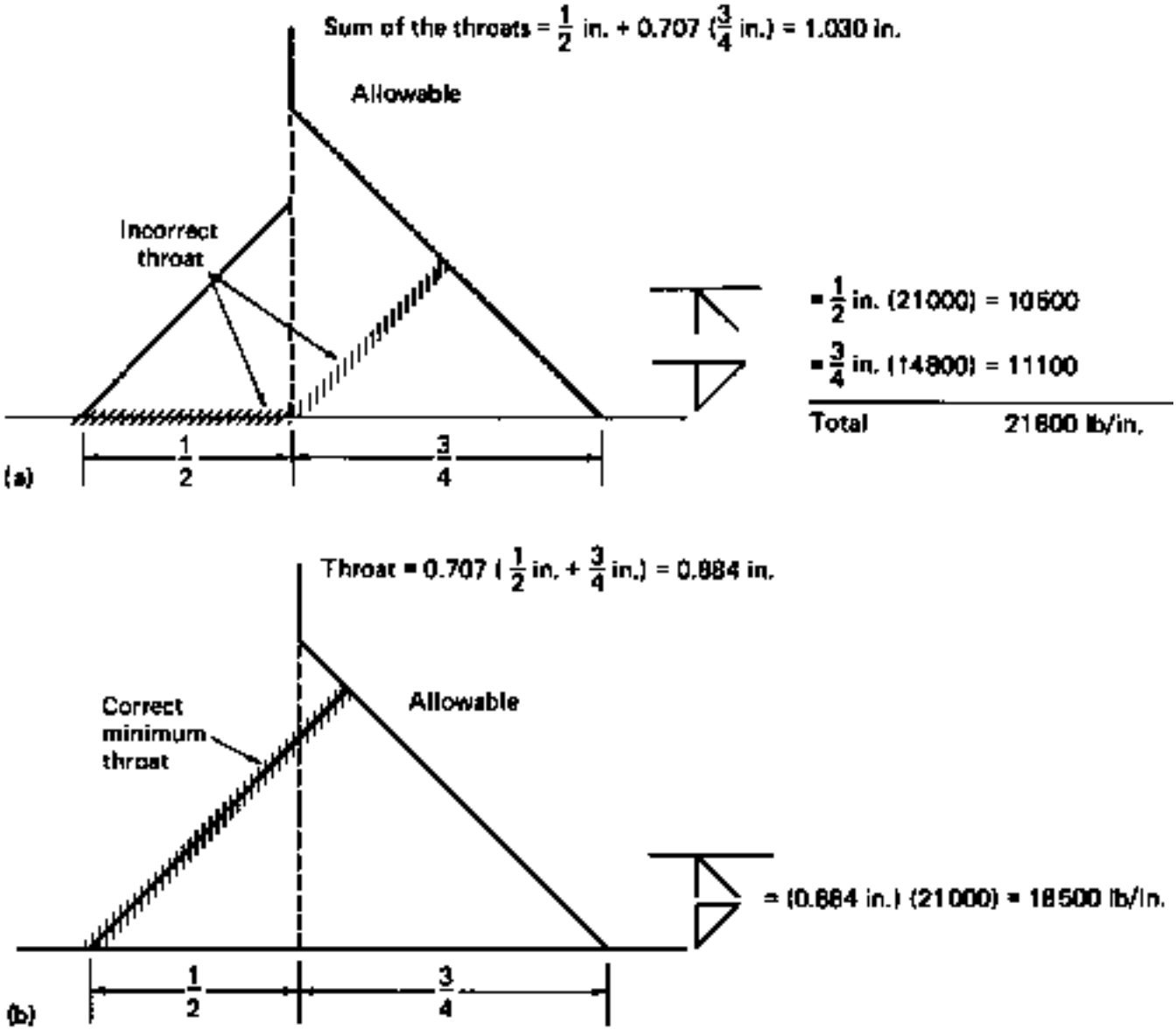


Fig. 11 Determination of allowable load on a combination weld. (a) Weld allowable load incorrectly figured adding each weld separately. (b) Weld allowable load correctly figured using minimum throat

Effect of Joint Design on Fatigue

The conventional approach to designing a multimember structure for fatigue resistance is to consider the members as first in importance and the joints as inevitable complications. However, of equal importance is the shape of the parts within the joint, because the shape and size of the members may multiply the external loading, or transform it into a different mode, causing possible stress concentrations. The importance of the effect of joint design on fatigue behavior cannot be overemphasized. Detail geometry and joint configuration

are the most important variables affecting the fatigue life of a structural detail. Discontinuities associated with the weld detail can behave as preexisting cracks, and therefore fatigue life is determined solely by fatigue crack propagation behavior. The fatigue strength of welded joints is relatively unaffected by the tensile strength of the material. This is because the majority of the fatigue life of a welded joint is spent in the propagation of fatigue cracks to critical sizes, not in the initiation of the fatigue crack. Although tensile strength may be affected by changes in alloys, heat treatment, or temper, the crack propagation rates are relatively insensitive to such changes.

The high stress concentration inherent in the geometry of a fillet weld lowers fatigue strength in such joints. The joint designs with poorest fatigue resistance are the nonsymmetrical types, such as strap, tee, and lap joints in which large secondary stresses are also developed. Poor fatigue resistance is also observed in non-load-carrying attachments or reinforcements to a plate strictly due to the stress concentration effect of those attachments. In general, the choice of filler metal and method of edge preparation exert little influence on welded joint fatigue behavior. Bevel welds are usually superior; lack of fusion and incomplete penetration problems often occur in fillet welds or square joints. Size effects must also be considered, because fatigue results obtained from small-scale specimens demonstrate distinctly higher fatigue limits than those obtained in full-scale sections (e.g., Ref 18).

The precise location and the magnitude of stress concentration in welded joints depend on the design of the joint and on the direction of the load. Some examples of stress concentrations in butt welds and fillet welds are given in Fig. 12. Concentration of stress at the toes of the weld may initiate cracking, even if the weld does not carry any load. Indeed, the weld toe is often the primary location for fatigue cracking in joints that have good root penetration. In situations where the root penetration is poor or the root gap is excessive, or in load-carrying fillet welds where the weld throat is insufficient, the root area can become the region of highest stress concentration. Fatigue cracks in these situations start from the root of the weld and generally propagate through the weld (Fig. 13).

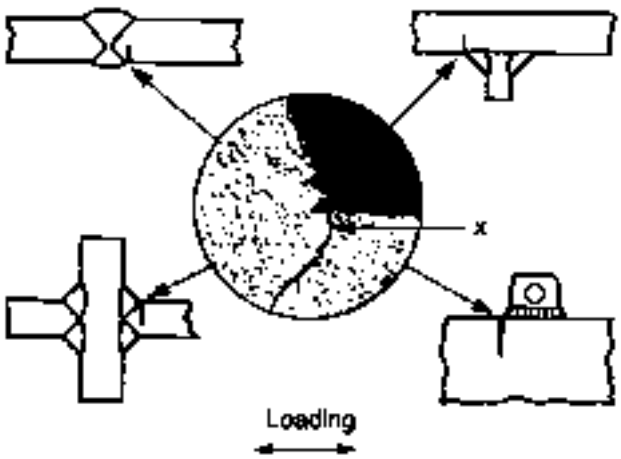


Fig. 12 Examples of stress concentrations in welded joints

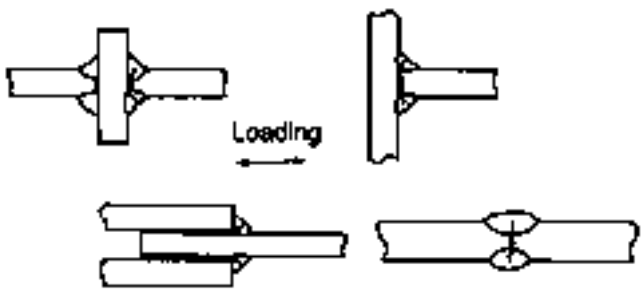


Fig. 13 Fatigue cracking from the weld root

The geometry/shape parameters that influence fatigue of welded joints by affecting the local stress concentration include plate thickness (T), attachment toe-to-toe length (L), attachment thickness (t), weld toe radius (r), weld angle (θ), and the profile of the weld surface (convex versus concave). It is generally found that

the fatigue strength of a welded joint decreases with increasing attachment length, increasing plate thickness, increasing weld angle, decreasing toe radius, and misalignment. By far, the most significant parameter is the weld toe radius. The weld toe stress-concentration factor, K_t, is a function of all these geometry variables, and a number of formulas are available for butt welds and cruciform joints subjected to bending or tensile load (Ref 19, 20).

Misalignment is another geometry-related parameter that can influence the fatigue performance of a welded joint. Misalignment can manifest itself in several ways, angular and axial misalignments being the most common. Generally, fatigue strength decreases with increasing misalignment. Formulas to calculate the local weld stress-concentration factor due to misalignment are now well established (Ref 4). The degree of influence of the geometry parameters cited previously can vary significantly with the loading condition (e.g., tension versus bending) and the joint type.

The following are a few examples of fatigue failures related to joint design or preparation.

Example 2: Fatigue Cracking of a Stainless Steel Elbow Assembly at a Welded Joint in a High-Stress Region. The welded elbow assembly shown in Fig. 14 was part of a hydraulic-pump pressure line for a jet aircraft. The other end of the tube was attached to a flexible metal hose, which provided no support and offered no resistance to vibration. The components of the elbow were made of American Iron and Steel Institute (AISI) type 321 stainless steel and were joined with ER347 stainless steel filler metal by gas tungsten arc welding. The line was leaking hydraulic fluid at the nut end of the elbow, and the assembly was returned to the manufacturer to determine the cause of failure.

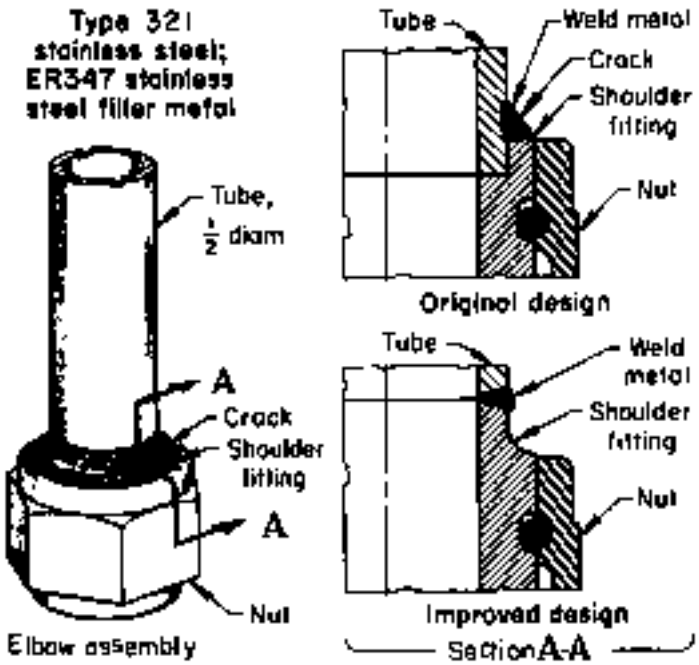


Fig. 14 Welded stainless steel elbow assembly that, as originally designed, cracked at the root of the weld under cyclic loading. The improved design moved the weld out of the high-stress area. Dimensions given in inches

Investigation. Visual examination of the elbow assembly revealed a crack in the fillet weld that joined the tube to the shoulder fitting. In assembly, the tube was inserted into a counterbore in the shoulder fitting (original design, section A-A, Fig. 14), then welded by the gas tungsten arc process.

The fitting was separated from the tubing so that the surfaces of the cracked area could be examined. Stains on the fracture surface indicated that fatigue cracks had initiated at the root of the weld, at that stress concentrator inherent in the design of the joint. Cracking in the weld metal had progressed partway around the joint, thus producing a leakage path through which pressurized hydraulic fluid had escaped.

Conclusion. Failure was by fatigue cracking initiated from a notch at the root of the weld and was propagated by cyclic loading of the tubing as the result of vibration and inadequate support of the hose assembly.

Corrective Measures. The joint design was changed from a cylindrical lap joint (tube inserted into a counterbore) to a square-groove butt joint (improved design, section A-A, Fig. 14). The new joint design

provided a more flexible joint by eliminating the notch and by moving the weld out of the high-stress area. An additional support was provided for the hose assembly to minimize vibration at the elbow.

Example 3: Fatigue Fracture of a Gas-Turbine Inner-Combustion-Chamber Case Assembly Because of Unfused Weld Metal and Undercuts. The case and stiffener of an inner-combustion-chamber case assembly failed by completely fracturing circumferentially around the edge of a groove arc weld joining the case and stiffener to the flange (Fig. 15a). The assembly consisted of a cylindrical stiffener inserted into a cylindrical case that were both welded to a flange. The case, stiffener, flange, and weld deposit were all of nickel-base alloy 718. It was observed that a manual arc weld repair had been made along almost the entire circumference of the original weld.

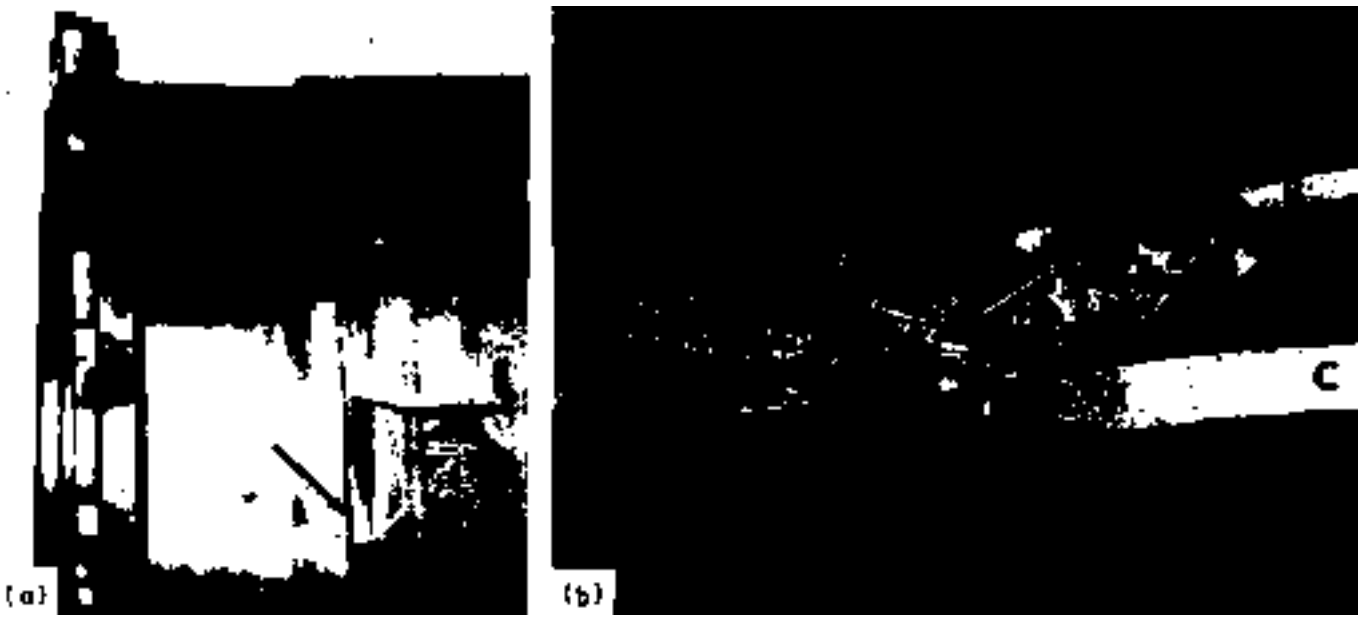


Fig. 15 Alloy 718 inner-combustion-chamber case assembly that fractured by fatigue in the weld joining the flange to the case and stiffener. (a) Exterior surface of the assembly showing the circumferential fracture of the case (arrow). 0.5×. (b) Section through the fracture showing the weld (region A) that originally joined the flange to the case (region B) and the stiffener (region C). The external repair weld (region D) had only partial fusion with the earlier bead. The arrow shows a film of oxide slag at the interface. Etched with 2% chromic acid plus HCl. 10×

Investigation. Microscopic examination of the fracture site revealed unfused weld metal surfaces and severe undercutting of the base material of the stiffener in many areas. The thickness of the case had also been undercut in several areas. Fatigue cracks had originated at multiple sites along the weld interfaces of the case and stiffener. Also, the groove weld contained areas of mismatch that were greater than those allowed by the specifications.

Metallographic specimens from sections of the fracture site showed the two weld beads that were deposited by a repair weld contained many regions of partial fusion as evidenced by oxide films and slag at the weld interface (Fig. 15b).

Conclusions. Failure was by fatigue from multiple origins caused by welding defects. The combined effects of undercutting the case wall, weld mismatch, and unfused weld interfaces contributed to high stress concentrations that generated the fatigue cracks. Ultimate failure was by tensile overload of the sections partly separated by the fatigue cracks.

Recommendation. Correct fit-up of the case, stiffener, and flange is essential, and more skillful welding techniques should be used to avoid undercutting and unfused interfaces.

Example 4: Fatigue Fracture of Welded Type 321 Stainless Steel Liners for a Bellows- Type Expansion Joint. The liners for bellows-type expansion joints in a duct assembly (Fig. 16a) failed in service. The duct assembly, used in a low-pressure nitrogen gas system, consisted of two expansion joints (bellows) connected by a 32 cm (12 in.) outside diameter (OD) pipe of ASTM A106, grade B, steel. Elbows of ASTM A234, grade B, steel, 180° to each other, were attached to each end of the assembly. A liner with an OD of 29 cm (11 in.) was welded inside each expansion joint. The liners were 1.3 mm (0.050 in.) thick and made of AISI type 321 stainless steel.

The upstream end of one liner was welded to an elbow, and the upstream end of the other liner was welded to the pipe; this allowed the downstream ends of the liners to remain free and permitted the components to move with expansion and contraction of the bellows.

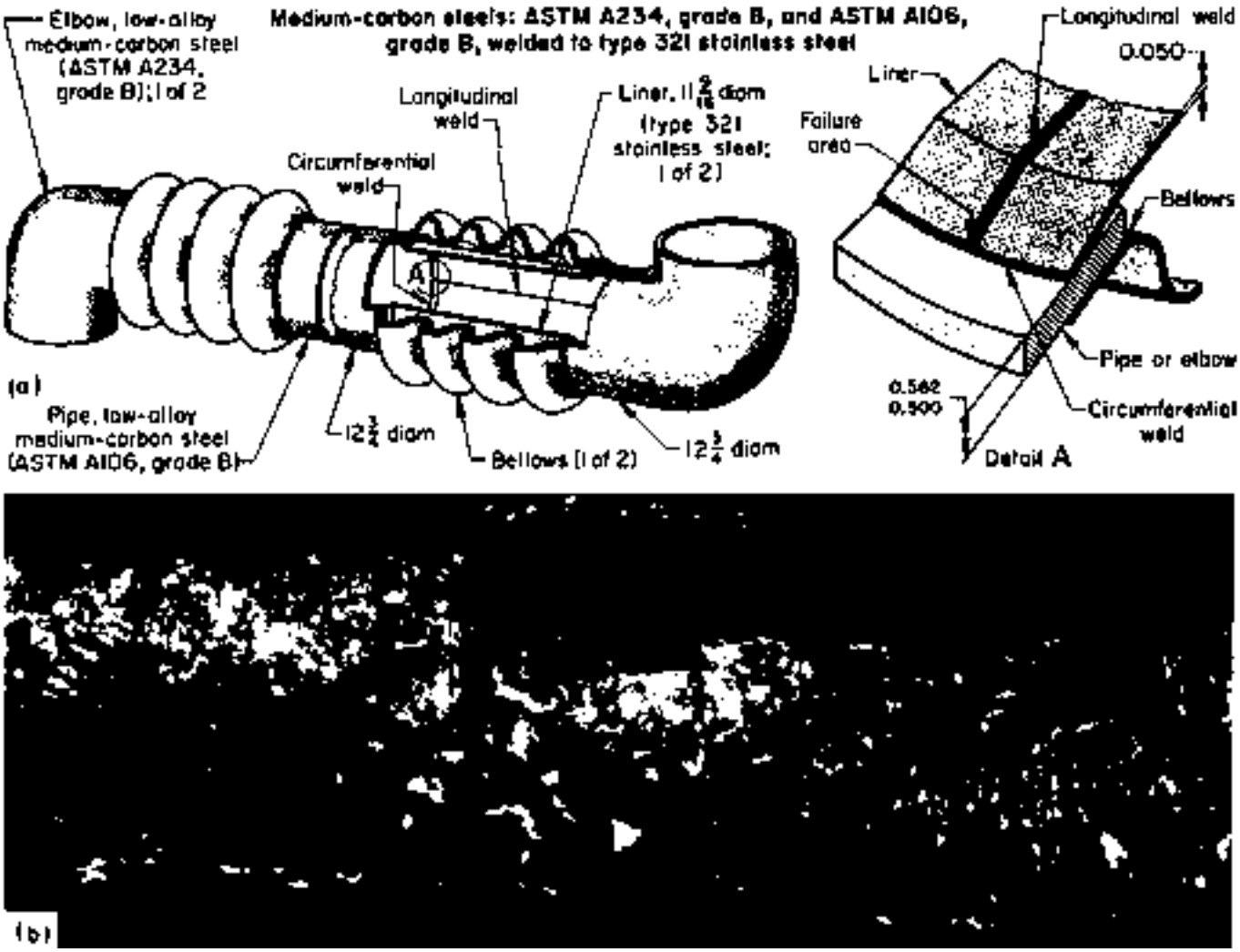


Fig. 16 Duct assembly of medium-carbon steels in which welded bellows liners of type 321 stainless steel fractured in fatigue. (a) Configuration and dimensions (given in inches). (b) Light fractograph showing fracture origin (top edge). 30×

Investigation. In field inspection, portions of the two failed liners were found downstream in the duct assembly. These were removed and sent to the laboratory for failure analysis. Many of the fracture surfaces had undergone extreme deformation, making it difficult to obtain detailed and complete information from the fracture surface. Fortunately, however, some fracture areas were protected from damage, making it possible to identify one fracture origin.

Laboratory inspection showed that the origin of failure in the portion of the fracture surface studied was in an area where two welds intersected—the longitudinal or seam weld forming the liner and the circumferential weld attaching the liner to the pipe. Figure 16(b) shows a fractograph of the fracture surface at the weld intersection, indicating the fracture origin.

Metallographic examination of a cross section containing the longitudinal weld revealed that the microstructure of the weld area contained no irregularities. Metallurgical evaluation of the cross section in the transverse plane of the fractured piece containing the intersection of the longitudinal and circumferential welds exhibited extreme banding within the microstructure. The banding was the result of chromium and nickel dilution in one area and the enrichment of the same elements in areas adjacent to that area. In addition, the intersection of the two welds contained cracks similar in appearance to those normally formed when variations in thermal expansion and contraction are encountered when welding dissimilar metals.

A microhardness traverse across the weld area revealed a base metal average hardness of 89 HRB. The hardness of the HAZ on both sides of the weld was 87 HRB, and the weld metal had a hardness of 93 to 98.5 HRB.

Conclusions. The liners failed in fatigue that was initiated at the intersection of the longitudinal weld forming the liner and the circumferential weld that joined the liner to the bellows assembly.

Because the liners were welded at one end and the other end was free, the components moved with expansion and contraction of the joints and were subjected to vibrational stresses imposed by the transfer of nitrogen gas. The stresses, created by cyclic loading, were concentrated mainly at the metallurgical notches formed at the weld metal/base metal interface. Differences in composition caused a variation in strength at the critical region of high residual stresses. Vibrational stresses were concentrated at the circumferential weld metal/base metal interface and at the intersection of the circumferential and longitudinal welds. The origin of the failure was at the intersection of the two welds—the region containing the most severe stress raisers. The design allowed the two liners to be cantilevered in a dynamic system subjected to both thermal stresses and vibrations from gaseous flow.

Corrective Measure. The thickness of the liners was increased from 1.3 to 1.9 mm (0.050 to 0.075 in.), which successfully damped some of the stress-producing vibrations.

Example 5: Fatigue Cracking of Headers for Superheated Water Because of Notches at Welds. A system of carbon steel headers, handling superheated water of 188 °C (370 °F) at 2 MPa (300 psi) for automobile-tire curing presses, developed a number of leaks within about 4 months after 2 to 3 years of leak-free service. All the leaks were in shielded metal arc butt welds joining 200 mm (8 in.) diameter 90° elbows and pipe to 200 mm (8 in.) diameter welding-neck flanges. A flange-elbow-flange assembly and a flange-pipe assembly that had leaked were removed for examination.

Investigation. Magnetic-particle tests confirmed the existence of the cracks in the welds through which leaking had occurred and revealed the presence of other cracks that had not yet penetrated to the weld surface. Radiographic inspection was conducted, with the source at the centerline of each flange and the film wrapped around the weld bead on the outside. The film showed the cracks indicated by the magnetic-particle inspection as well as additional, incipient cracks and regions of incomplete weld penetration that could not be correlated with the major crack locations.

The butt welds were sectioned, with some cuts being made through regions of pronounced magnetic-particle indication and others at regions where no indications were found. The polished-and-etched surfaces of two such sections from the flange-pipe weld are shown in Fig. 17.

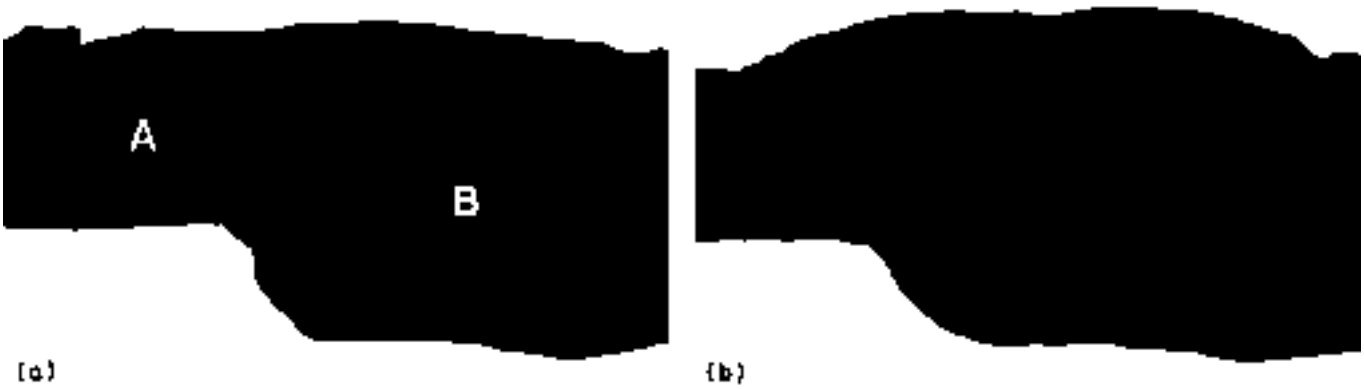


Fig. 17 Flange-to-pipe assembly of a carbon steel header, used for handling superheated water, that cracked by fatigue because of notches at welds. (a) Section through butt-welded joint showing crack (arrow A) that originated at toe of weld on inner surface, incomplete weld penetration (arrow B), and difference in thickness of flange (left) and pipe. Etched with 2% nital. Approximately 1.7×. (b) Section through same weld as in (a), but showing no crack indications. Etched with 2% nital. Approximately 1.7×

It was immediately apparent on sectioning that, although the OD of the flanges matched those of the elbows and pipe (nominal 220 mm, or 8 in.), there was a considerable difference in the inside diameters (IDs). Actually, only the elbows had been of schedule 80 wall (194 mm, or 7 in., nominal ID), as called for in the

specification for flanges, elbows, and pipe. The welding flanges that were used had been bored to 202.5 mm (7.975 in.) ID (for use with schedule 40 pipe), and the pipe had been of schedule 100, with a 189 mm (7.437 in.) nominal ID. The result was a disparity in ID of about 8.9 mm (0.350 in.) between flange and elbow and about 13.7 mm (0.538 in.) between flange and pipe. This was the reason for depositing an internal fillet weld, which created a stress-concentrating notch at the toe of the weld against the flange inner surface.

The welding procedure had been as follows. The edges were prepared by machining a 6.4 mm (0.25 in.) deep groove with a 75° included angle at the outside surface. The initial weld deposit was a root pass applied from the outside. Second and third passes were laid on top of this, the latter depositing a wide, weave-type bead. Finally, an attempt was made to reconcile the difference in wall thickness by depositing a fillet weld bead on the inside. In some instances, penetration was deep enough so that the fillet weld merged with the root pass of the exterior weld. In others, as in Fig. 17, this did not occur; as a result, a short span existed where there was incomplete penetration, and the original edges of the flange and elbow were still visibly in contact. Such a region of incomplete penetration is normally considered a site for stress concentration, but it can be seen in Fig. 17 that there was no crack growth at this point. All the major cracks found in either elbow or pipe welds originated at the toe of the fillet weld in the inner surface of the flange (Fig. 17a). As illustrated in Fig. 17(b), some fillet weld toes did not produce such cracks.

A hardness survey was taken across weld joint. The results showed that the weld beads and the HAZs were similar in hardness to the base metal and that there was no evidence of hard spots. The welds were sound, fine grained, and free from slag inclusions and porosity.

Chemical analysis of the five components of the two assemblies confirmed that these headers conformed to the requirements of the applicable specifications (flanges and elbow of ASTM A 105, grade II, and pipe of ASTM A 106, grade B; similarly numbered ASME SA specifications also apply).

A powdery, green deposit found in the interior of an iron-body gate valve of the system was determined by chemical analysis to consist of a trivalent chromium oxide or salt plus iron oxide or salt and 22% combined water. No organic matter was detected, and no copper was found.

Conclusions. The failures of the butt welds were the result of fatigue cracks caused by cyclic thermal stresses that initiated at stress-concentrating notches at the toes of the interior fillet welds on the surfaces of the flanges. Several of these penetrated to the exterior surface and caused leaks, but others existed that were still in an incipient growth stage. Incomplete penetration existed in some of the welds, but no evidence was found of cracks originating from these sites, and they were not a factor in the failures.

The powdery, green deposit was attributed to reduction of some of the chromate used as a corrosion inhibitor in the superheated water. It was not regarded as a contributing cause of the failures.

Recommendations. It was concluded that all other joints of the system possessing the same mismatch in wall thickness, and thus in all probability the same stress raiser at the toe of the interior fillet weld, could be expected to fail by the same mechanisms. Repair of such joints was not considered feasible, and it was recommended that they be replaced. Ultrasonic testing should be used to identify the joints by detecting the differences between the wall thicknesses of the flange and the pipe and of the flange and the elbow. Special attention to accuracy of fit-up in the replacement joints was also recommended to achieve smooth, notchfree contours on the interior surfaces.

Example 6: Fatigue Failure of an Inlet Header Because of Poor Welding Technique and Unfavorable Weld-Joint Design. While undergoing vibration testing, a type 347 stainless steel inlet header for a fuel-to-air heat exchanger cracked in the header tube adjacent to the weld bead between the tube and header duct (Fig. 18).

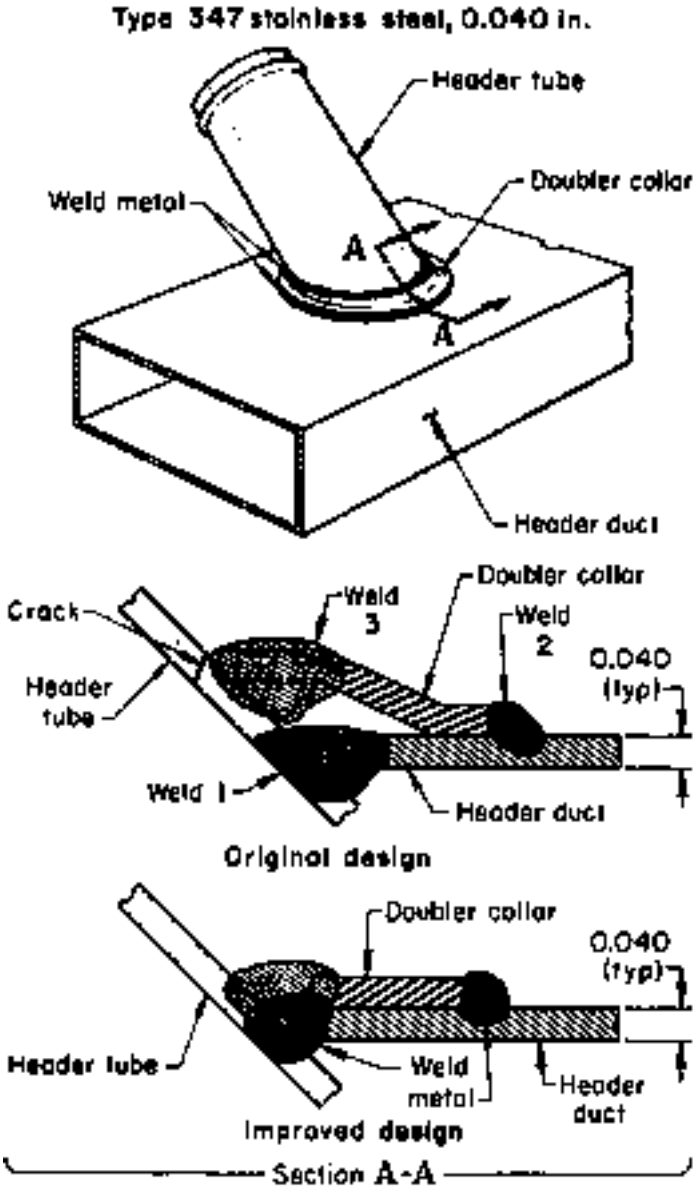


Fig. 18 Type 347 stainless steel inlet header for fuel-to-air heat exchanger that cracked due to poor welding technique and unfavorable joint design. Dimensions given in inches

Investigation. The weld was relatively massive, which was apparently necessary to fill large clearances between the mating parts. Collapse of the base metal in several areas adjacent to the weld indicated that there was plastic deformation from excessive heat. Liquid penetrant inspection of the weld between the header tube and the header duct revealed a 29 mm (1.14 in.) long crack in the weld on the tube side. Several weld bead undercuts were visible, and the crack appeared to propagate through each of these small imperfections.

The header was sectioned to open the crack and expose the fracture surface for examination. The tube and duct were initially welded together (weld 1, original design, section A-A, Fig. 18). A doubler collar, 16 mm (0.63 in.) wide and of the same material and thickness as the duct, was then welded to the duct along the outer edge of the collar (weld 2). A third weld joined the inner edge of the collar to the header tube, thus creating an area of twice the normal header thickness adjacent to the joint. This third weld, throughout much of the collar circumference, penetrated the original tube-to-duct weld (weld 1), but in the area of the fracture, there was a gap between the two sheets of stainless steel, and the third weld was separate (weld 3).

Examination of the underside of the joint showed incomplete fusion of the weld between the tube and duct. Distortion of the base metal adjacent to the weld was also visible from this side.

Examination of the surface of the opened crack revealed the origin of the crack at a weld undercut at the toe of weld 3. The crack then progressed through the HAZ of the base metal.

Conclusions. The crack in the header tube was the result of a stress concentration at the toe of the weld joining the doubler collar to the tube. The stress concentration was caused by undercutting from poor welding technique and an unfavorable joint design that did not permit a good fit-up.

Corrective Measures. The doubler collar was made so that it could be placed in intimate contact with the header duct (improved design, section A-A, Fig. 18). The two sheets were beveled, where necessary, to form a V-groove joint with the tube. Two weld passes in the V-groove were used to join the two sheets to the tube. This procedure resulted in a smaller, controlled, homogeneous weld joint with less distortion.

Other Failures Related to Joint Design

Discontinuities can be introduced during the welding process due to poor joint design. Joint design may contribute to excessive shrinkage stresses, increasing the risk of both solidification cracking and cold cracking, especially for high-strength materials. Special care must be taken to generate proper bead shape, preheat temperatures, and interpass temperatures in addition to correct electrode and flux combinations when welding high-strength materials. Care must also be taken to reduce hydrogen pick up in these welds. Dirt, grease, oil, moisture, and all hydrocarbons must be removed from the weld zone. A low-hydrogen welding process must be used.

The magnitude of the induced stresses also depends on the joint design, the degree of restraint imposed on the joint, the plate thickness, the orientation of the weld, and the size of the weld. Restrained corner or T-joints are the most susceptible to lamellar tearing, because the fusion boundary is roughly parallel to the plate surface (Fig. 19) (Ref 21) and through-thickness contraction stresses are high. Lamellar tearing in butt welds is rare. Thick-plate, high-imposed restraint (for example, rigid clamping) and large weld beads all increase residual stresses and the likelihood of this form of tearing.

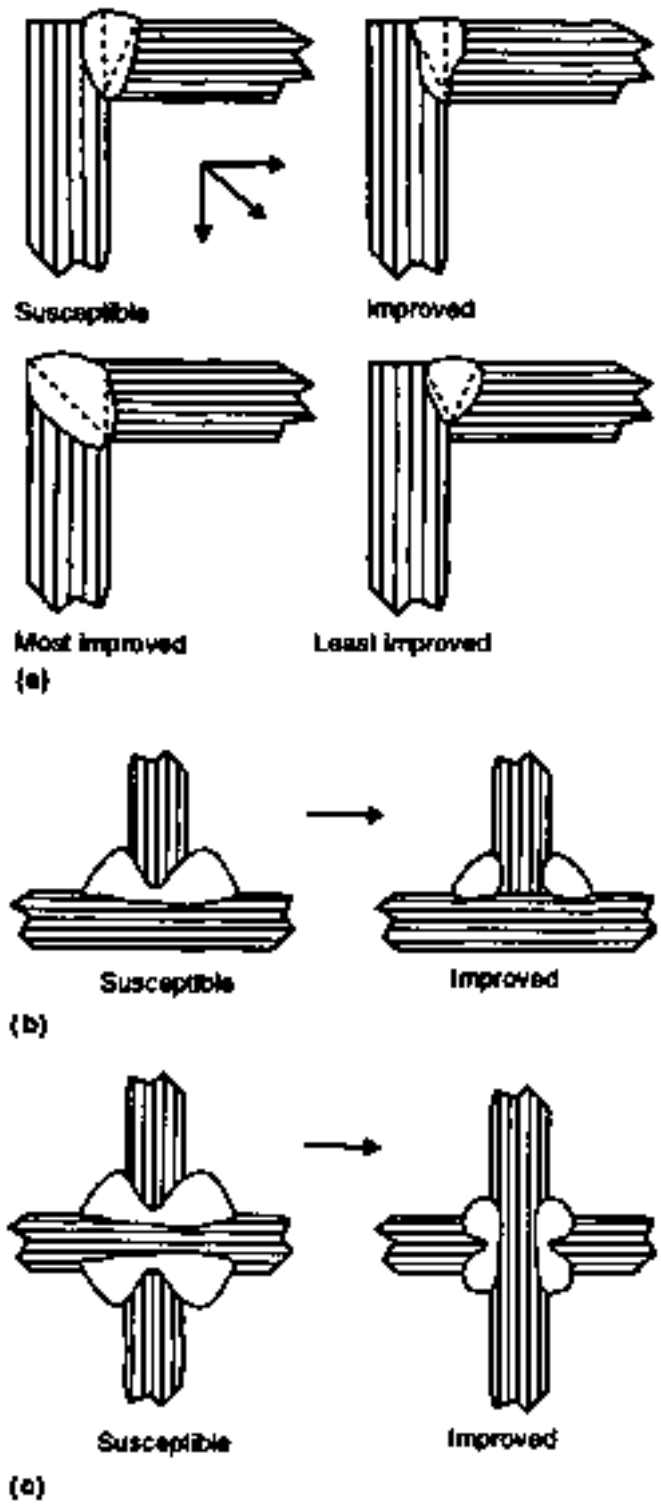


Fig. 19 Methods used to reduce lamellar tearing. (a) Angling the weld fusion line to avoid shrinkage stresses in the through-thickness direction. (b) Use of smaller partial-penetration welds to reduce joint restraint. (c) Placement of welding beads on the thinner plate (when welding plates of different thicknesses) so that smaller beads can be used. Source: Ref 21

Problems with joint penetration and incomplete root fusion can also be due to the use of a joint design that is not proper for the welding process being used, or to a disregard for the procedures that are intended to provide adequate penetration. A joint design developed for one metal is also not necessarily suitable for other alloys and should not be used unless it proves to be satisfactory, based on either experience or tests. Choosing the best process and design depends on many factors including:

- Material characteristics (sensitivity to porosity or cracking and the required mechanical properties)

- Details of the joint (plate thickness, joint design, restraint, welding position, heat input, and location)
- Weld economics (deposition rates and efficiencies, cost of labor, cost of consumables, capital cost of equipment, number of spare parts required, operator skills or training needed, setup, clean-up cost, and so on)

All these factors are considered in design and are covered in numerous other publications on welding, such as *Welding, Brazing, and Soldering,* Volume 6 of *ASM Handbook.* The following are just a few examples of failures related to weld design.

Example 7: Incomplete Fusion due to Poor Accessibility. A void caused by incomplete fusion is illustrated in Fig. 20 from a gas tungsten arc weld between a nickel alloy and an alloy steel. Special care must be taken when welding with nickel alloy filler metal in order to avoid this problem of incomplete fusion. The primary factor is weld-joint design. Molten nickel alloy weld metal is sluggish and has low fluidity; it does not wet or flow as well as steel weld metal. Therefore, joint accessibility must be sufficient to allow proper manipulation of the welding arc.

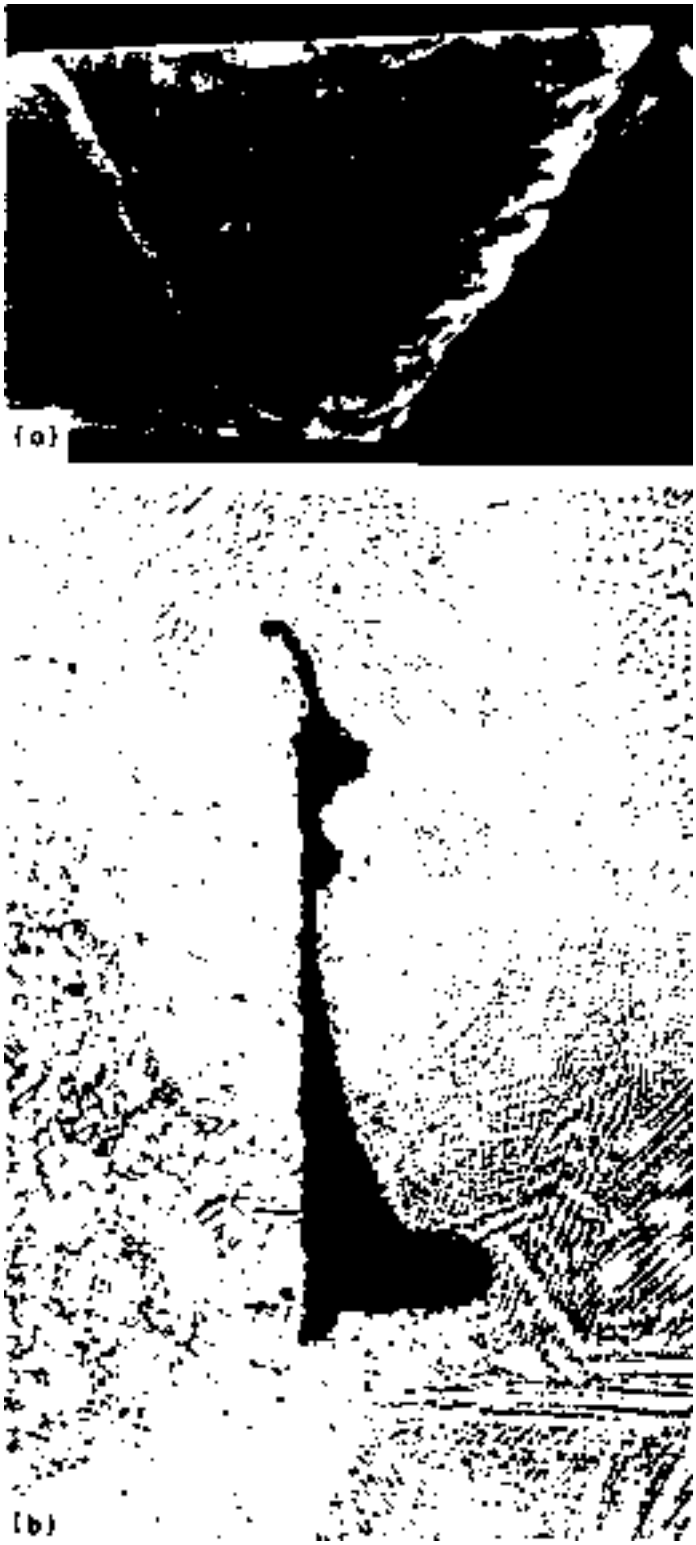


Fig. 20 Section through an automatic gas tungsten arc weld containing voids caused by incomplete fusion. (a) Base metal at left is Incoloy 800 nickel alloy, that at right is 2.25Cr-1.0Mo alloy steel. Filler metal was ERNiCr-3, used with cold wire feed. Macrograph. 1×. (b) Micrograph of the area circled in (a) showing void. 75×. Both etched with 2% nital

Example 8: Cracking of a Weld That Joined the Head to the Shell of a Steam Preheater Because of Poor Root Penetration. A weld that attached the head to the shell of a preheater containing steam at 1.4 MPa (200 psi) and was used in the manufacture of paper cracked in service. The length of service was not reported. A section of the failed preheater, plus a section of proposed new design, were selected for examination.

Investigation. The preheater was approximately 0.9 m (3 ft) in diameter and approximately 1.2 m (4 ft) long. The original joint (Fig. 21a) contained a 6.4 by 50 mm ($\frac{1}{4}$ by 2 in.) backing ring that had been tack welded to

the inside surface of the shell in a position to project approximately 16 mm ($\frac{5}{8}$ in.) beyond the fully beveled top edge of the shell. The ring served the dual purpose of backing up the weld and positioning the head; the projecting edge of the ring fitted against a 9.5 mm ($\frac{3}{8}$ in.) wide, 3.2 mm ($\frac{1}{8}$ in.) deep undercut on the inner corner of the rim of the head. The internal 90° angle in this undercut was sharp, with almost no fillet. A bevel from the lower edge of the undercut to the outside of the head completed the groove for the circumferential attachment weld.

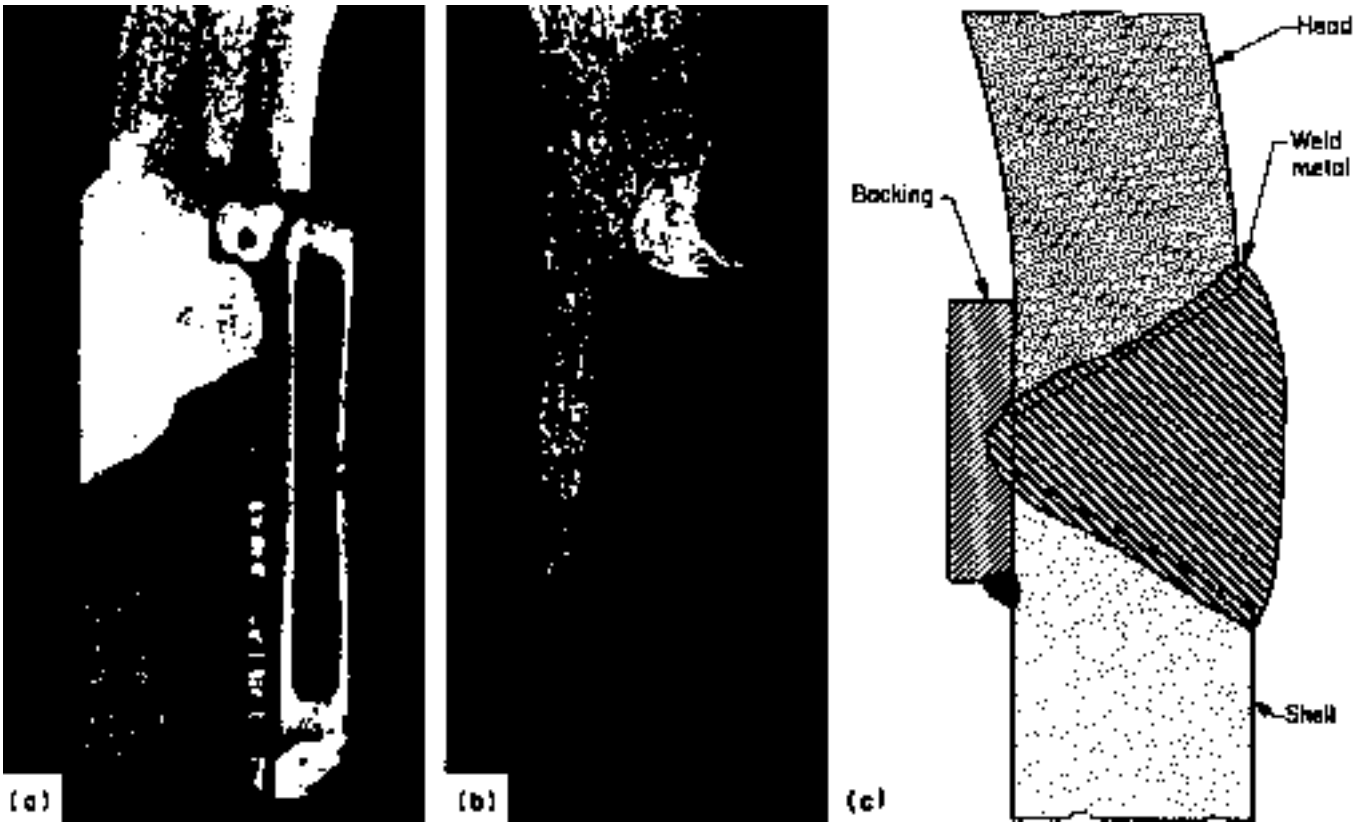


Fig. 21 Weld attaching the head to the shell of a steam preheater that cracked because of poor root penetration in original and first replacement joint designs. (a) and (b) Sections taken through the head-to-shell joint. Etched in hot 50% hydrochloric acid. Actual size. The section in (a) is the original design; (b) shows the first replacement design. (c) The final design, which achieved full root penetration

The head was joined to the shell by arc welding in three passes. In all sections that were taken through the welded joint and examined, the fusion was found to range from poor to none at the root of the weld and at the backing ring. The weld did not penetrate to the undercut rim of the head, and in some locations, the weld bead was standing free (not fused with the backing ring). In all sections of the joint examined, there was a crack beginning at the sharp, internal 90° angle of the undercut and extending toward the outer surface of the head. This crack had to penetrate only 50% of the full thickness of the head before final fracture occurred.

The first proposed replacement design (Fig. 21b) was more or less the reverse of the failed joint. A deep offset had been machined in the exterior of the rim of the head, reducing the wall thickness by 65%. The outer portion of the offset had been beveled to receive the weld deposit; below the offset, a horizontal step, approximately 2.5 mm (0.10 in.) deep, had been cut to align the head against the square-cut shell edge. Below the step was a long taper almost to the inner surface of the head. A number of sections through the proposed replacement joint showed incomplete root penetration of the weld, and in one section, a crack was discovered extending from the outer edge of the positioning step through the root-pass weld bead.

Conclusions. Cracking occurred in the HAZ in the head of the original design, originating in the sharp corner of the undercut, which was an inherent stress raiser. In the first replacement design, beveling of the head and shell was insufficient to permit full root-weld penetration, thus creating an inherent stress raiser. In both designs, the stress raisers would not have been present had full root-weld penetration been obtained.

Corrective Measures. To ensure full root penetration, the joint design shown in Fig. 21(c) was adopted. This joint retained the full thickness of both the head and the shell and, with the gap between the root edges, made it possible to obtain full root penetration, thereby reducing any notch effect to a minimum. A suggested alternative that would avoid the possibility of crevice corrosion was the use of a single-V-groove joint without a backing ring, with welding being done from the outside in two or more passes, ensuring that the root pass attained full penetration.

Example 9: Fracture of a Carbon Steel Pipe in a Cooling Tower. Joint design was responsible for mechanical notches at welds in a heat-exchanger component (Fig. 22). A 455 mm (18 in.) diameter, 8 mm ($\frac{5}{16}$ in.) wall carbon steel discharge line for a circulating-water system at a cooling tower fractured in service; a manifold section cracked where a Y-shaped connection had been welded. The steel pipe was made to ASTM A 53 specifications.

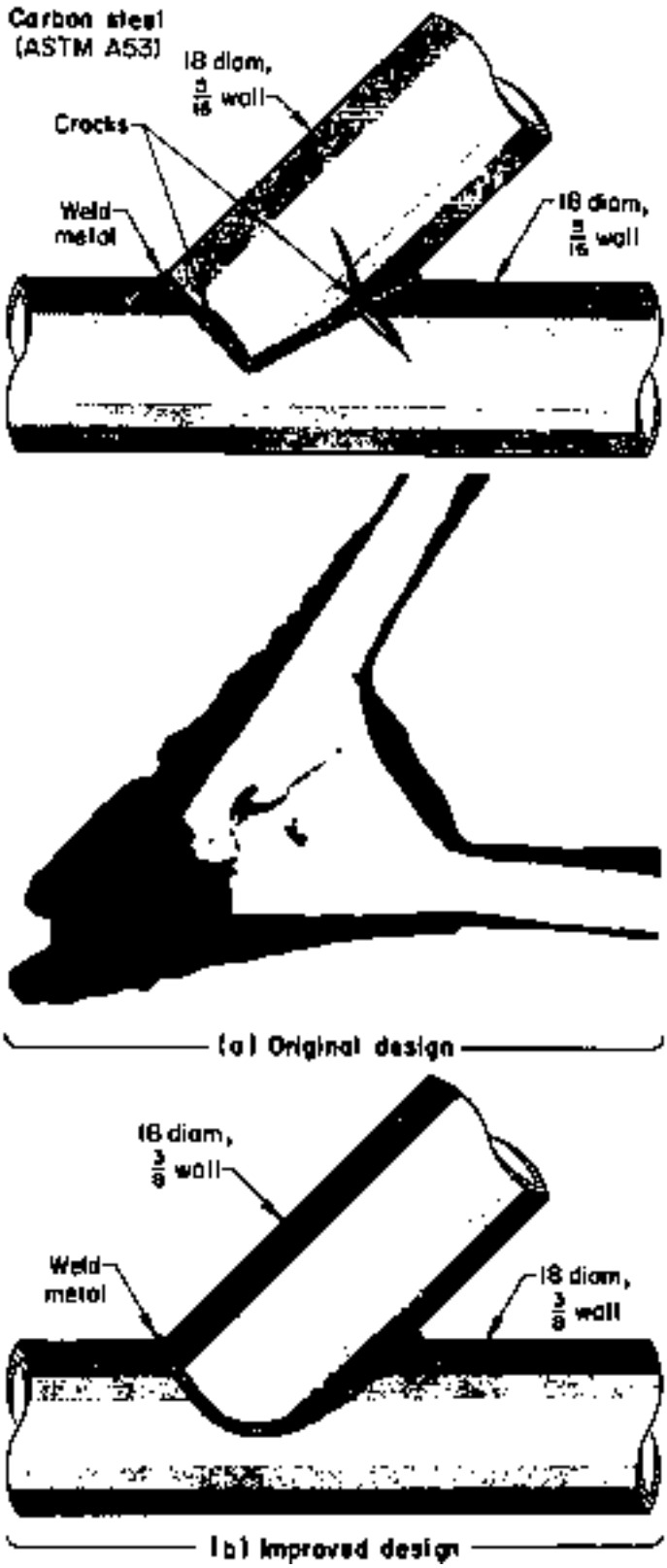


Fig. 22 Carbon steel discharge line at a cooling tower that failed because of poor fit-up at Y-joint and poor-quality welds. (a) Original joint design of pipe connection and location of cracks. Photograph is an oblique view of a section through the weldment, showing the abrupt intersection of pipe walls and the voids and crevices in the weld metal. (b) Improved joint design. Dimensions given in inches

Investigation. Examination of the pipe revealed that cracking occurred in the HAZ both transverse and parallel to the weld, as shown in Fig. 22(a). The end of the branch pipe was prepared for welding by making a transverse cut; a second cut was then made at about 15° to the longitudinal axis of the pipe, resulting in a sharp corner (Fig. 22a). The main pipe was then notched to fit the end of the branch pipe. Fitting of the saddle connection was poor, and no backing strips were used even though the pipe wall was thin; as a result, the

condition of the root-weld bead was below standard, and numerous deep crevices and pits were in the weld area exposed to the inner surfaces of the manifold (sectional view, Fig. 22a).

Conclusions. The pipe failed by fatigue. Cracks originated at crevices and pits in the weld area that acted as stress raisers, producing high localized stresses because of the sharp-radius corner design. Abnormally high structural stresses and alternating stresses resulting from the pump vibrations contributed to the failure.

Corrective Measures. The joint design was changed to incorporate a large-radius corner (Fig. 22b), and fitting of the components was improved to permit full weld penetration. Backing strips were used to increase weld quality. Also, the pipe wall thickness was increased from 8 to 9.5 mm ($\frac{5}{16}$ to $\frac{3}{8}$ in.).

Example 10: Failure of Welds in an Aqueduct Caused by Poor Welding Techniques. A 208 cm (82 in.) ID steel aqueduct fractured circumferentially at two points 152 m (500 ft) apart in a section above ground. A year later, another fracture occurred in a buried section 6.4 km (4 miles) away. Both pipes fractured during January at similar temperatures and pressures. The pipe had a 24 mm ($\frac{15}{16}$ in.) wall thickness, and the hydrostatic head was 331 m (1085 ft). The air temperature was approximately −13 °C (9 °F), the water temperature approximately 0.6 °C (33 °F), and the steel temperature approximately −4 °C (25 °F). The fractures occurred at bell-and-spigot slip joints (Fig. 23a) that, at the two initial fractures, had been fillet welded both inside and out; at the third break, the joints had been fillet welded only on the inside. The pipe had been shop fabricated of ASTM A572, grade 42, type 2, steel in 12 m (40 ft) lengths, then shop welded into 24 m (80 ft) lengths. Field assembly was with the bell-and-spigot joints.

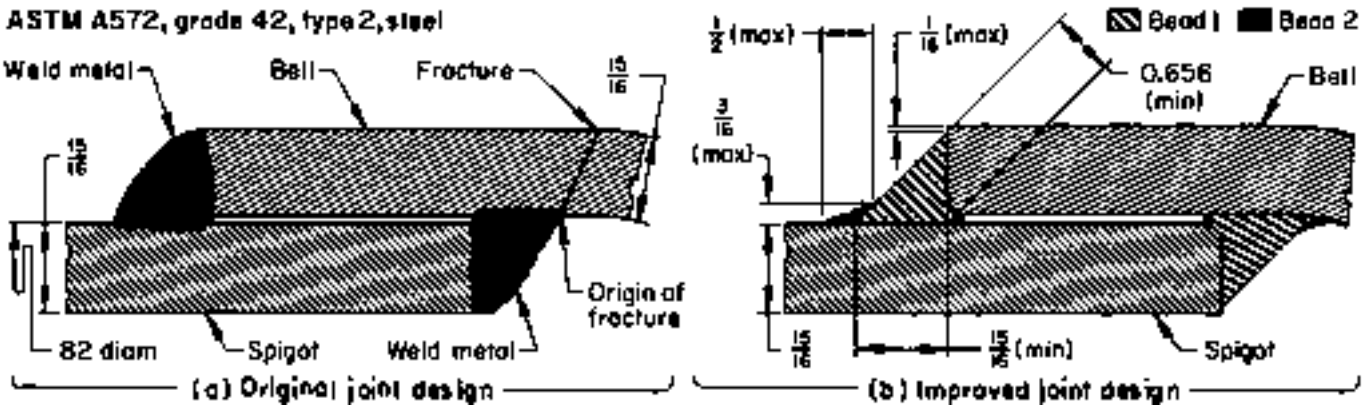


Fig. 23 Bell-and-spigot joint used in an aqueduct of steel pipe. (a) The original design cracked because of poor welding technique and poor choice of metal. (b) Improved design showing modification of weld beads. Dimensions given in inches

Investigation. All three fractures occurred at a bell section and initiated at the toe of the inside fillet weld (Fig. 23a). The welds in the first two fractures had been made in the field; the third weld that failed had been made in the shop. Also, both of the original inside welds at the early fractures had been gouged out to permit relocation of the pipe to maintain the supporting ring girder legs in a vertical position. The joints had then been rewelded on the inside. This had not been done in the third weld that failed later.

Charpy V-notch impact tests on the steel used for this pipeline revealed a ductile-to-brittle transition temperature that was unfavorably high—approximately 10 °C (50 °F). At the temperatures existing at the time of the failures, the impact strength of the steel was approximately 13.5 J (10 ft · lbf), which is considered to be dangerously brittle.

Inspection of the interior field welds showed that instead of using continuous circumferential stringer beads, the welder had used a wash-pass technique, depositing a heavy bead from toe to crown. This permitted faster welding but gave inconsistent penetration and a considerable number of slag inclusions and imperfections, such as deep valleys, grooves, and undercuts, which served as notches and stress raisers, at the toe of the weld.

Examination of the third fracture at the shop weld disclosed that areas in the HAZ and at the toe of the weld had hardness levels as high as 34 HRC. When test welds were made on metal preheated to 95 °C (200 °F) before welding, the hardness of the HAZ was 22 HRC.

Conclusions. Brittle fracture of the aqueduct pipe was attributed to a combination of stress concentrations at the toes of the fillet welds due to poor welding technique, including shop welds made without preheat, and a brittle condition of the steel at winter temperatures caused by a relatively high ductile-to-brittle transition temperature.
Corrective Measures. The following changes were made to eliminate welding and material problems:

- Any area to be welded was preheated to at least 65 °C (150 °F).
- In any weld having an undercut or abrupt transition to the base metal at the toe, an additional contour weld bead was deposited at the toe (Fig. 23b).
- Tensile loadings of above-ground sections were reduced by the installation of expansion joints.
- Where the pipe was buried and pipe temperatures varied considerably, butt welds were used.
- All ASTM A572 steel plate used was rolled from fully killed ingots made with a grain-refining practice and normalized after rolling, which provided a ductile-to-brittle transition temperature of approximately -32 °C (-25 °F).

References cited in this section

4. S.J. Maddox, "Fitness-for-Purpose Assessment of Misalignment in Transverse Butt Welds Subject to Fatigue Loading," Document XIII-1180–85, International Institute of Welding, 1985

17. O.W. Blodgett, Joint Design and Preparation, *Welding, Brazing, and Soldering,* Vol 6, *Metals Handbook* (9th ed.), 1983, p 60–72

18. S.J. Maddox, "Scale Effect in Fatigue of Fillet-Welded Aluminum Alloys," presented at Sixth Int. Conf. Aluminum Weldments (Cleveland), American Welding Society, 1995

19. V.I. Makhnenko and R.Y. Mosenkis, Calculating the Coefficient of Concentration of Stresses in Welded Joints with Butt and Fillet Welds, *Automat. Weld.,* Aug 1985

20. K. Iida, Y. Kho, J. Fukakura, M. Nihei, T. Iwadate, and M. Nagai, "Bending Fatigue Strength of Butt Welded Plate with Uranami Bead," IIW/IIS-XIII-1202–86, International Institute of Welding, 1986

21. "Welded Steel Construction (Metal Arc Welding)," CSA W59, Canadian Standards Association, 1989

Failure Origins in Arc Welds

The general types of process-related discontinuities in arc welds include:

- Porosity formed by gas entrapment during solidification
- Inclusions: slag inclusions, tungsten inclusions, oxide inclusions
- Lack of fusion (LOF) and lack of penetration (LOP)
- Geometric discontinuities, such as poor weld contours, undercut, underfill, overlap, excessive reinforcement, fillet shape, melt-through, misalignment, and distortion
- Shrinkage voids formed during solidification
- Craters or depressions that form at the termination of a weld bead or in the molten weld pool
- Spatter that forms when metal particles are expelled during welding
- Arc strikes (or arc burns) that form from any localized remelted metal, heat-affected metal, or change in the surface profile of any part of a weld or base metal resulting from an arc
- Metallurgical discontinuities, such as cracks, fissures, segregation, and lamellar tearing

In addition, localized embrittlement in welded regions can occur by a number of mechanisms. The three most common mechanisms are hydrogen embrittlement, strain aging, and temper embrittlement of steels. Each of these mechanisms can reduce the toughness of the material, either soon after welding or during service.

Most discontinuities may be produced by any of the welding processes, although some are found only in welds made by a particular process. The observed occurrence of discontinuities and their relative amounts depend largely on the welding process used, the inspection method applied, the type of weld made, the joint design and fit-up obtained, the material used, and the working and environmental conditions.

Weld acceptance standards also are used when a discontinuity has been clearly located, identified, sized, its orientation determined, and its structural significance evaluated. For example, the API standard 1104 (Ref 22) includes criteria for acceptance of steel welds based on a number of imperfections, including inadequate penetration, incomplete fusion, internal concavity, burn-through, slag inclusions, and porosity.

Acceptance criteria depends on the inspection method and the effects of the discontinuity on service fitness. As previously noted, moderate amounts of volumetric discontinuities, such as slag inclusions and porosity, typically do not reduce the fatigue strength of the weld sufficiently to warrant their classification as potential causes of fracture. In fact, fatigue cracks that initiate at the toe of a weld will propagate around areas of gas porosity or slag inclusions as often as through them (Fig. 24). Nonetheless, welders must still be constantly encouraged to make sound (perfect) welds independent of prevailing acceptance standards.

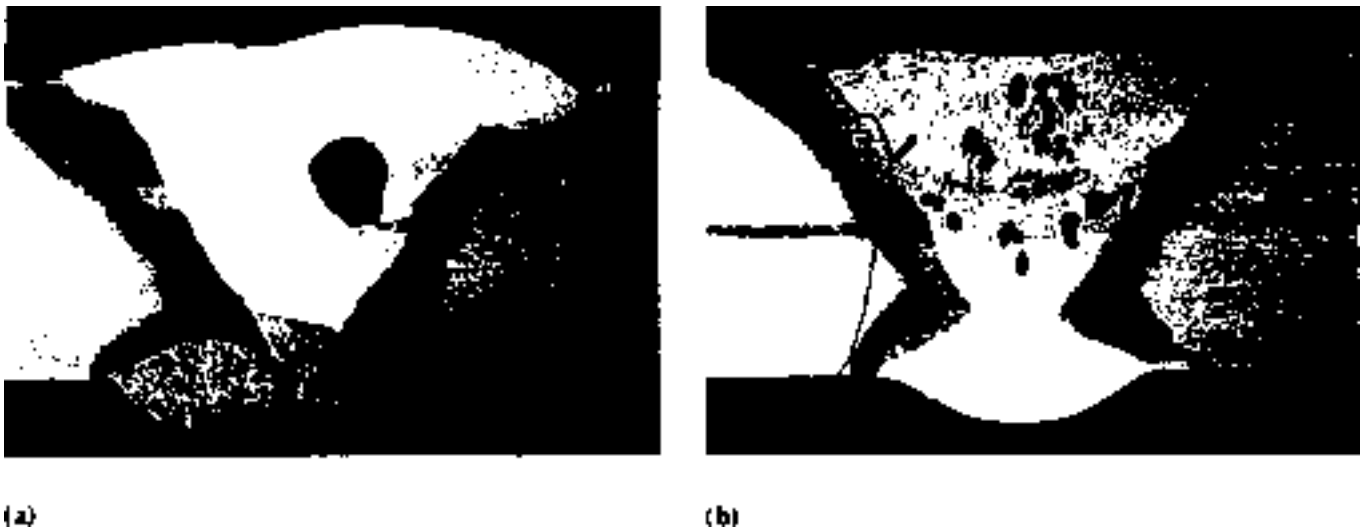


Fig. 24 Cracking from weld toe around volumetric discontinuities in butt welds. (a) Slag inclusion. (b) Porosity

Discontinuities such as hot (liquation or solidification) cracks, cold (hydrogen-induced) cracks, and lamellar tears are much more serious, because they are severe stress raisers and affect fatigue strength to a degree that cannot be compensated for in design. Hot cracks and lamellar tears normally occur during or immediately after welding and can be detected by nondestructive testing and eliminated from finished weldments. However, the formation of cold cracks resulting from the presence of hydrogen may be significantly delayed. For this reason, inspection is usually delayed at least 48 h (and, in many cases, several weeks) after welding.

This section briefly reviews the general types of process-related discontinuities and metallurgical-related cracking of arc welds. Failure origins related to other types of welding processes are described in subsequent sections of this article. Additional information on the engineering performance and properties of weldments for specific types of common structural alloys (e.g., carbon and low-alloy steels, stainless steels, aluminum alloys, titanium alloys) is in Ref 23.

Porosity

Porosity consists of cavities or pores that form in the weld metal as a result of entrapment of gases evolved or air occluded during the welding process, as evidenced by Fig. 25 and 26. Porosity in welds is caused by gas entrapment in the molten metal, by too much moisture on the base or filler metal, or by improper cleaning of the joint during preparation for welding. Elements such as sulfur, lead, and selenium in the base metal, and foreign contaminants such as oil, grease, paint, rust, and moisture in the welding area also can increase porosity in the weld metal. Porosity in aluminum is primarily caused by the entrapment of hydrogen during the solidification process from sources such as filler metal, base metal, surface contamination, and shielding gas.

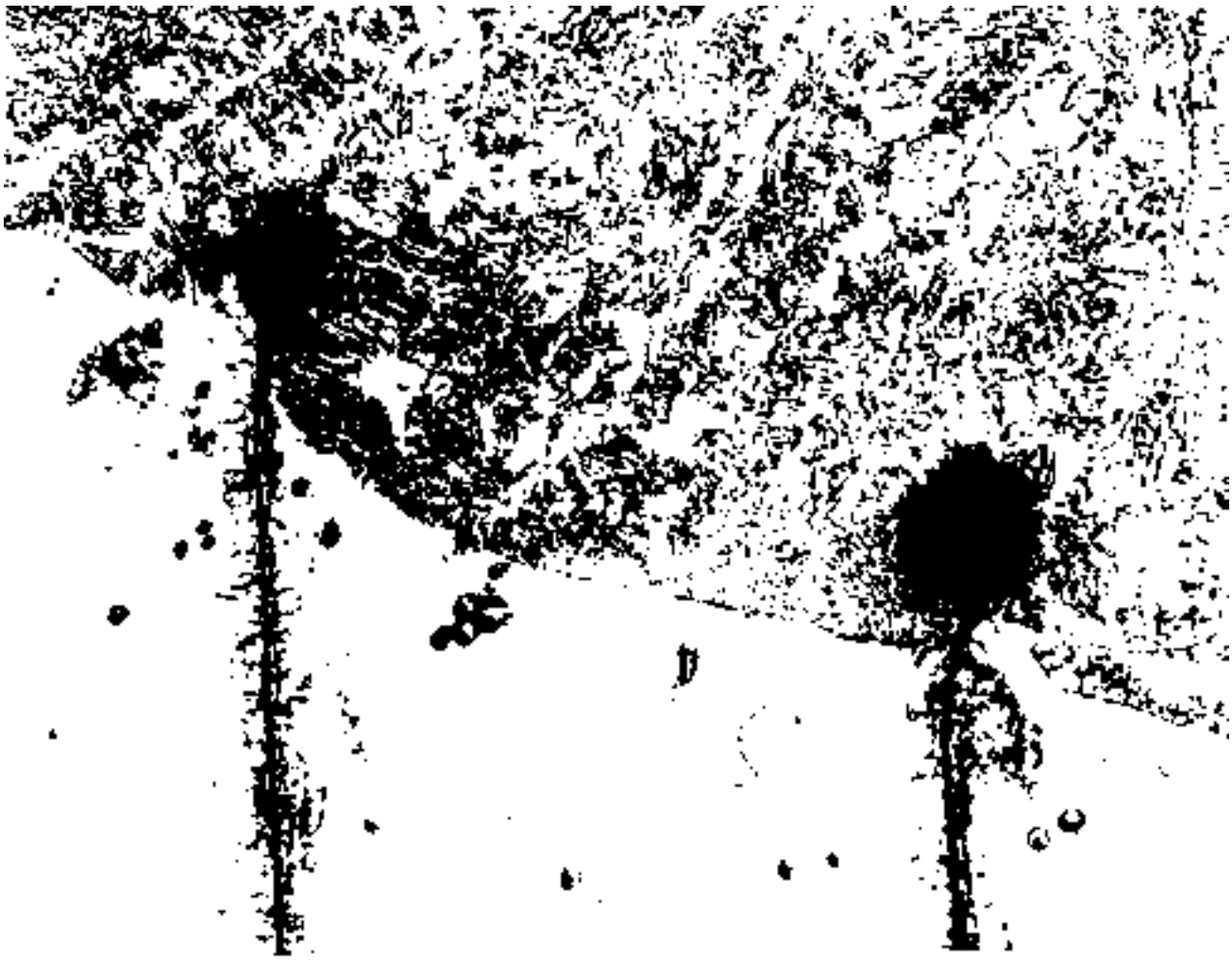


Fig. 25 Porosity in GMAW core-plated silicon steel laminations. 100×

Fig. 26 Pulsed GMAW spot weld showing porosity in dissimilar metal weldment; a copper-nickel alloy to a carbon-manganese steel using an ERNiCu-7 (Monel 60) electrode. Etchant, 50% nitric-50% acetic acid. 4×

Pores can take many shapes and sizes—from pinholes to large voids, from spherical to elongated or pear-shaped, with constrictions or expansions. Gas porosity can occur on or just below the surface of a weld and have a rounded or elongated teardrop shape with or without a sharp point. Distribution of porosity in welds may be aligned, clustered, or uniformly scattered. Pores can be uniformly distributed throughout the weld or isolated in small groups; they can also be concentrated at the root or toe of the weld.

Gas porosity is usually well distributed in overhead-position welds. It is frequently found along the top edge of welds made in the horizontal position and is least likely to be encountered in welds made in the vertical-up position unless there is a problem with shielding the molten puddle. Linear porosity is most often found in welds where the root-pass penetration is inadequate. Aligned or layered porosity exerts a much greater effect on mechanical properties than does uniformly distributed porosity. Microporosity, too fine to be detected by radiographic procedures, typically occurs in layers along the weld fusion line and at weld-pass interfaces. This type of porosity can lower mechanical properties appreciably, if it is extensive enough to lower static strength. Microporosity is typically found in aluminum alloy weldments.

Two common types of porosity are wormholes and blowholes. Wormholes are elongated voids with a definite worm-type shape and texture (Fig. 27). Blowholes may form on the surface of the weld if porosity cannot fully escape before the weld metal solidifies. “Herringbone” porosity (Fig. 28a) can form when the gas shield in an automated GMAW process is interrupted. Its appearance is due to the formation of highly elongated pores inclined to the direction of welding.

Fig. 27 Wormhole porosity in a weld bead. Longitudinal cut. ~20×

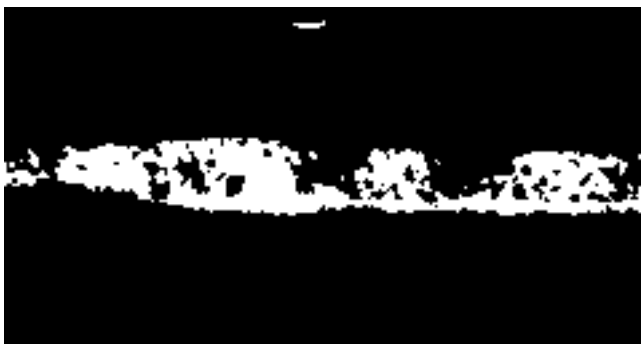


Fig. 28 Radiograph showing herringbone porosity in automatic weld due to disruption of gas shield

Causes of Weld Porosity. The most common cause of porosity is the result of gas-producing contaminants in the joint, but several other factors can influence the occurrence of porosity. For example, porosity may be associated with welding practice. Scattered surface porosity may result in submerged arc welding when the arc flashes through, due to insufficient flux. Insufficient flux also is more likely to occur in circumferential welds

than flat welds. Arc blow also can cause porosity, particularly in high-speed welds on light-gage sheet. Wherever there is an opportunity for flux to become trapped, there is the possibility of porosity.

Also, even though base metal composition generally has little effect on porosity, there are circumstances when composition of the base metal may influence porosity. For example, very high or very low carbon contents in steel can cause porosity. Segregation of impurities, such as sulfur in steel, may also be a factor. Instances of severe segregation, which appear as laminations, may cause large holes in welds (see Example 11 in this article).

The interrelationship of factors that may cause or control porosity is illustrated in Fig. 29. To minimize porosity:

- Remove scale, rust, paint, grease, oil, and moisture from the surface of the joint.
- Eliminate moisture from the flux or the shielding gas.
- Use short arc length (particularly with low-hydrogen electrodes).
- Keep the weld molten long enough (use preheating) to allow the gas to escape.
- Provide sufficient shielding (flux or gas).
- Improve fit-up.
- In steel, remove slag residue that may cause porosity, particularly in butt welds (fluxes on certain types of electrodes are worse than others in this respect). To control slag residue, allow weld penetration into the backup strip or maintain a clearance of at least 4 mm ($\frac{5}{32}$ in.) above the backup.
- Use low-hydrogen electrodes when welding steels that are low in carbon and manganese and high in sulfur and phosphorus. Control admixture of the weld and the workpiece by using low current and fast travel speed to obtain less penetration.
- Keep the sulfur content low in the base metal, thus minimizing formation of hydrogen sulfide (H_2S).
- Decrease surface tension and viscosity of both the liquid weld metal and the molten flux.
- Prevent contamination of GMAW filler-metal wire.

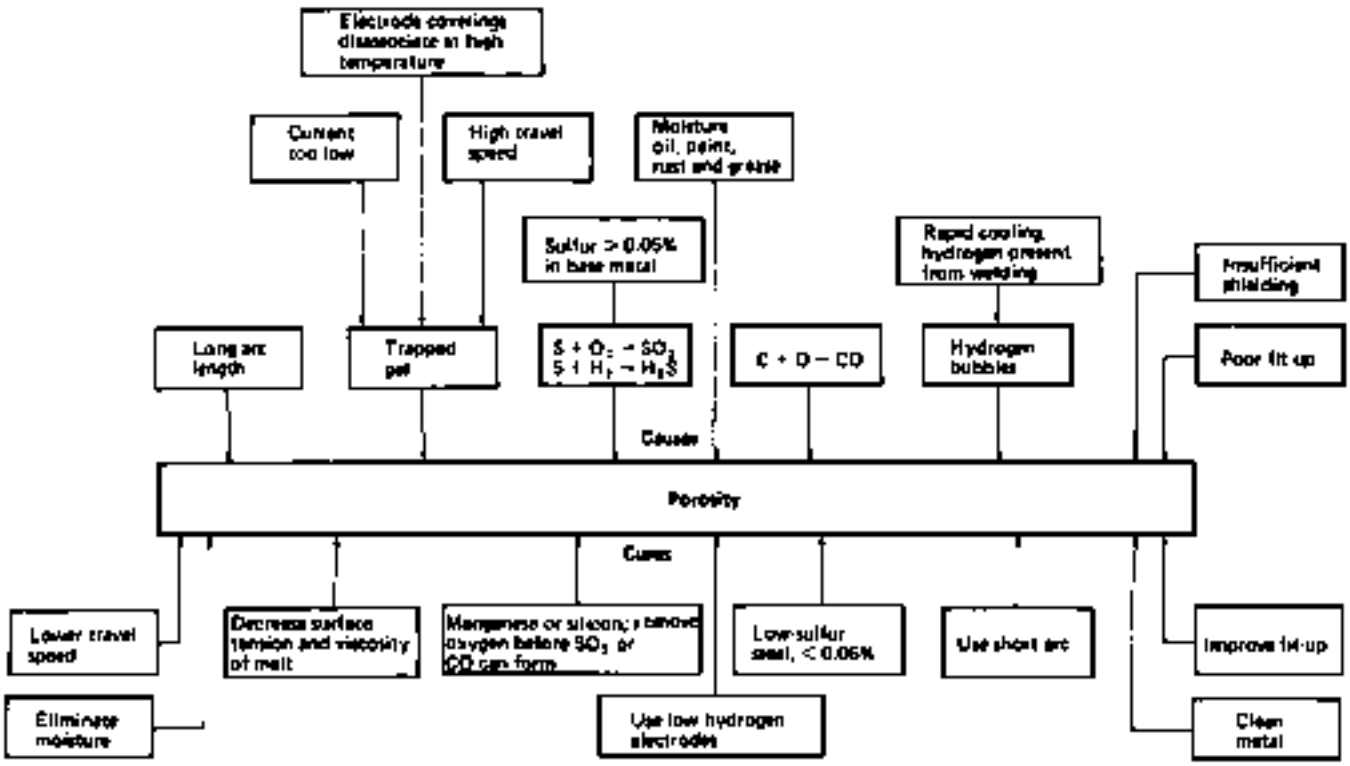


Fig. 29 Factors influencing porosity in weldments

The occurrence of porosity in aluminum alloy welds is almost always caused by entrapment of hydrogen gas. Hydrogen has high solubility in molten aluminum, but its solubility in solid aluminum is only a small fraction of its solubility in solid steel or solid titanium. Thus, hydrogen dissolves very rapidly in molten aluminum and

can become trapped during solidification. For aluminum, significant amount of porosity can even occur from an extremely limited source of hydrogen from, for example, the moisture of a fingerprint.
Weld-pool fluidity (or viscosity) is also a major controlling factor. For example, pores form most readily in aluminum when it is welded with filler metal such as aluminum-magnesium that produces a weld pool of a low fluidity. It is much less of a problem in welds made with an aluminum-silicon filler metal that produces a highly fluid weld pool. In GMAW of wrought aluminum alloys, contamination of the filler-metal surface is the predominant source of porosity because of the large surface area. When powder metallurgy parts are gas metal arc welded, the primary controlling factor in the occurrence of pores is the base metal itself.
The possible sources of hydrogen, which must be eliminated in order to produce sound aluminum welds, include:

- Poor quality of electrode or filler wire
- Lubricant on the base metal or filler because of careless storage or unsatisfactory cleaning of base metal surfaces
- Contamination of welding room by paint spray or oil drips
- Water condensate on base metal
- Moisture on the surface or in the form of a hydrated oxide on the base or filler metals (perhaps caused by failure to clean adequately the surfaces to be welded)
- Moisture leaks or condensation through poor torch seals inside the nozzle of water-cooled welding torches
- Leakage of moist air into the shielding gas through defective hoses
- Use of a shielding gas with a high dewpoint

Example 11: Fracture in HAZ due to Sulfide Laminations. In rolled carbon steel plate, the presence of flattened inclusions, such as silicates or manganese sulfides, can cause difficulties, particularly in welding T-joints. The cooling stresses may cause rupture by decohesion between the inclusions and the steel matrix, producing a fracture with a laminated appearance (Fig. 30a). The nature of fracture through the coarse-grain reaustenitized structure of the HAZ is shown in profile in Fig. 30(b). The fine details of the fracture can be seen in the scanning electron microscopy (SEM) fractograph in Fig. 30(c), which displays the dimpled rupture of the low-carbon steel matrix and the exposed surface of silicate and sulfide platelets.

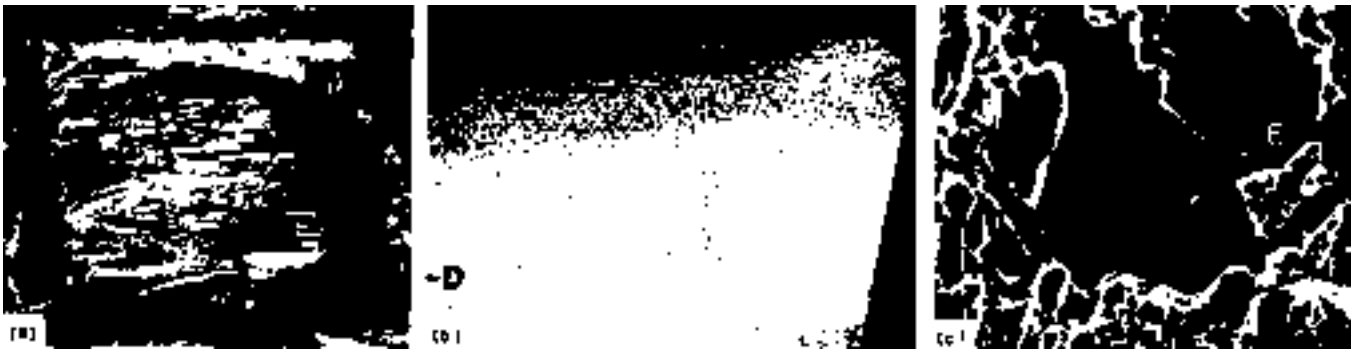


Fig. 30 Lamellar tear beneath a T-joint weld that joined two low-carbon steel plates. (a) Fractograph of lamellar tear showing separation that has followed flattened inclusions. Approximately 0.3×. (b) Section through fracture (top), which occurred in the coarse-grain reaustenitized region (A) of the HAZ. Region B is a fine-grain reaustenitized area. Region C is a partly reaustenitized region of ferrite and refined pearlite. Region D is an unaffected region of ferrite and pearlite. Approximately 8×. (c) SEM fractograph showing inclusion platelets (arrow E) and dimples (arrow F). 1300×

Effects of Weld Porosity. From the standpoint of fitness-for-purpose, porosity is regarded as the least harmful type of discontinuity. Low-carbon steels are generally easy to weld and are quite tough and ductile above the DBTT; therefore, the existence of porosity and slag inclusions in shielded metal arc welds is usually not critical, because their notch effect is minor compared to those of the joint details. However, excessive amounts can be detrimental to the static strength of the joint by reducing the load-bearing area of the weld. Extensive

porosity also can interfere with detection of the more harmful planar discontinuities, such as undercut, LOF, and LOP, by obscuring their presence.
The influence of porosity is evaluated according to its location, that is, surface or internal. Surface porosity has a detrimental effect on fatigue strength, especially when weld reinforcement is removed or the weld is undermatched. In high-cycle fatigue, porosity is least harmful, as long as weld buildup is not removed. If it is removed, the rate of decrease in fatigue strength initially is appreciable (up to 50% at 8% porosity); after reaching about 8% porosity, fatigue strength reduction rate decreases. To improve fatigue life of welds with surface porosity, the weld root should be dressed to minimize stress concentration. Fatigue failure originates at surface porosity rather than at buried slag inclusions. Pores in the surface are more harmful than internal pores of the same size and amount over a given weld length. Shape of the porosity is relatively unimportant. The presence of porosity in butt welds is structurally more significant than in fillet welds, because butt welds usually see more critical applications than fillet welds, and the stress concentrations at the toes of fillet welds are generally quite large.

Inclusions

Nonmetallic inclusions form as a result of reactions between metallic alloy elements and nonmetallic tramp elements, or by mechanical entrapment of nonmetallic slag or refractory particles. Extraneous inclusions may include:

- Oxides
- Sulfides
- Nitrides
- Carbides
- Other compounds
- Multiple phases

Among these nonmetallic inclusions, oxides and complex oxides occur most frequently in the size range known to influence weld metal microstructure. In some cases, alloys contain intermetallic compounds and/or second phases for strength and hardness, such as precipitation-hardened aluminum alloys. These microconstituent particles are integral parts of the alloys and are not considered as extraneous inclusions.
Metallic inclusions also occur. Metallic inclusions consist principally of particles of tungsten and copper. In gas tungsten arc welding, for example, the use of excessive current for a specific electrode size causes melting and deposition of tungsten in the weld. When present as fine, uniformly scattered particles, tungsten has little effect on the mechanical properties of the weld. Many codes, however, require that the welds be essentially free of such inclusions.
Copper inclusions may be found in gas metal arc welding if the electrode suffers burn-back to the contact tube. Such inclusions produce a brittle weld and can be a serious corrosion hazard; therefore, they should be removed from the weld deposit. Both tungsten and copper inclusions can be identified visually, and x-ray inspection detects tungsten particles because of their high density.
A third type of metallic inclusion can be caused by wire-brush bristles that become caught in the weld groove in preweld cleaning or in cleaning between weld passes.
Slag inclusions are nonmetallic materials formed by the slag reaction that are trapped in the weld (Fig. 31). Entrapment can occur when slag becomes mixed in the weld pool or when surface slag is not completely removed and is covered by subsequent weld passes. Slag may originate as pieces of unfused fluxes, which are trapped in a joint or allowed to flow ahead of the arc and be covered by the weld. Slag may also occur from heavy mill scale that has not been removed from a joint prior to welding. The presence of slag or inclusions can be anticipated when welding over highly crowned or rough welds, because their edges are extremely difficult to clean between passes or to penetrate during subsequent welding.

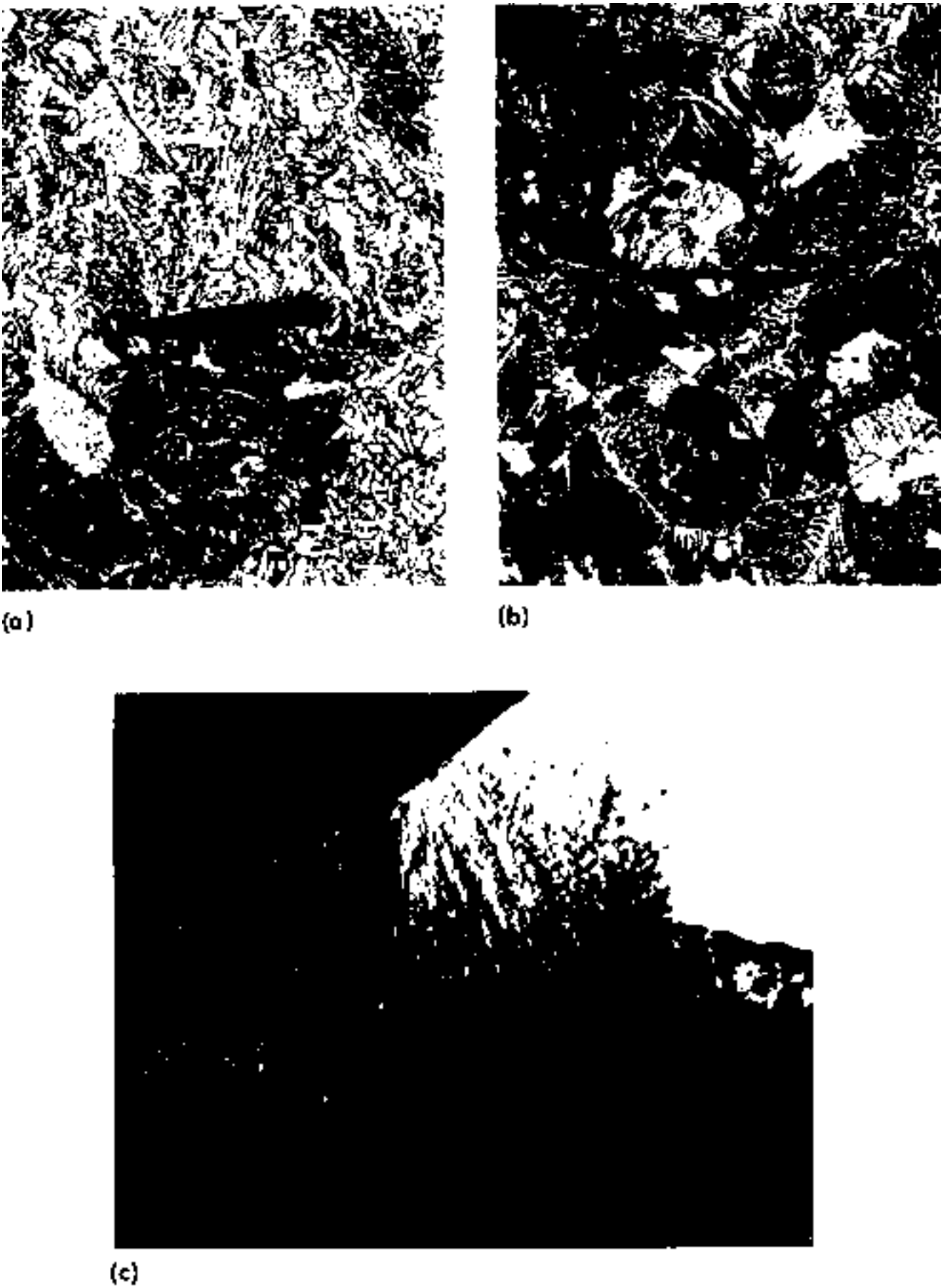


Fig. 31 Slag in steel metal. (a) Oblong. Etchant, 2% nital. 165×. (b) Lamellar. Etchant, 2% nital. 165×. (c) Jagged irregular. Fatigue fractograph as-is. ~15×

Slag inclusions can be found near the surface and in the root of a weld, between weld beads in multiple-pass welds, and at the side of a weld near the root. Multiple-pass welds are more prone to slag inclusions than single-pass welded joints. In multiple-pass welds, slag may be entrapped in any number of places in the weld between passes. This type of discontinuity can occur randomly as isolated foreign particles or as continuous or intermittent lines of inclusions parallel to the axis of the weld.

Slag inclusions are generally oriented along the direction of welding. They frequently are nonplanar in shape and can range in size from very small globules to large bands extending for several inches along the axis of the weld. In location, elongated slag deposits are often found between the first root pass and subsequent passes or along the weld-joint interface, while spherical slag inclusions can be distributed anywhere in the weld.

Slag inclusions occur in steel welds made by welding processes producing a flux—namely, SMAW, SAW, GMAW, and FCAW. Slag results from chemical reactions among the elements present in the molten weld pool and the protective flux. The problem is most common with the SMAW process, because it can be aggravated by poor manipulative techniques on the part of the welder. Slag may spill ahead of the welding arc and subsequently be covered by the weld pool because of poor joint fit-up, incorrect electrode manipulation, or forward arc blow. Slag trapped in this manner is generally located near the root. Radical motions of the electrode, such as wide weaving, may also cause slag entrapment on the sides or near the top of the weld after the slag spills into a portion of the joint that has not been filled by the molten pool.

Incomplete removal of the slag from the previous pass in multiple-pass welding is another common cause of entrapment. The general factors that affect slag entrapment are:

- *Cooling rate:* The faster the rate of solidification, the greater the chance of slag entrapment.
- *Temperature:* The lower the temperature of the weld pool, the less likely that slag will rise to the surface of the weld pool.
- *Viscosity:* The more viscous the molten weld, the slower the coalescence of foreign particles and their rise to the surface.
- *Stirring:* The more vigorous the agitation of the liquid metal, the higher the probability of slag entrapment.
- *Geometry:* The more undesirable the shape of the bead and the joint, the more difficult the removal of slag.

To avoid slag inclusions, the slag must be removed from the weld surface by chipping, wire brushing, grinding, or air arc gouging. To eliminate slag that becomes entrapped within the weld bead, extra time must be allowed for the slag to rise to the surface. This can be accomplished by adjusting cooling rate, weld-pool temperature, melt viscosity, weld-pool quiescence, and the geometry of the bead and joint. Figure 32 indicates the relationship among many of the factors that cause or prevent slag inclusions.

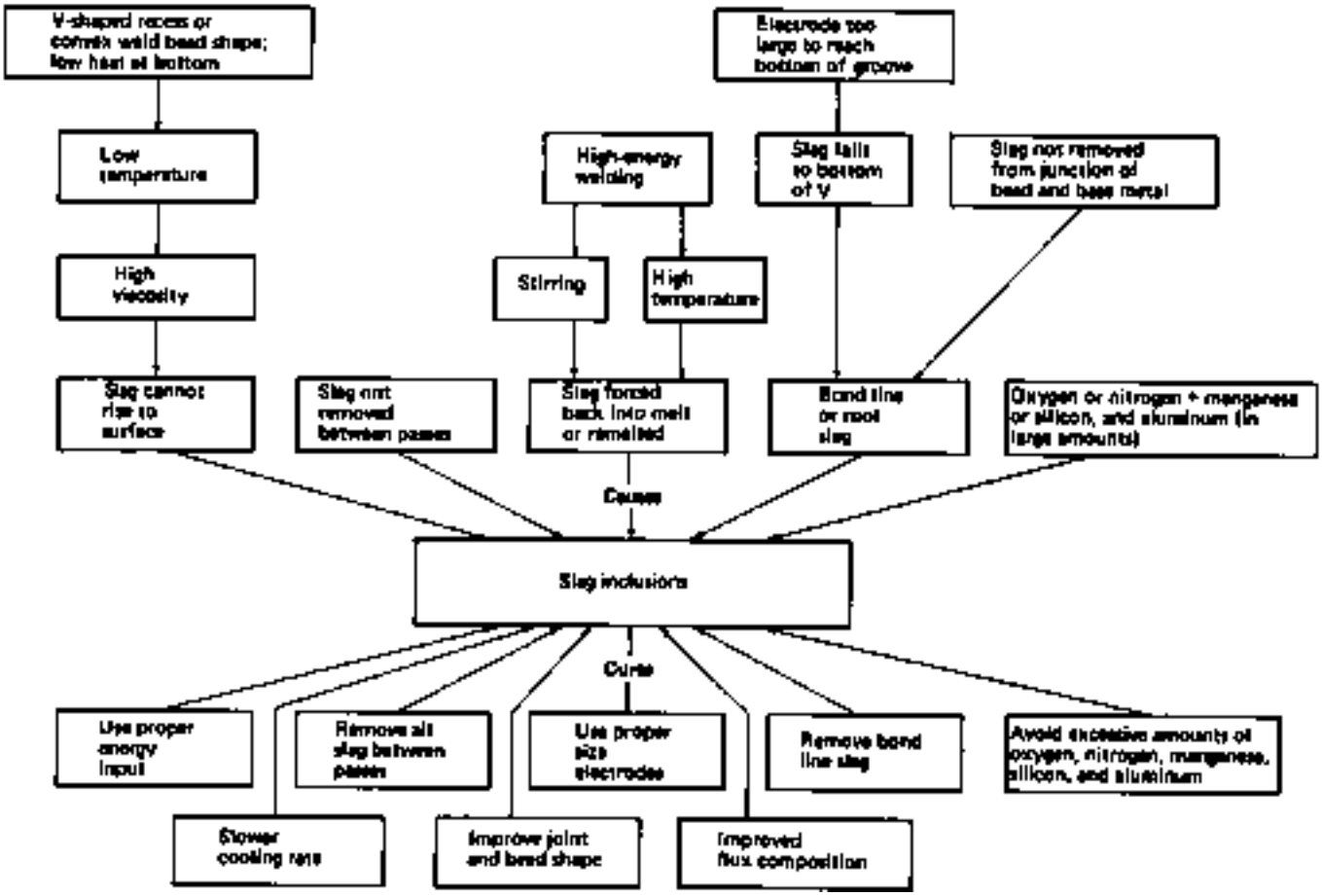


Fig. 32 Causes and remedies of slag inclusions in weldments

Fluxes exert a pronounced influence on the properties and soundness of the resultant weld. The type of flux used influences the quantity and type of foreign particles that may form in the weld bead. Basic fluxes promote low oxygen and/or sulfur contents in the weld metal, thereby minimizing the formation of slag inclusions.

Effect of Slag Inclusions. Acceptable limits on the size and prevalence of slag inclusions depend on the product specifications or standards. Slag inclusions between weld beads reduce the strength of the weld. The structural importance of the presence of slag inclusions in a weld is evaluated in terms of their size, shape, amount, distribution, and location. Tensile strength of the weld is reduced in proportion to the projected area of the slag; yield strength is affected to a lesser extent. Tensile ductility is reduced appreciably. Slag inclusions occurring along the fusion line, which is the weakest part of most welds, are more significant than porosity.

Quality band fatigue curves (e.g., Fig. 33) are usually plotted on a stress-number of cycles (*S-N*) diagram and are structured in order of increasing length of slag inclusions, where *S* denotes the stress range and *N* signifies endurance (number of cycles). Each band corresponds to a specific size range of discontinuity, which is permitted in conformance with the design criteria for a given structure. For a given length of slag, the number of cycles decreases as the stress range increases. Quality categories can be established on the basis of as-welded and stress-relieved structures. The methods used for the detection of slag below the surface of single-pass or multiple-pass welds are limited primarily to, radiographic and ultrasonic inspection, as described further in *Nondestructive Evaluation and Quality Control*, Volume 17, and *Welding, Brazing, and Soldering*, Volume 6 of *ASM Handbook.*

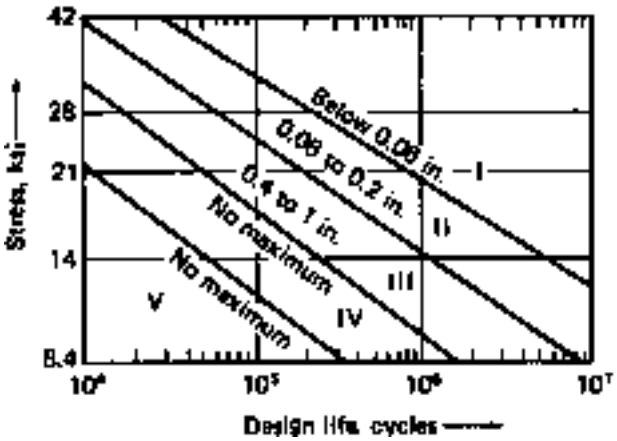


Fig. 33 Quality bands for slag inclusions representing as-welded carbon-manganese steel

Tungsten inclusions are particles found in the weld metal from the nonconsumable tungsten electrode used in GTAW. These inclusions are the result of:

- Exceeding the maximum current for a given electrode size or type
- Letting the hot tip of the electrode make contact with the molten puddle
- Using an excessive electrode extension
- Inadequate gas shielding or excessive wind drafts, which result in oxidation
- Using improper shielding gases such as argon-oxygen or argon-CO_2 mixtures, which are used for GMAW

Tungsten inclusions, which are not acceptable for high-quality work, can only be found by internal inspection techniques, particularly radiographic testing. Tungsten inclusions or a very high-density (barium-containing) slag appear as white spots.

Oxide inclusions are particles of surface oxides that have not melted and are mixed into the weld metal. These inclusions occur when welding metals that have surface oxides with very high melting points, such as aluminum and magnesium. Oxide inclusions weaken the weld and can serve as initiation points for cracking. The best method of prevention is to clean the joint and weld area thoroughly before welding.

Oxidation of shielded arc welds occurs only when there has been improper or inadequate shielding. Titanium, for example, is sufficiently reactive to require shielding of both the face and the root sides of the weld as well as of the tungsten electrode (if gas tungsten arc welded) to prevent oxygen and nitrogen contamination and embrittlement of both the weld and the HAZ. Such additional shielding is not mandatory for the root side of steel or aluminum welds, but also can be used to improve the quality of any weld.

Another effect of the shielding material may be the formation of compounds, such as oxides or nitrides, in the molten puddle from reaction with minute impurities, such as oxygen or nitrogen, in the shielding gas. Such compounds have been shown to lower the toughness of the weld compared to that of welds made with pure inert gas. Considerable advances have been made with shielding-gas mixtures for GMAW. These new mixtures have overcome some of the problems of lack of toughness.

Example 12: Fatigue Cracking of Welded Tubular Posts in a Carrier Vehicle Because of the Presence of Inclusions That Acted as Stress Raisers. Two tubular steel posts in a carrier vehicle failed by cracking at the radius of the flange (Fig. 34) after five weeks of service. The posts were two of four that supported the chassis of the vehicle high above the wheels so that the vehicle could straddle a stack of steel or lumber for pickup and transport. Over a period of approximately four years, 47 other posts had failed in nine vehicles. The latest failures occurred in posts of an improved design (Fig. 34) that had been used in an attempt to halt the pattern of cracking.

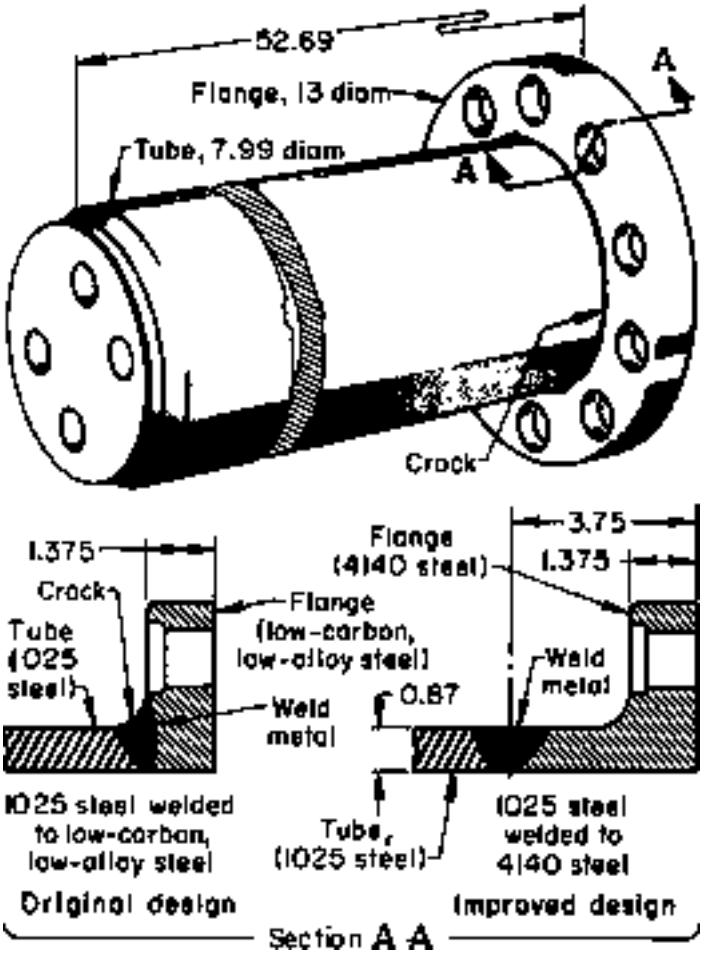


Fig. 34 AISI 1025 steel tube post for a carrier vehicle. The post failed in fatigue because of improper design and choice of flange metal. Dimensions given in inches

The original design involved a flat flange of low-carbon low-alloy steel that was welded to an AISI 1025 steel tube with a wall thickness of 22 mm ($\frac{7}{8}$ in.) as shown in Fig. 34. The fillet portion of the weld was machined to a smooth radius, but this region was highly stressed, and fatigue accounted for the numerous failures of posts.

In an effort to improve the strength of the joint, an improved design was adopted. The improvement was associated with the flange, which was machined from a 100 mm (4 in.) thick plate of normalized low-carbon low-alloy steel so that the welded joint would be approximately 50 mm (2 in.) away from the flange fillet. However, after the brief service of five weeks, the two posts showed the same fatigue cracks in the fillet as the original design.

Investigation. Chemical analysis of the flange metal showed that it met specification requirements. Examination of a metallographic specimen from the fracture area revealed a banded structure in the direction of rolling and aluminum oxide stringers at the fillet surface, which undoubtedly served as stress raisers, in the direction of crack extension.

Conclusions. The failures in the flanges of improved design were attributed to fatigue cracks initiating at the aluminum oxide inclusions in the flange fillet. It was concluded that a higher-strength flange was necessary.

Corrective Measures. The design of the flange with the weld approximately 50 mm (2 in.) from the fillet was retained, but the metal was changed to a forging of AISI 4140 steel, oil quenched and tempered to a hardness of 241 to 285 HB. Preheating to 370 °C (700 °F) before and during welding with AISI 4130 steel wire was specified. The weld was subjected to magnetic-particle inspection and was then stress relieved at 595 °C (1100 °F), followed by final machining. No further failures of posts were reported in four years of service after the changes.

Geometric Discontinuities

The most common geometric weld discontinuities are those associated with imperfect shape or unacceptable weld bead contour. Geometric discontinuities shown schematically in Fig. 35 include:

- *Overlap:* The protrusion of weld metal beyond the toe, face, or root of the weld
- *Undercut:* A groove melted into the base metal adjacent to the toe or root of a weld and left unfilled by weld metal
- *Underfill:* A depression on the face of the weld or root surface extending below the surface of the adjacent base metal

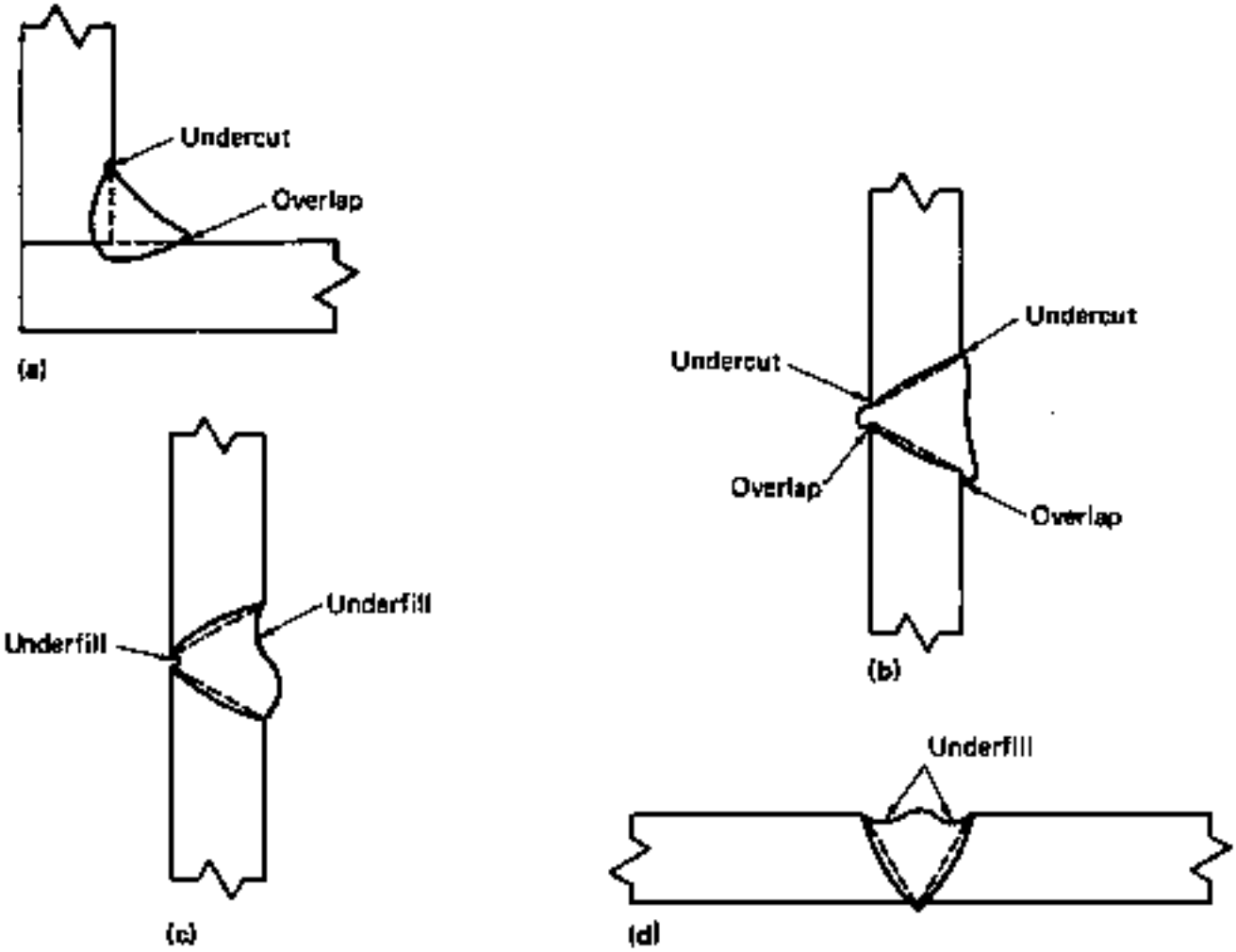


Fig. 35 Weld discontinuities affecting weld shape and contour. (a) Undercut and overlapping in a fillet weld. (b) Undercut and overlapping in a groove weld. (c) and (d) Underfill in groove welds

Another type of geometric discontinuity is melt-through, where the arc melts through the bottom of a joint welded from one side. To prevent melt-through, the welding current and the width of the root opening should be reduced, and travel speed should be increased. Excessive reinforcement is the opposite of underfill. It too is an imperfection because its shape leads to a high stress concentration.
Overlaps are the protrusion of the weld metal over the edge or toe of the weld bead. It occurs when the weld puddle is too cold for fusion or when the puddle is too hot and solidifies after protruding beyond the toe or root of the weld. This defect can cause an area of incomplete fusion, which creates a notch and can lead to crack initiation. To minimize overlapping:

- Use a higher travel speed.
- Use a higher welding current.
- Reduce the amount of filler metal added.
- Change the electrode angle so that the force of the arc does not push molten weld metal over unfused sections of the base metal.

Underfill is defined as a depression or loss of metal on the face of a weld extending below the surface of the adjacent base metal. Underfill reduces the cross-sectional area of the weld below the designed amount and therefore is a point of weakness and creates an increased stress concentration where failure may initiate. To minimize underfill, the voltage, travel speed, and root opening should be reduced. Figure 36 shows underfill in a pulsed gas metal arc spot weld joining carbon steel to a 90Cu-10Ni alloy using ERNiCu-7 filler metal.

Fig. 36 Pulsed GMAW spot weld showing a lack of fill-in. Etchant, 50% nitric-50% acetic acid. 4×

Undercutting (Fig. 37a) is a notching of the base metal located at the toe of a weld that occurs when the weld metal does not completely fill in the molten surface of base metal at the toe to form a smooth junction at the weld toe. Undercutting is particularly troublesome, because it produces stress raisers that create problems under impact, fatigue, or low-temperature service. Fatigue strength is markedly reduced by even small undercut depths. The critical size of a geometric discontinuity, such as undercut in fillet welds for a given leg length, decreases as the plate thickness increases.

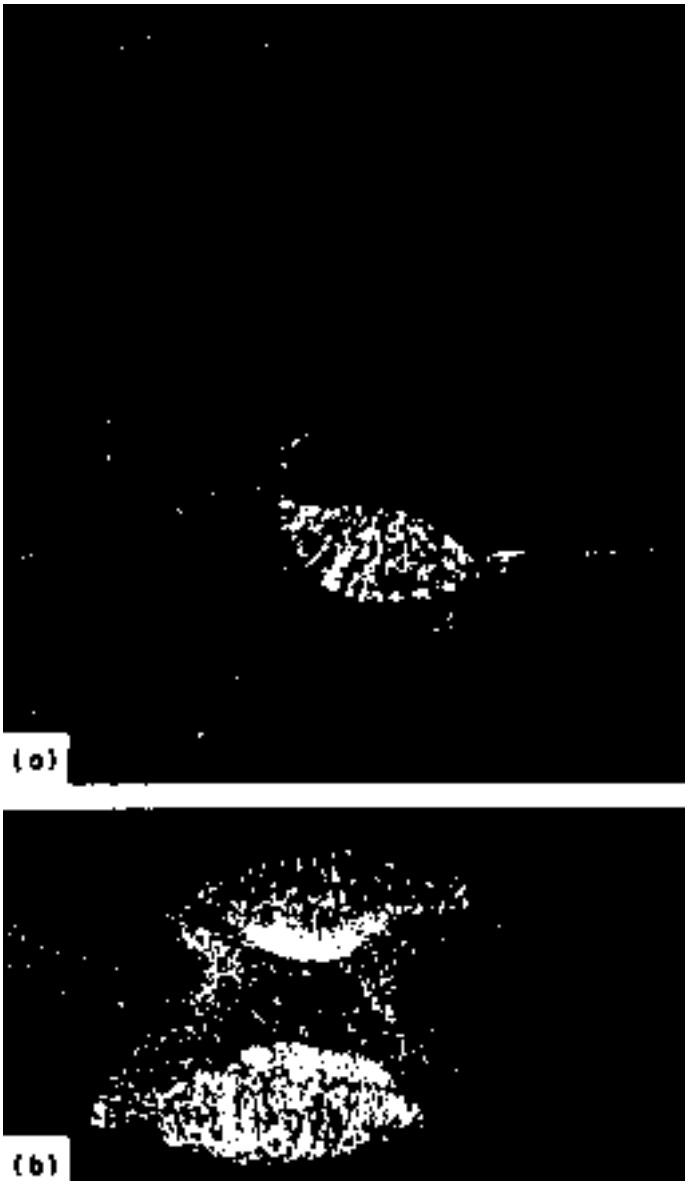


Fig. 37 Two types of poor contours in arc welds. (a) Fillet weld showing two forms of undercut plus weld spatter and uneven leg length. (b) Butt weld showing a high, sharp crown

To minimize undercut:

- Decrease current, travel speed, and electrode diameter, which controls weld-pool size.
- Change electrode angle so that the arc force holds the weld metal in the corners.
- Avoid weaving.
- Maintain constant travel speed.
- Change the surface tension and viscosity of the molten weld metal where possible.

Undercut actually refers primarily to the base metal adjacent to the weld, whereas imperfect shape is an imperfection of the weld itself. This can include such things as excessive reinforcement on the face of a weld, which can occur on groove welds as well as fillet welds. There is also the problem with excessive reinforcement on the root of the weld, primarily open-root groove welds. Excessive reinforcement is economically unsound. It can also be a stress raiser and is objectionable from an appearance point of view. It is normally a factor involved with fit-up, welder technique, welding current, and type of electrode.
Undercut results from too high a welding current or too slow a welding speed. Conversely, excessive reinforcement (Fig. 37b) may occur from too low a current or too fast a travel speed. In addition to proper control of welding current, the use of a weave type of bead may be helpful in contour control. However, the

weave technique must be used with caution, because it increases the possibility of undercut and only welding discontinuities and adverse effects may result.

Example 13: Fatigue Fracture of a Shaft for an Amusement Ride Because of Undercuts in Welds. An amusement ride in a shopping center failed when a component in the ride parted, permitting it to fly apart. The ride consisted of a central shaft supporting a spider of three arms, each of which was equipped with an AISI 1040 steel secondary shaft about which a circular platform rotated. Each platform carried six seats, each seat being capable of holding up to three persons. The whole assembly could be tilted to an angle of 30° from horizontal. The main shaft rotated at about 12 rpm and the platforms at a speed of 20 rpm.

The accident occurred when one of the secondary shafts on the amusement ride broke, allowing the platform it supported to spin off the spider. The point of fracture was adjacent to a weld that attached the shaft to a 16 mm ($\frac{5}{8}$ in.) thick plate, which in turn bore the platform support arms.

Investigation. The fracture surface of the shaft was found to possess beach marks characteristic of fatigue fracture (Fig. 38). It was apparent that the fillet weld had undercut the shaft slightly, creating points of stress concentration that served as nuclei for the fatigue cracks. Crack propagation occurred from two directions, penetrating all of the shaft section except for a roughly triangular region (Fig. 38) that was the zone of final fast fracture.

Fig. 38 AISI 1040 steel shaft for an amusement ride. Fatigue fracture originated at weld undercuts. Two sets of beach marks and a triangular final-fracture zone are visible. Approximately 0.4×

The shaft at the location of the fracture was 108 mm ($4\frac{1}{4}$ in.) in diameter; at the bearing raceway, 64 mm ($2\frac{1}{2}$ in.) below the weldment, it was 98 mm ($3\frac{7}{8}$ in.) in diameter. Stress analysis indicated that cross-sectional area was not the determining factor in the stress level. A section through the fracture region revealed a small secondary crack (which did not contribute to fracture but which arose from the same causes) and clearly defined HAZs. The structure confirmed a report from the manufacturer that there had been no postweld heat treatment.

Conclusions. A likely cause for the fatigue failure was the combination of residual stresses generated in welding and centrifugal service stresses from operation that were accentuated by areas of stress concentration at the undercut locations. Without the excessive residual stress, the shaft dimensions appeared ample for the service load.

Recommendations. The fillet weld should be applied with more care to avoid undercutting. The residual stresses should be minimized by preweld and postweld heat application.

Example 14: Fatigue Failure of a Carbon Steel Water-Wall Tube Because of an Undercut at a Welded Joint. In a new stationary boiler, furnace water-wall tubes were welded to the top of a dust bin for rigid support (Fig. 39a). The tubes measured 64 mm (2.5 in.) outside diameter by 3.2 mm (0.125 in.) wall thickness and were made of carbon steel to ASME SA-226 specifications. Shortly after start-up, one of the tubes broke at the welded joint.

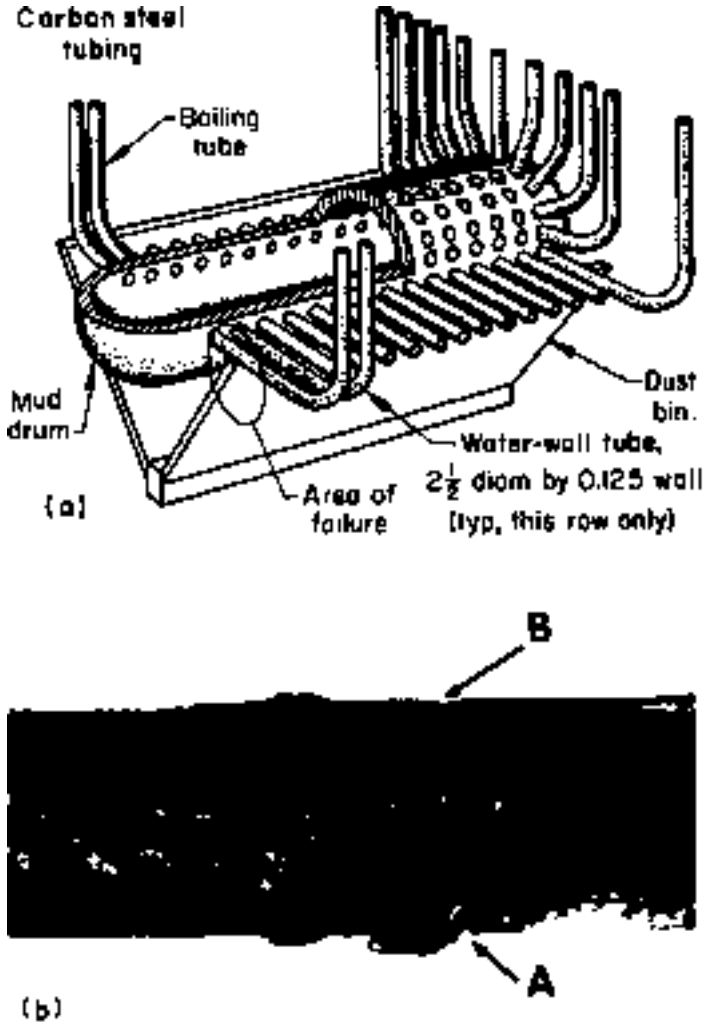


Fig. 39 Stationary boiler in which a carbon steel water-wall tube failed by fatigue fracture at the weld joining the tube to a dust bin. (a) Illustration of a portion of the boiler showing location of failure. Dimensions given in inches. (b) Photograph of fractured tube. Fatigue crack is at arrow A, and ductile fracture is at arrow B.

Investigation. The portion of the fracture surface adjacent to the attachment weld around the bottom half of the tube had a flat, brittle appearance. Over the remainder of the tube wall, the fracture was ductile, with a noticeable shear lip and necking. Beach marks were evident on the fracture surface, indicating that fatigue was the fracture mechanism. The origin was identified as a small crevicelike undercut at the edge of the weld bead, shown by arrow A in Fig. 39(b).

Conclusion. Cracking began by fatigue from a severe stress raiser. As the fracture progressed out of the region of high restraint where the tube was welded to the dust bin, the fracture changed from fatigue to ductile overload, as evidenced by necking opposite the origin. It was concluded that the crevicelike undercut was likely the primary cause of the fracture and that the source of the necessary fluctuating stress was tube vibration inherent in boiler operation.

Recommendations. The remaining water-wall tubes in the row were inspected by the magnetic-particle method; several other tubes were found to have small fatigue cracks adjacent to the welds attaching the tubes to the dust bin. The broken tube was replaced, cracks in other tubes were repaired by welding, and the attachment welds

were blended smoothly into the tube surfaces for the entire length of the dust bin. No further failures occurred after the severe undercut was eliminated.

Incomplete Fusion and Incomplete Penetration

Incomplete fusion (IF) (formerly known as lack of fusion, or LOF) refers to the degree to which the original base metal surfaces have been fused to the filler metal. Incomplete penetration (IP) (formerly known as lack of penetration, or LOP) refers to the degree to which the base metal has been melted and resolidified to result in a deeper throat that was present in the joint before welding. In effect, a joint can be completely fused but have incomplete root penetration to obtain the throat size specified. Based on these definitions, IF discontinuities are located on the sidewalls of a joint, and IP discontinuities are located near the root (Fig. 40). With some joint configurations, such as butt joints, the two terms can be used interchangeably.

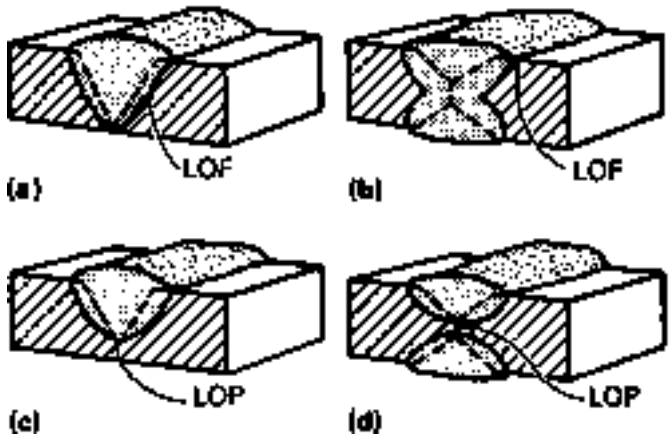


Fig. 40 Incomplete fusion in (a) a single-V-groove weld and (b) a double-V-groove weld. Incomplete penetration in (c) a single-V-groove weld and (d) a double-V-groove weld

Causes of IF and IP. Inadequate joint penetration and incomplete fusion in arc welds usually result from:

- Improper fit-up or groove design, where the joint design does not afford proper torch accessibility and there is an unsuitable angle of impingement between the base metal and the welding arc
- Too low a welding current
- Poor welding techniques, such as improper electrode manipulation and unfavorable torch angle for the travel speed and current being used
- Poor preparation of the base metal surface, such as inadequate wire brushing between weld passes or poor fit-up in preparing the joint

To prevent IF and IP, heat input must be increased to allow complete fusion between the weld deposits and the workpiece as well as between successive passes in multipass joints. The weldment surface must be cleaned of all contaminants. Figure 41 illustrates many of the causes of and methods of prevention for IF. Typical causes and methods of prevention for IP are illustrated in Fig. 42. Joint geometry must permit proper arc control and access for the electrode. If the gap is excessive, weave manipulation should be adjusted accordingly.

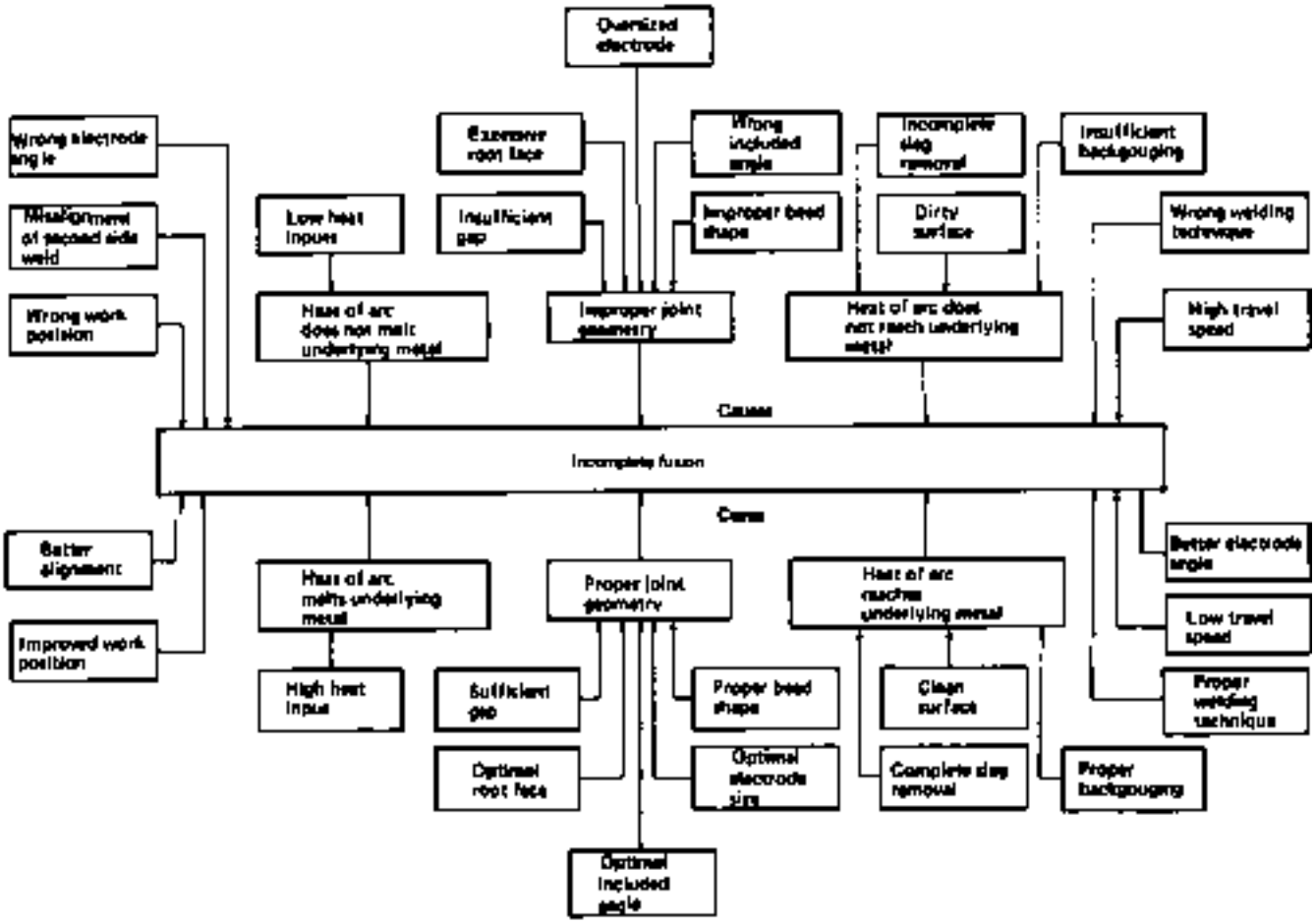


Fig. 41 Causes and cures of incomplete fusion in welds

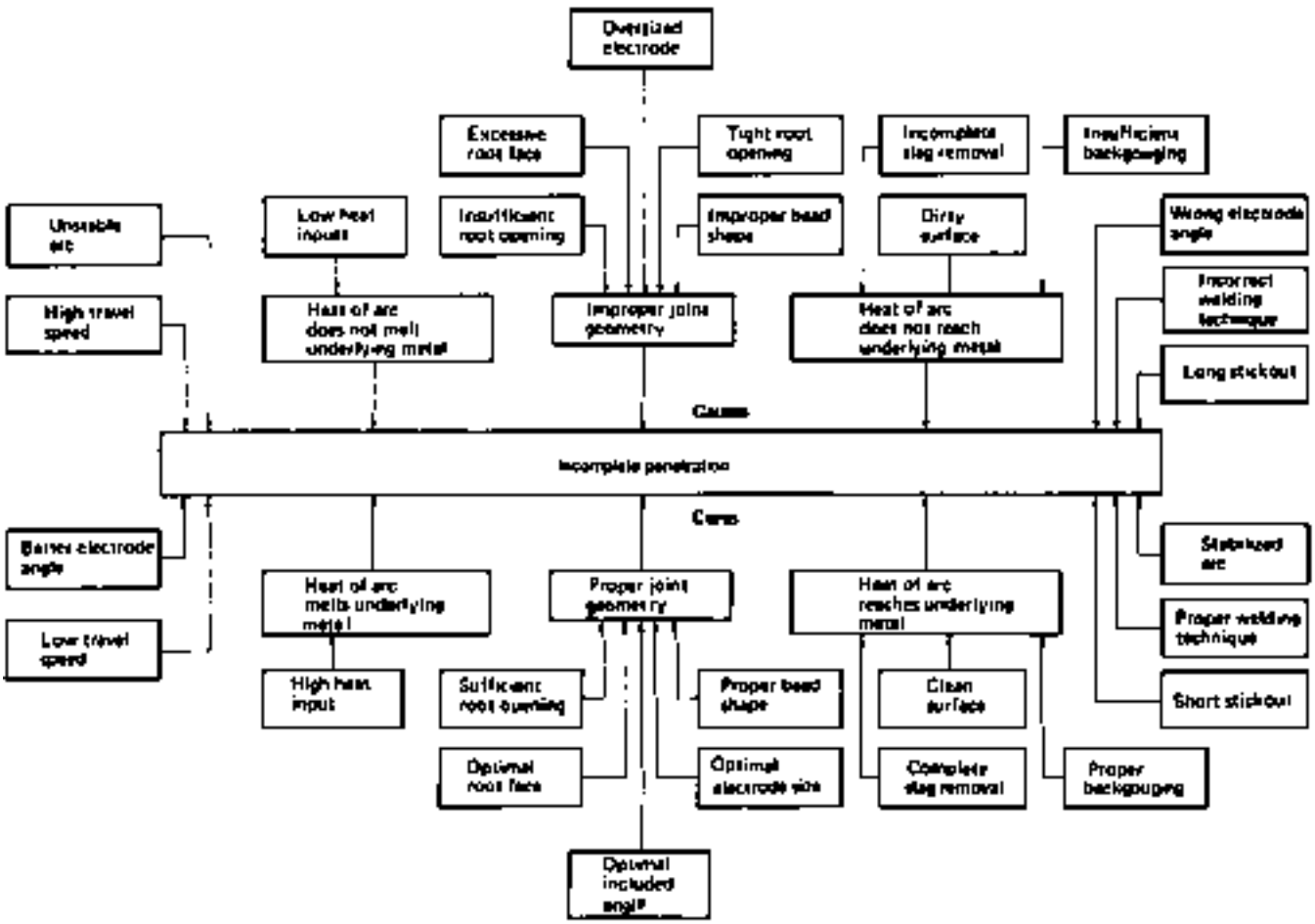


Fig. 42 Factors influencing incomplete penetration in welds

Figure 43 is an example of IF, where a portion of the filler metal is incompletely fused with the base material. Incomplete fusion can also form between weld beads in the case of multiple-pass welds. These essentially two-dimensional flaws occur when insufficient heat is absorbed by the underlying metal from the weld, causing incomplete melting at the interfaces of the weld and the base metal of successive passes. Incomplete fusion usually is elongated in the direction of welding, with either sharp or rounded edges, depending on the conditions of formation. The causes of IF (Fig. 41) include excessive travel speed, bridging, excessive electrode size, insufficient current, poor joint preparation, overly acute joint angle, improper electrode manipulation, and excessive arc blow.

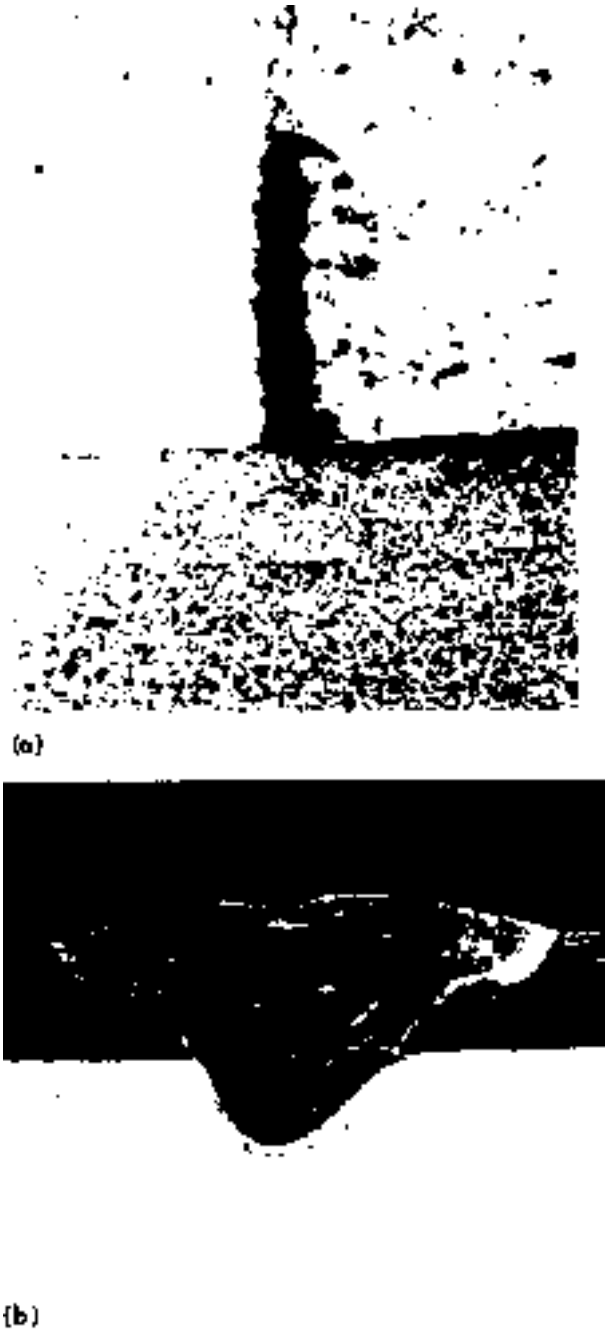


Fig. 43 Incomplete fusion in a pulsed gas metal arc spot weld involving ERNiCu-7 (Monel 60), 0.89 mm (0.035 in.) diameter filler metal, copper-nickel to steel weldment. Etchant, 50% nitric-50% acetic acid. (a) View showing IF flaw. 30×. (b) View showing that IF was eliminated by tapering the circular joint. 4×

Figure 44 is an example of IP, where the weld metal has not penetrated to the root of the weld joint. In some cases, IP may be detected visually through the absence of a root bead. Incomplete penetration is caused by

incorrect welding technique or by improper root gap. Incorrect welding techniques (shown graphically in Fig. 42) include:

- Low amperage
- Oversized electrode diameter
- High travel speed
- Wrong electrode angle
- Incorrect joint cleaning, where surface contaminants such as oxide, oil, or dirt prevent full melting of the underlying metal
- Improper work position
- Misalignment of back side weld
- Excessive inductance in the dip transfer mode of GMAW
- Insufficient backgouging
- Too large a flat land in a groove weld, and undesirable arc manipulation by the welder, particularly in the vertical position

Fig. 44 Incomplete penetration in a butt welded joint in steel. Original plate thickness, 19 mm ($\frac{3}{4}$ in.). As-polished. 1.4×

Radiographic methods may be unable to detect IF and IP discontinuities in certain cases, because of the small effect they have on x-ray absorption, although lack of sidewall fusion is readily detected by radiography. Ultrasonically, both types of discontinuities often appear as severe, almost continuous, linear porosity because of the nature of the unbonded areas of the joint. Except in thin sheet or plate, these discontinuities may be too deep-lying to be detected by magnetic-particle inspection.

Effects of IF and IP. The presence of incomplete fusion and incomplete penetration can be critical, because, in addition to reducing the effective cross-sectional area, it can produce discontinuities that can be almost as sharp as cracks and can even display crack-extended tips. In welding some alloys, incomplete fusion and incomplete penetration can result from an oxide layer on the joint sidewall.

The fitness-for-purpose analysis of IF and IP, when present in carbon steels, is typically related to the ductile-to-brittle transition temperature, because the significance of these discontinuities can change. When material is below its transition temperature, IF and IP are potentially harmful and can trigger brittle fracture. If, on the other hand, the steel is above its transition temperature, these discontinuities may be only slightly worse than slag inclusions or porosity. Consequently, it is essential to define the exact transition temperature and conditions that will be encountered in service.

The effect of IF and IP can be masked by weld metal overmatching and reinforcement. The direction of the applied alternating stress with respect to IF and IP is also important. When the applied load is parallel to these discontinuities, the effect is minimal; the reverse is true when the direction of stress is perpendicular. The significance of IF and IP, however, also depends on their “aspect ratio,” which indicates the sharpness of the discontinuity (ratio of width-to-depth), and location.

Few experimental studies have been conducted to examine the effect of IF on fatigue strength. However, fatigue behavior of weldments with IF discontinuities can be estimated using the mathematical models for fatigue crack propagation. In an investigation of the significance of IF and IP in an Al-4.4Mg alloy welded by GMAW, it was found that both IF and IP were more serious than undercut, surface irregularity, and macro- and microporosity (Ref 24). The fatigue strength of this aluminum alloy in a 12 mm (0.47 in.) thick test sample was reduced 60% by the presence of IF and/or IP.

Example 15: Fracture of a Chip-Conveyor Pipe at a Flange Weld as a Result of Poor Fit- Up. A pipe in a chip conveyor cracked at the toe of an exterior fillet weld connecting a flange to the pipe.

The chip conveyor consisted of several spool sections. Each section was made up of a length of 560 mm ($22\frac{1}{16}$ in.) OD by 546 mm ($21\frac{1}{2}$ in.) ID low-alloy steel pipe and two 713 mm ($28\frac{1}{16}$ in.) OD by 562 mm ($22\frac{9}{64}$ in.) ID flanges of 13 mm ($\frac{1}{2}$ in.) thick low-carbon steel, which were welded to the two ends of the pipe. The wall thickness of the pipe was 7 mm (0.275 in.). The composition specified for the pipe steel was 0.25C-0.98Mn-3.52Ni-1.34Cr-0.24Mo, which approximates a 9300 steel with high molybdenum. The hardness of the pipe was reported as 495 HB, and of the flange as 170 HB. The flanges were joined to the pipe by SMAW, using E7018 electrodes.

Investigation. It was found that one end of the pipe had been cut before assembly at an angle of 76° and 39 minutes to its axis, which presented an ellipse (Fig. 45) over which the flange was forced to fit by hydraulic pressure. A 90° angle would have given a radial clearance between the pipe and flange of approximately 1.1 mm ($\frac{3}{64}$ in.), but the long OD of the elliptical pipe was 576 mm (22.7 in.) compared to the ID of the flange of 562 mm (22.1 in.), which placed both components in a condition of severe stress.

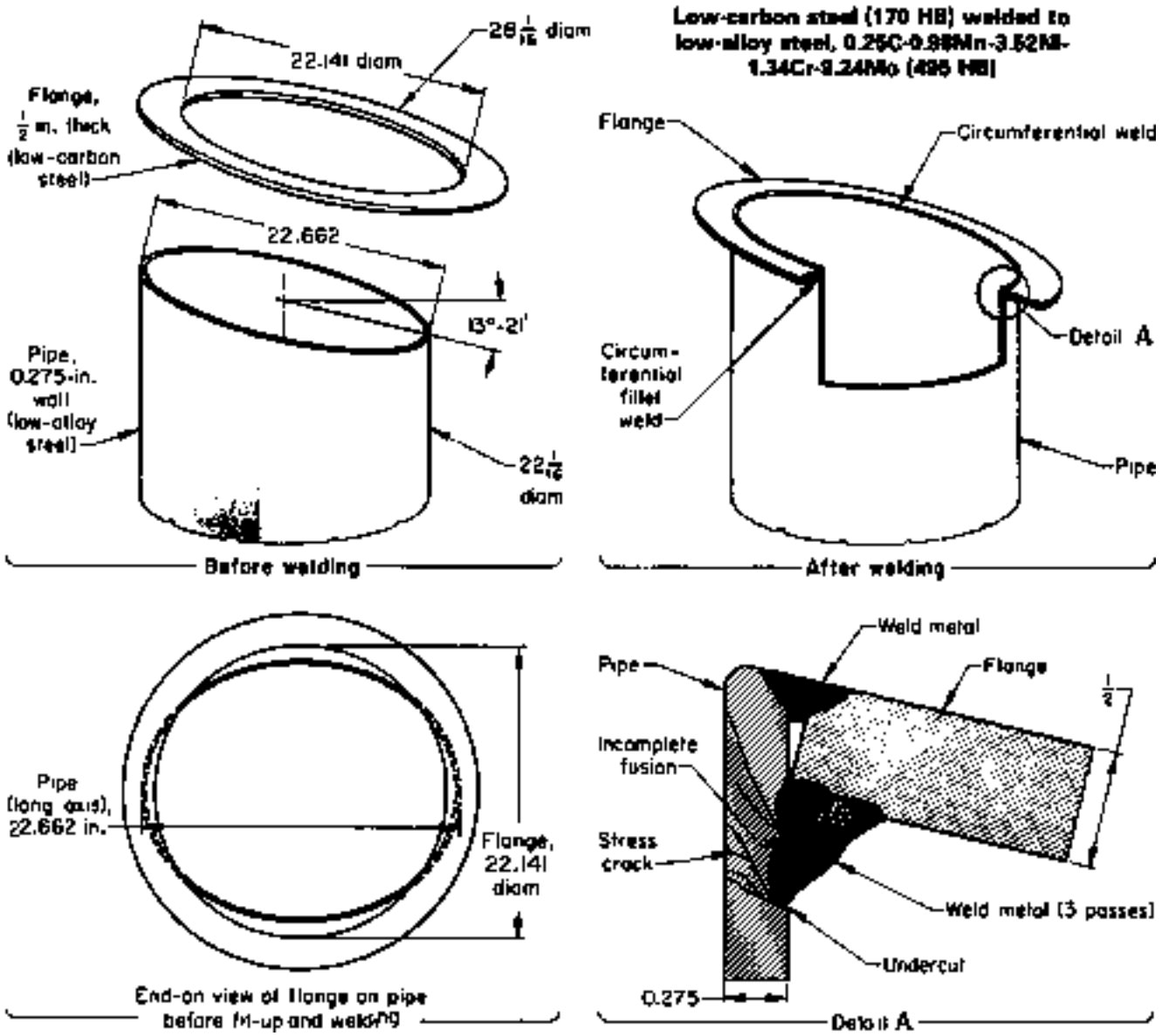


Fig. 45 Low-alloy steel conveyor pipe that cracked at fillet welds securing a carbon steel flange because of poor fit-up. Dimensions given in inches

A schematic illustration of a section cut through the weld in the area of maximum interference is shown in detail A in Fig. 45. Underbead cracking began at the toe on the pipe wall of the first fillet-weld pass at a point where there was incomplete fusion and inadequate penetration of the third and final pass. Several cracks initiated at the toe of the third pass on the pipe wall, where there was a slight undercut.

Conclusions. The conveyor pipe failed by brittle fracture, which was attributed to the stresses induced in forcing the circular flange over the elliptical section of the pipe. The toe of the weld and the adjacent undercut were stress raisers that determined the point of major crack origin. Under residual stress, the internal point of incomplete fusion also initiated additional cracks.

Recommendation. A proper fit between an elliptical flange and pipe end would eliminate the cracking.

Metallurgical Discontinuities of Welds

Metallurgical discontinuities include a variety of weld cracks that are severe stress concentrators characterized by sharp tips and high length-to-width ratios. Weld and base metal cracks (as described in the section “Weld Cracks” in this article) are generally considered to be defects, especially during manufacturing, although some, such as fisheyes and microfissures, are perhaps better characterized as anomalies or imperfections. The types of weld cracking include:

- *Fisheye:* A discontinuity found on the fracture surface of a weld in steel that consists of a small pore or inclusion surrounded by a bright, round area
- *Microfissures:* Small, cracklike discontinuities with only a slight separation (opening displacement) of the fracture surfaces
- *Lamellar tearing:* A type of cracking that occurs in the base metal or HAZ of restrained weld joints that is the result of inadequate ductility in the through-thickness direction of steel plate
- Hydrogen-induced cold cracking
- Hot cracking

Weldability is often influenced by the susceptibility of a material to cracking. The weldability of martensitic steels is greatly affected by hardenability that can result in cold cracking. Welded joints in ferritic steels may have low ductility as a result of grain coarsening. The weldability of austenitic steels is governed by their susceptibility to hot cracking. With precipitation-hardening steels, weldability is related to the mechanisms associated with the precipitation-hardening transformation.

Cold cracking is usually not encountered in aluminum alloy welds, except when a bead has been laid down that is too small to resist the shrinkage cooling stresses. However, both crater cracks and longitudinal hot cracks can occur in aluminum alloy welds if the welding technique is not favorable. In wrought heat treatable aluminum alloys, the relation between working and weld direction influences cracking. Welds made in the short-transverse direction often develop grain-boundary microcracks in the HAZ.

Fisheyes are cracks surrounding porosity or slag in the microstructure of a weld. Fisheyes are a form of hydrogen embrittlement cracking caused by the presence of hydrogen that collects at these locations and embrittles the surrounding metal in a manner similar to (but less severe than) cold cracking. What distinguishes a fisheye is the characteristically shiny appearance when viewed optically (Fig. 46 and 47); it only appears as the result of tensile fracture, such as in a tensile test.

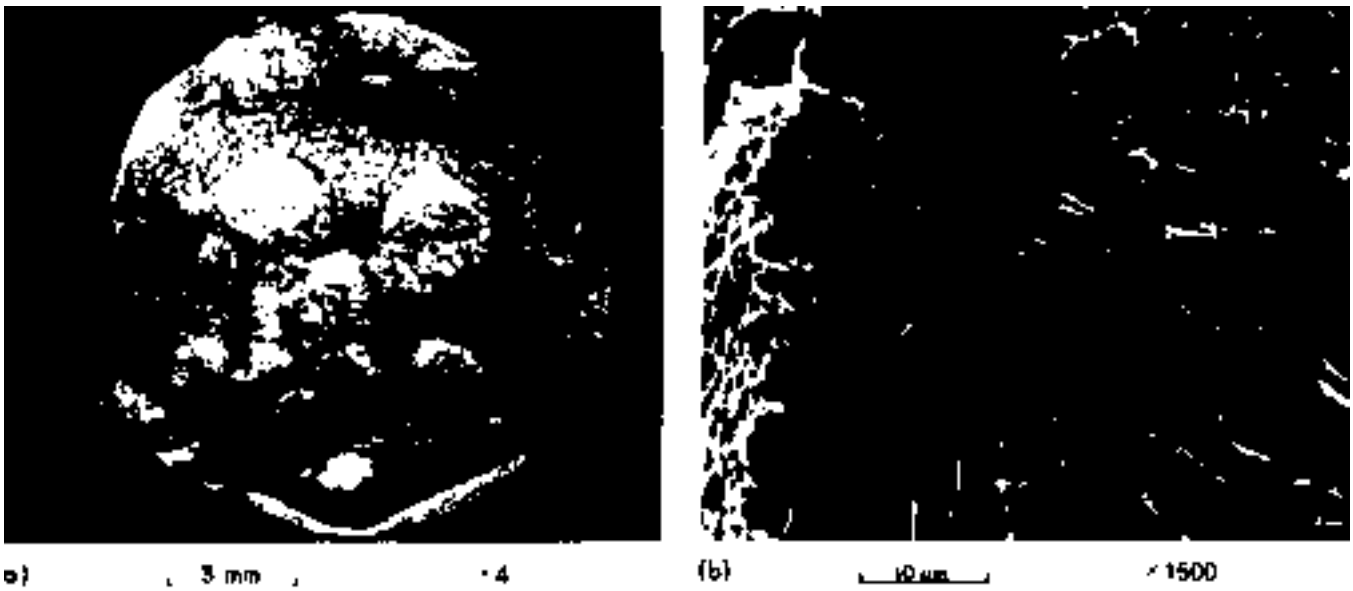


Fig. 46 Fisheyes in E7018 weld metal. (a) Fisheyes in as-welded tensile specimen tested at room temperature. Optical macrograph. (b) Mixed-mode fracture in as-welded three-point bend test specimen at room temperature (SEM). Optical fractograph

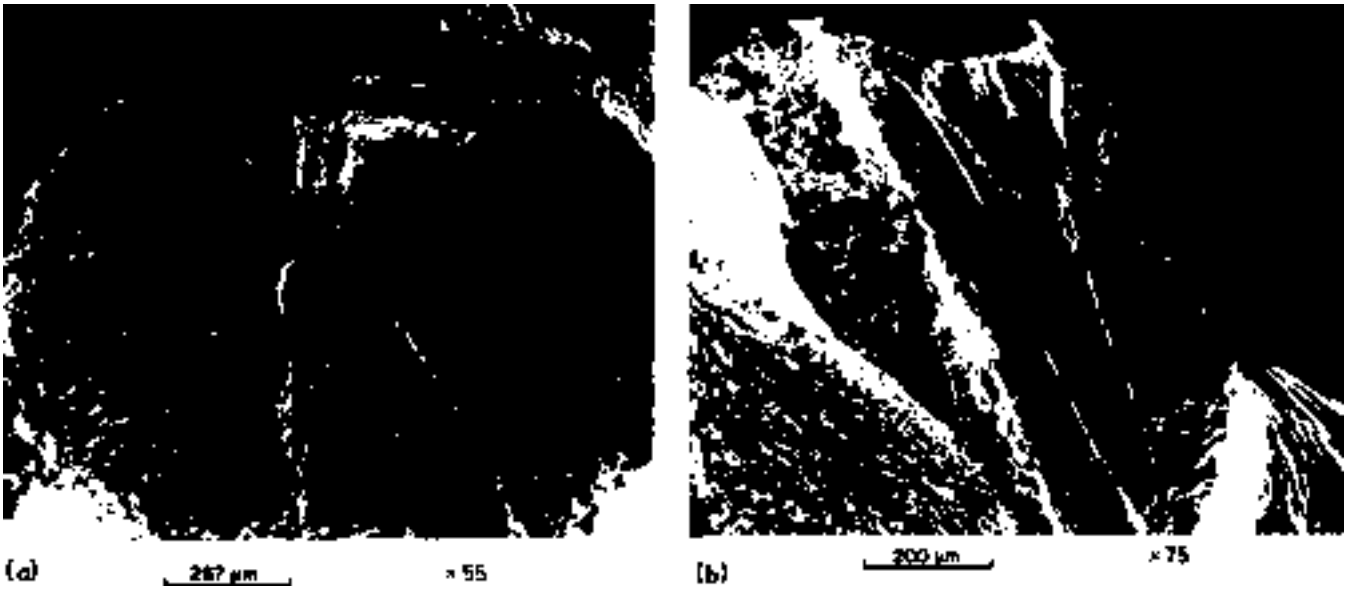


Fig. 47 Formation of fisheyes. (a) At pore, in as-welded E6010 tensile specimen tested at room temperature. (b) Slag/IF in as-welded E11018 tensile specimen strained at 1.0% at room temperature (SEM)

Fisheyes, also termed *flakes, blisters,* and *halos,* are predominantly found in ferritic steels, welded with high-hydrogen electrodes. Complete understanding of the conditions for the formation of this type of crack is still evolving. Because they were first observed during bend testing, it is frequently assumed that stressing of the weldment beyond its yield strength is required to form fisheyes, but this has not been confirmed.

The overall fracture appearance of fisheyes usually indicates a mixed or quasicleavage fracture mechanism, as seen in Fig. 46(b). When a halo is examined at high magnification, a fracture surface with a pattern radiating from a nucleating discontinuity is revealed. Small, buried microfissures also have been noted to nucleate fisheyes in as-welded E6010 weld metal. Elevated-temperature aging or postweld heat treatment can effectively eliminate fisheyes. Other factors that help prevent the occurrence of this type of discontinuity include high preheat, good joint design, low-hydrogen electrodes, and reduction in restraint. Note that use of low-hydrogen electrodes alone is apparently insufficient to eliminate fisheye formation. However, low-hydrogen electrodes are susceptible to moisture pick up; other low-high welding processes can suffer from shielding difficulties. In Charpy V-notch testing, fisheyes did not form, apparently because the deformation rate is too high to permit the required diffusion of hydrogen.

Microfissures are metallurgically related flaws formed during weld metal solidification. They are frequently found in austenitic stainless steel welds where the weld metal has an inadequate ferrite content. Microfissures can be of the hot or cold type of cracking, induced by small strain. Applied strain is needed to produce cracking; the amount needed to initiate cracking depends on material toughness. Because material is always subject to some type of strain, either from external or internal sources, the occurrence of cracking is therefore controlled by the conditions of welding, during which material toughness may be decreased. A micrograph example of microcracking (microfissures) is shown in Fig. 48.

Fig. 48 Micrograph of weld metal microcrack. Etchant, 2% nital. 215×

As an example of hot fissuring, a transfer line in a pressure vessel system carrying 73% sodium hydroxide developed a leak at a weld joining a cast Monel coupling to wrought Monel pipe. Metallographic examination disclosed many small microfissures in the weld metal, but no microfissures were found in the HAZs of the wrought Monel. This section of line contained high thermal stresses that resulted from injection of steam a short distance upstream from the coupling. Chemical analysis revealed that the cast Monel coupling had a silicon content of 1.97%, whereas the wrought Monel piping had a silicon content of 0.10%. Castings of high-silicon Monel generally have poor weldability and are often hot short, or sensitive to cracking in the weld HAZ. Fissures in the cast coupling were formed during welding and subsequently penetrated the section when the coupling was subjected to the thermal shock of steam injection. In this case, a weldable-grade casting should be specified to ensure freedom from hot fissuring.

Detection of microfissures is difficult by conventional nondestructive techniques; usually, metallographic preparation is necessary. However, if microfissures are produced by forming operations, and if such microfissures lie close to or in the HAZ, they may grow to macrofissures that can be detected by visual inspection or by examination using magnetic-particle, liquid penetrant, or radiographic methods.

Lamellar tearing is a form of base metal cold cracking that occurs parallel to the plate surface adjacent to the welds. This is shown schematically in Fig. 49 for a T-joint that is fillet welded on both sides. It may occur in the base metal or the HAZ of restrained weld joints and is characterized by a steplike crack parallel to the rolling plane. The crack originates internally because of tensile strains produced by the cooling contraction of the weld zone.

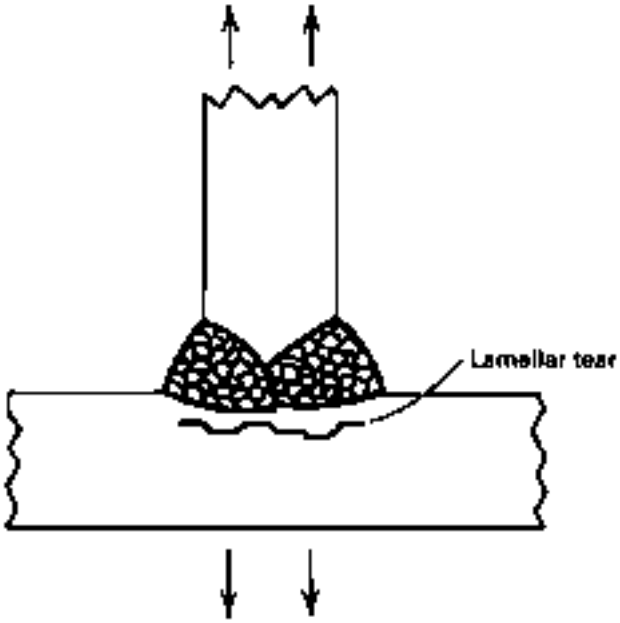


Fig. 49 Lamellar tear caused by thermal contraction strain

Lamellar tearing is associated with nonmetallic inclusions such as oxides, sulfides, and silicates that are elongated in the direction of rolling. The net result of these inclusions is a decrease in the through-thickness ductility. Other factors that affect lamellar tearing are design details, plate thickness, magnitude, sign and distribution of induced stresses, type of edge preparation and properties, as well as base metal fabrication. Figure 50 shows lamellar tearing in the HAZ of a corner joint in a carbon-manganese steel.

Fig. 50 Lamellar tearing in the HAZ of a carbon-manganese steel corner joint. Etched with 2% nital

Lamellar tears can result rapidly either in service or during fabrication. In the previously described case of a catastrophic brittle fracture of a pressure vessel (Ref 16), the cause of failure was associated with lamellar tears that originated at nonmetallic inclusions on planes parallel to the plate surface, caused by weld shrinkage stresses. These cracks caused by the root weld pass were not by themselves large enough to initiate fracture, but subsequent thermomechanical cycling during the remainder of the welding process caused dynamic strain aging and localized embrittlement at the crack tips. Another example of fracture caused by lamellar tearing follows.

Example 16: Failure During Fabrication of an Armature Because of Lamellar Tearing. During the final shop welding of a large armature for a direct-current motor (4475 kW, or 6000 hp), a large bang was heard, and the

welding operation stopped. When the weld was cold, nondestructive evaluation revealed a large crack adjacent to the root weld.

Investigation. Figure 51(a) shows the design of the weld and the region where cracking occurred. The 100 mm (4 in.) thick plate had a chemical composition of 0.21C-0.77Mn; composition of the 125 mm (5 in.) plate was 0.19C-0.76Mn, giving carbon equivalents (CE) of 0.34 and 0.32, respectively (CE = C + Mn/6 + (Cr + Mo + V)/5 + (Ni + Cu)/15); these CE levels are indications of good weldability and resistance to weld cracking. The welds were made with a preheat of 65 °C (150 °F), with the heat applied only on the outside wall at point B. The weld was deposited using a single-head submerged arc process, and the failure occurred when 32 mm (1.25 in.) of weld depth had been laid. Sections through the failure revealed that the cracking had initiated at the root of the weld at the intersection of the plate and backing bar. The crack had then propagated at right angles to the fusion zone through the HAZ in a series of steps. The main crack had propagated parallel to the fusion boundary along the subcritical HAZ and was associated with long stringers of type II manganese sulfide (MnS) inclusions (Fig. 51b).

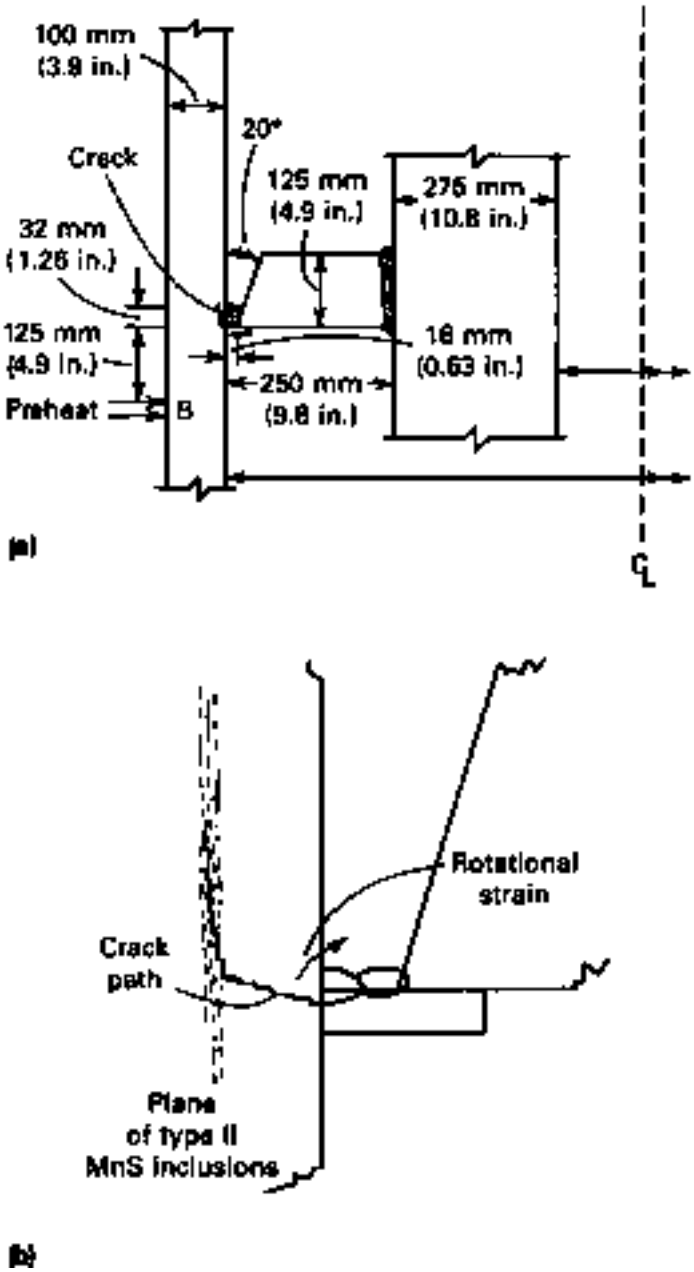


Fig. 51 Armature for a large electric motor that failed during fabrication. The cause of failure was lamellar tearing. (a) Schematic diagram of the weld preparation and the position of the failed weld. (b) Schematic of crack initiation and propagation from the root of the weld

The morphology of the cracking is typical of lamellar tearing. This can occur in fillet or groove welds where welding stresses are effectively applied across the thickness of at least one of the plates being joined. If the through-thickness ductility is very low, as in this case because of the presence of large stringers of type II manganese sulfides, then cracking results. The presence of hydrogen in the weldment can also enhance the likelihood of cracking. In this case, the deposition of an unbalanced groove weld where rotation can occur around the root, together with the superimposition of thermal strains from the one-sided application of preheat, produces a high enough strain to initiate and propagate the lamellar tear.

Some of the major solutions available for preventing lamellar tearing—for example, changing the design or using prior destructive testing, such as a through-the-thickness tensile test—cannot be used in this particular case. The high level of restraint in this particular welding application may also have given problems even with a low-risk plate, for example, vacuum degassed sulfide shape controlled plate. However, some benefit can be obtained from the use of welding procedures that reduce the effect of restraint. In this case, the use of a balanced welding technique, together with a higher preheat of 150 °C (300 °F) uniformly applied to both plates, helps reduce the restraint.

Conclusion. The weld failed by lamellar tearing as a result of the high rotational strain induced at the root of the weld caused by the weld design, weld sequence, and thermal effects.

Corrective Measure. The immediate solution to the problem was to use a “butter” repair technique. This involved removing the old weldment to a depth beyond the crack and replacing this with a softer weld metal layer before making the main weld onto this butter layer. In this manner, the peak stresses are generated in a ductile region of buttered weld metal, and the high contractional strains are withstood. A successful repair was effected.

Hydrogen-Induced Cold Cracking. Cold cracking is the term used for cracks that occur after the weld has solidified and cooled; it occurs in either the HAZ or the weld metal of low-alloy and other hardenable steels. Because these cracks occur under conditions of restraint, they are often referred to as restraint cracks. Cracking may occur several hours, days, or weeks after the weld has cooled; consequently, the term *delayed cracking* is also used. On the basis of location, cracks are often described as toe cracking, root cracking, or underbead cracking.

For cold cracks to occur in steels, three principal factors must be present: atomic hydrogen, HAZ or portion of the weld metal that is susceptible to hydrogen embrittlement, and a high tensile stress resulting from restraint. Controlling one or more of these factors may reduce the occurrence of cold cracking. The basic relationships among the variables responsible for cold cracking and the methods of controlling these variables are summarized in Fig. 52 and 53.

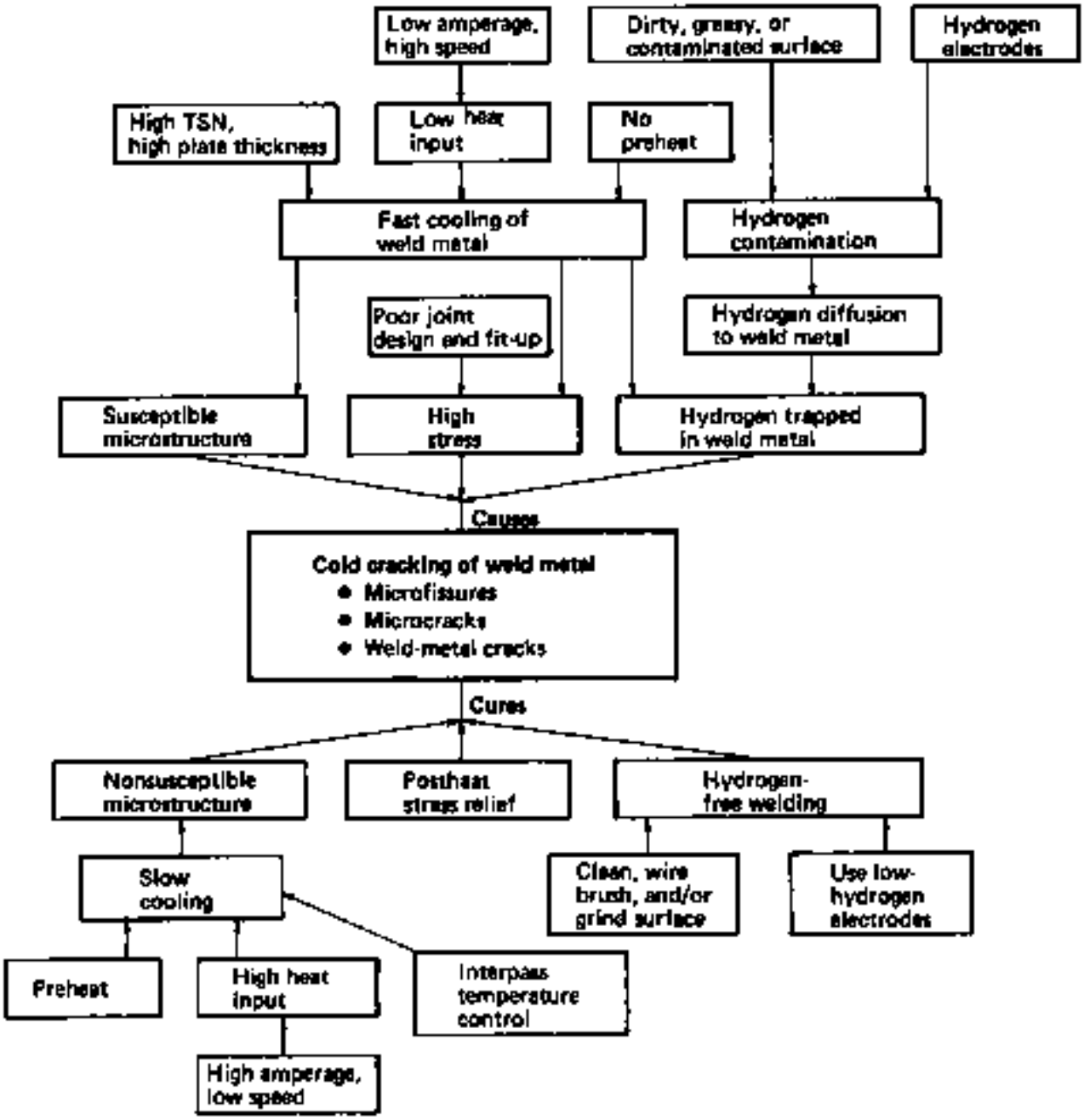


Fig. 52 Causes and cures of cold cracking in weld metal. Thermal severity number (TSN), which is four times the total plate thickness capable of removing heat from the joint, is thus a measure of the member's ability to serve as a heat sink.

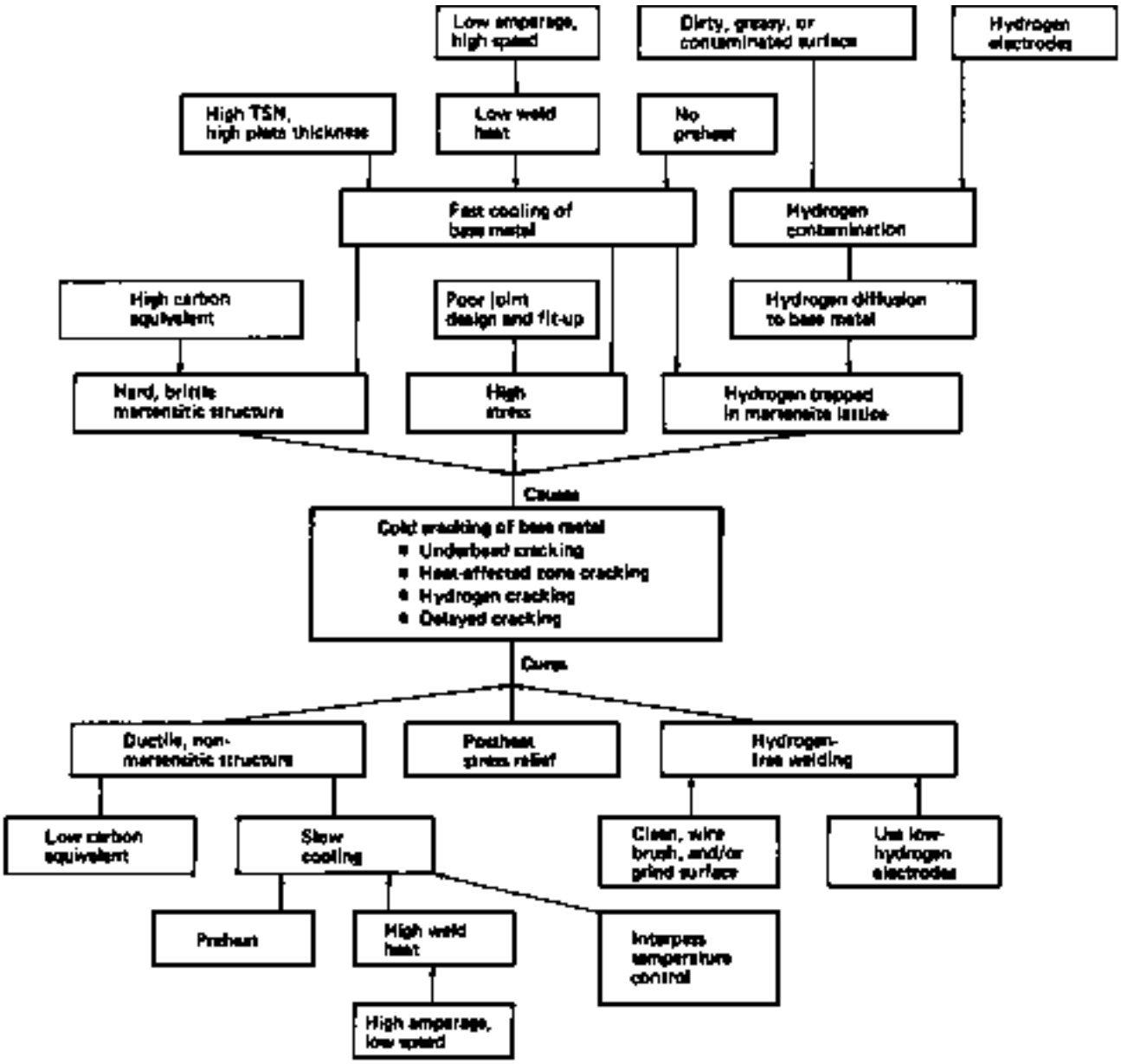


Fig. 53 Causes and cures of cold cracking in base metal

In steels, cracking in the base metal is often attributed to high carbon, alloy, or sulfur content. Control of this cracking requires the use of low-hydrogen electrodes, high preheat, sufficient interpass temperature, and greater penetration through the use of higher currents and larger electrodes. The susceptibility of the microstructure to cold cracking relates to the solubility of hydrogen and the possibility of supersaturation. Austenite, in which hydrogen is highly soluble, is least susceptible to cold cracking, and martensite, in which the solubility of hydrogen is lower, is most susceptible, because the rapid cooling necessary for the austenite-to-martensite transformation traps the hydrogen in a state of supersaturation in the martensite. For this reason rapid cooling rates must be avoided.

Cold cracks, or hydrogen cracks, form in welds after solidification is complete. In steel, cold cracks depend on the presence of a tensile stress, a susceptible microstructure, and dissolved hydrogen (Ref 25, 26). The stress may arise from restraint by other components of a weldment or from the thermal stresses created by welding in a butt, groove, or T-joint. The process of HAZ cracking in steels is usually described as follows (Ref 27):

- Hydrogen is introduced into the hot weld metal and dissolves into the weld pool as atomic or ionized hydrogen.
- As the weld metal cools, it becomes supersaturated with hydrogen, and some hydrogen diffuses into the HAZ.

- Under rapid cooling, hydrogen has insufficient time to diffuse from the HAZ. Also, rapid cooling increases the chance for the HAZ to transform from austenite to martensite.
- Atomic hydrogen is insoluble in martensite and seeks rifts and discontinuities in the lattice, where it collects.
- External stress caused by thermal contraction, combined with the effects of the trapped hydrogen, causes discontinuities to enlarge to crack size.
- In time, hydrogen diffuses to the new crack root and causes the crack to enlarge, thus the occurrence of delayed cracking.

This process repeats itself until the stress in the weld has been relieved. Hydrogen and stress also play an important role in weld metal cracking. Microcracks in the weld metal are often associated with small, nonmetallic inclusions that provide points of initiation.

Hydrogen cracking can occur in either the HAZ or the weld metal and can be either transverse or longitudinal to the weld axis. The level of preheat or other precautions necessary to avoid cracking depends on which region, weld metal or HAZ, is the more sensitive. In the older carbon-manganese medium-strength structural steels, the HAZ was usually the more critical region, and weld metal rarely caused a problem. Modern steels with low carbon are now very resistant to hydrogen cracking; therefore, the weld metal may be the controlling factor.

Hydrogen cracking in the weld metal depends on the same fundamental factors as in the HAZ, that is, hydrogen content, microstructure, and stress. The controlling variables in practice are usually strength level, hydrogen content, restraint, and tensile stress level. In single-pass welds and the root run of multiple-pass welds, the root gap provides a stress concentration to the transverse stress; this leads to longitudinal cracks in the weld metal.

The higher dilution of the root run may result in a harder weld bead that is more likely to crack. This is the predominant form of cracking in such applications as pipelines. In cases where transverse notches exist—for example, an unspliced backing bar (that is, where the backing bar has not been splice-welded into one continuous piece)—the longitudinal stress becomes the controlling factor, and transverse cracks may occur. In heavy multiple-pass welds, cracking is generally in the transverse direction, either normal to the surface or at an angle of about 45°. The latter type has been called chevron cracking. The solution to these cracking problems rests with proper control of hydrogen through such practices as handling of electrodes, drying of fluxes and electrodes, and adequate preheat.

The presence of hydrogen in a weld is generally due to moisture that is introduced in the shielding gas, dissociated by the arc to form elemental hydrogen, and dissolved by the molten weld puddle and by the adjacent region in the HAZ. In the supersaturated state, the hydrogen diffuses to regions of high stress where it can initiate a crack. Continued diffusion of the hydrogen to the region of stress concentration at the crack tip extends the crack. This behavior means that hydrogen-induced cold cracking is time dependent, that is, time is needed for hydrogen diffusion, and the appearance of detectable cracks can be delayed until long after the weld has passed inspection. The redistribution of hydrogen, however, does not depend only on the thermal history of the weld and the diffusion coefficient of hydrogen. The stresses in the weld area also affect the redistribution behavior and thus the cold-cracking tendency.

The remedy is to eliminate as far as possible the sources of hydrogen—water, oils, greases, waxes, paint, and rust that contains hydrogen or hydrates. Preweld and postweld heating promote the escape of hydrogen by diffusion and reduce the level of residual tensile stress that may be present. Hydrogen damage to weld deposits (either as porosity or as cracking) can be avoided simply by proper control of hydrogen through handling of electrodes, drying of fluxes and electrodes, and adequate preheat. For low-hydrogen shielded metal arc electrodes, this means storing in holding ovens to prevent moisture from reentering the electrode coating from the air. Normally, such ovens should be controlled between 120 and 175 °C (250 and 350 °F). Electrodes may require baking to remove moisture at temperatures as high as 425 °C (800 °F) before storing, but excessive baking can oxidize ferroadditions in the electrode coating and affect deoxidation of the weld pool.

The most common approach for eliminating hydrogen cold cracking—after ensuring that a low-hydrogen process is being used and that the base metal and filler metals are properly cleaned—is to employ preheating. The use of carbon equivalent (CE) formulas in conjunction with the degree of restraint help to define the level of preheat required when joining different steels. As both the CE and degree of restraint increase, the need for and the level to which preheat is applied also increase. For steels, when preheat does not act as an effective way to stop hydrogen cracking—and if increasing the preheat temperature does not work—applying a postheat for 3

to 4 h after welding has finished may be effective. Postheating in conjunction with preheating is far more effective than preheating alone in solving cold cracking problems.

Hot cracking may be caused by joint design and the restraint imposed on the weld, by bead shape, or by the presence of low-melting constituents that extend the temperature range of low hot strength and low ductility to temperatures below that of the alloy. In steels, these constituents are compounds, such as phosphides and sulfides, and elements, such as copper, which can segregate at grain boundaries and cause grain-boundary tearing under thermal-contraction stresses. Hot cracks can be minimized or eliminated through residual-element control and control of the weld pool shape. Weld beads with a high depth-to-width ratio can promote the buildup of low-melting phases at the pool centerline and thus lead to centerline hot cracking.

The same type of hot crack is a crater crack. The material in the crater region is richer in solutes than the remainder of the weld, because of the directional nature of the weld metal solidification, and is very prone to cracking. The best method of control is to use welding materials with minimal contents of residual elements, such as phosphorus, sulfur, and copper in steel, thus reducing the possibility of segregation. If this is not feasible, the next best procedure is to avoid the slow cooling rates that favor segregation between the liquidus and solidus temperatures and accelerate cooling until the crater has solidified. In some hot-cracking situations, a reduction in travel speed can alter the solidification directions in the weld deposit and thus reduce susceptibility to hot cracking.

Hot cracking is the most common form of weld metal cracking in steels, especially stainless steels. The term encompasses cracks that form while a weld is solidifying or when a weldment is reheated. As noted, it is often caused by excessive sulfur, phosphorus, or lead content in the base metal. It can also be caused by an improper method of breaking the arc or in a root pass when the cross-sectional area of the weld is small, compared to the mass of the base metal. Solidification cracking, liquation cracking, ductility-dip cracking, and reheat cracking are types of hot cracking. Such cracks are normally intergranular and interdendritic in nature and can occur in either the weld metal or the base metal HAZ.

Solidification cracking occurs within a few hundred degrees of the nominal liquidus temperature of the weld metal and is induced by welding stresses and the presence of low-melting-point constituents that form as a result of segregation during the liquid-to-solid-phase transformation process.

Liquation cracking (often called hot tearing or HAZ burning) is another type of high-temperature cracking that may occur in welds of all metals. Localized liquation of low-melting-point constituents, such as foreign inclusions, carbide phases, or local segregations of certain alloying elements in the solid, may form liquid grain-boundary films. Welding shrinkage strains may, in turn, be sufficiently high to open liquation cracks.

Ductility-dip cracking may occur in either the weld metal or the HAZ of austenitic stainless steels and nickel-base superalloys. This type of crack forms while the weld cools through a range of temperatures where the ductility of the particular metal is inherently low. The ductility-dip temperature for austenitic stainless steel falls just below the recrystallization temperature. If sufficient strain is present as the metal cools through the ductility-dip temperature range, cracking occurs.

Reheat cracking or stress-relief cracking may form in the HAZ of welds in ferritic and austenitic steels. Postweld heat treatment may be specified for weldments to reduce (or eliminate) residual stresses and/or to restore toughness in the HAZ and weld metal. Reheat cracking may take place during this postweld heat treatment as a result of pre-postweld heat treatment microstructure of the HAZ brought about by welding. Welding heat input induces an elevated temperature solution treatment in the HAZ, which leads to dissolution of metallic carbides. Upon reheating, interstitial carbon precipitation hardening may strengthen the grains sufficiently so that relaxation of internal stresses may occur only by grain-boundary sliding, with little or no intergrain deformation; hence, reheat cracking occurs. Reheat cracking has also been linked to temper embrittlement and the presence of trace elements, such as boron, arsenic, antimony, and tin, in steels. Remedies for reheat cracking include changes in metal composition (use of less-susceptible steel), reduction in thickness, decrease in sulfur content, slower heating rate to avoid thermal shock, controlled cooling rate to prevent precipitation of carbides, higher creep ductility, good joint design, decrease in constraint, high preheat, elimination of stress concentration at welds, and use of weld metal with low hot strength.

Causes and Prevention of Hot Cracking. The factors affecting the formation of hot cracks include metallurgical properties, welding variables, and mechanical properties. The basic relationships among these factors and methods for controlling hot cracking in steels are shown in Fig. 54 and 55.

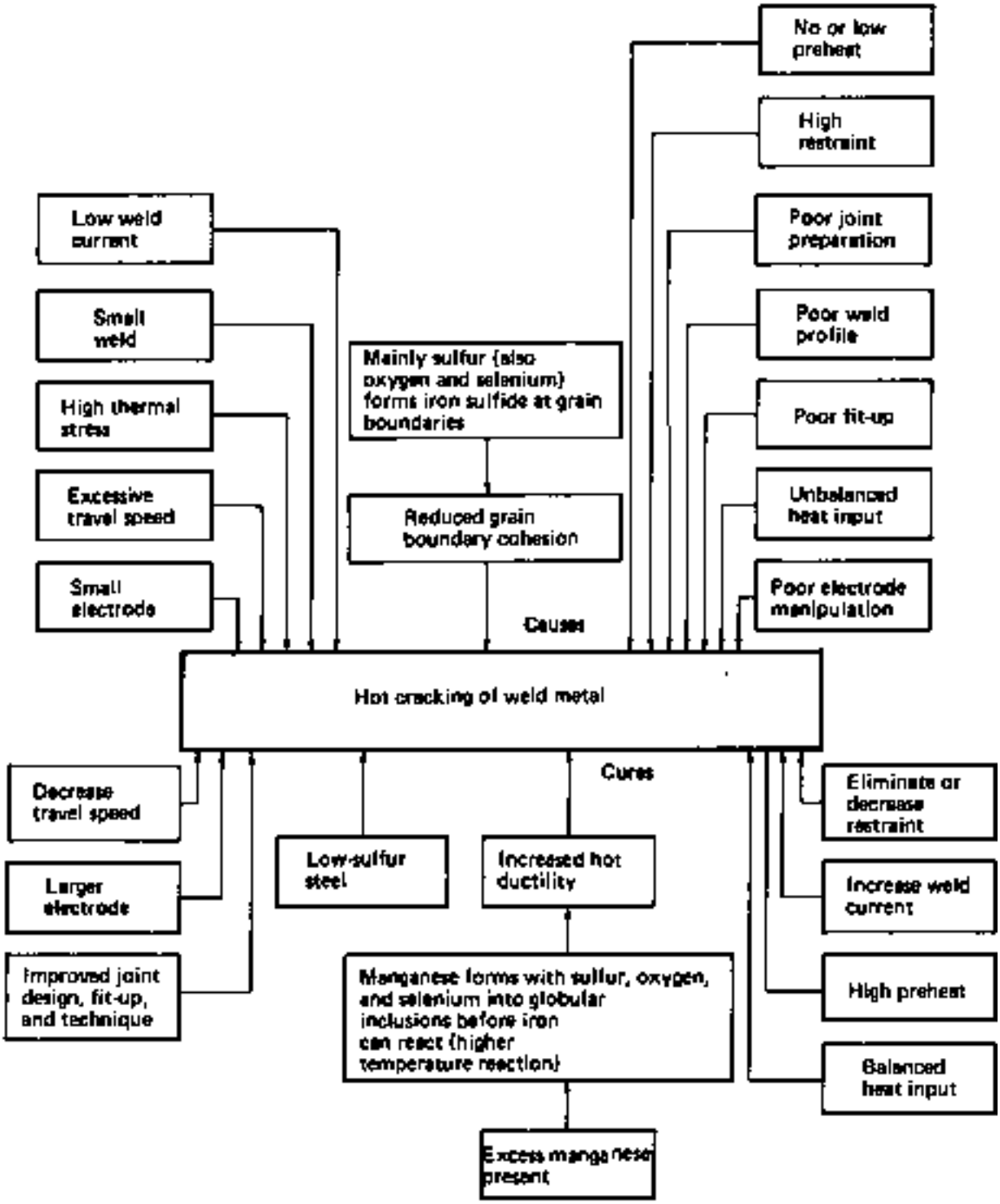


Fig. 54 Factors affecting hot cracking in weld metal

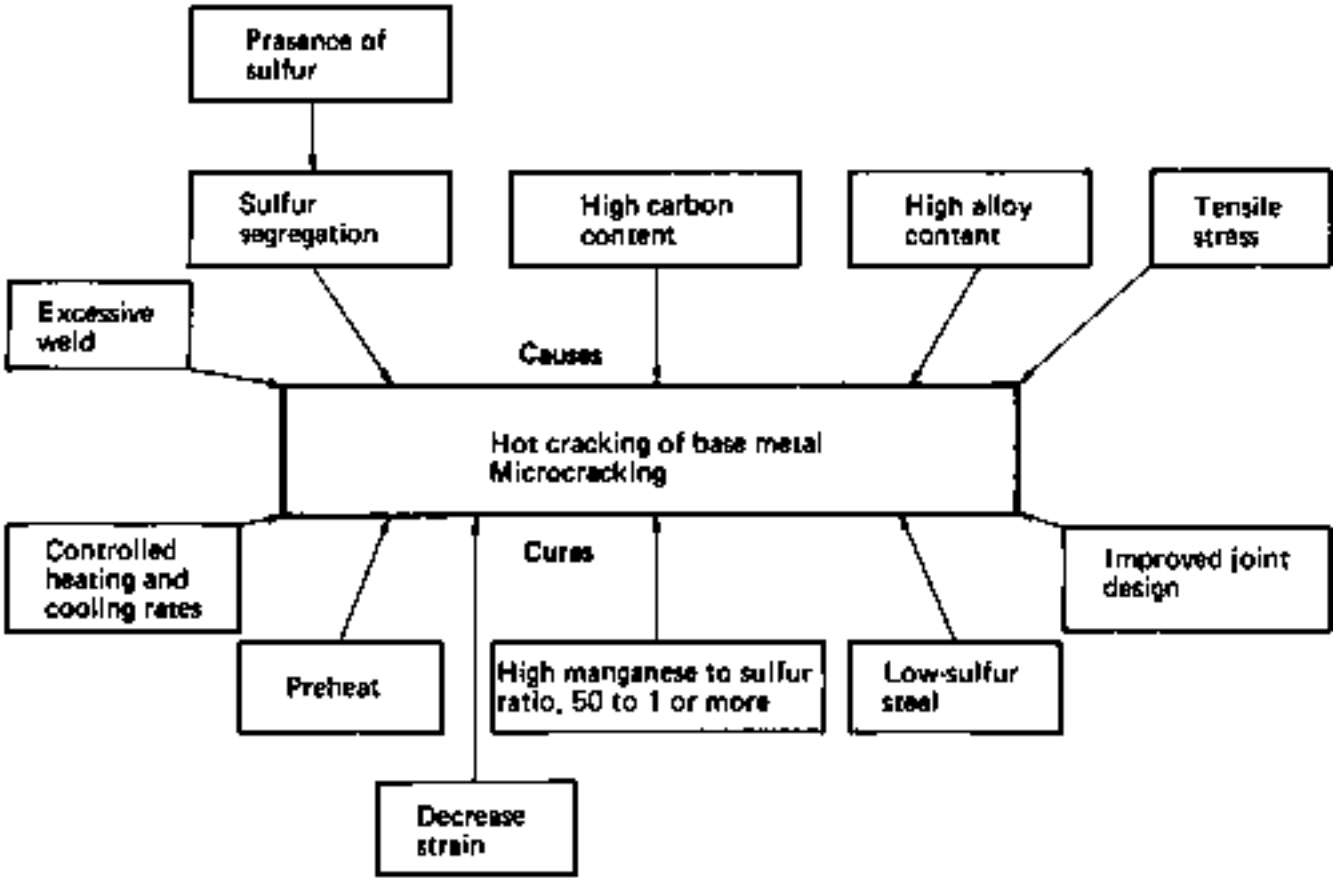


Fig. 55 Factors affecting hot cracking in the base metal HAZ

During solidification of weld metal, several compounds form whose presence, in their liquid phase, at grain boundaries contributes to hot cracking. Sulfur is considered the most detrimental element, because it reacts to form many low-freezing-temperature compounds such as iron sulfide.

Sulfur compounds have the greatest effect on hot cracking in weldments stressed in the through-thickness direction of the base metal. Therefore, sulfur content in the base and filler metals should be kept low. Carbon is another detrimental element, because it affects the liquid phase of the weld metal and tends to decrease high-temperature ductility of the weld metal. The level of carbon present in the base metal cannot be altered to any great extent, but the effects of carbon can be offset by high manganese-to-sulfur ratios. With a carbon content of 0.06 to 0.11 wt% in the weld metal, hot cracks can be eliminated completely when the manganese-to-sulfur ratio is >22. For carbon contents between 0.11 to 0.13%, a ratio of over 30 is required to avoid hot cracking. When the carbon content exceeds 0.13%, the effect of the manganese-to-sulfur ratio is nullified. For this reason, the carbon content of most deposited steel weld metal is kept low (that is, 0.06% C or below).

Silicon and phosphorus do not affect the liquid phase of the weld metal directly; however, they promote segregation of the sulfur and thereby aid sulfur reactions. Nickel has a more direct effect on the liquid phase, but its effects on crack formation are not completely understood.

No matter how pronounced the liquid phase in the weld metal is, hot cracks do not form unless tensile stresses are imposed on the weld. Unfortunately, stresses are impossible to avoid. Consequently, the greater the stresses imposed during solidification or reheating, the more severe the cracking.

The size and thickness of the base metal, the joint design and its restraints, and the size and shape of the weld bead influence the mechanical stresses in a weld. Furthermore, different welding processes produce different amounts of heat input, thus producing variations in microstructural changes and levels of residual stresses. Joint design should provide a good fit-up. Excessive amounts of weld metal should be avoided. Welding procedures that minimize weld restraint should be specified.

Failure Origins Related to Specific Arc Welding Methods

Shielded Metal Arc Welding. The condition of the electrode coating is important to the quality of shielded metal arc welds. There are two general types of coatings: one is a cellulose-type, which breaks down in the arc to produce hydrogen; the other is a cellulose-free coating that produces very low hydrogen levels. The first type of electrode coating is used for welds on noncritical structures in which hydrogen in moderate amounts is acceptable or in situations where its fast-freeze and good penetration characteristics are desirable and hydrogen is controlled by preheat requirements. The second type of electrode coating is used for welds on structures for which even small amounts of hydrogen are not tolerable.

The low-hydrogen electrode coatings are hygroscopic and pick up moisture from the atmosphere, particularly if the relative humidity is high. This occurs especially if the electrodes are exposed to the air for more than 2 to 4 h. Moisture or other hydrogen-containing compounds in the coating are dissociated by the arc temperature, and the resulting atomic hydrogen is readily dissolved in the weld puddle. Water vapor may also be directly dissolved by the molten steel with very little dissociation. The solubility for hydrogen in the steel decreases abruptly and drastically (as much as 4 to 1) when the molten phase freezes (Ref 28). A further, significant decrease in solubility occurs when the austenite transforms to either ferrite or martensite. During freezing, the hydrogen may escape by diffusion or formation of bubbles, but a portion remains in the austenite in a supersaturated state. Any dissolved water is largely ejected as bubbles during the freezing process, which may form a “picture frame” of wormholes if the quantity of dissolved water and the rate of freezing are both moderate. If a great deal of vapor has been dissolved, the metal puffs seriously, resulting in an extremely porous structure.

Flux Cored Arc Welding. In this method, the operating characteristics of the electrodes vary with the core composition and the shielding gas (if used). Care should be taken to remove the slag, in some cases, and to maintain the low-hydrogen characteristics of the tubular electrodes. Adequate shielding must be maintained through either an external source or a self-shielding type of electrode.

Gas Metal Arc Welding. Typical problems that can occur in this type of welding are IF defects caused during short-circuit transfer and porosity from shielding-gas problems. Adequate deoxidizers are required in the electrode to prevent porosity in the weld. The low-hydrogen nature of the process is lost if the filler metal or shielding gas are contaminated. Electrodes must be properly stored and cleaned.

Submerged Arc Welding. Factors that contribute to the formation of discontinuities in other arc welding processes generally also prevail for the submerged arc process. However, in SAW, flux is separate from the electrode, and the quantity of the flux layer can affect results. For example, if the flux layer is too deep, it produces a ropey reinforcement, whereas too shallow a layer contributes to the formation of porosity and gas pockets. Also, a flux layer that is too narrow causes a narrow reinforcement. Flux must be kept dry to maintain low-hydrogen characteristics.

Gas Tungsten Arc Welding. A potential but infrequent problem in GTAW is the pickup of tungsten particles from the electrode, caused by excessive welding current or physical contact of the electrode with the weld puddle. Tungsten inclusions are extremely brittle and fracture readily under stress and can thus act as failure origins. Tungsten inclusions can be caused by excessively high welding currents, incorrectly sized or shaped electrodes, and oxygen contamination of the shielding gas. Tungsten particles are easily detected in radiographs.

Plasma arc welding is closely related to GTAW. A principal difference between these two welding processes is the keyhole effect produced by the constricted arc column of a plasma arc weld. This effect is created by displacement of molten metal by the forceful plasma jet, permitting the arc column to pass completely through the workpiece. If the welding current is turned off abruptly at the end of a welding pass, the keyhole does not close, and a hole or a large region of subsurface porosity is produced. This is not a problem, however, if the weld can be terminated on a runoff tab.

References cited in this section

16. A.R. Cuiffreda and R.D. Merrick, Brittle Fracture of a Pressure Vessel: Study Results and Recommendations, *Proceedings of the American Petroleum Institute,* Vol 62, 1983

22. “Welding Pipelines and Related Facilities,” API STD 1104, 19th ed., American Petroleum Institute, 1999

23. *Weld Integrity and Performance,* ASM International, 1997

24. G. Screm et al., The effects of Weld Defects in Reducing the Fatigue Strength of Welded Joints, *Alluminio,* Vol 44 (No. 3), 1975

25. B.A. Graville, *The Principles of Cold Cracking Control in Welds,* Dominion Bridge, Hanover, 1975

26. F. Coe, *Welding Steels Without Hydrogen Cracking,* The Welding Institute, Abington, Cambridge, U.K., 1973

27. R.D. Stout, *Weldability of Steels,* 3rd ed., Welding Research Council, 1978

28. D.J. Carney, Gases in Liquid Iron and Steel, *Gases in Metals,* American Society for Metals, 1953, p 69–118

Other Welding Processes

Previous sections have outlined some of the general origins of failure associated with the inherent discontinuity of welds and the imperfections that might be introduced from arc welding processes. The remaining portion of this article describes failure origins in other welding processes. More detailed information on the these welding processes is provided in *Welding, Brazing, and Soldering,* Volume 6 of *ASM Handbook.*

Failure Origins in Electroslag Welds

Electroslag welding is most commonly used for joining relatively thick sections of low-carbon steels, but it can also be used to weld medium-carbon steels, such as AISI 1045 and 1050, and, to a lesser extent, for welding high-strength structural steels, high-strength alloy steels (such as D-6AC), stainless steels, and nickel alloys.

Because an arc is used to start the welding in the electroslag process and is used until a pool of molten flux (slag) has been formed that is deep enough to quench the arc, the discontinuities typical of arc welds are usually present in the initial deposit. These potential discontinuities, however, can be isolated from the welded joint by the use of a starting lug that can be cut away later, leaving only metal deposited by electrical resistance heating of the molten flux. This heat serves to melt both the filler metal and the surface of the base metal that are in contact. Once the metal has been deposited by resistance heating of the flux, welding must be continuous until the joint has been completely formed, if welding defects are to be avoided. A halt at any intermediate stage in completing the joint weld probably results in a region containing porosity, nonmetallic inclusions, cracks, and undercuts when welding is restarted. In welds completed without interruption, discontinuities can be detected visually or by nondestructive testing.

Inclusions. Slag inclusions derived from the flux pool may be found at the weld interface. These inclusions can be best avoided by making certain that the guide tube (if used) for the electrode wire is centered between the base metal surfaces and between the dams. Other inclusions may result from melting of nonmetallic laminations in the steel being welded. When located by nondestructive evaluation, the laminations should be gouged to a depth greater than the expected weld penetration, and the cavities filled by repair welds using low-hydrogen electrodes.

Porosity. Because electroslag welding is basically a low-hydrogen process, gas porosity is usually not encountered. However, if gas porosity should be detected and if the flux used is of the composite or bonded type, the flux should be prebaked. Electroslag welding establishes directional solidification in the welds, which is favorable for the production of welds that are free of shrinkage porosity.

Poor Weld Contour. The weld-face contours are formed by the copper dams or shoes that bridge the gap between the plates being welded. If the clamping pressure on the shoes is too great, the resulting flexure causes an improper weld contour. In some circumstances, the weld face may contain some undercutting, which should be easily detected by visual, magnetic-particle, or liquid penetrant inspection and repaired by grinding or by arc welding plus grinding.

Excessive Grain Growth. Because of their slow cooling rates, electroslag welds usually have a large grain size in the weld fusion zone. This can lead to lower toughness, particularly at low temperatures, and may make electroslag welds more sensitive to contamination than other types of weldments.

Failure Origins in Electrogas Welds

Many aspects of electrogas welding resemble electroslag welding, and similar failure origins may occur. Poor weld contour, cold cracking, IF, IP, porosity, and hot cracking are similar in nature. Because a flux is not employed in electrogas welding (except in flux cored electrodes or perhaps a small amount to start the process), this possible source of inclusions is frequently absent in electrogas welds. Nonmetallic inclusions introduced by base metal laminations are similar to that in electroslag welding.

Gas porosity in electrogas welding is almost always the result of an inadequate volume of shielding gas or because air was aspirated into the shielding gas hose. A large gas flow and an increased height to gas coverage should correct this condition. Highly directional solidification exists in electrogas welding to the same degree as in the electroslag process. Thus, the welds should be free of shrinkage porosity.

The occurrence of hot cracks usually can be related to the alloy content of the electrode filler metal and can be eliminated by selection of the proper combination of filler metal and shielding gas. A filler metal should be similar to the base metal in composition, tensile strength, and elongation. Selection of the shielding gas depends somewhat on the composition of the base metal.

Failures in Resistance Welds

The general term *resistance welding* includes resistance spot welding, resistance seam welding, and projection welding, which are all closely related. Flash welding (a resistance butt welding process) is dealt with separately in this article, predominantly because the extensive arcing and plastic upsetting that characterize this process make it appreciably different from conventional resistance welding processes.

The qualities of resistance welds are affected by many variables, including the properties of the material to be welded, the surface smoothness and cleanness, the presence of any surface coatings, the electrode size and shape, and the welding machine settings that determine welding time, pressure, and current. Successful welding depends on consistent weld properties, which in turn require uniform welding conditions. Experience has shown that a change in any single variable of more than 10% is sufficient to make the weld unacceptable. Unacceptability may represent failure to meet a specified property limit, such as a minimum tensile or impact strength, or it may indicate actual fracture of the weld. Some possible causes of resistance weld failures are described subsequently.

Inclusions. The sources of inclusions in a resistance weld include the surfaces of the parts being joined and their internal structures. Surface contaminants that may cause inclusions are dirt, rust, scale, certain types of coatings, and sometimes oil and grease. These may be between the surfaces to be joined or may be between a part and the electrode contacting the part. Whether or not the inclusions cause failure depends on their quantity, size, location within the weld microstructure, and such properties as melting point or softening range. An inclusion that is likely to become fluid enough at the nugget-fusion temperature to penetrate a grain boundary would be particularly damaging.

Porosity. Some inclusions may cause weld porosity, which is generally considered undesirable but is totally unacceptable if certain size limitations are exceeded. Other inclusions may generate blowholes, usually a cause for immediate rejection. Improper machine settings, particularly those causing excessive current and insufficient pressure, can cause porosity or blowholes. These defects at the weld surface can cause part rejection, but moderate porosity near the center of the weld nugget is usually acceptable.

Inadequate Penetration. Except in aircraft-quality spot welding, limits of nugget penetration in welded parts are commonly not specified. Even for aircraft quality, the range may be from 20 to 80%. Inadequate penetration results from too low a current density in the weld, perhaps from a malfunction or improper setting of the machine or other cause, such as mushroomed electrodes, partial shunt of welding current through adjacent welds, excess pressure, or improper material.

The effect of inadequate current density was observed when a bench-type resistance spot welder was being used to join two crossed wires by direct-current welding for an electronic assembly. The measured weld strength was 4 kg (9 lb). After maintenance on the weld head, the cables to the electrodes were accidentally reversed, and the

subsequent welds showed a strength of 1.4 to 2 kg (3 to 4 lb). This was due to the effect of polarity in welding dissimilar metals with direct current. This effect can also be observed in very short-time welds (1 to 2 cycles) made with alternating current.

In another instance in which inadequate current density prevented satisfactory welding of a crossed-wire junction, a tinned-copper wire was to be welded to a nickel wire. The welding operation, in which power input was 25 watt-seconds, was unsuccessful in producing a weld. Inspection of the wires for cleanness with emery paper revealed a copper color beneath the surfaces of both wires, and a hand magnet showed that both wires were nonmagnetic. Another tinned-copper wire had been accidentally substituted for the nickel wire; a power input as high as 65 watt-seconds would have been necessary to weld together the two copper wires.

Insufficient current density resulted in inadequate joint penetration when three spot welds were made in attaching a small angle to a sheet; both the angle and the sheet were of cold rolled steel 1.6 mm ($\frac{1}{16}$ in.) thick. The spot welds were made 13 mm ($\frac{1}{2}$ in.) apart. However, welds 2 and 3 were often weak. Because shunting was suspected as the cause of the weak welds, tests were performed, which showed that weld 1 was always good, while welds 2 and 3, when made in that order, were both weak, because part of their current passed through the earlier-made welds. Conversely, it was shown that if weld 3 was made after weld 1, they were both strong, whereas if weld 2 was made last, it was extremely weak. It was decided to eliminate weld 2 and to use two welds 25 mm (1 in.) apart, which would minimize shunting, ensure adequate current density, and provide two strong welds consistently.

Poor weld shape encompasses a variety of defects that may result from part configurations causing undersized nuggets that can lead to failure. Examples of these conditions include poor fits, inadequate flange width or contacting overlap, edge bulges, cracks, burn-throughs, and distortion. In many such instances, the defects are often the result of carelessness on the part of the machine operator. Less frequently, the fault may be the result of an improper design, such as the insufficient flange width of the bracket shown in Fig. 56.

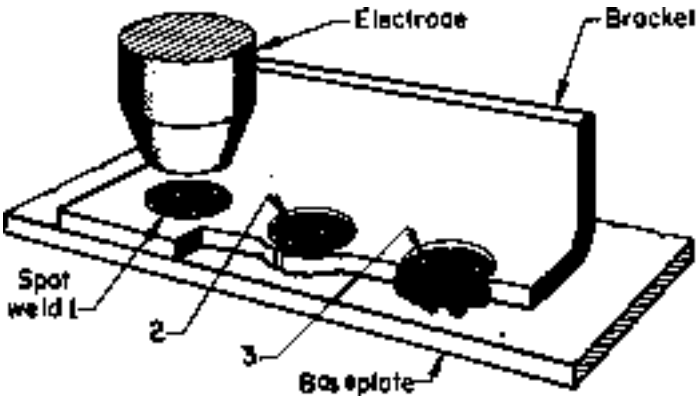


Fig. 56 Schematic of the failure of spot welds to join a bracket to a baseplate assembly. The failure was caused by inadequate welding current density resulting from extension of the electrode beyond the edge of the bracket flange.

The bracket shown in Fig. 56 had been difficult to spot weld to the baseplate. The size of the welding electrode was determined by the bracket thickness, which required that the electrode be of a specific size that related to the electrode-face diameter and to the electrode-shank diameter. The narrowing flange caused the electrode used to extend over the edge of the bracket, producing a bulge and crack at spot weld 2 and an unsound weld at position 3 because of metal expulsion. The operator could have selected a smaller-diameter electrode, but then the spot welds, although sound, would have been smaller and perhaps unable to carry the load on the bracket. The correct solution is to redesign the bracket with a wider flange or to emboss the flange with projections to concentrate the current flow at proper distances from the flange edge.

Poor weld surfaces that can cause weld failure, usually because of unsatisfactory nugget formation, are often the result of poor machine operations or adjustments that lead to weld spatter, blowholes, electrode pickup, or excessive indentation.

Cracks. Welds containing cracks usually result from overheating, improper loading during the welding cycle, or the use of welding programs that are unfavorable for crack-sensitive metals. Hot cracks are uncommon in

resistance welding, because the time at high temperature is so brief. Cold cracks may occur, because the weld metal froze under insufficient pressure and was therefore forced to undergo thermal contraction relative to the surrounding matrix that demanded more deformation than the metal could tolerate. Such cracks weaken the weld considerably if they are at the weld rim but are innocuous if at the weld center. These cracks may be avoided by proper pressure control, especially by application of forging pressure at the end of the weld cycle.

Unfavorable fabrication procedures or inadequate bend radii sometimes cause cracks that propagate in parts to induce failures that can be confused with weld-crack failures. There is a clear need to determine the true crack origin before assigning the cause of failure in such circumstances. Water quenching during resistance seam welding can lead to quench cracks from the transformation to martensite and thermal contraction stresses.

Inadequate Weld Properties. Conditions that affect nugget formation include surface coatings, preweld cleaning, electrode overlap, spot-weld-to-edge distance, sheet thickness, and wall thickness of embossed projections. Any variation in properties that influences the electrical resistance between the parts influences the weld quality. Intermittent weld failures can result from variations in surface conditions that are not machine controlled. Examples include the presence of rust or of die casting flash or mold-release compound in layers of varying thickness. Careful surface cleaning and inspection before welding are the only preventatives.

Parts made of hardenable alloys may be hardened by the high heat and rapid quench of a typical resistance welding cycle. Such parts may have inadequate ductility or toughness, and the welds can easily be broken unless they are tempered, either as an added portion of the weld cycle or in a separate operation outside the welding machine.

Failure Origins in Flash Welds

Flash welding is a variation of resistance welding, in which arcing (flashing) is used to supply the major portion of the initial heat for the formation of a butt weld. Preweld cleaning of the workpieces at the die contact area is important to ensure proper current flow and prevent local overheating of the workpiece surface. No surface coating (plated metal, conversion coating, or anodized coating) should be present before welding, because it results in weld-area contamination or die burns. Oxides and foreign particles can cause small pits in the surface as they overheat and burn into the workpiece. During the flashing action, craters are formed that contain molten metal and possibly oxides. If the energy input that produces the flashing is properly controlled, these oxides should be expelled with the flashing molten-metal particles that give the process its name. When the upsetting force is then applied, most of the impurities not expelled by the flashing are expelled with plastically deformed upset metal. Any nonmetallics that are not expelled usually remain at the fusion line, with little apparent depth back into the base metal on either side. Most static and fatigue failures that occur in flash welds originate at such discontinuities. These discontinuities normally have little effect on static strength but can measurably reduce fatigue life. Excessive energy input and insufficient upsetting force or travel speed are the most common causes of such inclusions.

Porosity. Insufficient upsetting force or travel may leave porous areas of cast metal in the weld. Excessive electrical energy input can also lead to porosity by the formation of large craters by the expulsion of molten metal. Where large-diameter pieces are to be joined, one end face should be slightly chamfered so as to start the flashing at the center of the cross section. This helps to avoid trapping of particles in the weld.

Incomplete Fusion. Inadequate heating results in incomplete fusion, but it may also result from an upsetting travel that is so slow that the metal does not become sufficiently plastic to forge properly during upsetting. Another cause of inadequate heating can be the shunting of electrical current around the closed side of a ring to be flash welded; sufficient current may be bypassed to make flashing difficult, even resulting in a melt-through of the closed side of the ring.

Poor Weld Contours. Whenever heating is not uniform over the intended joint, an unfavorable temperature gradient is established that contributes to misalignment after welding. If parts are not properly aligned by the dies, they may slip past each other during upsetting, creating a lap instead of a proper weld. Workpieces of different cross sections that have not been adjusted in design to achieve a good heat balance, such as tubes of different diameters or wall thicknesses, may slide (telescope) over each other and create a poor-quality weld.

The internal flash and upset metal formed in flash welding of tubes or pipe are normally not removed (often, they cannot be), which leaves a reinforcement with two sharp notches adjacent to the fusion line. These notches act as stress raisers that, under cyclic loading, reduce the fatigue strength. They also restrict fluid flow and can serve as barriers to streamline fluid movement concentration-cell sites for corrosion.

Hot Cracks. Alloys that possess low ductility over a temperature range below the melting point may be susceptible to hot cracking. Such alloys are more difficult to flash weld but can usually be welded if the most favorable welding conditions are selected. Tension obviously should be avoided, and because the rim of the upset metal is in circumferential tension as upsetting progresses, it may be necessary to keep the upsetting force to the minimum acceptable. In such an operation, some cracks are likely to form that are shallow enough to allow complete removal when the upset metal is machined away. Precise coordination of current cessation and upsetting travel are important, and the joint should be under moderate compression during cooling.
Cold Cracks. Insufficient heat during upsetting or excessive upsetting travel causes colder metal to be forced into the weld zone. The metal cracks transversely to the weld line in the upset zone. This effect is also encountered in the flash welding of hardenable steels, because of the mass quench provided by the remainder of the workpiece. The use of a slow cooling rate, as well as heat treatment following welding, alleviates transformation stresses that could cause cracking under service loads.

Failure Origins in Upset Butt Welds

Upset resistance butt welding is similar to flash welding but is less widely used. The advantages of resistance butt welding are:

- There is little weld spatter, because there is no flashing.
- Less metal is used in the weld upset.
- The upset is usually smooth and symmetrical, although it may be pronounced.

Upset butt welding is used in joining small ferrous and nonferrous strips, wires, tubes, and pipes end to end.
Because there is no flashing of abutting workpieces, joint preparation for upset butt welding is more critical than for flash welding. Shearing to provide clean, flat, parallel ends is the usual procedure for small parts and wire. Beveled or other special shapes, such as spherical, are used to restrict the area of initial contact in butt welding larger parts, especially if they have oxidized ends. Wires are sometimes pinch cut in wedge shapes and the opposing ends rotated 90°.
Unlike a flash weld, for which the proper upset slope is 45 to 80°, the slope of the burr of an upset butt weld is normally 80 to 90°, quite thin, and often split into two, three, or even four petal-shaped segments (Fig. 57); the burr diameter should be at least twice the wire diameter. A split burr with a diameter three to four times the wire diameter is a better burr. The burr can be easily removed by hand or by cutting pliers.

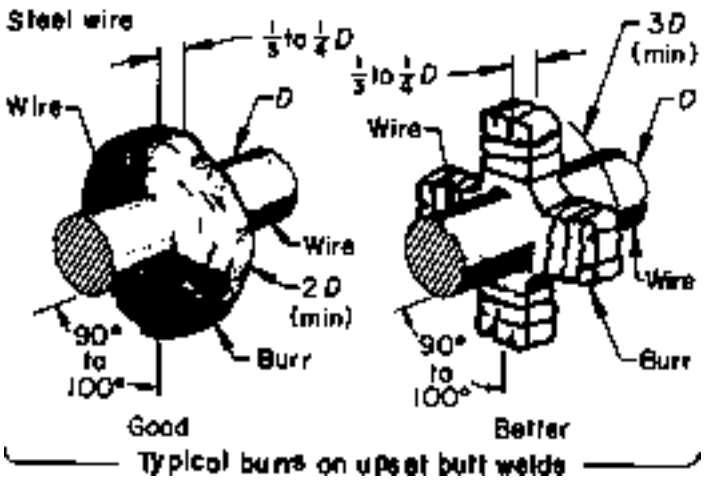


Fig. 57 Upset butt welded steel wire showing typical acceptable burrs on the welds. Dimensions given in inches

Failure Origins in Friction Welds

Welding by the direct conversion of mechanical energy to thermal energy at the interface of the workpieces without heat from other sources is called friction welding. The types of defects that may lead to failure of friction welds are described subsequently.

Center Defects. A small unwelded area that can occur in friction (inertia) welding at the center of the interface of the workpieces is called a center defect. It can range from 0.25 to 6.4 mm (0.010 to 0.250 in.) in diameter, depending on the size of the workpiece. Center defects are more common in low-carbon and medium-carbon steels than in high-carbon or high-strength alloy steels. Their cause is insufficient friction at the center of the interface, giving rise to heat generation that is inadequate to achieve a complete weld. An increase in peripheral velocity or a decrease in axial pressure avoids creation of a center defect. Another remedy is to provide either a center projection or a slight chamfer to one of the workpieces.

Restraint Cracks. A type of crack that is most commonly encountered in friction welding a bar to a larger bar or plate is called a restraint crack. This type of crack, which is usually observed in medium-carbon, high-carbon, and alloy steels, occurs during cooling and is caused by thermal stresses, particularly if the HAZ is thicker at the center than at the periphery. Decreasing the peripheral velocity or increasing the axial pressure to gain a thinner HAZ at the center eliminates this type of crack, but care must be exercised to avoid overcompensation, which would create a center defect.

Weld-Interface Carbides. Carbide particles may be present in a friction weld interface in tool steels or heat-resisting alloys. In both metals, the particles impair weld strength. If a friction weld of tool steel is fractured, circumferential shiny spots are seen in the weld interface. The formation of these spots is attributed to the lack of adequate time and temperature of the process to take the carbide particles into solution. In welding of heat-resisting alloys, precipitation of carbide particles at the interface at welding temperature can be prevented by increasing the amount of upset.

Poor Weld Cleanup. If the exterior surface of a friction weld does not clean up completely in machining to final diameter, it is the result of hot tearing at or near the surface. In ferrous metals, this usually happens when welding two alloys of different forgeability. The best procedure to avoid this is to increase the section size of the workpiece that has the better forgeability. If the workpieces must be of the same size, an increase in surface velocity and in axial pressure is helpful. For instance, in welding of a low-alloy steel engine-valve stem to a heat-resisting alloy head (using the flywheel method), a stepped pressure cycle has proved useful. The second pressure is applied to the end of the weld cycle to stop the flywheel rapidly, thereby eliminating the hot tearing. In friction welding of nonferrous alloys, poor weld cleanup is usually caused by hot tearing at the back edge of the HAZ, especially in copper alloys. An increase in speed helps avoid this condition. For aluminum alloys, an increase in speed plus an increase in axial pressure normally eliminates such hot tearing.

Hot-Shortness Cracks. Brittleness in the hot-forging temperature range requires special precautions if friction welds free of hot-shortness cracks are to be made in a hot-short metal. The welding conditions must be adjusted to keep the weld temperature to a minimum, using high axial pressures and very low surface velocities.

Trapped oxide occurs occasionally. Oxide entrapment is usually caused by a concave initial weld interface or a remnant of a hole for a machine center in one or both workpieces. As welding starts, the pieces behave like tubes, with flash flowing inward toward the axis. In normal circumstances, the trapped air is not able to escape and remains either under pressure as gas pores or combines with the metal as oxide and nitride. If pieces with machine-center holes must be used, the cavities should be large enough so that the air is not compressed more than 90%. Concave surfaces should be avoided.

Porosity. If cast workpieces that contain shrinkage or gas porosity at or near the weld interface are friction welded, the porosity is normally evident in the final interface. These defects are usually gray or black, but differ in appearance from center defects, which have a metallic appearance. Porosity cannot be eliminated, but careful inspection of the workpieces reveals the defect.

Failure Origins in Electron Beam Welds

Electron beam welding (vacuum and nonvacuum) is used to weld essentially any metal that can be arc welded, and the weld quality in most metals is either equivalent to or better than that achieved in the best gas tungsten arc or plasma arc welding. The advantages of vacuum electron beam welding include the capability of making deeper, narrower, and less tapered welds with less total heat input than is possible with arc welding and a superior control of depth of penetration and other weld dimensions and properties. Discontinuities that may cause difficulties in electron beam welds are described subsequently.

Inclusions are minimal in vacuum electron beam welds because of the protection from oxidation and because many impurities are vaporized at high temperature and very low pressure. If foreign metal is inadvertently positioned in the path of the electron beam, it is added to the melt in a dispersed form. If this occurs, that portion of the weld should be gouged out and rewelded.

Porosity. Electron beam welds are particularly susceptible to the formation of porosity resulting from the release of gases dissolved or trapped in the base metal or from impurities remaining on the metal surface because of inadequate cleaning. To a limited extent, welding conditions, such as the power input, can influence the degree of porosity created. In the welding of carbon steels, for example, aluminum-killed or silicon-killed grades should be selected to minimize internal porosity. The porosity that can occur in an electron beam weld in a rimmed low-carbon steel is illustrated in Fig. 58(a). The material was AISI 1010 steel welded at a speed of 25 mm/s (1 in./s) at 150 kV · A with a current of 4.3 mA.

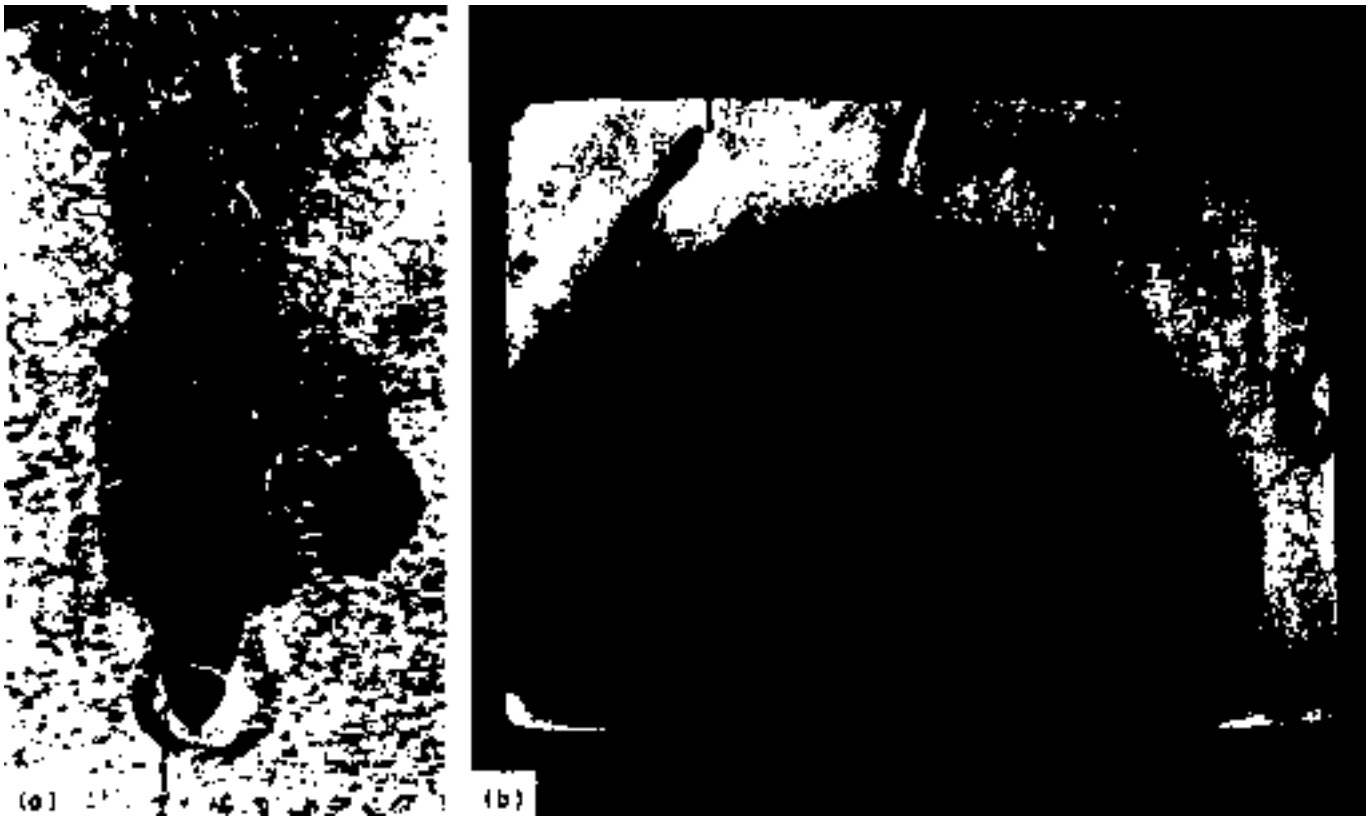


Fig. 58 Gas porosity in electron beam welds of low-carbon steel and titanium alloy. (a) Gas porosity in a weld in rimmed AISI 1010 steel. Etched with 5% nital. 30×. (b) Massive voids in weld centerline of 50 mm (2 in.) thick titanium alloy Ti-6Al-4V. 1.2×

An unusual variety of porosity called massive voids is sometimes found in electron beam welded joints that are thicker than 25 mm (1 in.). These cavities are similar to wormhole porosity in steel and may result from gas evolution from the base metal during welding or from the presence of an inverted V-shape gap in the joint before final welding. This inverted V-shape (joint edges in contact at weld-face surface, separated by perhaps 0.8 mm, or 0.03 in., at root surface) may be caused by poor fit-up or may be produced when partial-penetration tack welds or locking passes are used before full-penetration passes in 50 mm (2 in.) thick material. The appearance of massive voids in a centerline section of an electron beam weld in 50 mm (2 in.) thick titanium alloy Ti-6Al-4V is shown in Fig. 58(b). These defects are readily detected by x-ray inspection. Corrective measures to avoid their occurrence include more careful check of joint fit-up and the procedures cited previously for general porosity.

Incomplete fusion and inadequate penetration are caused either by insufficient power input or by misalignment between the electron beam and the seam. Two types of missed-seam defects include angular misalignment and positional misalignment, where a beam parallel to the seam is sufficiently displaced to one side to miss the seam. Such a defect may be eliminated by rewelding the area with increased power or a more accurate alignment.

A different type of incomplete-fusion defect, called spiking, is often associated with the irregularly shaped melting pattern at the root of a partial-penetration electron beam weld. Labels such as cold shuts, root porosity, and necklace porosity have been used for defects of this type. They are typically conical in shape and appear similar to linear porosity when examined radiographically in a direction parallel to the electron beam orientation. Therefore, they have the characteristics of solidification and shrinkage defects in the weld root at the tips of a portion of the spikes. An illustration of spiking that occurred in a 75 mm (3 in.) thick titanium alloy Ti-6Al-4V plate is shown in Fig. 59. A sound weld existed to a depth of 55 mm (2.15 in.) below the weld face.

Fig. 59 A spiking defect in a centerline section at the root of an electron beam weld. The partial-penetration weld was made in 75 mm (3 in.) thick Ti-6Al-4V titanium alloy plate. 1.6×

The peak-to-valley depths can vary over a wide range, depending on the welding conditions employed; these depths can sometimes amount to 25% of the depth of penetration. The primary procedure for preventing spiking is to defocus the beam so that beam crossover is near the surface instead of being buried 38 mm ($1\frac{1}{2}$ in.) deep in the metal. Also helpful is a reduction in travel speed to aid fusion of the root spike. Usually, the depth of a partial-penetration pass in titanium alloys cannot exceed 35 mm ($1\frac{3}{8}$ in.) without the occurrence of some spiking, although the exact depth is a function of the capability of the equipment.

Poor Weld Contours. Poor contours of the face of the weld in electron beam welds may be improved by using cosmetic weld passes. Relatively poor root-surface contours are inherent in the process at the beam exit, especially as base metal thickness is increased. Thin sections (less than 3.8 mm, or 0.15 in., thick) can be welded by partial penetration of the thickness, which entirely avoids the rough root side at the beam exit of a full-penetration weld.

Hot Cracks. Craters and crater cracks are typically produced in weld-stop areas on full-penetration passes. Use of runoff tabs or procedures for slope-down and pull-out of the beam are helpful in solving this problem.

Hot cracks in the weld may mean that the base metal alloy is not weldable (because of its metallurgical characteristics) or that a foreign metal has contaminated the melt. The use of filler metal (skin stock) to provide weld metal that is less brittle than the base metal (or metals) can prevent cracking; this is especially helpful in joining dissimilar metals if a filler metal is chosen that has a composition that is compatible with both members of the joint.

In welding some materials, such as the heat treatable aluminum alloys, the grain direction of the base metal relative to the weld can affect the susceptibility to cracking and should be critically reviewed when planning welding. Welds made or stressed in the short-transverse direction, for instance, often show a tendency to develop grain-boundary microcracks in the HAZ.

Bursts. A disk-shaped solidification imperfection that occurs at the electron beam weld centerline is called a burst. It is usually found near a change in mass, for example, near corners or near adjacent lugs or stiffeners. Occasionally, burst lengths to 25 mm (1 in.) have occurred. Bursts can be prevented by reducing the travel speed as the beam nears the change in mass or by welding on the runoff tabs to reduce the change in mass.

Failure Origins in Laser Beam Welds

Laser beam welding is used in the automotive, electronics, consumer products, and aerospace industries to join wire to wire, wire to sheet, tube to sheet, and for stud welding. Problems experienced in continuous-wave high-power laser beam welding are very similar to those in electron beam welding processes. Several discontinuities that may cause difficulties in pulsed laser beam welds are discussed subsequently.

Inclusions. Because the laser equipment itself does not contact the workpieces, the presence of any inclusions that could affect the quality of a laser beam weld would be due to foreign materials either on the surface or within the structure of the metal being joined. In most instances, any foreign substances on the workpiece surfaces are vaporized by the high heat of welding. Careful selection, cleaning, and preparation of materials are the best means of minimizing failures from inclusions of any type.

Porosity. Weld pulse-time schedules must be very carefully developed to avoid the formation of voids and porosity. Cleanness of workpiece surfaces and avoidance of low-boiling alloy constituents, such as zinc, are also important in minimizing voids.

Inadequate Penetration. Generally, the penetration that can be achieved with a laser beam increases with the square root of the available pulse time and of the thermal diffusivity of the metal being welded. However, at the same time, high surface reflectivity may be a cause of reduced energy input and inadequate penetration. Simply raising the energy level to compensate for high reflectivity may exceed the intensity limits and produce vaporization and holes. One technique to overcome this difficulty is to apply a suitable thin, absorptive coating, such as an anodize or black oxide, that decreases the surface reflection without affecting the metallurgy of the weld. Metal surfaces can sometimes be roughened to obtain lower reflectivity. An ideal technique when the workpiece shape permits, however, is to arrange the surfaces to be joined so that the weld is in a crevice or hole that can act as a black body, which is a perfect absorber. In this case, the high reflectivity is an advantage, guaranteeing that melting occurs only at the intended area.

Poor Weld Shape. The most favorable joint configuration for pulsed laser beam welding is probably the lap joint, although sound welds can also be made with other types of joints. In any configuration, the important factors are process variations, because these have a definite influence on weld quality. At the desired laser energy output, for example, a 10% change in output produces a 4% change in weld strength. Focusing distance is likewise important to weld strength. For a 50 mm (2 in.) focal length, variations in position of up to ±1.3 mm (±0.05 in.) still produce acceptable welds. Even though no separation between wires in a lap joint is recommended, some degree of separation can be tolerated.

Poor welds can also be caused by variations in alignment of the axis of the joint with respect to the axis of the laser beam; weld failures may thus be attributed to the use of fixtures that were inadequate to position and hold the workpieces with the necessary precision. The use of tandem laser beams can help prevent this problem.

A further factor influencing weld shape and weld strength is the environment in which the laser functions. The laser beam is modified by any condition that absorbs radiant energy, such as smoke, weld spatter, water vapor, or condensates on the ruby, lens, or mirror surfaces of the equipment. Clean, deionized cooling water through the optical cavity, dry nitrogen gas flow through the laser chamber, an exhaust fan for the welding chamber, and regular replacement of the glass slide over the objective lens all represent desirable procedures to maintain constant laser beam input energy to the weld joint.

Base Metal Cracking. When laser beam energy is put into the weld at too high a rate, the surface may not only suffer from vaporization, but base metal cracking may also occur from thermal shock. This relates to the limitations of thickness that can be welded, which in turn depend on the thermal properties of the metal, its vaporization characteristics, and reflectivity. Generally, longer pulse times are needed to weld the lower-thermal-conductivity metals, and higher pulse-energy densities are necessary in welding high-reflectivity and high-melting-point alloys.

Failure Origins in High-Frequency Induction Welds

Pipes for high-pressure service, pipes for well casings, and structural beams have been welded by the high-frequency induction process. Hook cracks are a type of discontinuity associated with these welds. They are difficult to detect and require extremely careful inspection for assurance that they are not present. Delamination of the pipe wall just outside of the narrow weld zone can be the cause of hook cracks. The slight upset of the metal at the weld acts to turn the ends of these usually very small subsurface cracks in the direction of the metal flow (Fig. 60). Because the metal flow of the upset is toward both surfaces, the “hook” of the hook crack may turn toward either the outside or the inside surface, depending on which is closer to the delamination and in

which direction the greatest amount of upset occurs. Generally, upset is greater toward the outer surface, thus causing hooking to turn outward. Use of the cleanest steel available and the lowest possible heat input greatly reduces the possibility of hook cracks.

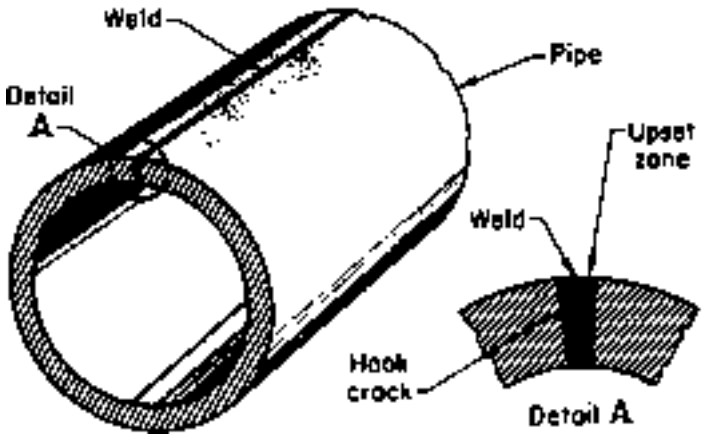


Fig. 60 Schematic of a hook crack in a pipe caused by pipe-wall delamination after high-frequency welding. The "hook" has turned outward to follow the direction of metal flow in the outer portion of the upset weld zone.

Depending on the materials and welding parameters used, high-frequency weldments are susceptible to other forms of cracking. In one case, weld-centerline cracking was encountered during flaring and flattening tests of high-frequency induction-welded type 409 stainless steel tubing for automobile exhaust systems. Simulation of the welding process and metallographic examinations of simulated and actual welds from several different heat compositions demonstrated that the heat composition, which controlled the type of phase transformations occurring during cooling, determined whether weld cracking occurred. Compositions having a high nickel equivalent (Ref 29) or low chromium equivalent contained martensite at and adjacent to the weld, which strengthened the welds and forced plastic straining to occur in the ductile sheet outside the weld during testing. On the other hand, compositions having a lower nickel equivalent or higher chromium equivalent transformed to weaker ferrite at the weld centerline, which permitted bondline defects to initiate fracture.

Reference cited in this section

29. J. LeFevre, R. Tricot, and R. Castro, Nouveaux Aciers Inoxydables a 12% Chrome (New Stainless Steels with 12% Chromium), *Rev. Métall.*, Vol 70, April 1973, p 259–268

Acknowledgments

The following authors and contributors of previous editions provided substantial content used in this article. Appreciation is expressed to: Tarsem Jutla, Caterpillar Inc.; Alan G. Glover, Welding Institute of Canada; Daniel Hauser, Edison Welding Institute; Edward A. Metzbower, Naval Research Laboratory; O.W. Blodgett, Lincoln Electric; the ASM Committee on Weld Discontinuities (for *Welding, Brazing, and Soldering,* Volume 6 of *Metals Handbook* (9th ed.), 1983), and the ASM Committee on Failure Analysis of Weldments (for *Failure Analysis and Prevention,* Volume 10 of *Metals Handbook* (8th ed.), 1975).

References

1. *W. Moyer,* Mechanical Testing of Welded Joints, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* 2000, p 845–852

2. E.R. Barker, *Brittle Behavior of Engineering Structures,* John Wiley & Sons, 1957

3. J. Signes, R.G. Baker, J.D. Harrison, and F.M. Burdekin, Factors Affecting the Fatigue Strength of Welded High Strength Steels, *Br. Weld. J.*, Vol 14 (No. 3), 1967, p 108–116

4. S.J. Maddox, "Fitness-for-Purpose Assessment of Misalignment in Transverse Butt Welds Subject to Fatigue Loading," Document XIII-1180–85, International Institute of Welding, 1985

5. S.J. Maddox, Recent Advances in the Fatigue Assessment of Weld Imperfections, *Weld. J.*, July 1993, p 42–51

6. J.R. Gordon, "A Review of Currently Available Fracture Analysis Methods and Their Applicability to Welded Structures," paper presented at Sixth Annual North American Welding Research Conference, Edison Welding Institute and American Welding Society, Oct 1990

7. "Code of Practice for Fatigue Design and Assessment of Steel Structures," BS 7608, British Standards Institution, 1993

8. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," BS PD 6493:1991, British Standards Institution, Aug 1991 (superseded by BS 7910)

9. "Guide on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," BS 7910, British Standards Institute, 1999

10. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6-Rev. 3, Central Electricity Generating Board (United Kingdom), 1987

11. V. Kumar, M.D. German, and C.F. Shih, "An Engineering Approach for Elastic-Plastic Fracture Analysis," Technical Report NP 1931, Electric Power Research Institute, July 1981

12. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, STP 803, Vol II, ASTM, 1983, p 206

13. "The Method of Assessment for Defects in Fusion Welded Joints with Respect to Brittle Fracture," WES 2805, Japan Welding Engineering Society

14. "IIW Guidance on Assessment of the Fitness for Purpose of Welded Structures," Draft for Development, IIW/IIS-SST-1157-90, International Institute of Welding, 1990

15. "Recommended Practice for Fitness-for-Service of Refinery Equipment," API 579, American Petroleum Institute

16. A.R. Cuiffreda and R.D. Merrick, Brittle Fracture of a Pressure Vessel: Study Results and Recommendations, *Proceedings of the American Petroleum Institute*, Vol 62, 1983

17. O.W. Blodgett, Joint Design and Preparation, *Welding, Brazing, and Soldering*, Vol 6, *Metals Handbook* (9th ed.), 1983, p 60–72

18. S.J. Maddox, "Scale Effect in Fatigue of Fillet-Welded Aluminum Alloys," presented at Sixth Int. Conf. Aluminum Weldments (Cleveland), American Welding Society, 1995

19. V.I. Makhnenko and R.Y. Mosenkis, Calculating the Coefficient of Concentration of Stresses in Welded Joints with Butt and Fillet Welds, *Automat. Weld.*, Aug 1985

20. K. Iida, Y. Kho, J. Fukakura, M. Nihei, T. Iwadate, and M. Nagai, "Bending Fatigue Strength of Butt Welded Plate with Uranami Bead," IIW/IIS-XIII-1202–86, International Institute of Welding, 1986

21. “Welded Steel Construction (Metal Arc Welding),” CSA W59, Canadian Standards Association, 1989

22. “Welding Pipelines and Related Facilities,” API STD 1104, 19th ed., American Petroleum Institute, 1999

23. *Weld Integrity and Performance,* ASM International, 1997

24. G. Screm et al., The effects of Weld Defects in Reducing the Fatigue Strength of Welded Joints, *Alluminio,* Vol 44 (No. 3), 1975

25. B.A. Graville, *The Principles of Cold Cracking Control in Welds,* Dominion Bridge, Hanover, 1975

26. F. Coe, *Welding Steels Without Hydrogen Cracking,* The Welding Institute, Abington, Cambridge, U.K., 1973

27. R.D. Stout, *Weldability of Steels,* 3rd ed., Welding Research Council, 1978

28. D.J. Carney, Gases in Liquid Iron and Steel, *Gases in Metals,* American Society for Metals, 1953, p 69–118

29. J. LeFevre, R. Tricot, and R. Castro, Nouveaux Aciers Inoxydables a 12% Chrome (New Stainless Steels with 12% Chromium), *Rev. Métall.,* Vol 70, April 1973, p 259–268

Introduction

HEAT TREATING—of all the various steel processing methods—has the greatest overall impact on control of microstructure, properties, residual stresses, and dimensional control. This article provides an overview of the effects of various material and process-related parameters on residual stress, distortion control, cracking, and microstructure/property relationships as they relate to various types of failure. The subjects that are discussed include:

- Phase transformations during heat treating
- Metallurgical sources of stress and distortion during heating and cooling
- Effect of materials and quench-process design on distortion and cracking
- Quenchant selection
- Effect of cooling characteristics on residual stress and distortion
- Methods of minimizing distortion
- Tempering
- Effect of the heat treatment process on microstructure/property-related failures, such as cracking

Failures Related to Heat Treating Operations

G.E. Totten, G.E. Totten & Associates, LLC; M. Narazaki, Utsunomiya University (Japan); R.R. Blackwood and L.M. Jarvis, Tenaxol Inc.

Phase Transformation during Heating and Cooling

Steel Transformation

Properties such as hardness, strength, ductility, and toughness are dependent on the microstructural products that are present in steel. The first step in the transformation process is to heat the steel to its austenitizing temperature. The steel is then cooled rapidly to avoid the formation of pearlite, which is a relatively soft transformation product; to maximize the formation of martensite, a relatively hard transformation product; and to achieve the desired as-quenched hardness.

The most common transformational products that may be formed from austenite in quench-hardenable steels are, in order of formation with decreasing cooling rate: martensite, bainite, pearlite, ferrite, and cementite. The formation of these products and the proportions of each are dependent on the time and temperature cooling history of the particular alloy and the elemental composition of that alloy. The transformation products formed are typically illustrated with the use of transformation diagrams, which show the temperature-time dependence of the microstructure formation process for the alloy being studied. Two of the most commonly used transformation diagrams are time-temperature-transformation (TTT) and continuous cooling transformation (CCT) diagrams.

Time-temperature-transformation diagrams, also called isothermal transformation diagrams, are developed by heating small samples of steel to the temperature where austenite transformation structure is completely formed, that is, the austenitizing temperature (T_A), and then rapidly cooling to a temperature intermediate between the austenitizing and the martensite start (M_s) temperature and then holding for a fixed period of time until the transformation is complete, at which point the transformation products are determined. This is done repeatedly until a TTT diagram is constructed, such as that shown for an unalloyed steel (American Iron and Steel Institute, or AISI, 1045) in Fig. 1. Time-temperature-transformation diagrams can only be read along the isotherms. The temperature A_1 is where transformation to austenite begins, and temperature A_3 is where the transformation to austenite is complete.

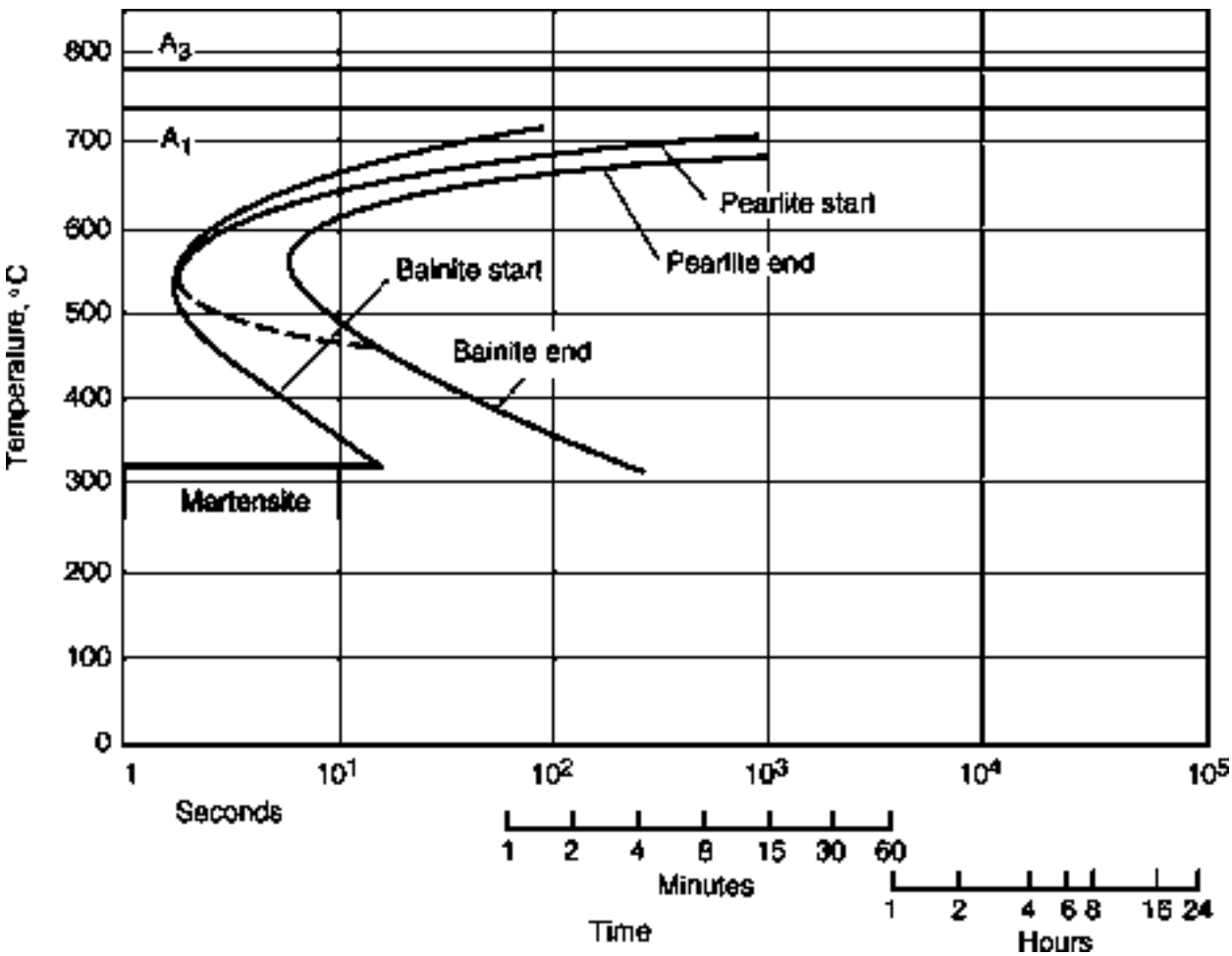


Fig. 1 Time-temperature-transformation diagram of an unalloyed steel containing 0.45% C. Austenitization temperature, 880 °C (1620 °F). The temperature A_1 is where transformation to austenite begins, and temperature A_3 is where the transformation to austenite is complete. Courtesy of Verlag Stahlessen mbH Dusseldorf

Continuous Cooling Transformation Diagrams. Alternatively, samples of a given steel may be continuously cooled at different specified rates, and the proportion of transformation products formed after cooling to various temperatures intermediate between the austenitizing temperature and the M_s may be determined in order to construct a CCT diagram, such as the one shown for an unalloyed carbon steel (AISI 1045) in Fig. 2. Continuous cooling transformation curves provide data on the temperatures for each phase transformation, the amount of transformation product obtained for a given cooling rate with time, and the cooling rate necessary to obtain martensite. The critical cooling rate is dictated by the time required to avoid formation of pearlite for the particular steel being quenched. As a general rule, a quenchant must produce a cooling rate equivalent to, or faster than, that indicated by the "nose" of the pearlite transformation curve in order to maximize martensite transformation product. Continuous cooling transformation diagrams can only be read along the curves of different cooling rates.

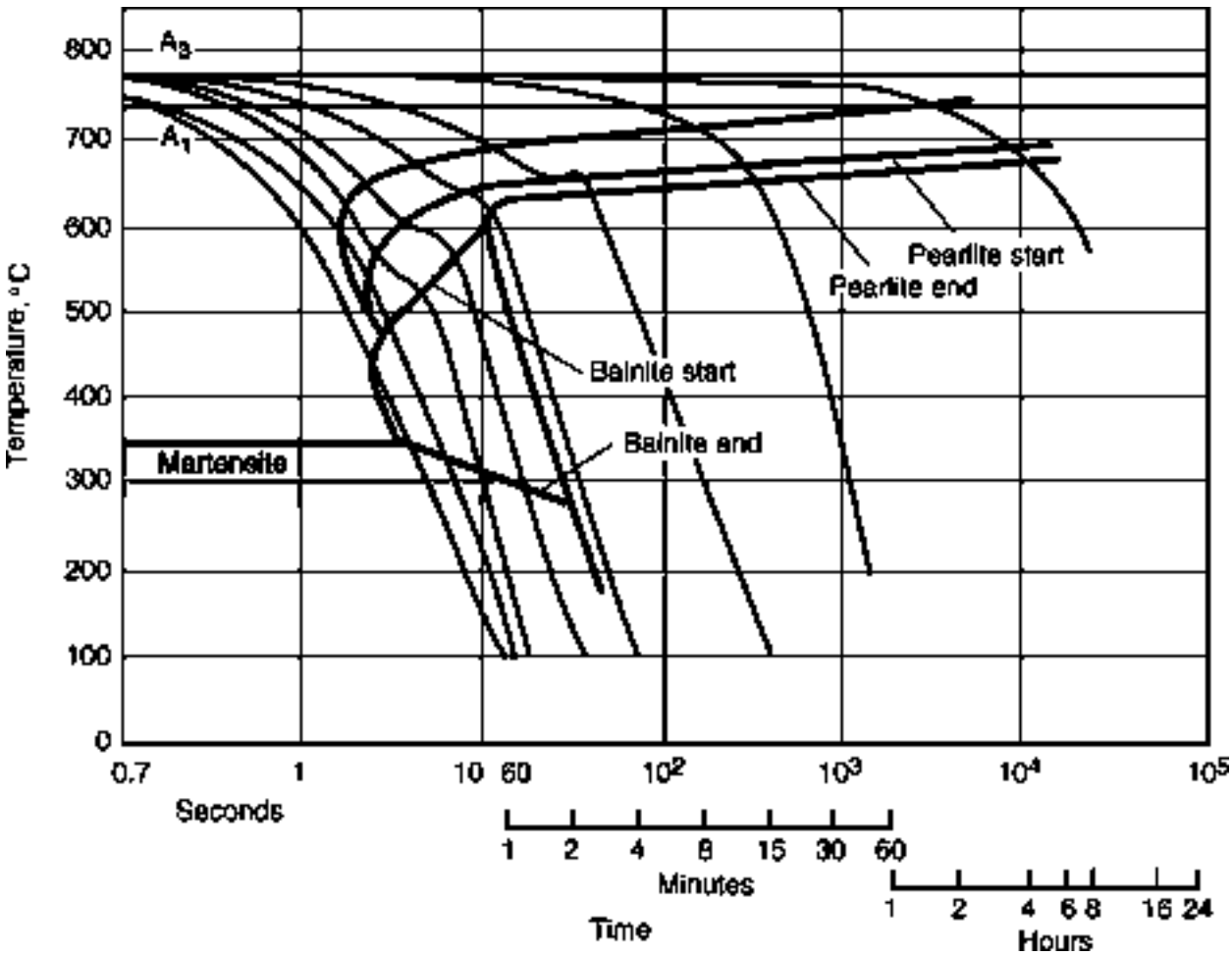


Fig. 2 Continuous cooling transformation diagram of an unalloyed steel containing 0.45% C. Austenitization temperature, 880 °C (1620 °F). Courtesy of Verlag Stahlessen mbH Dusseldorf

Caution: Although it is becoming increasingly common to see cooling curves (temperature-time profiles) for different quenchants, such as oil, water, air, and others, superimposed on either TTT or CCT diagrams, this is not a rigorously correct practice, and various errors are introduced into such analysis due to the inherently different kinetics of cooling used to obtain the TTT or CCT diagrams (described previously) versus the quenchants being represented. If a cooling curve is to be superimposed on a transformation diagram, a CCT, not a TTT, diagram should be used.

Metallurgical Crystal Structure

When steel is slowly cooled, it undergoes a crystal structure (size) change as it transforms from a less densely packed austenite (face-centered cubic, or fcc) to a more densely packed body-centered cubic (bcc) structure of ferrite. At faster cooling rates, the formation of ferrite is suppressed, and martensite, which is an even less

densely packed body-centered tetragonal (bct) structure than austenite, is formed. Illustrations of these crystal structures are provided in Fig. 3. This results in a volumetric expansion at the M_s temperature, as shown in Fig. 4.

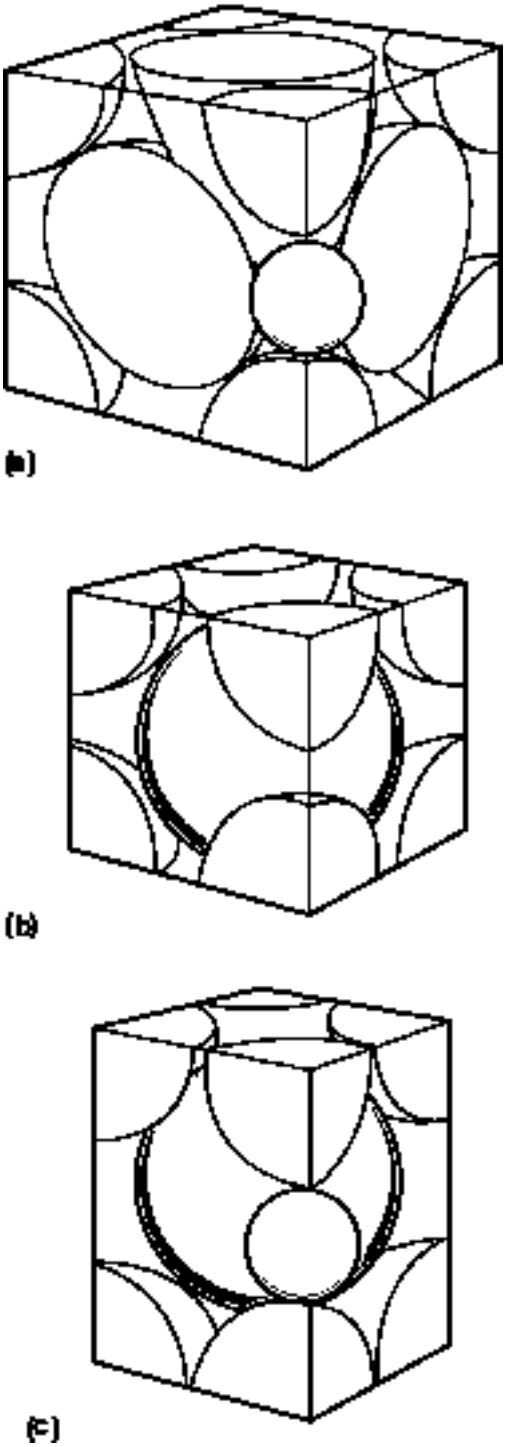


Fig. 3 Crystal structures. (a) Austenite (fcc). (b) Ferrite (bcc). (c) Martensite (bct)

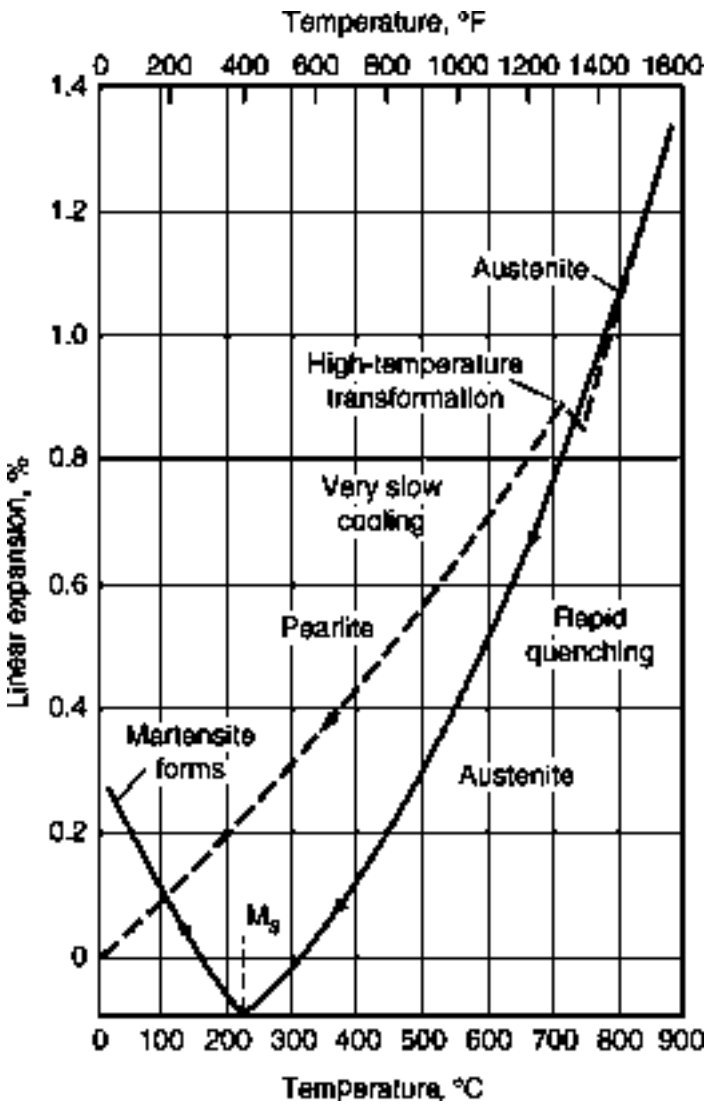


Fig. 4 Steel expansion and contraction on heating and cooling

Figure 5 shows that the crystal lattice of austenite expands with increasing carbon content (Ref 1). It has been reported that, typically, when a carbide-ferrite mixture is converted to martensite, the resulting expansion due to increasing carbon content is approximately 0.002 in./in. at 0.25% C and 0.007 in./in. at 1.2% C (Ref 1). The fractional increase in size when austenite is converted to martensite is approximately 0.014 in./in. for eutectoid compositions. This illustrates the effect of carbon structure and steel transformation on residual stresses and distortion leading to dimensional changes.

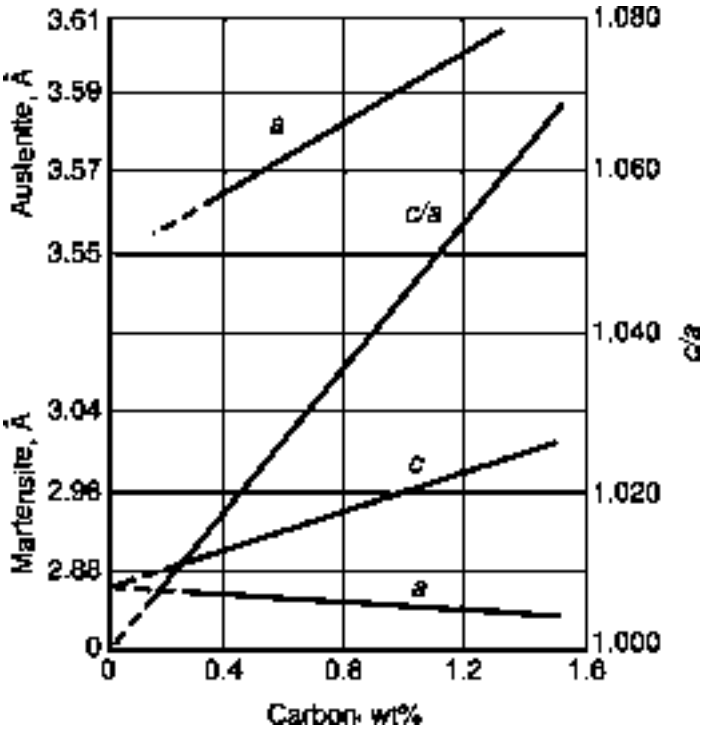


Fig. 5 Carbon content versus lattice parameters of (retained) austenite and martensite at room temperature. *a* at the top of the graph is the lattice parameter of fcc austenite. *a* and *c* in the lower half of the graph are the lattice parameters for tetragonal martensite. The ratio of *c*/*a* for martensite as a function of carbon content is also given.

Estimation of Volumetric Change due to Steel Transformation

In the previous discussion, it has been shown that there are various microstructures possible on heating and cooling of steel, and that the potential microstructural transformations that are possible for a given steel are illustrated by their CCT or TTT diagrams. Furthermore, dimensional changes are possible, depending on the carbon content and microstructural transformation product formed. Table 1 summarizes the atomic volumes of different microstructural components as a function of carbon content (Ref 2). Table 2 provides an estimate of volumetric changes as a function of carbon content for different metallurgical transformations (Ref 3, 4).

Table 1 Atomic volumes of selected microstructural constituents of ferrous alloys

Phase	Apparent atomic volume, $Å^3$
Ferrite	11.789
Cementite	12.769
Ferrite + carbides	11.786 + 0.163 C[(a)]
Pearlite	11.916
Austenite	11.401 + 0.329 C[(a)]
Martensite	11.789 + 0.370 C[(a)]

(a) C = % carbon

Table 2 Size changes during hardening of carbon tool steels

Reaction	Volume change, %	Dimensional change, in./in.
Spheroidite → austenite	-4.64 + 2.21 (%C)	0.0155 + 0.0074 (%C)
Austenite → martensite	4.64 - 0.53 (%C)	0.0155 - 0.00118 (%C)
Spheroidite → martensite	1.68 (%C)	0.0056 (%C)
Austenite → lower bainite [(a)]	4.64 - 1.43 (%C)	0.0156 - 0.0048 (%C)
Spheroidite → lower bainite [(a)]	0.78 (%C)	0.0026 (%C)

Reaction	Volume change, %	Dimensional change, in./in.
Austenite → aggregate of ferrite and cementite [(b)]	4.64 - 2.21 (%C)	0.0155 - 0.0074 (%C)
Spheroidite → aggregate of ferrite and cementite [(b)]	0	0

(a) Lower bainite is assumed to be a mixture of ferrite and epsilon carbide.
(b) Upper bainite and pearlite are assumed to be mixtures of ferrite and cementite.

Volumetric expansion occurring as a result of quenching can be estimated from (Ref 5):

$$\Delta V/V \times 100 = (100 - V_c - V_a) \times 1.68C + V_a(-4.64 + 2.21C) \qquad \text{(Eq 1)}$$

where $(\Delta V/V) \times 100$ equals the percentage change in volume, V_c equals the percentage by volume of undissolved cementite, $(100 - V_c - V_a)$ equals the percentage by volume of martensite, V_a equals the percentage by volume of austenite, and C equals the percentage by weight of carbon dissolved in austenite and martensite.

If the value of $(\Delta V/V)$ is known or can be computed, then internal stresses that are developed in a part due to temperature differences (ΔT) arising from either one-dimensional heating or cooling can be estimated from (Ref 6):

$$\sigma = E \cdot \varepsilon = E \cdot \frac{1}{3}(\Delta V/V) = E \cdot \alpha \cdot \Delta T \qquad \text{(Eq 2)}$$

where σ is stress, ε is strain, E (modulus of elasticity) = 2×10^5 N/mm^2, and α (coefficient of thermal expansion) = 1.2×10^{-5}. Relative volume changes due to phase transformation are illustrated in Fig. 6 (Ref 6).

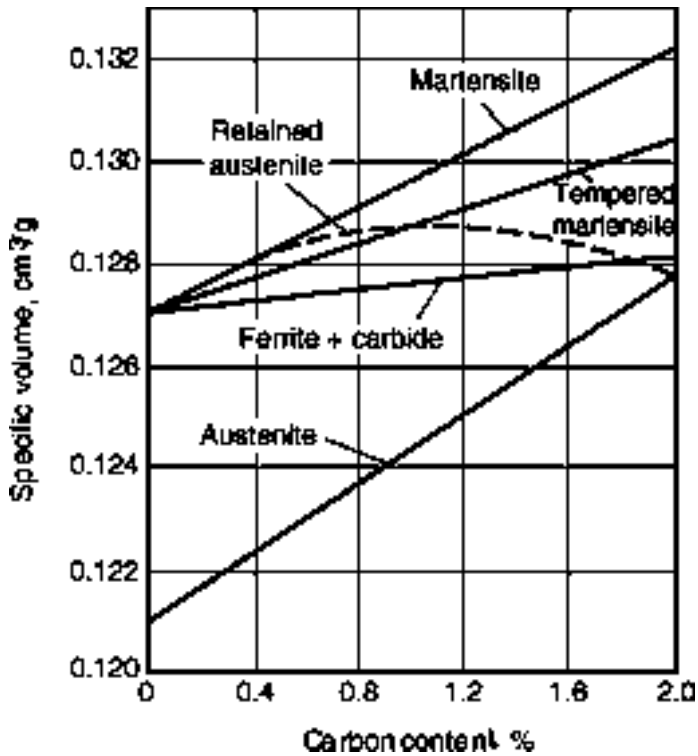


Fig. 6 Specific volume ($\Delta V/V$) of carbon steels relative to room temperature. Tempered martensite, <200 °C (390 °F)

Figure 7 shows that stresses, such as an increase in hydrostatic pressure, accelerate phase transformations (Ref 7). This occurs whether the stresses are tensile or compressive and results in accelerated austenite decomposition and increased M_s temperature. The strain of this process is often estimated as being equal to the volumetric expansion divided by 3 (Ref 8). Transformation plasticity is a process whereby a stress affects linear strain. This is illustrated in Fig. 8, where the effect of applied stress on martensitic transformation strain is shown (Ref 8). Generally, the M_s temperature is increased by tensile stress and decreased by hydrostatic pressure. Figure 9 shows that stress exhibits very large effects on the start and stop times for pearlitic transformation (Ref 7).

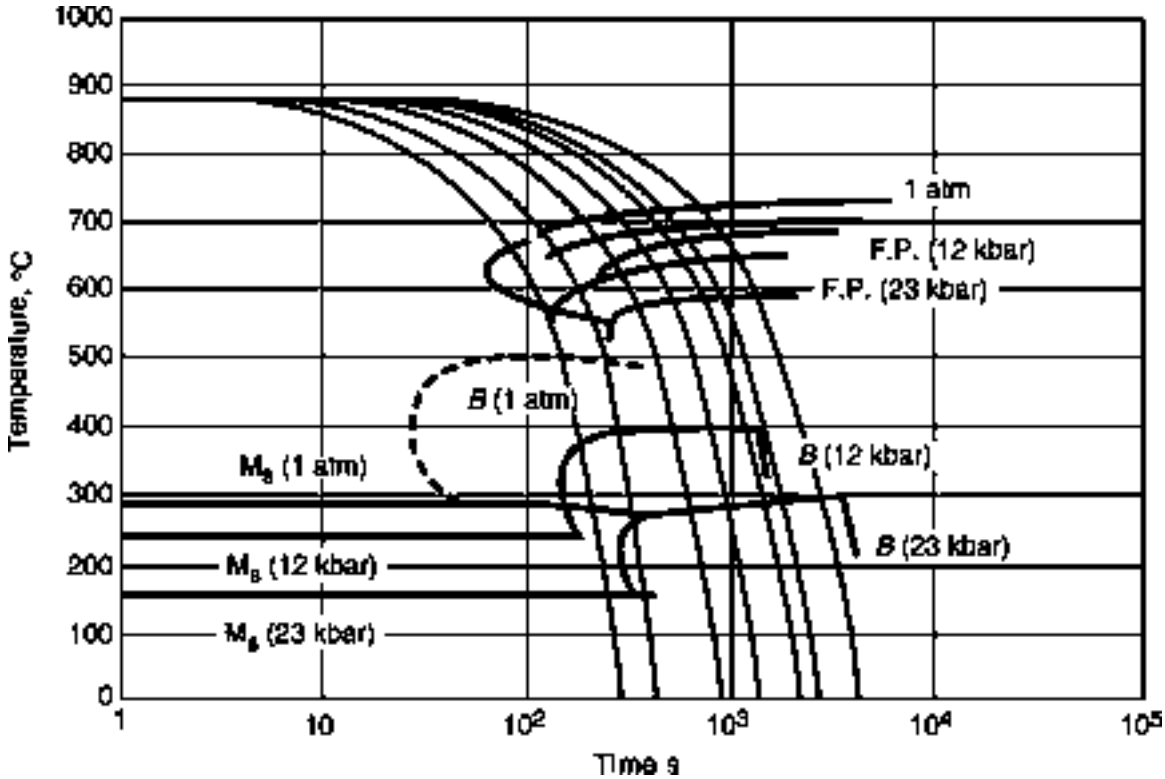


Fig. 7 Effect of hydrostatic pressure on the transformation kinetics of 50CV4 steel. *B*, bainite; F. P., ferrite-pearlite transformation; M_s, martensite start temperature. Source: Ref 7

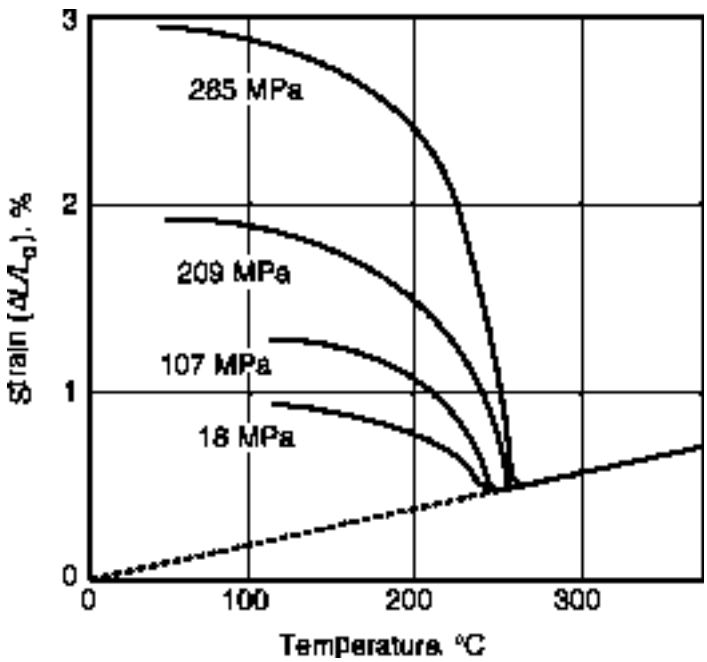


Fig. 8 Dilatometer curves for a 0.6% carbon steel (60NCD11) for different applied stresses

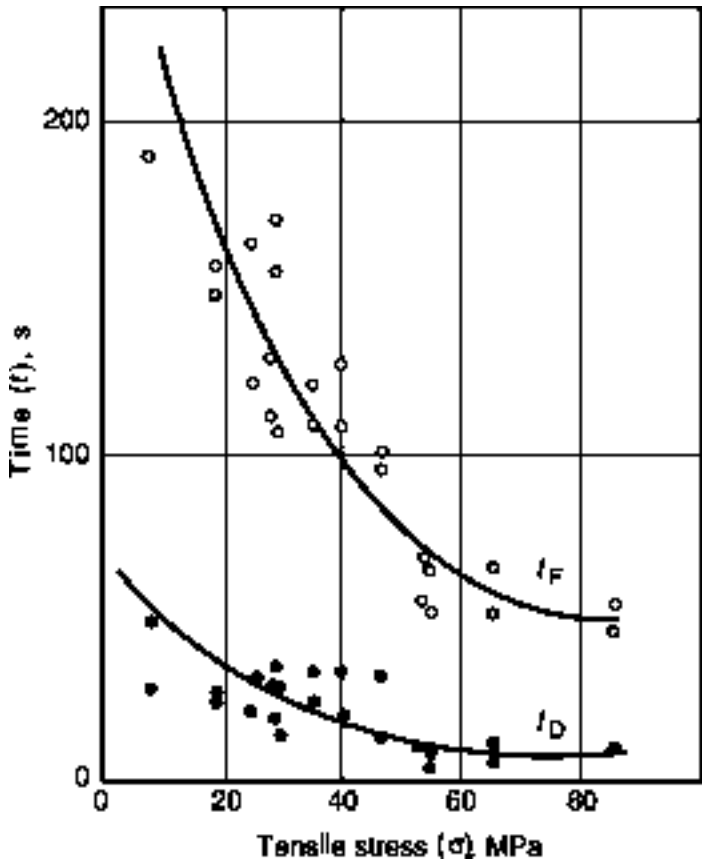


Fig. 9 Effect of tensile stress on pearlite transformation starting and ending times. Isothermal transformation at 673°C (1243 °F), eutectoid steel. The t_D and t_F times are transformation starting and ending times, respectively.

Cooling and Steel Metallurgical Transformation

Cooling without Transformation (Ref 6, 7). If steel is cooled sufficiently fast, cooling is not accompanied with microstructural transformation changes. Under these conditions, the surface of the component cooled much more quickly than the core at first, as illustrated in Fig. 10 (Ref 9). At this point, the specific volume in the core is greater than that of the surface, and the reduction in volume at the surface (due to lower temperature) is resisted by the greater volume in the core, resulting in the surface being in tension and the core in compression.

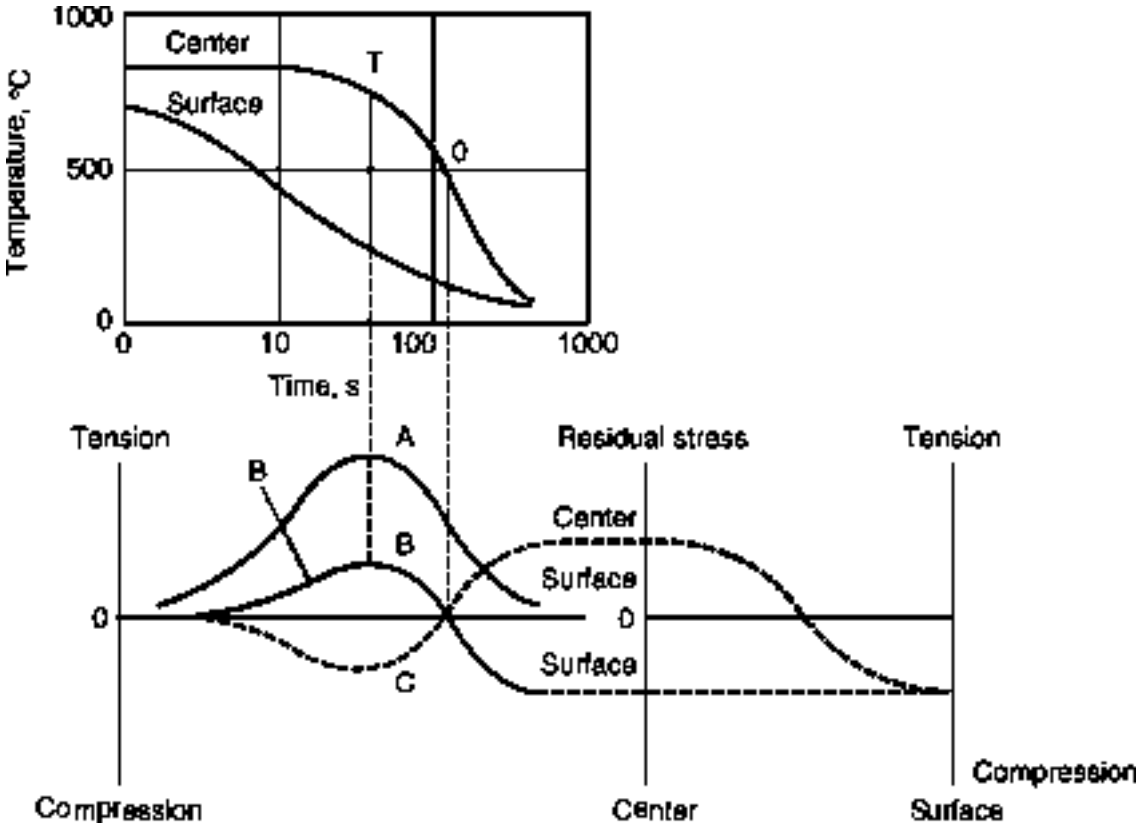


Fig. 10 Development of thermal stresses within steel on cooling. T, time instant at maximum temperature difference; 0, time instant of stress reversal; curve A, stress variation at the surface under elastic conditions. B and C are actual thermal stress variations at the surface and the core under elastic-plastic conditions.

At the point of maximum temperature difference between the surface and the core (point T), the core cools (shrinks) more quickly than the surface, leading to an elastic dimensional reduction of the surface until a point of stress conversion is obtained, at which point the surface is in compression relative to the core. After the cooling processes have been completed, the residual stress distribution between the surface and the core, shown at the bottom right in Fig. 10 (Ref 9), is obtained. If the surface stresses exceed the hot yield strength of the material, it plastically deforms, resulting in thermally induced dimensional changes.

Cooling with Transformation (Ref 6, 7). When steels that may undergo transformational changes are quenched, the possibility of the formation of both thermal and transformational stresses must be considered. Figure 11 illustrates three different examples of this process (Ref 8).

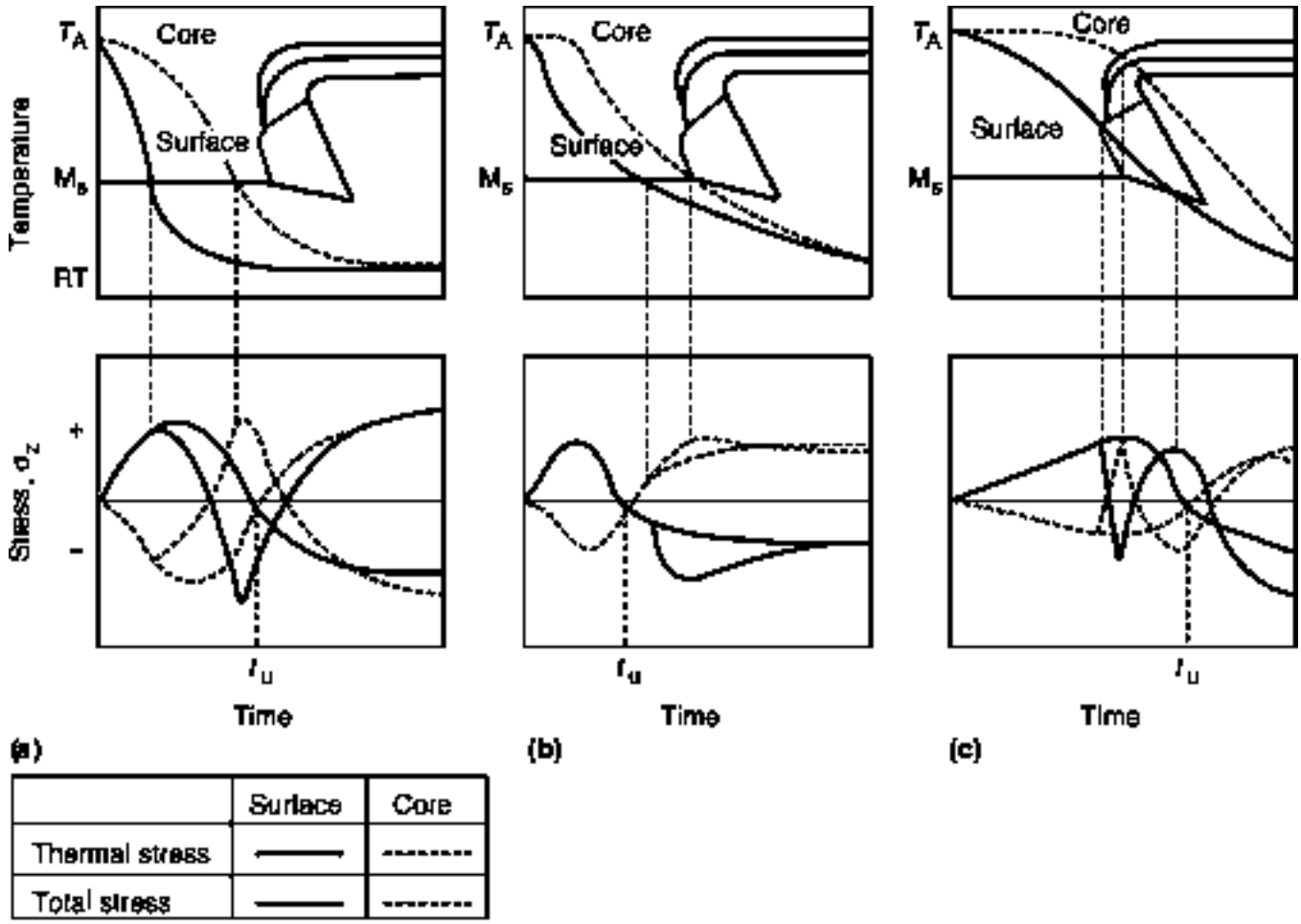


Fig. 11 Comparison of thermal and transformational stresses for three different quenching conditions. See text for details. t_u, time instant of stress reversal

Example 1, illustrated in Fig. 11(a), occurs when phase transformation of both the surface and the core occurs before the thermal stresses change sign. Above the M_s transformation temperature, the stresses that are formed are thermal. On further cooling, the stresses in the core exceed the yield strength, and plastic deformation (elongation) occurs. Subsequent martensitic transformation at the core provides a substantial stress component, due to volumetric increases from martensitic phase transformation. This causes the core to be in compression and the surface to be in tension simultaneously.

Example 2 (Fig. 11b) illustrates the case that begins after the thermal stresses change sign. The transformation-induced volume increase of the surface layer adds to the compressive stresses at the surface. Because the stresses are balanced, there is a corresponding increase in the tensile stresses in the core.

Example 3 is a case where, although the transformation of the core starts later, it finishes before the surface (Fig. 11c). During the cooling process, the sign of the stresses changes three times. There are important consequences whether the core transforms before or after the stress reversal, because thermal stresses may be counteracted, and tensile surface residual stresses may result when the core transforms after the surface and before the stress reversal (Ref 8).

This is illustrated, in Fig. 12. If the steel transformation occurs before the thermal stress maximum (sequence 4 in Fig. 12), the ferrite/pearlite structure of a cylindrical test specimen is distorted into a barrel shape (Ref 6). If the transformation occurs after the thermal stress maximum (sequence 1 in Fig. 12), the austenite is pressed into a barrel shape, followed by a volumetric increase due to martensitic transformation. This results in high tensile residual stresses on the surface. If the steel transformation occurs simultaneously with the maximum thermal stresses, transformation in the core occurs prior to surface transformation (sequence 3 in Fig. 12), and a barrel shape appears with high compressive stresses on the surface. If the surface is transformed prior to the core (sequence 2 in Fig. 12), the transformational stresses decrease or possibly even reverse the thermal stresses. If this occurs, a spool shape is formed.

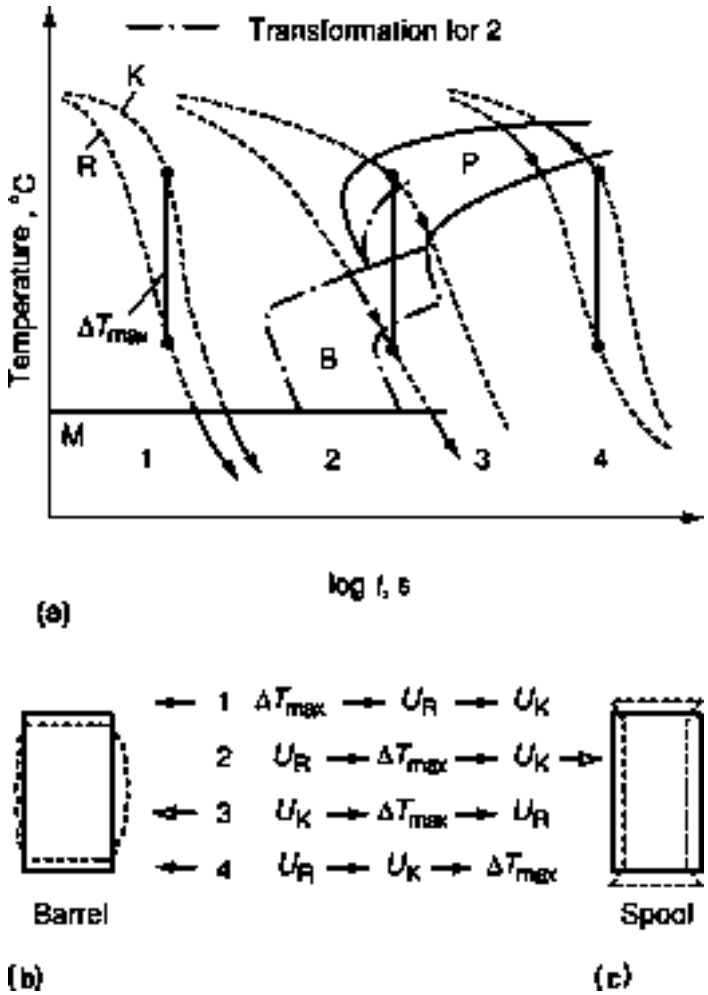


Fig. 12 Size change due to thermal changes and phase transformation. K, core; R, surface

These data show that the position of cooling curves for both the surface and the core for the quenching process must be considered with the appropriate TTT curve for the steel of interest, and that there are numerous mixtures of thermal and transformational stresses possible. Table 3 provides a summary of these processes (Ref 6).

Table 3 Size change and residual stress caused by heat treatment of prismatic parts

	(ΔT)[a]	**($\Delta V/V$)[b]**	**Dimensional dependency**	**($\Delta\varepsilon$)[c]**	**(σE)[d]**	**Example**
Change of microstructure	Small	≠0	$\varepsilon_1 = \varepsilon_{11} = \varepsilon_{111}$	0	0	Tempering, precipitation hardening, isotropic material
Thermal stress	Big	0	$\varepsilon_1 < \varepsilon_{11} < \varepsilon_{111}$	>0	<0	Quenching of austenitic steels
Thermal and transformation stress						
Transformation R + K ***after*** thermal stress maximum	Small	>0	$\varepsilon_1 \sim \varepsilon_{11} \sim \varepsilon_{111}$	0	>>0	In air; full hardness penetration
	Big	>>0	$\varepsilon_1 < \varepsilon_{11} < \varepsilon_{111}$	>0	>>0	In water; full hardness penetration
Transformation R ***before*** K ***after*** thermal stress maximum	Big	>0	$\varepsilon_1 > \varepsilon_{11} > \varepsilon_{111}$	<0	>0	Medium hardness penetration

	$(\Delta T)^{(a)}$	$(\Delta V/V)^{(b)}$	**Dimensional dependency**	$(\Delta\varepsilon)^{(c)}$	$(\sigma E)^{(d)}$	**Example**
Transformation R ***before*** K ***after*** thermal stress maximum	Big	≥0	$\varepsilon_1 << \varepsilon_{11} << \varepsilon_{111}$	>>0	<<0	Shallow hardness penetration (shell hardening)
Transformation ***before*** R + K thermal stress maximum	Small	~0	$\varepsilon_1 \approx \varepsilon_{11} \approx \varepsilon_{111}$	0	~0	Normalizing, quenching without hardening
	Big	~0	$\varepsilon_1 < \varepsilon_{11} < \varepsilon_{111}$	>0	<0	

R, surface; K, core.
(a) $\Delta T = (T_K - T_R)$.
(b) $\Delta V/V$, relative volume change.
(c) Convexity $\Delta\varepsilon = \varepsilon_k - \varepsilon_m$; ε_m, ε_k, dimensional change in middle of the plane.
(d) σ_E, residual stress at the surface (remaining stress)

In some steels, those with higher carbon content and alloy steels, the M_f temperature is below 32 °F (0 °C), which means that it is likely that at the conclusion of the heat treating process there is as much as 5 to 15% of austenite remaining (Ref 10). The amount of retained austenite exhibits significant effects on the magnitude of compressive stresses formed and, ultimately, on dimensional changes to be expected. Some of the factors affecting retaining austenite formation include chemical composition (which dictates the M_s temperature), quenching temperature, quenching cooling rates, austenitizing temperature, grain size, and tempering.

References cited in this section

1. S. Mocarski, Carburizing and Its Control, Part I: Basic Considerations, *Ind. Heat.,* Vol 41 (No. 5), 1974, p 58–70

2. A. Bavaro, Heat Treatments and Deformation, *Trait. Therm.,* Vol 240, 1990, p 37–41

3. F. Legat, Why Does Steel Crack during Quenching, *Kovine Zlitine Technol.,* Vol 32 (No. 3–4), 1998, p 273–276

4. R.W. Bohl, Difficulties and Imperfections Associated with Heat Treated Steel, Lesson 13, Heat Treatment of Steel, Materials Engineering Institute Course 10, ASM International, 1978

5. K.E. Thelning, *Steel and Its Heat Treatment,* Butterworths, London, U.K., 1985

6. H. Berns, "Distortion and Crack Formation by Heat Treatment of Tools," *Radex Rundsch.,* Vol 1, 1989, p 40–57

7. G. Beck, Internal Stresses Associated with Heat Treatment of Alloys, *Mém. Etud. Sci. Rev. Métall.,* June 1985, p 269–282

8. T. Ericsson and B. Hildenwall, Thermal and Transformational Stresses, *Sagamore Army Mater. Res. Conf., 28th Proc.,* 1981, p 19–38

9. C. Jensen, Dimensional Changes during Heat Treatment of Hot Work Steel, *Ind. Heat.,* March 1993, p 33–38

10. P. Mayr, "Dimensional Alteration of Parts due to Heat Treatment," *Residual Stresses in Science and Technology,* Vol 1, Garmisch-Partenkirschen, FRG, 1986, DGM Informationsgesellschaft mbH, Adenauerallee 21, D-6370, Oberursel, FRG 1987, p 57–77

Tempering

Steel parts are often tempered by reheating after quench hardening to obtain specific values of mechanical properties. Tempering of steel increases ductility and toughness of quench-hardened steel and also relieves quench stresses and ensures dimensional stability. The tempering process involves heating hardened steel to some temperature below the eutectoid temperature for the purposes of decreasing hardness and increasing toughness. In general, the tempering process is divided into four stages, which are summarized in Table 4 (Ref 9). These include (Ref 11, 12):

1. Tempering of martensite structure
2. Transformation of retained austenite to martensite
3. Tempering of the decomposition products of martensite and at temperatures above 480 °C (900 °F)
4. Decomposition of retained austenite to martensite

Table 4 Metallurgical reactions occurring at various temperature ranges and related physical changes of steel during tempering

Stage	Temperature range		Metallurgical reaction	Expansion/contraction
	°C	°F		
1	0–200	32–392	Precipitation of ε-carbide. Loss of tetragonality	Contraction
2	200–300	392–572	Decomposition of retained austenite	Expansion
3	230–350	446–662	ε-carbides decompose to cementite	Contraction
4	350–700	662–1292	Precipitation of alloy carbides. Grain coarsening	Expansion

Source: Ref 9

Figure 13 illustrates the effect of microstructural variation during tempering on the volume changes occurring during the tempering of hardened steel (Ref 9).

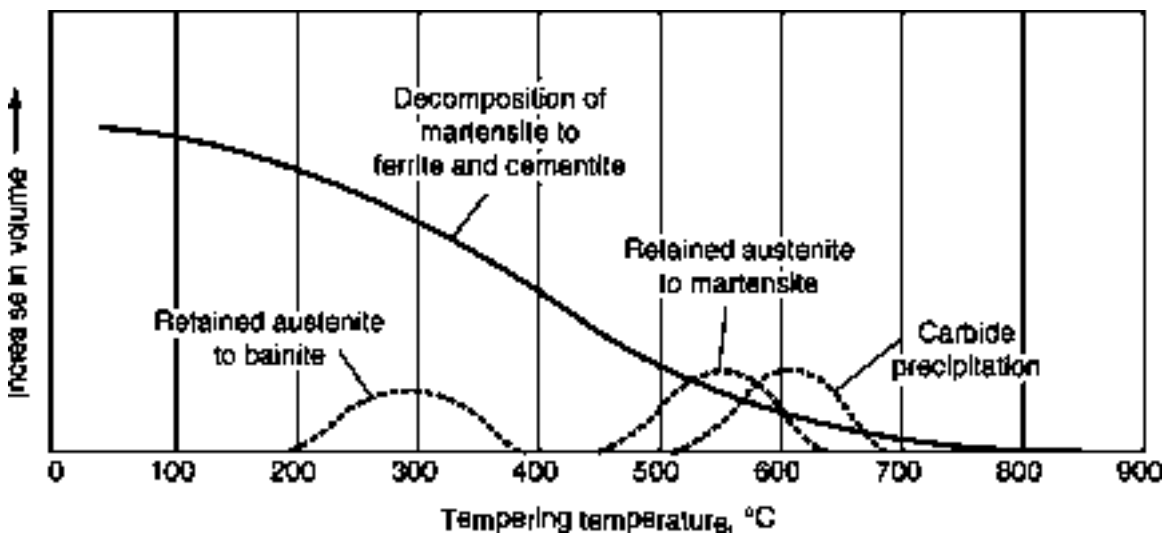


Fig. 13 Effect of microstructural constitutional variation on volume changes during tempering

Figure 14 illustrates the effect of dimensional variation and retained austenite content of a bearing steel (100Cr6) as a function of tempering temperature (Ref 10). Tempering may also lead to dimensional variation due to relaxation of residual stresses and plastic deformation, which is due to the temperature dependence of yield strength (Ref 11).

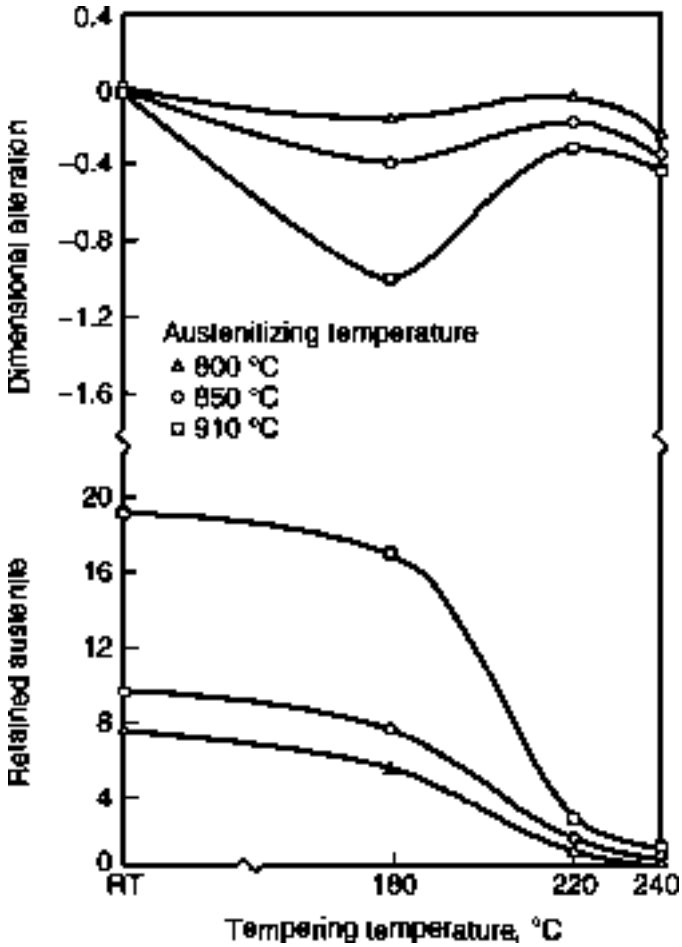


Fig. 14 Dimensional variation and retained austenite content of 100Cr6 steel as a function of tempering temperature

In addition to dimensional change by microstructural variation, tempering may also lead to dimensional variation due to relaxation of residual stresses and plastic distortion, which is due to the temperature dependence of yield strength (Ref 13). Figure 15 shows the distortion of round steel bars (200 mm, or 8 in., in diameter and 500 mm, or 20 in., in length) by quenching and by stress relieving by tempering. A medium-carbon steel bar (upper diagrams) and a hardenable steel bar (lower diagrams) are used in this experiment. Figures 15(a) and (d) are the results of quenching from 650 °C (1200 °F) without phase transformation. The distortion in each case is almost the same, regardless of the different quenchants and the different chemical compositions. These convex distortions are caused by nonuniform thermal contraction and resultant thermal stress during cooling. Figures 15(b) and (e) are the results of quenching from 850 °C (1560 °F) with phase transformation. The distortion in Fig. 15(e) (hardenable steel) shows a convex configuration, but the distortion in Fig. 15(b) (medium-carbon steel of poor hardenability) shows a configuration that combines convex and concave distortion. In addition, water quenching has a greater effect on distortion than oil quenching. Figures 15(c) and (f) show the configurations after tempering. These results show that tempering after quenching results in not only volumetric changes but also convex distortions. Such distortions seem to be related to relieving of residual stresses by tempering.

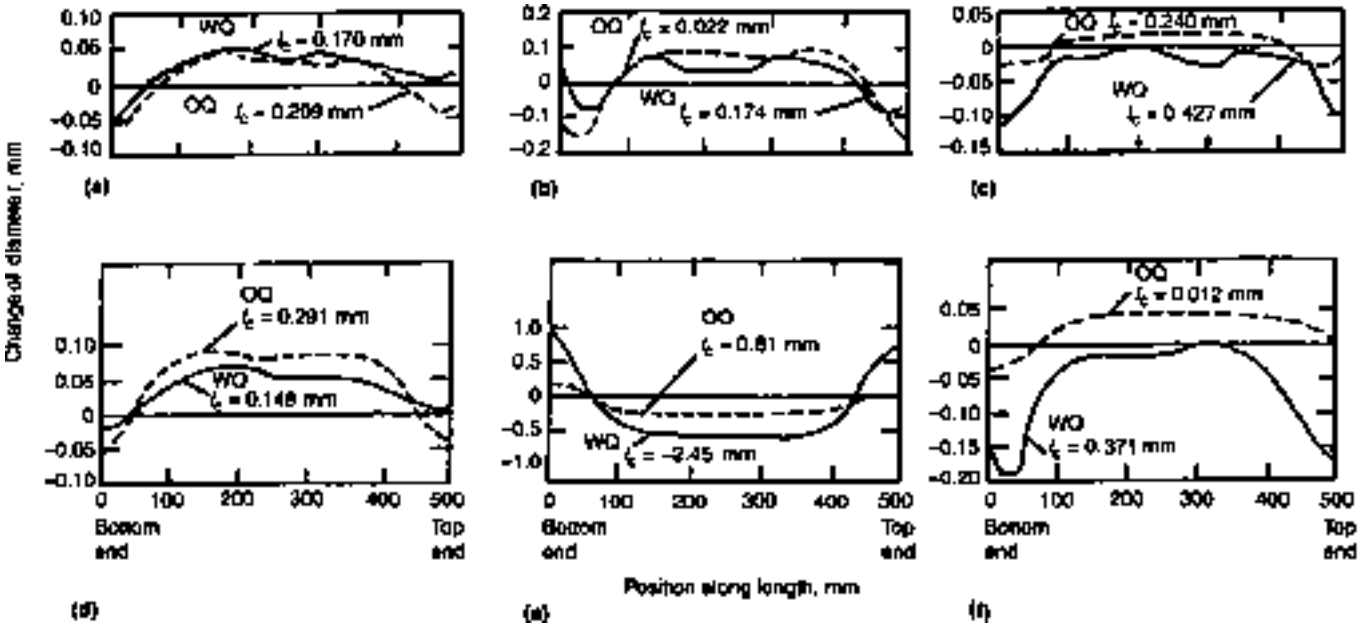


Fig. 15 Deformation of medium-carbon and hardenable steel bars by quenching from below and above the transformation temperature and by stress relieving. l_c, change of length; WQ, water quench; OQ, oil quench. (a) to (c) JIS S38C steel (0.38% C). (d) to (f) JIS SNCM 439 steel (0.39% C, 1.8% Ni, 0.8% Cr, 0.2% Mo). (a) and (d) Quenched from 650 °C (1200 °F). (b) and (e) Quenched from 850 °C (1560 °F). (c) and (f) Tempered at 650 °C (1200 °F)

Figure 16 and 17 (Ref 14) show the example of stress relief by tempering. A solid cylinder, 40 mm (1.6 in.) in diameter and 100 mm (4 in.) in length, was examined for analyses and experiments of tempering performed after water quenching. Calculated residual stress distributions after water quenching are illustrated in Fig. 16. Open and solid circles in the figure correspond to measured stresses on the surface of the cylinder by x-ray diffraction technique. Residual stress distributions after tempering at 400 °C (750 °F) are shown in Fig. 17(a) and (b) for typical elapsed times of 2 and 50 h, with measured values on the surface. These results show that the stresses in all directions decrease with elapsed time in tempering.

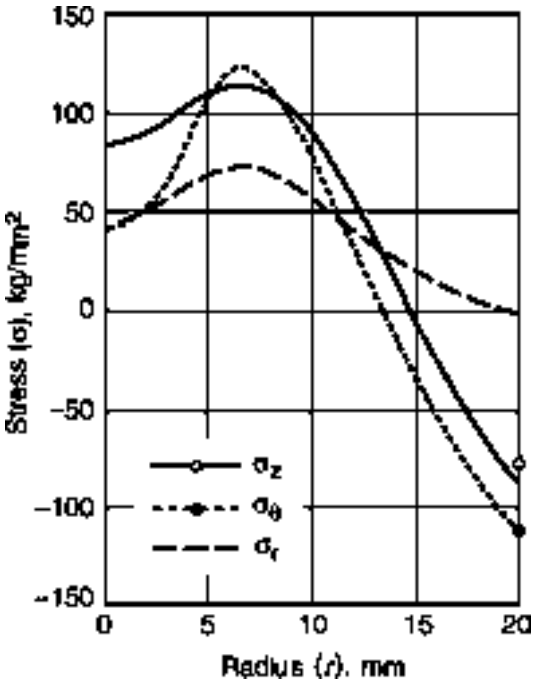


Fig. 16 Stress distribution in a cylinder after quenching

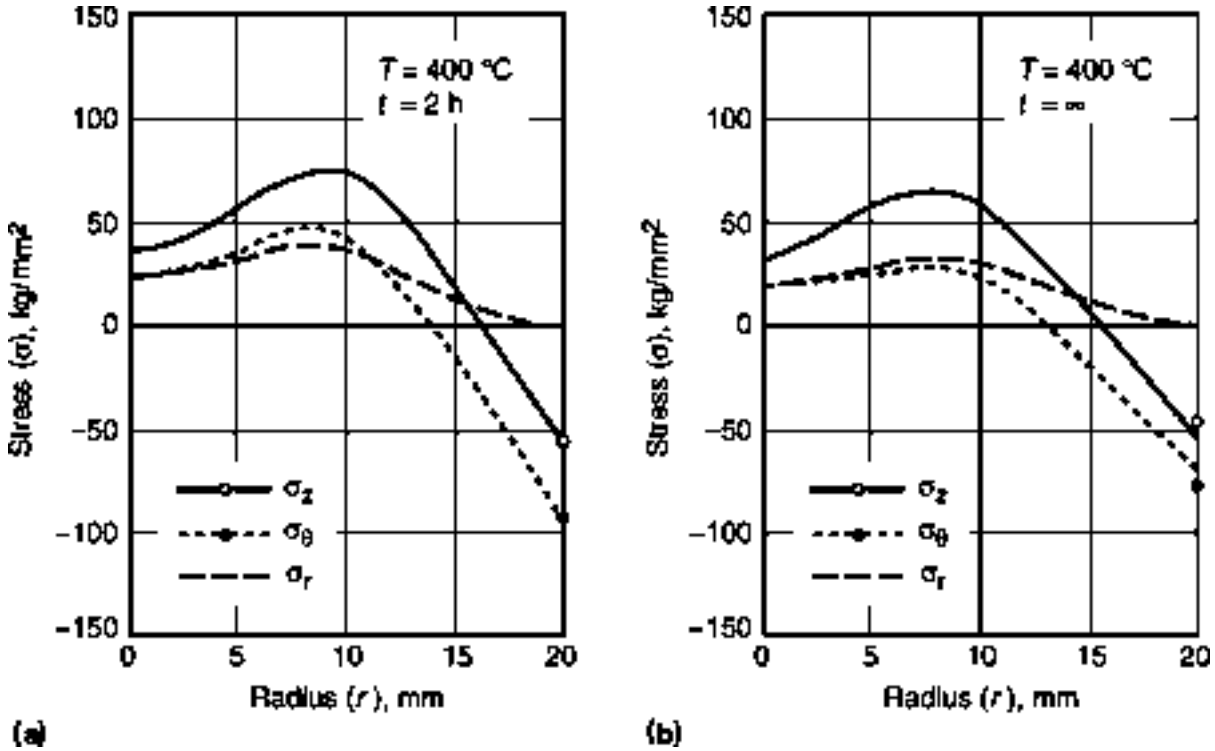


Fig. 17 Stress distribution in a cylinder after tempering. (a) *t* = 2 h. (b) Assume *t* = ∞.

The following are recommendations with respect to the tempering process (Ref 4):

- Generally, the higher the tempering temperature, the greater the resulting ductility and toughness. However, this is at the expense of strength and hardness.
- High-carbon steels that contain appreciable amounts of retained austenite undergo an increase in hardness when tempered at approximately 230 °C (450 °F) as a result of decomposition of the austenite.
- Steel tempered at approximately 260 °C (500 °F) undergoes a loss of toughness (Fig. 18) that is associated with the formation of carbide films around prior martensite plates.
- Alloy steels undergo temper brittleness or a reduction in toughness that occurs when the steel is tempered in (or slowly cooled through) the temperature range of 510 to 593 °C (950 to 1100 °F). This phenomenon does not occur in plain carbon steels, but the degree of embrittlement is enhanced by nickel, manganese, and chromium content. Tempering should be carried out below this range for steels susceptible to this embrittlement or above the range, followed by rapid quenching through the range. Molybdenum additions minimize the susceptibility of steels to this form of embrittlement.
- Highly alloyed steels, for example, high-speed steels, many contain large amounts of retained austenite that is stable during tempering but transforms to brittle martensite on cooling from tempering. A second temper is then necessary to temper this brittle secondary martensite.

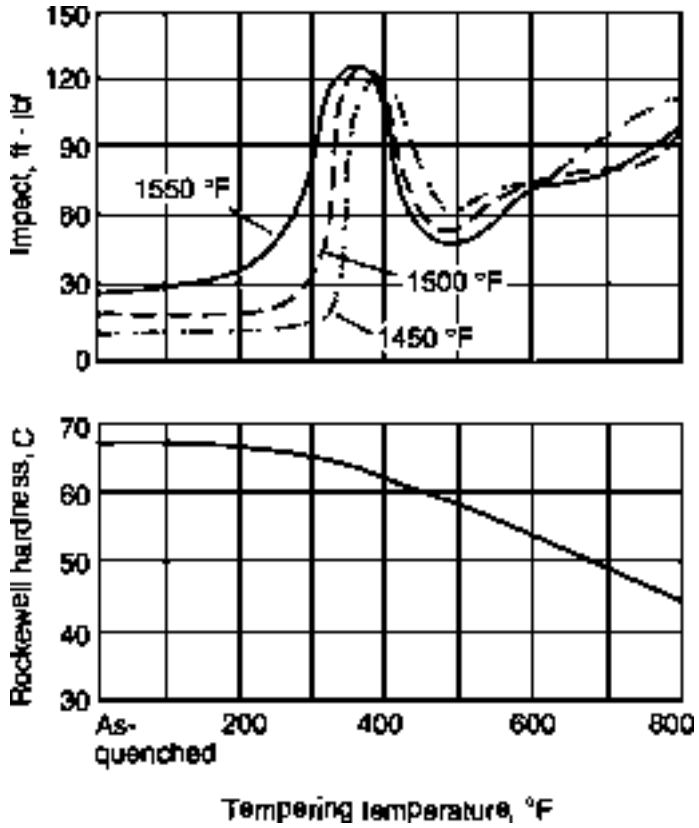


Fig. 18 The 260 to 315 °C (500 to 600 °F) impairment in torsion toughness in very hard steels. Note: Reduction in toughness is not detected by hardness measurements. Source: Ref 4

References cited in this section

4. R.W. Bohl, Difficulties and Imperfections Associated with Heat Treated Steel, Lesson 13, Heat Treatment of Steel, Materials Engineering Institute Course 10, ASM International, 1978

9. C. Jensen, Dimensional Changes during Heat Treatment of Hot Work Steel, *Ind. Heat.,* March 1993, p 33–38

10. P. Mayr, "Dimensional Alteration of Parts due to Heat Treatment," *Residual Stresses in Science and Technology,* Vol 1, Garmisch-Partenkirschen, FRG, 1986, DGM Informationsgesellschaft mbH, Adenauerallee 21, D-6370, Oberursel, FRG 1987, p 57–77

11. R.W. Edmonson, Dimensional Changes in Steel during Heat Treatment, *Met. Treat.,* Vol 20 (No 6), 1969, p 3–19

12. M. Wisti and M. Hingwe, Tempering of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 121–136

13. Y. Toshioka, Heat Treatment Deformation of Steel Products, *Mater. Sci. Technol.,* Vol 1, 1985, p 883–892

14. T. Inoue, K. Haraguchi, and S. Kimura, Analysis of Stresses due to Quenching and Tempering of Steel, *Trans. Iron Steel Inst. Jpn.,* Vol 18, 1978, p 11–15

Metallurgical Sources of Stress and Distortion during Reheating and Quenching

Basic Distortion Mechanism. Shape and volume changes during heating and cooling can be attributed to three fundamental causes (Ref 15):

- Residual stresses that cause shape change when they exceed material yield strength. This occurs on heating when the strength properties decline.
- Stresses caused by differential expansion due to thermal gradients. These stresses increase with the thermal gradient and cause plastic deformation as the yield strength is exceeded.
- Volume changes due to transformational phase change. These volume changes are contained as residual stress systems until the yield strength is exceeded.

Relief of Residual Stresses. If a part has locked-in residual stresses, these stresses can be relieved by heating the part until the locked-in stresses exceed the strength of the material. A typical stress-strain curve obtained from a tension test is shown in Fig. 19 (Ref 15). Initial changes in shape are elastic, but under increased stress, they occur in the plastic zone and are permanent. On heating, the stresses are gradually relieved by changes in the shape of the part due to plastic flow. This is a continuous process, and as the temperature of the part is increased, the material yield stress decreases, as shown in Fig. 20 (Ref 16). It is a function not only of temperature but also of time, because the material creeps under lower applied stresses. It is apparent that the stresses can never be reduced to zero, because the material always possesses some level of yield strength below which residual stresses cannot be reduced.

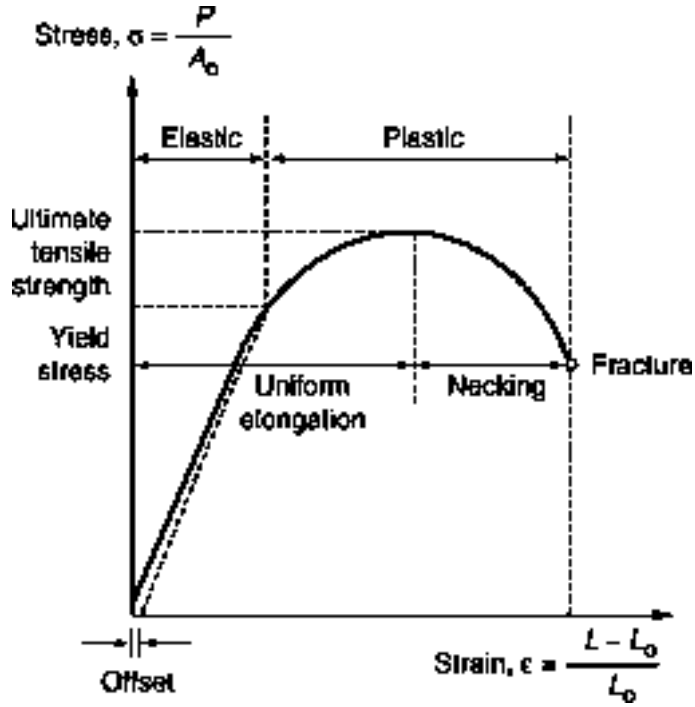


Fig. 19 Various features of a typical stress-strain curve obtained from a tension test

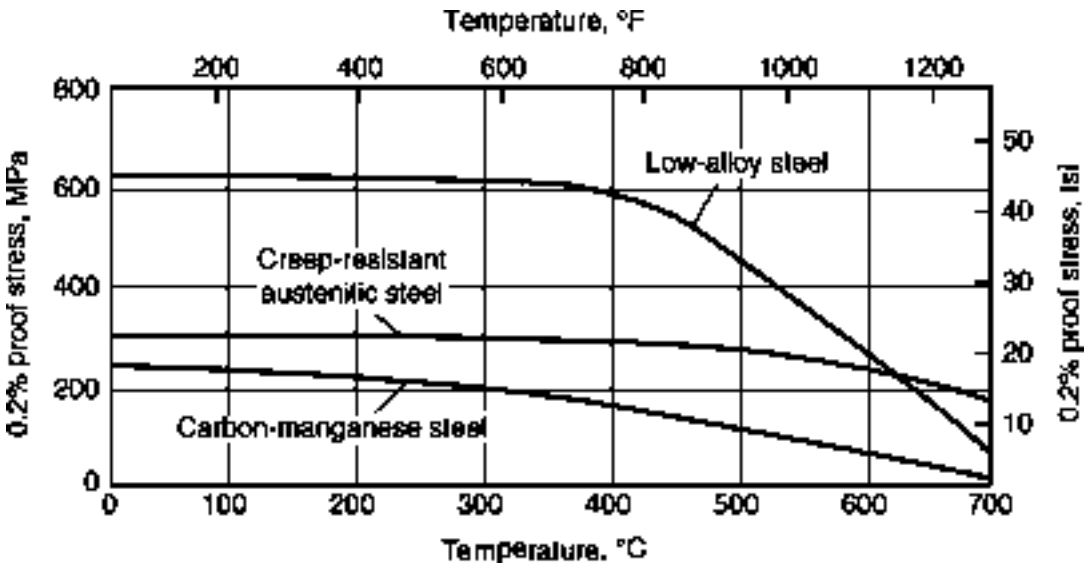


Fig. 20 Variation of yield strength with temperature for three generic classes of steel

Volume Changes during Phase Transformations. When a steel part is heated, it transforms to austenite, with an accompanying reduction in volume, as shown in Fig. 21 (Ref 17). When it is quenched, the structure transforms from austenite to martensite, and its volume increases. If these volume changes cause stresses to be set up that are constrained within the strength of the material, a residual stress system is created. If the stresses cannot be contained, then material movement occurs, which causes cracking under extreme conditions. The expansion is related to the composition of the steel. Figure 21 shows the relative volume increase of two steels as a function of austenitizing temperature and specimen dimensions.

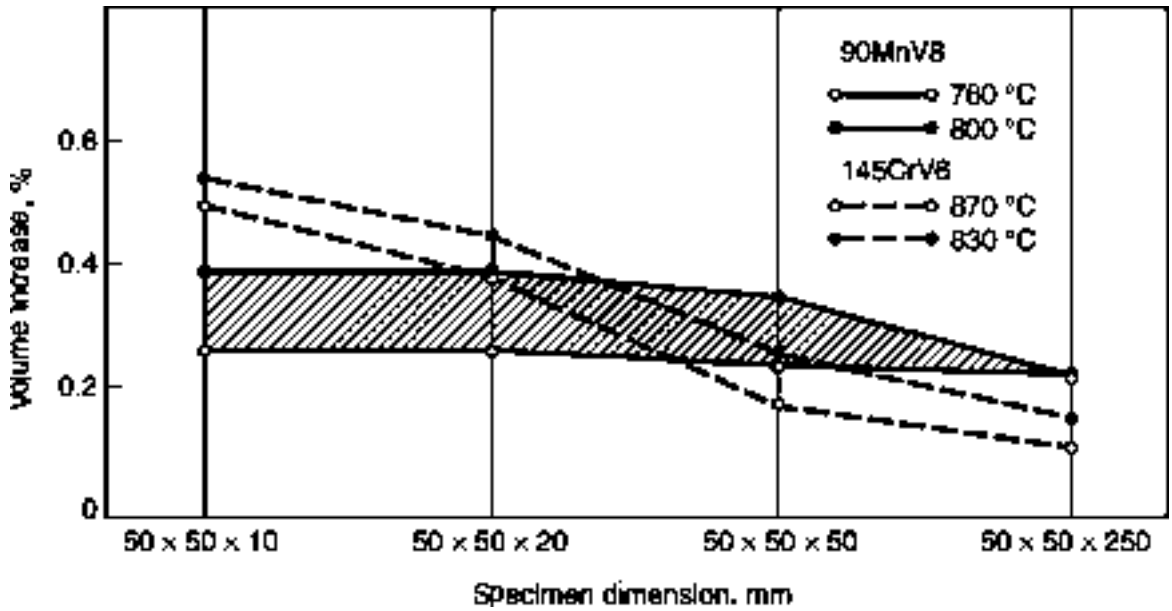


Fig. 21 Volume increase of steels 90MnV8 and 145CrV6 as a function of austenitization temperature and specimen dimensions

While these physical changes are well known, the situation is made more complex when all three events occur simultaneously. In addition, other events, such as heating rate, quenching, and inconsistent material composition, further complicate the process.

When parts are heated during heat treatment, a thermal gradient exists across the cross section of the component. If a section is heated so that a portion of the component becomes hotter than the surrounding material, the hotter material expands and occupies a greater volume than the adjacent material and is thus exposed to applied stresses that cause a shape change when they exceed material strength. These movements can be related to heating rate and section thickness of the component.

References cited in this section

15. G.E. Totten and M.A.H. Howes, Distortion of Heat Treated Components, *Steel Heat Treatment Handbook,* G.E. Totten and M.A.H. Howes, Ed., Marcel Dekker Inc., 1997, p 292

16. D.A. Canonico, Stress-Relief in Heat Treating of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 33–34

17. C.E. Bates, G.E. Totten, and R.L. Brennan, Quenching of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 67–120

Effect of Materials and Process Design on Distortion

Quenchant selection and quenching conditions are critically important parameters in quench system design. For example, one study compared the distortion obtained when a 0.4% medium-carbon plain steel bar 200 mm (8

in.) in diameter by 500 mm (20 in.) long is quenched in water or oil from 680 °C (1260 °F) (Ref 18, 19). The results, Fig. 22(a) and (c), show that essentially equivalent variation in diameter and length with both cooling processes was obtained, which was due to thermal strains within the steel. Interestingly, the well-known diameter variations at the end of the bar, known as the "end-effect," were observed, which is attributable to heat extraction from both the sides and ends of the bar.

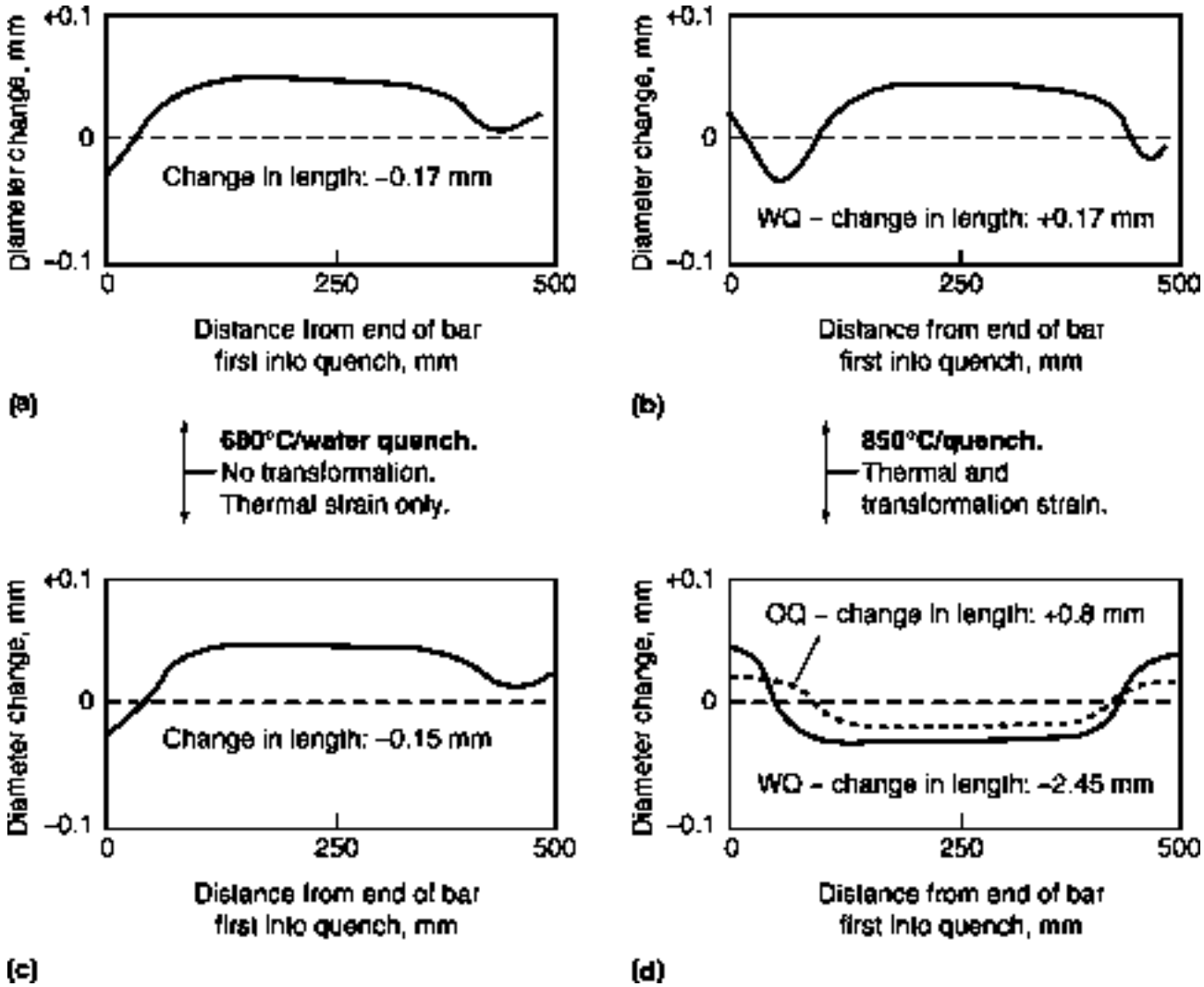


Fig. 22 Dimensional variation of a medium-carbon (0.4%) steel bar (200 mm, or 8 in., diam × 500 mm, or 20 in.) after the indicated heat treatments. These bars were quenched vertically with one end down (marked "0" in the figure). (a) and (c) show no transformation, thermal strain only after water quenching from 680 °C (1260 °F). (b) and (d) show thermal and transformation strains after quenching from 850 °C (1560 °C). OQ, oil quench; WQ, water quench

If the same steel bars of the same dimensions are heated to 850 °C (1560 °F) to austenitize the steel and then quenched in water or oil, the results shown in Fig. 22(b) and (d), respectively, were obtained (Ref 18, 19). Considerably greater dimensional variation and lengthening of the bar (for the oil quench) was obtained due to both thermal and transformational strains with the steel.

The dimensional changes of a 70 mm (2.75 in.) steel (0.15% C, 1% Mn, 0.75% Cr, 0.85% Ni) cube were modeled after austenitizing and then quenching in water and oil (Ref 18, 20). The results of this work are shown in Fig. 23. As seen in the figure, the edges and faces shrink (becoming concave), and the effects are greater when the steel is quenched in water than when it is quenched in oil.

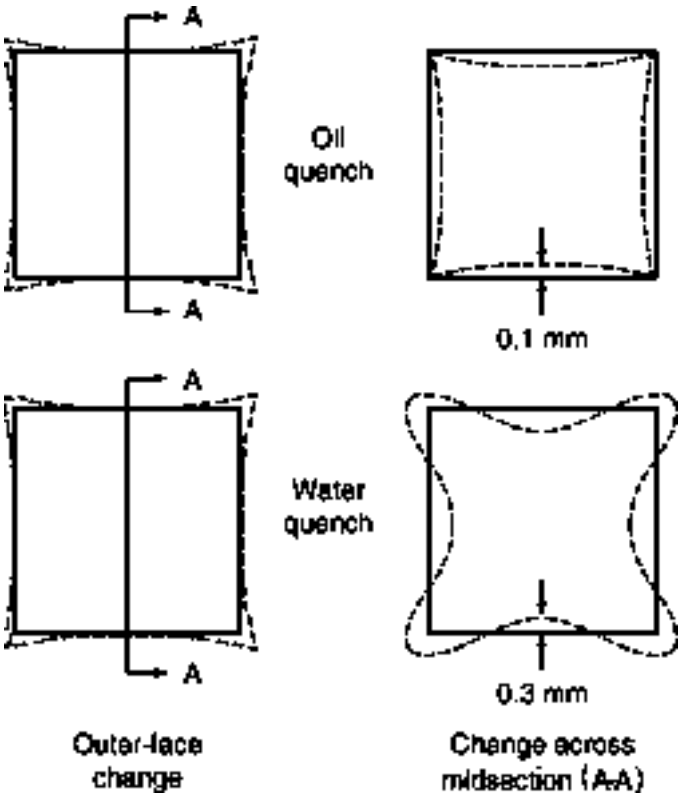


Fig. 23 Dimensional changes in a 70 mm (2.75 in.) steel (0.15% C, 1% Mn, 0.75% Cr, 0.85% Ni) bar after austenitizing and then quenching in water or oil

There are various factors that can affect distortion and growth of steel in heat treating. These include component design, steel grade and condition, machining, component support and loading, surface condition, heating and atmosphere control, retained austenite, and the quenching process (Ref 21).

Component Design. One of the overwhelming causes of steel cracking and unacceptable distortion control is component design. Poor component design promotes distortion and cracking by accentuating nonuniform and nonsymmetrical heat transfer during heating and cooling. Component design characteristics that are common to distortion and cracking problems include (Ref 22, 23):

- Parts that are long, with thin cross sections. Long and thin parts are defined as those greater than $L = 5d$ for water quenching, $L = 8d$ for oil quenching, and $L = 10d$ for austempering, where L is the length of the parts, and d is the thickness or diameter. Long and thin parts may be quenched using a support mechanism, such as that illustrated in Fig. 24.
- Parts that possess large cross-sectional area (A) and are thin (t), which are defined as $A = 50t$.

Fig. 24 Die quenching system. Courtesy of Gleason Tooling Products Group

Parts that exceed these dimensions often must be straightened or press quenched to maintain dimensional stability (Ref 23). If possible, materials with sufficient hardenability should be oil or salt quenched. A schematic illustration of a press quenching system is provided in Fig. 25.

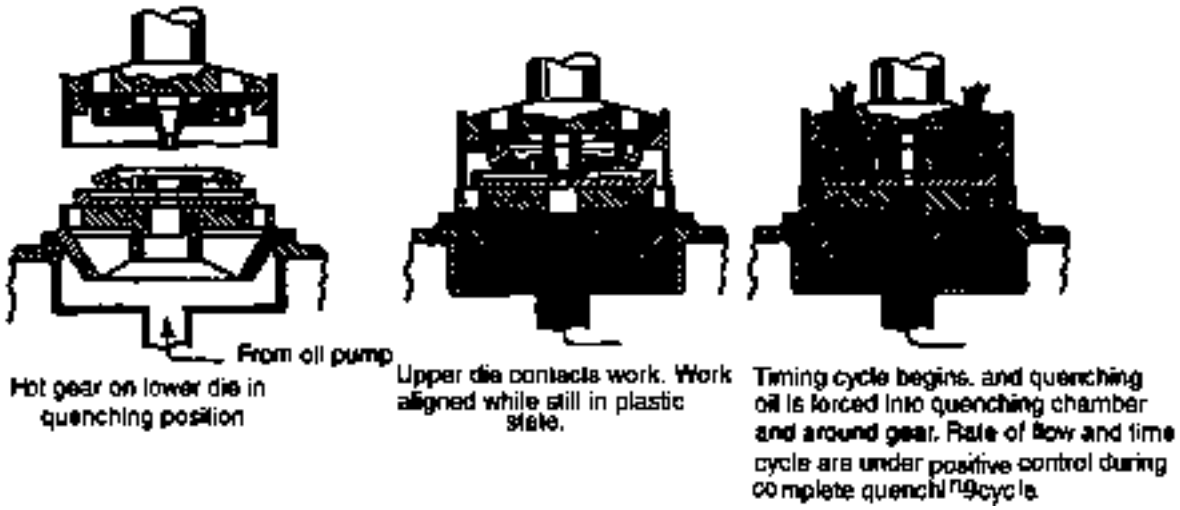


Fig. 25 Schematic of a press quench system. Courtesy of Gleason Tooling Products Group

The following are additional guidelines (Ref 4):

- Balance the areas of mass.
- Avoid sharp corners and reentrant angles.
- Avoid sharp corners between heavy and thin sections.
- Avoid single internal or external keys, keyways, or splines.
- Provide adequate fillet or radius at the base of gear teeth, splines, and serrations.
- Do not have holes in direct line with the sharp angles of cutouts.
- Avoid a sharp corner at the bottom of small openings, such as in drawing or piercing dies, because spalling or flaking is likely to result at these points.
- Keep hubs of gears, cutters, and so on as near the same thickness as possible, because dishing is likely to occur.

Additional design considerations to reduce the probability of distortion, cracking, and soft spots include (Ref 4):

- Order stock large enough to allow for machining to remove decarburized surfaces and surface imperfections, such as laps and seams.
- Do not drill screw holes closer than 6.35 mm (0.25 in.) from edges of die blocks or large parts, where possible. Cracking may be avoided by using steel that may be hardened by using lower quench severity, or if possible, pack the bolt hole to reduce thermal stresses arising due to quenching.
- Avoid blind holes, if possible.
- All parts should be designed with round corners and fillets wherever possible.
- Use air-hardening or high-carbon (oil- and air-hardening) tool steel on unbalanced and intricately shaped dies.
- Add extra holes, if possible, on heavy, unbalanced sections to allow for faster and more uniform cooling when quenched.
- Do not machine knife blades to a sharp cutting edge before hardening.
- Avoid deep scratches and tool marks.
- On long, delicate parallels, shafts, and so on, rough out and have pieces annealed to remove stresses before finish machining.
- Always use the grade or composition of steel most suitable for the work that the part has to perform.

A well-designed part is the best ensurance against breakage in hardening.
Design symmetry is also an important variable to minimize distortion. For example, the unsymmetrical gear design shown in the upper diagram of Fig. 26 typically may undergo distortion, as shown in the lower diagram of Fig. 26 (Ref 22). (The load on a gear tooth increases as a power of 4.3 with the taper [Ref 22]). The solution to the gear design problem shown in Fig. 26 is to provide greater symmetry, as shown in Fig. 27. If this is not possible, press quenching or tooth-by-tooth induction hardening may be the best alternative (Ref 22, 23).

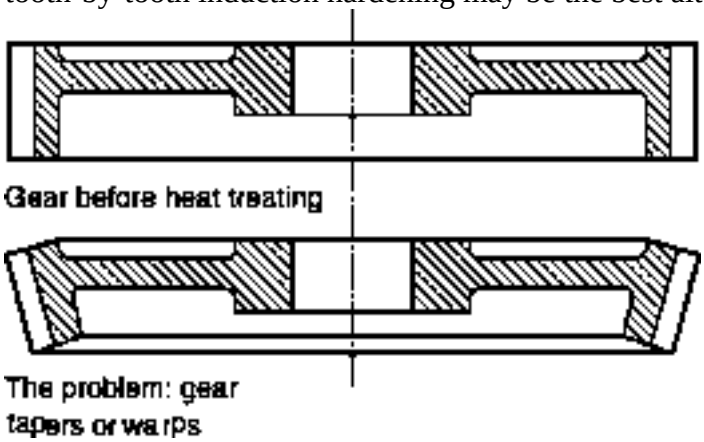


Fig. 26 Schematic of a gear that is difficult to harden without the distortion shown

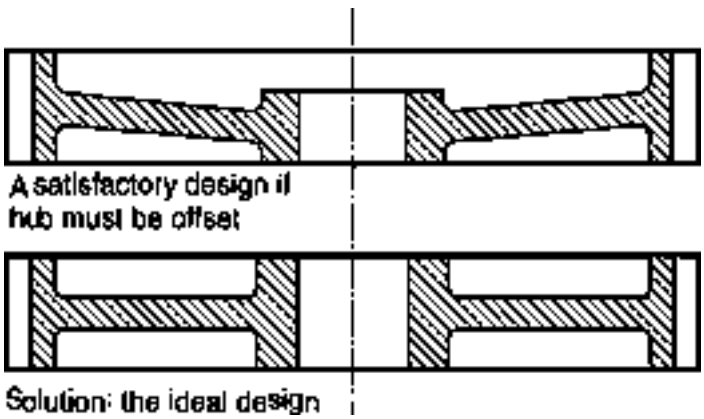


Fig. 27 Design solutions to the distortion problem shown in Fig. 26

Another common design problem occurs with parts having holes, deep keyways, and grooves. One illustration of this problem is hardening of a shaft over a lubrication cross hole, as illustrated in Fig. 28 (Ref 22). Preferred

alternative designs are also shown in Fig. 28. If a radial cross hole is mandatory, the use of carburized steel with oil quenching would be preferred.

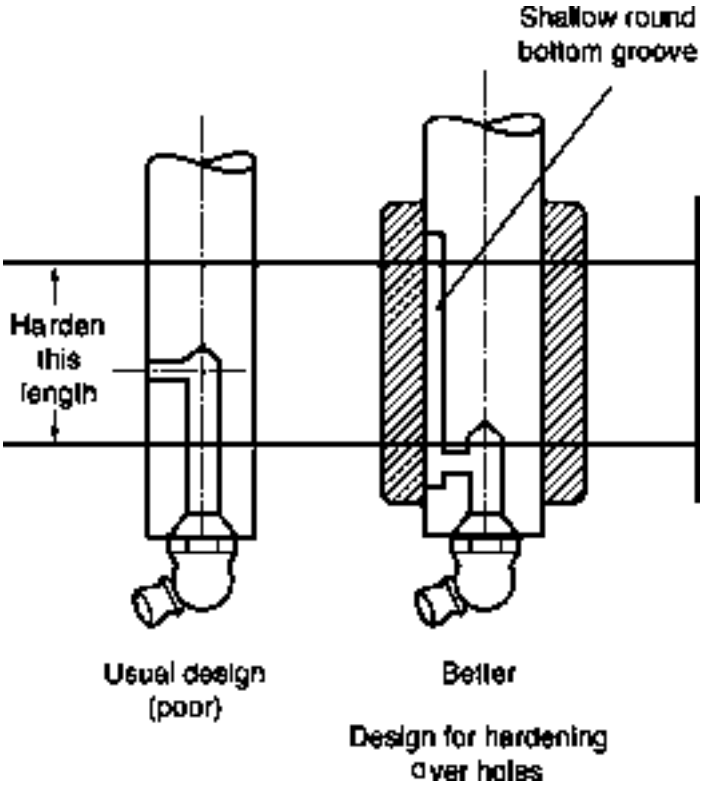


Fig. 28 Design solutions to the quench-cracking problem often encountered in shaft hardening over a cross hole

The distortion encountered when quenching a notched part, such as a shaft with a milled slot, is illustrated in Fig. 29 (Ref 23). In this case, nonuniform heat transfer results. The metal within the notch is affected by the shrinkage of the metal around it, due to slower cooling within the slot caused by vaporization of the quenchant. Therefore, on cooling, the metal on the side with the shaft is “too short,” pulling the shaft out of alignment. A general rule for solving such quench distortion problems is that the “short side is the hot side,” which means that the inside of the bowed metal was quenched more slowly than the opposite side (Ref 23).

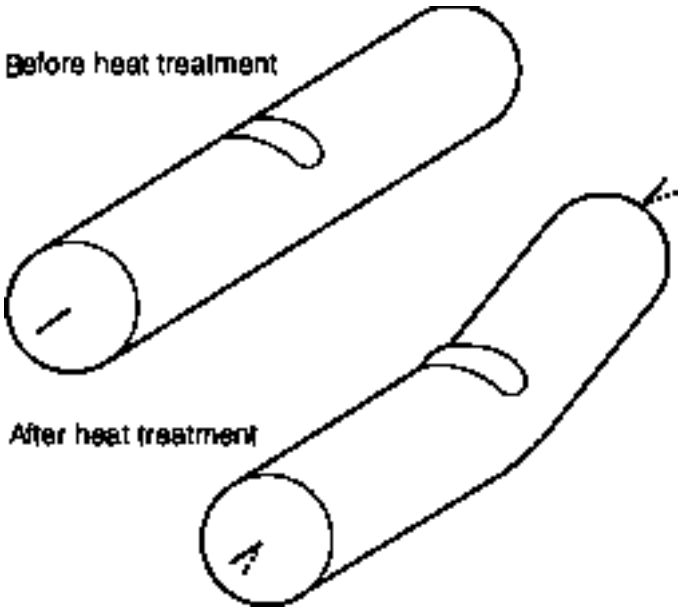


Fig. 29 Distortion often encountered when quenching a notch

Figure 30 provides an illustration of recommended design corrections that reduce the possibility of quench-cracking (Ref 4). If it is unavoidable that quench crack-sensitive designs be used, then two possible design modifications should be considered. The first is to consider a two-piece assembly that is mechanically joined or shrunk-fitted together. The second is to manufacture the part from steel that can be hardened by martempering or by air cooling. These alternative design strategies produce lower stress and minimize cracking.

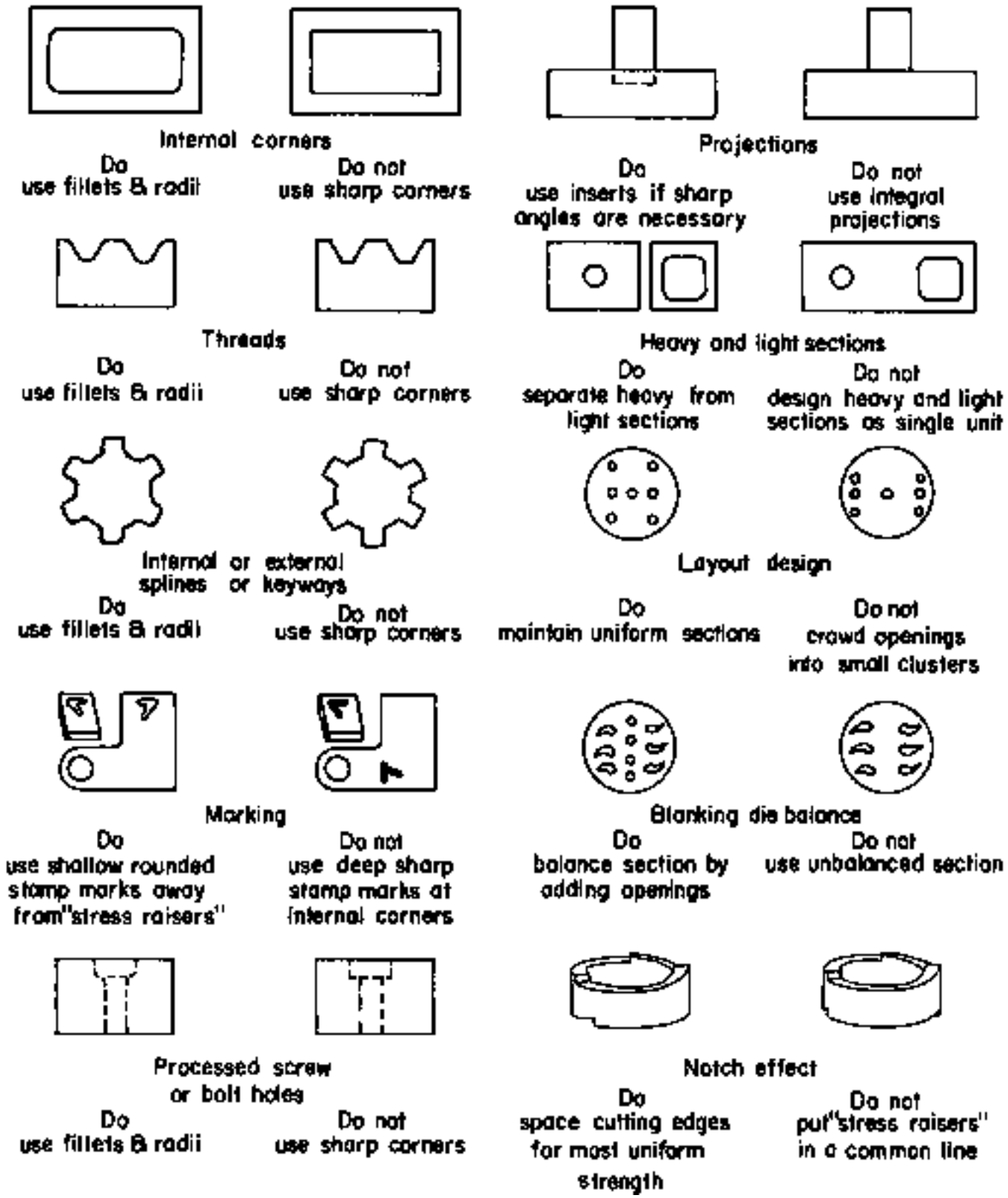


Fig. 30 Part design recommendations for minimal internal stresses. Source: Ref 4

Steel Grade and Condition. Although steel cracking is most often due to nonuniform heating and cooling, material problems may be encountered. Some typical material problems include (Ref 22):

- The compositional tolerances should be checked to ensure that the alloy is within specification.
- Some alloys are particularly problematic. For example, some steel grades must be water quenched when the alloy composition is on the low side of the specification limit. Conversely, if the alloy composition is on the high side, cracking is more common. Steel grades that exhibit this problem include: 1040, 1045, 1536, 1541, 1137, 1141, and 1144. As a rule, steels with carbon contents and hardenability greater than 1037 are difficult to water quench (Ref 22).
- Some steel grades with high manganese are prone to microsegregation of manganese, gross segregation of chromium, and thus are prone to cracking. These include 1340, 1345, 1536, 1541, 4140, and 4150. If possible, it is often a good choice to replace the 4100 series with the 8600 series of steels (Ref 24).
- "Dirty" steels, those containing greater than 0.05% S (such as 1141 and 1144), are more prone to cracking. One reason for this is the greater alloy segregation in dirty steels, leading to alloy-rich and alloy-lean regions. Another reason is that these steels typically have a greater number of surface seams, which act as stress raisers. Also, dirty steels and steels with higher sulfur levels are often manufactured to coarse-grain practice for improved machinability; this also imparts greater brittleness and propensity for cracking.
- Decarburization of up to 0.06 mm (0.0025 in.) per 1.6 mm (1/16 in.) diameter may be present.

It is well known that cracking propensity increases with carbon content. Therefore, the carbon content of the steel is one of the determining factors for quenchant selection. Table 5 summarizes some average carbon-content concentration limits for quenching steel into water, brine, or caustic (Ref 24).

Table 5 Suggested carbon-content limits for water, brine, and caustic quenching

Shapes	Maximum carbon content, %
Furnace hardening	
General use	0.30
Simple shapes	0.35
Very simple shapes (e.g., bars)	0.40
Induction hardening	
Simple shapes	0.50
Complex shapes	0.33

Higher-carbon steels undergo greater expansion on hardening, and because the M_s temperature decreases with carbon content, expansion that does occur during martensitic transformation does so at a lower temperature, where the cracking of the steel is more likely than distortion. Therefore, cracking is more likely when hardening high-carbon steels than when heat treating low- or medium-carbon steels. As a rule of thumb, plain carbon steels with less than 0.35% C rarely crack on hardening, even under severe quenching conditions (Ref 4). The carbon content of a steel should never be greater than necessary for the specific application of the part.

It is well known that regions containing high concentrations of coarse carbide microstructure as a result of improper forging may become the initiation point for subsequent quench cracking, particularly with parts of complex shape (Ref 25). It is important to provide a sufficient forging reduction ratio to allow the carbide formation to become fine and uniform (Ref 26).

Because part manufacture, such as gear production, often requires machining, the condition of the steel that is going to be machined is critically important. Some workers have recommended that normalized and subcritical annealed steel is the ideal condition (Ref 21). Subcritical annealing is performed to relieve stresses incurred during normalization without softening or homogenizing the steel. The subcritical annealing process reduces the carbon content and alloy carbide content in the austenite, allowing the production of more lath martensite in the microstructure, which provides higher fracture toughness and higher impact toughness (Ref 25).

Improper Steel Chemistry. Steel hardenability is determined by its chemistry. The quench conditions required to obtain the desired properties are a function of the hardenability. Therefore, if the steel chemistry is incorrect, the selected quench process conditions may, if too severe, lead to cracking. Unfortunately, this problem is not uncommon.

An example of this problem is quench cracking that occurred in AISI 1070 steel bearing raceways. Metallographic analysis confirmed the presence of quench cracking. However, the steel chemistry (Table 6) was incorrect for a plain AISI 1070 steel (Ref 27). The higher-carbon, higher-hardenability steel with a high

manganese content, which was found in this example, would be more susceptible to quench cracking, using the normally specified quenchant for the 1070 steel bearing raceways.

Table 6 Comparison of steel obtained and specification range of steel chemistry of AISI 1070 steel used for bearing raceway problem

Element	Specification range for AISI 1070, %	Content in steel obtained, %
Carbon	0.65–0.75	0.74
Manganese	0.60–0.90	0.97
Phosphorus	0.11	0.04
Sulfur	0.026	0.05
Silicon	0.10–0.20	0.23
Nickel	…	0.07
Chromium	…	0.11
Molybdenum	…	0.22
Copper	…	0.10

Prior Steel Structure. The structure of the steel prior to hardening, for example, extruded, cast, forged, cold formed, and so on, may enhance the potential for cracking during the quench. Each as-formed structure requires a specific time and temperature cycle to condition the material for proper hardening. For example, cast structures must be homogeneous, cold-formed structures require normalization and annealing, and forged structures must be grain-refined by normalization.

Cracking may be caused by microstructures, resulting in nonuniform heating or cooling. In Fig. 31 (Ref 27), microstructures obtained on forged AISI 403 stainless steel valve stems exhibited longitudinal cracking after quenching. Microstructural analysis suggested that cracking was caused by thermal stresses during forging or during heating prior to forging. Figure 31 also shows a coarse-grain condition associated with high-temperature surface oxidation. Further examination revealed evidence of high- and low-thermal oxidation within the crack profile. The presence of this condition suggests cracking occurred prior to or during forging.

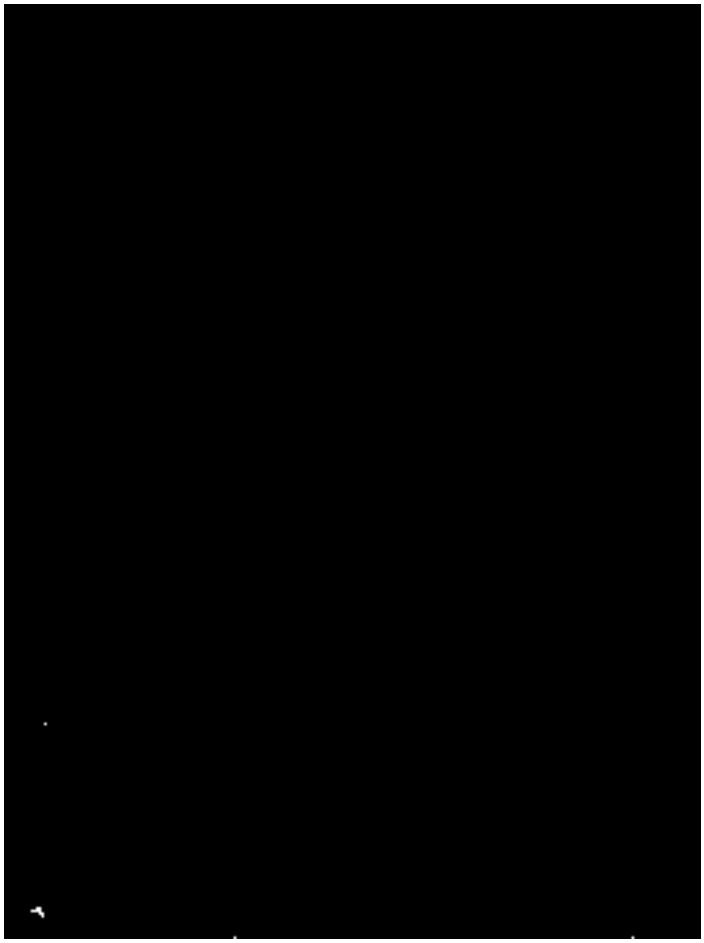


Fig. 31 Micrograph of type 403 stainless steel as-forged. The microstructure is predominantly a mixture of carbide particles in a matrix of ferrite. No evidence of quenching and tempering was observed. High- and low-temperature oxidation can be observed on the surface of the sample and within the crack profile. 100×; Villela's reagent. Source: Ref 27

Excessive overheating is called “burning,” which refers to incipient melting of low-melting constituents into liquid films that concentrate embrittling components in the grain boundaries, as illustrated in Fig. 32 (Ref 4). The problem of the brittle nature of these films is exacerbated by accompanying void formation, which is caused by contraction shrinkage when the liquid film cools. Although burning usually occurs during rolling or forging, it may not be observed until after heat treatment.

Fig. 32 Severely overheated 1038 steel showing initial stage of burning. Ferrite (white) outlines prior coarse austenite grain boundaries; matrix consists of ferrite (white) and pearlite (black). Source: Ref 4

A mixed-grain-size microstructure may be formed when a steel was heated in its coarsening range during austenitization. The hardening response of this mixed microstructure, which is illustrated in Fig. 33, is unpredictable (Ref 4).

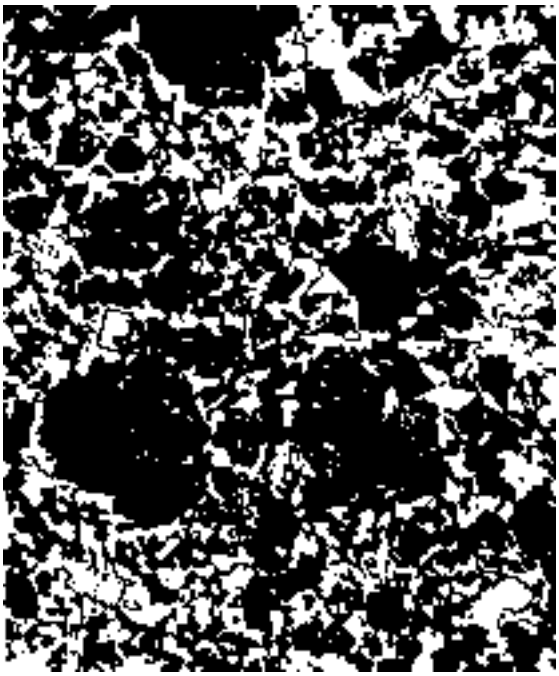


Fig. 33 AISI 1040 steel bar austenitized at 913 °C (1675 °F) for 30 min, then cooled slowly in a furnace. White areas are ferrite; dark areas are pearlite. Mixed grain size due to heating into the coarsening range is also observed. Source: Ref 4

When the carbon content of a steel is greater than the eutectoid concentration, it is usually quenched from below the Ac_{cm} temperature (the temperature at which cementite completes solution in austenite), resulting in the presence of undissolved carbides just prior to quenching. This reduces the amount of retained austenite that often accompanies excessive heating temperatures. As is shown subsequently, problems of increased distortion and cracking increase with the retained austenite content. It is important that undissolved carbides be in the dispersed form of spheroids and not in the form of grain-boundary films, which produce a brittle structure after tempering (Ref 4).

The steel austenitizing temperature is a compromise between achieving rapid solution and diffusion of carbon, and minimizing grain growth. A fine-grained steel may be heated to a temperature for rapid austenitization with little danger of grain growth. However, quenched structures formed from large austenite grains exhibit poor toughness and are crack sensitive. The quench cracking illustrated in Fig. 34 occurred because of excessively large austenite grains due to an excessively high austenitizing temperature (Ref 4).

Fig. 34 Quench cracks due to excessively large grain boundaries resulting from excessively high austenitizing temperature. Note cracking patterns associated with prior coarse austenite grain boundaries. Source: Ref 4

The optimal austenitizing time is determined by dividing the total time in the furnace between heating time (time necessary to bring the part to the austenitizing temperature) and transformation time (the time required to produce the desired microstructural transformation or to complete the desired diffusion process). Excessive heating times to provide for transformation may result in undesirable grain growth.

Heating and Atmosphere Control. Localized overheating is particularly a potential problem for inductively heated parts (Ref 28, 29). Subsequent quenching of the part leads to quench cracks at sharp corners and areas with sudden changes in cross-sectional area (stress raisers). Cracking is due to increases of residual stresses at the stress raisers during the quenching process. The solution to the problem is to increase the heating speed by increasing the power density of the inductor. The temperature difference across the heated zone is decreased by continuous heating or scanning of several pistons together on a single bar (Ref 29).

For heat treating problems related to furnace design and operation, it is usually suggested that (Ref 28):

- The vestibules of atmosphere-hardening furnaces should be loaded and unloaded with purging. Load transfer for belt and shaker hearth furnaces should only occur with thorough purging to minimize atmosphere contamination.
- Hardening furnaces typically contain excessive loads prior to quenching. If the steel at quenching temperature is greater than 20% of the distance from discharge to charge door, it is too much. Either the production rate can be increased or some of the burners can be turned off.

One source of cracking, which may appear similar to quench cracking, can also occur due to excessive heating rate to the austenitizing temperature. This is illustrated for AISI 4140 steel in Fig. 35 (Ref 27). In this case, surface oxidation and decarburization within the crack, which would not have been obtained if the cracking occurred during the quench, was observed.

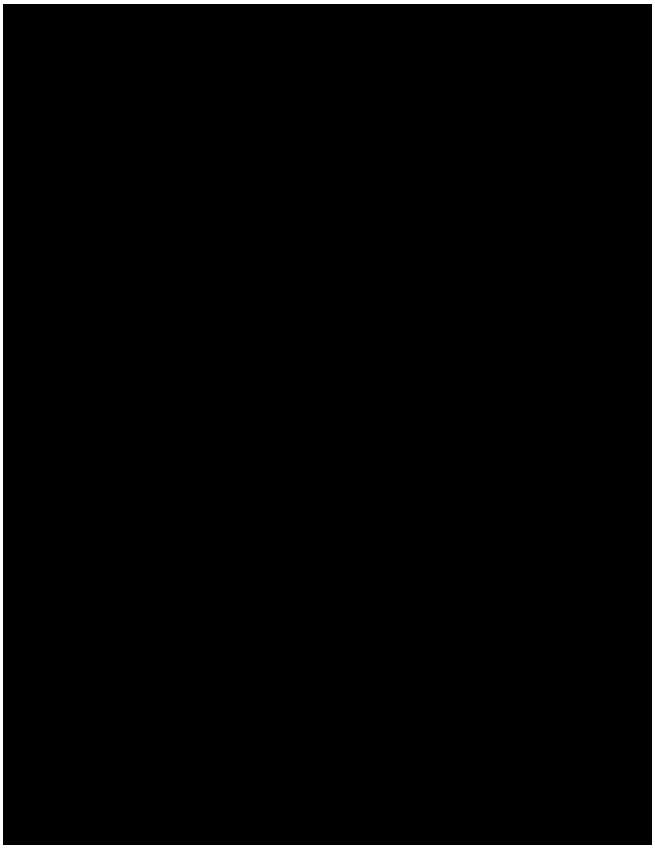


Fig. 35 Micrograph of AISI 4140 steel as quenched and tempered. The microstructure is tempered martensite with evidence of decarburization and high-temperature oxidation on the surface of the crack profile. 50×; 2% nital etch. Source: Ref 27

Cracking may also be due to localized overheating (nonuniform heating), shown in Fig. 36, which is a microstructure of AISI 4140 tube end sections (Ref 27). Circumferential cracking at a mid-thickness location of the sample was reported on the tube end. The tubes were reportedly austenitized and then spray quenched. Microstructural examination of the steel revealed a coarse-grain condition due to overheating during austenitization prior to quenching. The cracking occurred along the coarse-grain boundary, as illustrated in Fig. 36. However, microstructural analysis of samples from other regions of the tube indicated fine-grain martensite. Taken together, these data suggest that the austenitization furnace contained hot spots that caused localized overheating and grain coarsening. The overheated locations cracked in the presence of quench stresses. Cracking occurred in the midthickness locations due to the inherent weakness of the material centers carried over from the original billet or casting.

Fig. 36 Micrograph of AISI 4140 steel as quenched and tempered. The microstructure is tempered martensite with intergranular quench cracking along the prior austenite grain boundaries. 100×; 2% nital etch. Source: Ref 27

An important source of steel distortion and cracking is nonuniform heating and not using the appropriate protective atmosphere. For example, if steel is heated in a direct-gas-fired furnace with high moisture content, the load being heated may adsorb hydrogen, leading to hydrogen embrittlement and subsequent cracking, which would not normally occur with a dry atmosphere (Ref 22, 28).

Component Support and Loading. Many parts may sag and creep under their own weight when heat treated, which is an important cause of distortion. An example of a type of component that is susceptible to such distortion is a ring gear. The dimension limits by which ring gears are classified are provided in Fig. 37 (Ref 21). (A general dimensional classification of various distortion-sensitive shapes is provided in Fig. 38 [Ref 21].) Proper support when heating is required to minimize out-of-flatness and ovality problems, which may result in long grinding times, excessive stock removal, high scrap losses, and loss of case depth (Ref 21). To achieve adequate distortion control, custom supports or press quenching may be required.

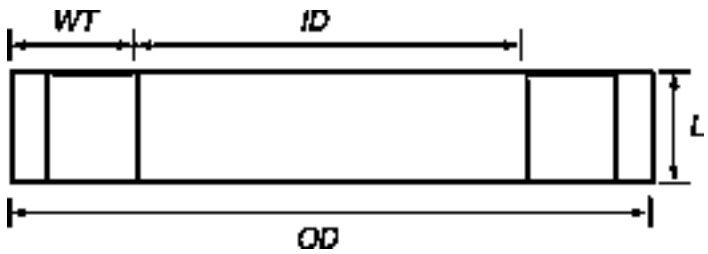


Fig. 37 Dimensions of a ring gear shape. Shape limitation: length/wall thickness ≤ 1.5; inside diameter (ID)/outside diameter (OD) > 0.4. Minimum wall thickness (WT) is defined by WT ≥ 2.25 × module + [0.4 × 5 (mod × *L* × OD^3)$^{1/2}$]

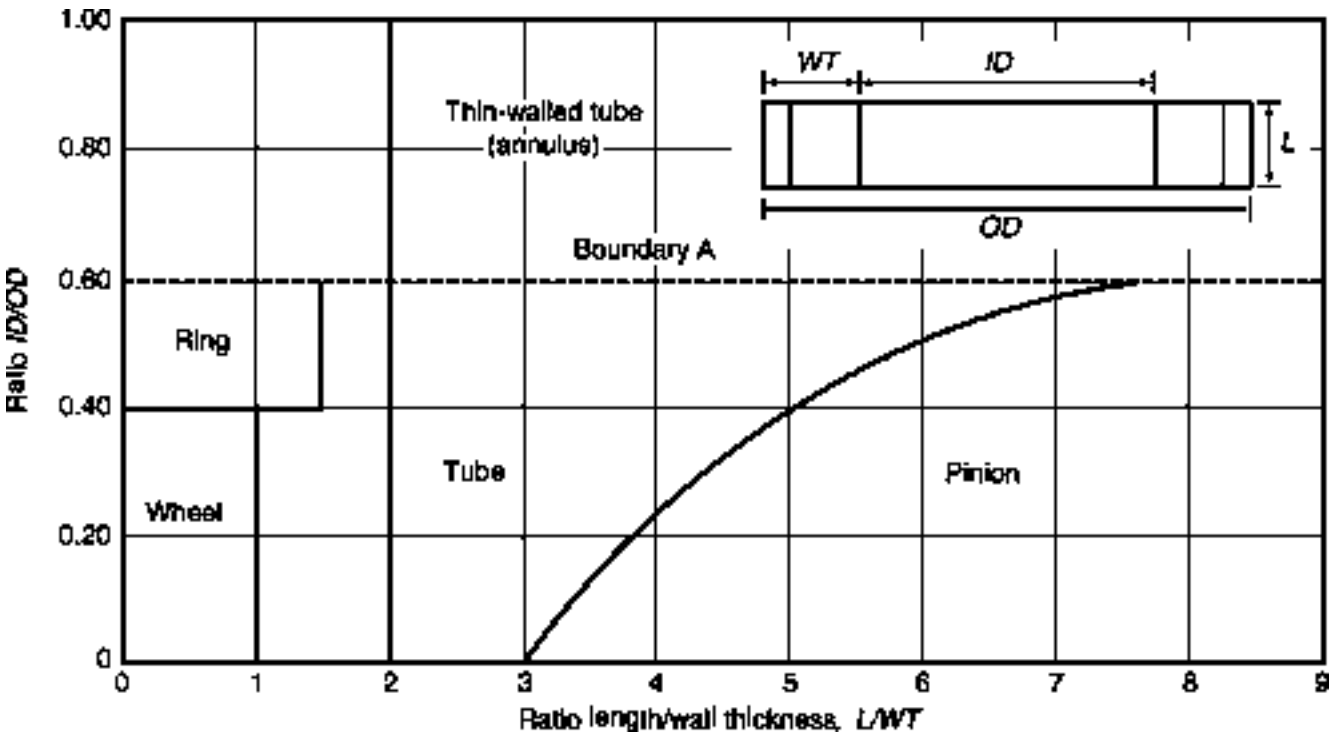


Fig. 38 Classification of distortion-sensitive shapes

Pinion shafts, as defined in Fig. 39, are susceptible to banding along their length, if they are improperly loaded into the furnace, as shown in Fig. 40 (Ref 21). When this occurs, the pinion shafts must then be straightened, which adds to production cost.

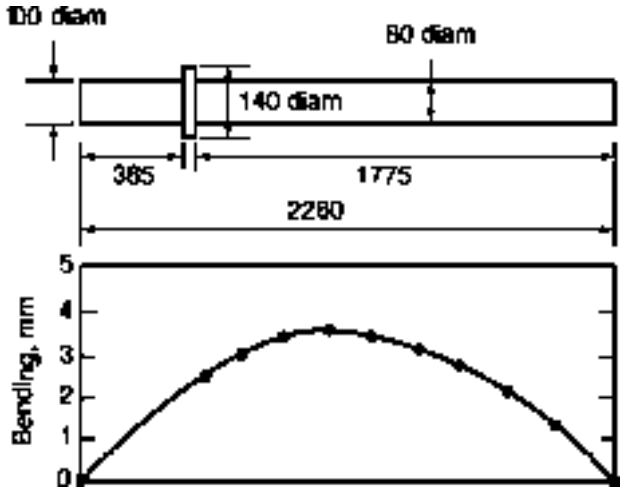


Fig. 39 Distortion of JIS SCM 440 (0.4% C, 1.05% Cr, 0.22% Mo) steel pinion shafts after oil quenching from 850 °C (1560 °F) while vertically suspended and tempering at 600 °C (1110 °F)

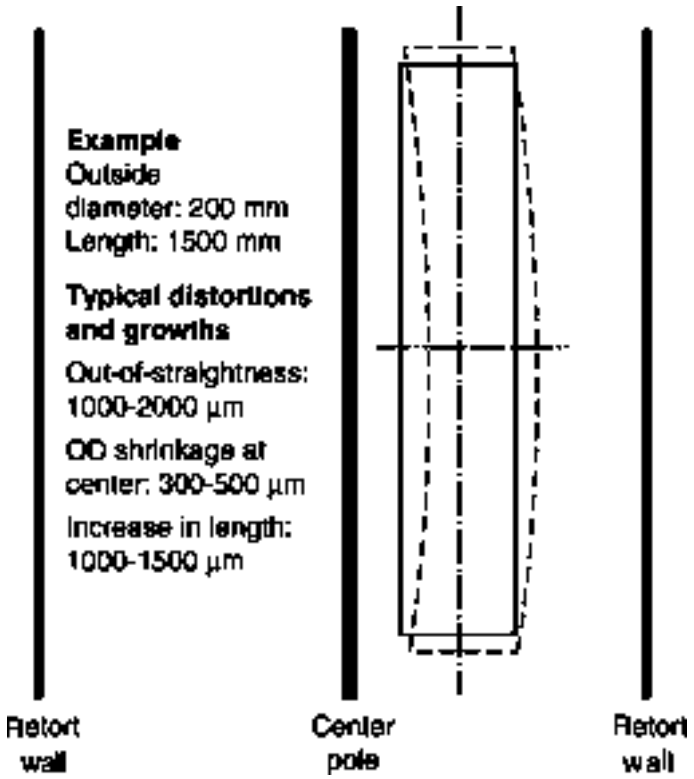


Fig. 40 Example of pinion shaft distortion due to furnace loading

Surface Condition. Quench cracking may be due to various steel-related problems that are only observable after the quench, but the root cause is not the quenching process itself. These prior conditions include prior steel structure, stress raisers from prior machining, laps and seams, alloy inclusion defects, grinding cracks, chemical segregation (bonding), and alloy depletion (Ref 27).

One prior-condition problem is tight scale, which is encountered with forgings hardened from direct-fired gas furnaces with high-pressure burners (Ref 23, 24). The effect of tight scale on the quenching properties of two steels, 1095 carbon steel and 18-8 stainless steel, is illustrated in Fig. 41 (Ref 25). These cooling curves were obtained by still quenching into fast oil. A scale of not more than 0.08 mm (0.003 in.) increases the rate of cooling of 1095 steel as compared to the rate obtained on a specimen without scale. However, a heavy scale (0.13 mm, or 0.005 in.) deep retards the cooling rate. A very light scale, 0.013 mm (0.0005 in.) deep, also increased the cooling rate of the 18-8 steel over that obtained with the specimen without scale.

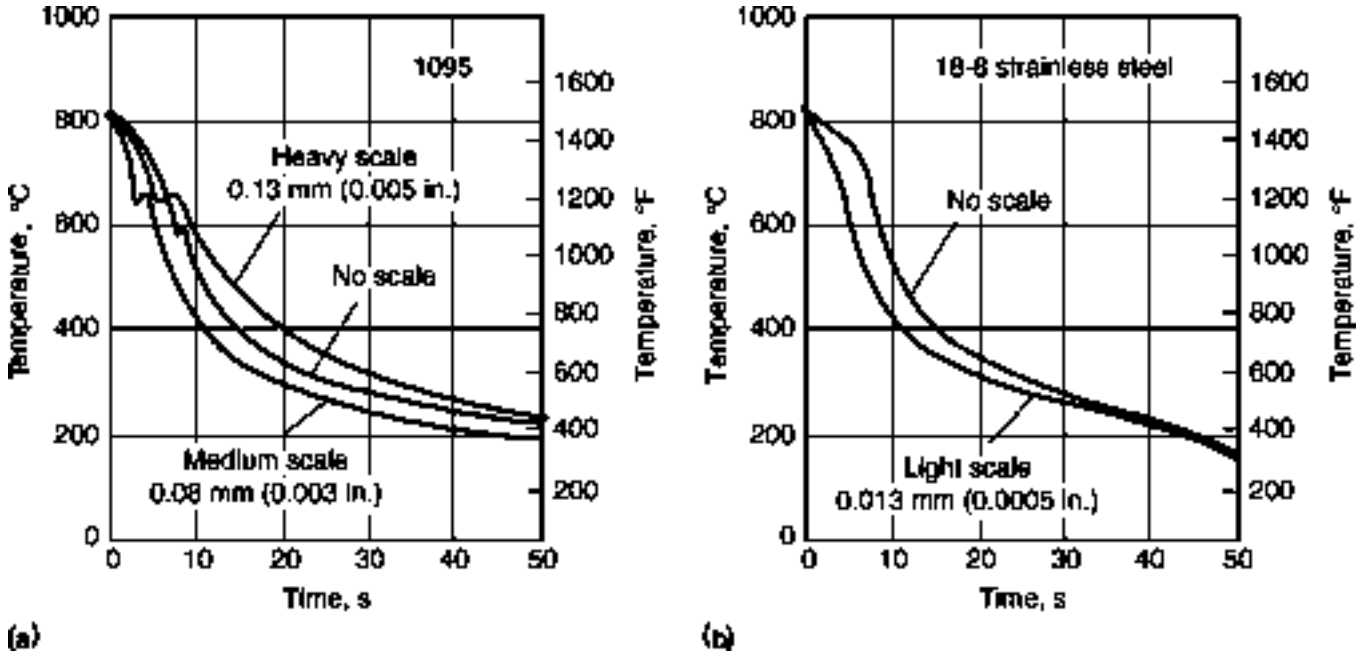


Fig. 41 Centerline cooling curves showing the effect of scale on the cooling curves of two different steels quenched in fast oil without agitation. (a) 1095 steel; oil temperature, 50 °C (125 °F). (b) 18-8 stainless steel; oil temperature, 25 °C (75 °F). Test specimens were 13 mm (0.5 in.) diam by 64 mm (2.5 in.) long

In practice, the formation of tight scale varies in depth over the surface of the part, resulting in thermal gradients due to differences in cooling rates. This problem may yield soft spots and uncontrolled distortion and is particularly a problem with nickel-containing steels. Surface oxide formation can be minimized by the use of an appropriate protective atmosphere.

The second surface-related condition is decarburization, which may lead to increased distortion or cracking (Ref 26). At a given depth within the decarburized layer, the part does not harden as completely as it would at the same point below the surface if there were no decarburization. This leads to nonuniform hardness, which may contribute to increased distortion and cracking, because (Ref 23):

- The decarburized surface transforms at a higher temperature than the core (the M_s temperature decreases with carbon content). This leads to high residual tensile stresses at the decarburized surface or a condition of unbalanced stresses and distortion.
- Because the surface is decarburized, it exhibits lower hardenability than the core. This causes the upper transformation products to form early, nucleating additional undesirable products in the core. The decarburized side is softer than the side that did not undergo decarburizing, which is harder. The greater amount of martensite leads to distortion.

The solution to this problem is to restore carbon into the furnace atmosphere or to remove the decarburized layer by machining.

Another surface-related condition that may lead to cracking or material weakening is the formation of surface seams or nonmetallic inclusions, which may occur in hot-rolled or cold-finished material. The presence of these defects prevents the hot steel from welding to itself during the forging process, for example, creating a stress raiser. To prevent this problem with hot-rolled bars, stock should be removed before heat treatment. Recommendations are provided in Table 7 (Ref 24).

Table 7 Minimum recommended material removal from hot-rolled steel products to prevent surface seam and nonmetallic stringer problems during heat treatment

Condition	Minimum material removal per side[(a)]	
	Nonresulfurized	Resulfurized
Turned on centers	3% of diameter	3.8% of diameter

	Minimum material removal per side[a]	
Centerless turned or ground	2.6%	3.4%

(a) Based on bars purchased to special straightness, i.e., 3.3 mm (0.13 in.) in 1.5 m (5 ft) maximum

A seam or nonmetallic depth of 0.025 mm (0.001 in.) per 3.3 mm (0.13 in.) diameter maximum is usually acceptable for cold-finished bars (Ref 24). If the seam depth is excessive, it is recommended that the bars be magnaflux inspected prior to heat treatment.

Stress Raisers from Prior Machining, Laps, and Seams. Surface conditions from prior machining conditions act as stress raisers, which are areas of dimensional changes (Fig. 42) (Ref 27). Examples of such stress raisers include fillets (Fig. 43), thread and gear roots and machining marks (Fig. 44 and 45 respectively), rolling seams (Fig. 46 and 47), and forging laps (Fig. 48 and 49). Forging laps are due to concentrations of oxides that are folded in during the forging process. The presence of these oxides prevents the hot steel from welding to itself during the forging process. This leads to cracking, as shown. (Note: The sample is first viewed in the unetched condition to locate the crack. Then, the sample is etched, if desired, for microstructural identification.)

Fig. 42 Micrograph of AISI 4140 steel as quenched and tempered. The microstructure is tempered martensite with quench cracking in the area of dimensional change. 91×; 2% nital etch. Source: Ref 27

Fig. 43 Micrograph of AISI 4142 steel as quenched and tempered. The microstructure is tempered martensite with quench cracking at the fillet radius. 100×; 3% nital etch. Source: Ref 27

Fig. 44 Micrograph of AISI 4140 steel as quenched and tempered. The microstructure is tempered martensite with quench cracking initiating from a machine groove. 100×; 2% nital etch. Source: Ref 27

Fig. 45 Micrograph of AISI 4118 carburized steel as quenched and tempered. The microstructure is tempered martensite (unetched) with a quench crack propagating from a machining burr. 200×. Source: Ref 27

Fig. 46 Micrograph of AISI 8630 steel as quenched. The microstructure is martensite, where cracking initiated from a rolling seam. Source: Ref 27

Fig. 47 Micrograph of type 403 stainless steel as quenched and tempered. The microstructure is predominantly tempered martensite, with cracking promoted by the seam. 100×; Vilella's reagent. Source: Ref 27

Fig. 48 Micrograph of AISI 1030 steel as direct-forge quenched and tempered. The microstructure is tempered martensite (unetched) with forged-in scale adjacent to cracking. 100×. Source: Ref 27

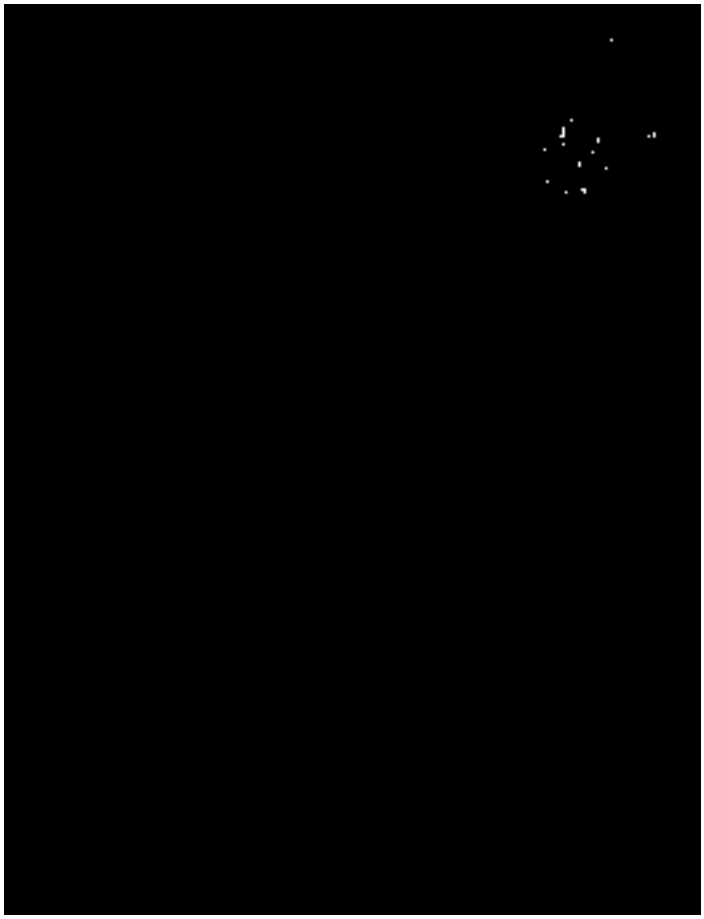


Fig. 49 Micrograph of AISI 1045 as-forged steel illustrating a forging lap. 27×; 2% nital etch. Source: Ref 27

Several types of internal voids may be present in steel. The first type is known as a pipe void, which is an imperfection formed after casting and cooling (see the article “Failures Related to Metalworking”, in this volume). The pipe void is a cavity formed by contraction during solidification of the last portion of liquid metal in the ingot and may even survive rolling or forging, but, because the ends of the ingot are closed, it is not detected until the bar stock is cut, at which time it is readily visible.

Another type of internal void is a forge or roll burst. The pipe void is readily distinguishable from the forge or roll burst by the more rounded nature of the pipe void compared to the thin, cracklike appearance of the forge or roll burst. Steel porosity also produces voids in steel castings. These voids are caused by trapped gases. Porosity is another source of potential steel cracking. This is illustrated in Fig. 50 (Ref 27).

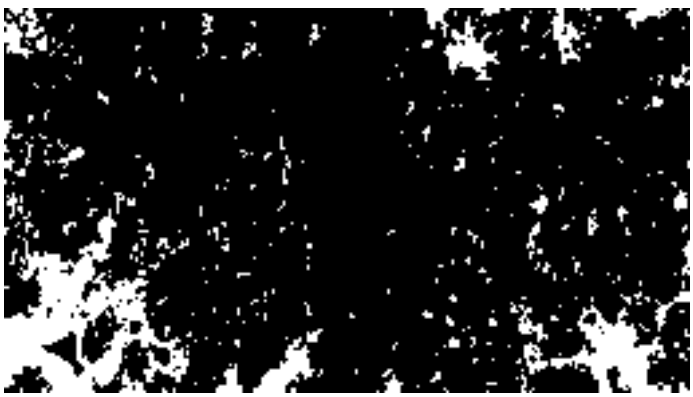


Fig. 50 Micrograph of AISI 8630 cast steel as quenched and tempered. The microstructure is tempered martensite, pearlite, and ferrite, showing a potential cracking condition. 91×; 3% nital etch. Source: Ref 27

Nonmetallic Inclusions. All steels contain numerous nonmetallic inclusions, but the cleaner grades have fewer large or significant inclusions than do the conventional grades. There are two types of nonmetallic inclusions. Exogenous nonmetallic inclusions occur when particles or large lumps of refractory fall into the molten steel during the steelmaking process. These contaminants provide surfaces in which indigenous inclusions may nucleate and grow. Indigenous inclusions arise from reactions, such as deoxidation products Al_2O_3 and SiO_2 or desulfurization products such as MnS (Ref 30).

Nonmetallic inclusions may exhibit degraded transverse toughness and ductility. They may also reduce fatigue strength, although soft inclusions are less harmful than are hard inclusions of the same size. In carburized parts, although nonmetallic inclusions may not significantly affect the fatigue limit, they could result. Figure 51 illustrates a combination of nonmetallic inclusions and banding due to microsegregation (discussed subsequently) (Ref 30). If a void is present at the site of a nonmetallic inclusion, possibly due to forging, the degree that the defect acts as a stress raiser increases. However, by some mechanism, such as shearing, the stress-raising effect may decrease, as illustrated by Fig. 52, where the nonmetallic inclusion appears as a “butterfly” after etching (Ref 30).

Fig. 51 Nonmetallic inclusions and banding in a microsegrated 1% C alloy steel showing retained austenite. Source: Ref 30

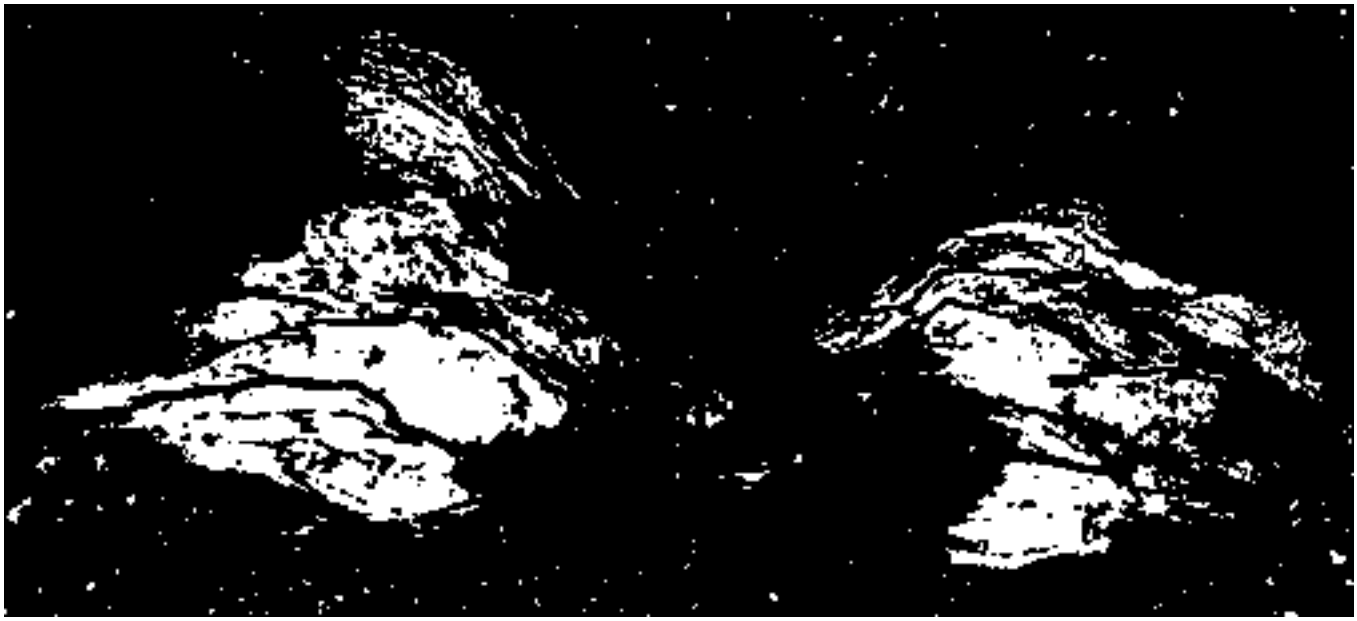


Fig. 52 White "butterfly" etching developed at a nonmetallic inclusion as a result of contact loading. 675×. Source: Ref 30

Machining. Material removal during machining can result in high residual stress levels and, ultimately, unacceptable distortion (Ref 21). When excessive machining stresses are imparted, the process may require modification to include a rough machining stress relieving followed by fine machining. Tables 8 and 9 are the suggested AISI design minimum allowable tolerances for machining (Ref 4).

Table 8 Minimum allowance for machining flat and square bars

Specified thickness, in., incl[(a)]		**Specified width, in., incl**										
		0 to $\frac{1}{2}$	**Over $\frac{1}{2}$ to 1**	**Over 1 to 2**	**Over 2 to 3**	**Over 3 to 4**	**Over 4 to 5**	**Over 5 to 6**	**Over 6 to 7**	**Over 7 to 8**	**Over 8 to 9**	**Over 9 to 12**
Hot-rolled square and flat bars												
0 to $\frac{1}{2}$	A	0.025	0.025	0.030	0.035	0.040	0.045	0.050	0.055	0.060	0.060	0.060
	B	0.025	0.035	0.040	0.050	0.065	0.080	0.095	0.105	0.120	0.130	0.140
Over $\frac{1}{2}$ to 1	A	…	0.045	0.045	0.050	0.055	0.060	0.065	0.070	0.075	0.075	0.075
	B	…	0.045	0.050	0.060	0.075	0.095	0.115	0.130	0.150	0.155	0.155
Over 1 to 2	A	…	…	0.065	0.065	0.070	0.070	0.075	0.080	0.080	0.095	0100
	B	…	…	0.065	0.070	0.085	0.105	0.125	0.145	0.165	0.170	0.170
Over 2 to 3	A	…	…	…	0.085	0.085	0.085	0.085	0.090	0.100	0.100	0.100
	B	…	…	…	0.085	0.100	0.120	0.135	0.155	0.170	0.190	0.190
Over 3 to 4	A	…	…	…	…	0.115	0.115	0.115	0.115	0.125	0.125	0.125
	B	…	…	…	…	0.115	0.125	0.140	0.170	0.190	0.190	0.190
Cold-drawn square and flat bars												
0 to $\frac{1}{2}$	A	0.025	0.025	0.030	0.035	0.040	0.045	…	…	…	…	…
	B	0.025	0.035	0.040	0.050	0.065	0.080	…	…	…	…	…
Over $\frac{1}{2}$ to 1	A	…	0.045	0.045	0.050	0.055	0.060	…	…	…	…	…
	B	…	0.045	0.050	0.060	0.075	0.095	…	…	…	…	…
Over 1 to 2	A	…	…	0.065	0.065	0.070	…	…	…	…	…	…
	B	…	…	0.065	0.070	0.085	…	…	…	…	…	…
Forged square and flat bars												

Specified thickness, in., incl[(a)]		Specified width, in., incl										
		0 to $\frac{1}{2}$	Over $\frac{1}{2}$ to 1	Over 1 to 2	Over 2 to 3	Over 3 to 4	Over 4 to 5	Over 5 to 6	Over 6 to 7	Over 7 to 8	Over 8 to 9	Over 9 to 12
Over $\frac{1}{2}$ to 1	A	…	…	0.060	0.065	0.065	0.075	0.080	0.085	0.090	0.100	0.110
	B	…	…	0.072	0.084	0.100	0.120	0.144	0.168	0.200	0.200	0.200
Over 1 to 2	A	…	…	0.090	0.090	0.090	0.100	0.110	0.115	0.125	0.140	0.150
	B	…	…	0.090	0.100	0.108	0.124	0.148	0.172	0.200	0.200	0.200
Over 2 to 3	A	…	…	…	0.120	0.120	0.125	0.130	0.135	0.150	0.160	0.175
	B	…	…	…	0.120	0.136	0.140	0.148	0.172	0.200	0.200	0.200
Over 3 to 4	A	…	…	…	…	0.150	0.150	0.160	0.180	0.200	0.210	0.225
	B	…	…	…	…	0150	0.150	0.160	0.180	0.200	0.210	0.225
Over 4 to 5	A	…	…	…	…	…	0.180	0.180	0.190	0.210	0.225	0.250
	B	…	…	…	…	…	0.180	0.180	0.190	0.210	0.225	0.250
Over 5 to 6	A	…	…	…	…	…	…	0.210	0.225	0.125	0.250	0.250
	B	…	…	…	…	…	…	0.210	0.225	0.225	0.250	0.250
Over 6 to 7	A	…	…	…	…	…	…	…	0.250	0.250	0.250	0.250
	B	…	…	…	…	…	…	…	0.250	0.250	0.250	0.250

(a) Minimum allowance, inches per side, for machining prior to heat treatment. A, allowance on thickness, in. per side; B, allowance on width, in. per side

Table 9 Minimum allowance per side for machining rounds, hexagons, and octagons prior to heat treatment

Ordered size, in., incl	Minimum allowance per side, in.			
	Hot rolled	Forged	Rounds rough turned	Cold drawn
Up to $\frac{1}{2}$	0.016	…	…	0.016
Over $\frac{1}{2}$ to 1	0.031	…	…	0.031
Over 1 to 2	0.048	0.072	…	0.048
Over 2 to 3	0.063	0.094	0.020	0.063
Over 3 to 4	0.088	0.120	0.024	0.088
Over 4 to 5	0.112	0.145	0.032	…
Over 5 to 6	0.150	0.170	0.040	…
Over 6 to 8	0.200	0.200	0.048	…
Over 8 to 10	0.200	0.072	…	…

References cited in this section

4. R.W. Bohl, Difficulties and Imperfections Associated with Heat Treated Steel, Lesson 13, Heat Treatment of Steel, Materials Engineering Institute Course 10, ASM International, 1978

18. W.T. Cook, A Review of Selected Steel-Related Factors Controlling Distortion in Heat-Treatable Steels, *Heat Treat. Met.*, Vol 26 (No. 2), p 27–36

19. Y. Toshioka, Heat Treatment Deformation of Steel Products, *Mater. Sci. Technol.*, Vol 1, Oct 1985, p 1883–1892

20. A. Thuvander and A. Melander, Calculation of Distortion during Quenching of a Low Carbon Steel, *First ASM Heat Treatment and Surface Engineering Conference in Europe (Proc.),* Trans Tech Publications, Part 2, 1992, p 767–782

21. P.C. Clarke, Close-Tolerance Heat Treatment of Gears, *Heat Treat. Met.,* Vol 25 (No. 3), 1998, p 61–64

22. R.F. Kern, Thinking Through to Successful Heat Treatment, *Met. Eng. Q.,* Vol 11 (No. 1), 1971, p 1–4

23. R. Kern, Distortion and Cracking, Part II: Distortion from Quenching, *Heat Treat.,* March 1985, p 41–45

24. R. Kern, Distortion and Cracking, Part III: How to Control Cracking, *Heat Treat.,* April 1985, p 38–42

25. H.-J. Chen and Z.W. Jiang, Microstructure Improvement and Low-Temperature Quenching of Dies Made of Cr12-Type Steel, *Jinshu Rechuli,* No. 8, 1992, p 39–41

26. H.-H. Shao, Analysis of the Causes of Cracking of a 12% Cr Steel Cold Die during Heat Treatment, *Jinshu Rechuli,* No. 11, 1995, p 43

27. R.R. Blackwood, L.M. Jarvis, D.G. Hoffman, and G.E. Totten, Conditions Leading to Quench Cracking other than Severity of Quench, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 575–585

28. F. Legat, "Why Does Steel Crack during Quenching," *Kovine Zlitine Technol.,* Vol 32 (No. 3–4), 1998, p 273–276

29. X. Cheng and S. He, Analysis of Quenching Cracks in Machine-Tool Pistons under Supersonic Frequency Induction Hardening, *Heat Treat. Met. (China),* No. 4, 1991, p 51–52

30. G. Parrish, *Carburizing: Microstructure and Properties,* ASM International, 1999

Quenching

Quenchant Selection and Severity

Quenchants must be selected to provide cooling rates capable of producing acceptable microstructure in the section thickness of interest. However, it is not desirable to use quenchants with excessively high heat-removal rates. Typically, the greater the quench severity, the greater the propensity to increased distortion or cracking. Although a reduction of quench severity leads to reduced distortion, it may also be accompanied by undesirable microstructures. Therefore, it is difficult to select an optimal quenchant and agitation. Cooling power (quench severity) of quenchant should be as low as possible while maintaining a sufficiently high cooling rate to ensure the required microstructure, hardness, and strength in critical sections of the steel parts.

Quench severity is defined as the "ability of a quenching medium to extract heat from a hot steel workpiece, expressed in terms of the Grossmann number (*H*)" (Ref 31). A typical range of Grossmann *H*-values (numbers) for commonly used quench media are provided in Table 10, and Fig. 53 provides a correlation between the *H*-value and the ability to harden steel, as indicated by the Jominy distance (*J*-distance) (Ref 23). Although Table 10 is useful to obtain a relative measure of the quench severity offered by different quench media, it is difficult to apply in practice, because the actual flow rates for "moderate," "good," "strong," and "violent" agitation are unknown.

Table 10 Grossmann *H*-values for typical quenching conditions

Quenching medium	Grossmann value

	(*H*)
Poor (slow) oil quench, no agitation	0.20
Good oil quench, moderate agitation	0.35
Very good oil quench, good agitation	0.50
Strong oil quench, violent agitation	0.70
Poor water quench, no agitation	1.00
Very good water quench, strong agitation	1.50
Brine quench, no agitation	2.00
Brine quench, violent agitation	5.00
Ideal quench	[a]

(a) It is possible, with high-pressure impingement, to achieve *H*-values greater than 5.00.

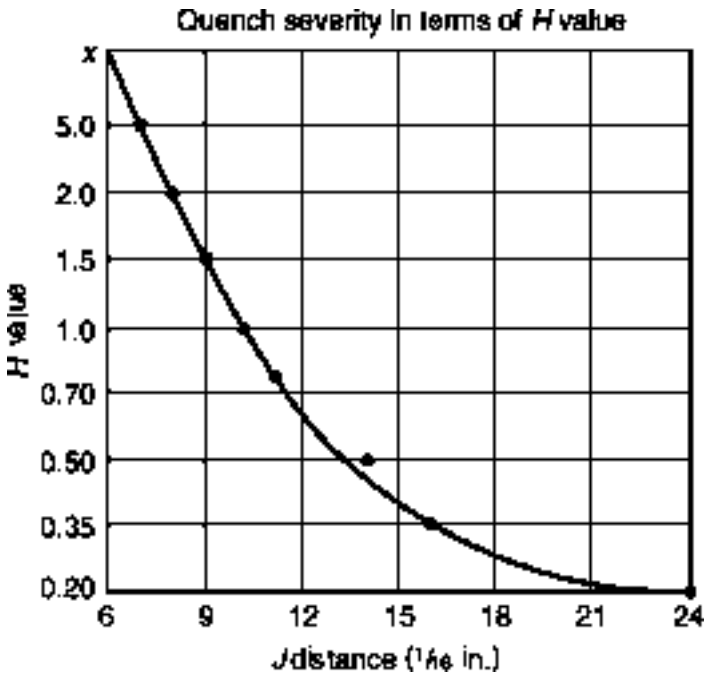


Fig. 53 Quench severity in terms of Grossman (*H*) values. *J*, Jominy distance. Source: Ref 23

Alternatively, the measurement of actual cooling rates or heat fluxes provided by a specific quenching medium does provide a quantitative meaning to the quench severity provided. Some illustrative values are provided in Table 11 (Ref 32).

Table 11 Comparison of typical heat transfer rates for various quenching media

Quench medium	Heat transfer rate ($W \cdot m^{-2} K^{-1}$)
Still air	50–80
Nitrogen (1 bar)	100–150
Salt bath or fluidized bed	350–500
Nitrogen (10 bar)	400–500
Helium (10 bar)	550–600
Helium (20 bar)	900–1000
Still oil	1000–1500
Hydrogen (20 bar)	1250–1350
Circulated oil	1800–2200
Hydrogen (40 bar)	2100–2300
Circulated water	3000–3500

Typically, the greater the quench severity, the greater the propensity of a given quenching medium to cause distortion or cracking. This usually is the result of increased thermal stress, not transformational stresses. Specific recommendations for quench media selection for use with various steel alloys is provided by standards such as Aerospace Material Specification (AMS) 2759. Figure 54 illustrates a quench crack due to an excessively high cooling rate (Ref 27).

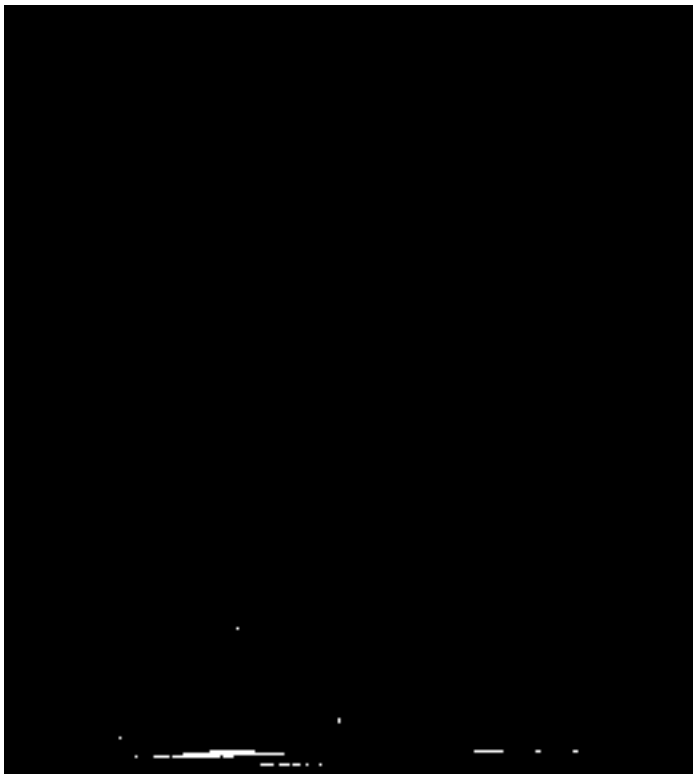


Fig. 54 Micrograph of AISI 4340 quenched and tempered steel illustrating a macroetched pure quench crack. Source: Ref 27

Some additional general comments regarding quenchant selection include (Ref 23, 28):

- Most machined parts made from alloy steels are oil quenched to minimize distortion. Most small parts or finish-ground larger parts are “free” quenched. Larger gears, typically those greater than 20 cm (8 in.), are fixture (die) quenched to control distortion. Smaller gears and parts, such as bushings, are typically plug quenched on a splined plug, which is usually constructed from carburized 8620 steel.
- Although a reduction of quench severity leads to reduced distortion, it may also be accompanied by undesirable microstructures, such as the formation of upper bainite (quenched pearlite) with carburized parts.
- Quench speed may be reduced by quenching in hot (570 to 750 °C, or 1060 to 1380 °F) oil. When hot oil quenching is used for carburized steels, lower bainite, which exhibits properties similar to martensite, is formed.
- Excellent distortion is typically obtained with austempering, quenching into a medium just above the M_s temperature. The formation of retained austenite is a significant problem with austempering processes. Retained austenite is most pronounced where manganese and nickel are major components. The best steels for austempering are plain carbon steels and chromium and molybdenum alloy steels (Ref 23).
- Aqueous polymer quenchants may often be used to replace quench oils, but quench severity is still of primary importance and appropriate quench system design is necessary.
- Gas or air quenching provides the least distortion and may be used if the steel has sufficient hardenability to provide the desired properties.
- Low-hardenability steels are quenched in brine or vigorously agitated oil. However, even with a severe quench, undesirable microstructures, such as ferrite, pearlite, or bainite, can form.

Quenchant Uniformity

Quench nonuniformity is one of the greatest contributors to quench cracking. Quench non-uniformity can arise from nonuniform flow fields around the part surface during the quench or non-uniform wetting of the surface (Ref 23, 33, 34, 35). Both lead to nonuniform heat transfer during quenching. Nonuniform quenching creates large thermal gradients between the core and the surface of the part. The effect of nonuniform quenching is illustrated with forged AISI 1045 crankshafts in Fig. 55 and 56 (Ref 27). Microstructural examination showed a mixture of nonuniform cross-sectional microstructures. Areas of tempered martensite adjacent to pearlite, bainite, acicular ferrite, and ferrite at prior austenite grain boundaries were observed.

Fig. 55 Micrograph of AISI 1045 steel as quenched and tempered. Microstructure shows bands with banded tempered martensite and some bainite. The crack profile revealed evidence of tempering oxide and secondary cracking. 200×; 2% nital etch. Source: Ref 27

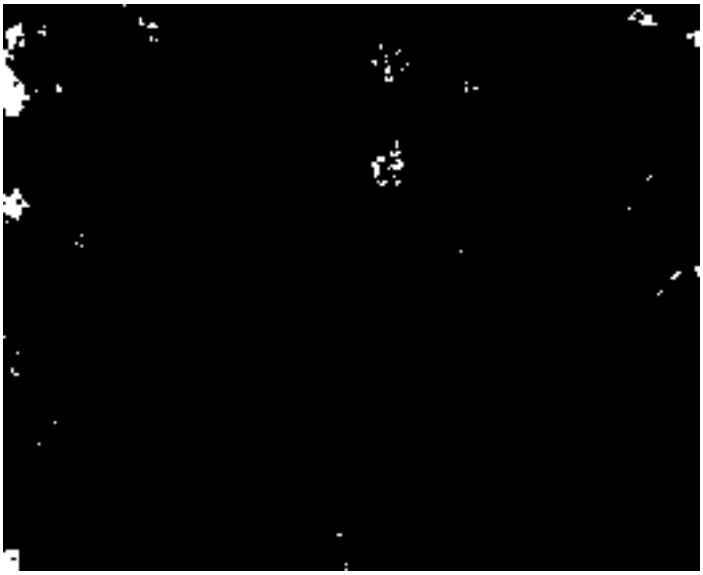


Fig. 56 Micrograph of AISI 1045 steel as quenched and tempered; representative of underheated microstructure adjacent to cracking. 376×; 2% nital etch. Source: Ref 27

Poor agitation design is a major source of quench nonuniformity. The purpose of the agitation system is not only to take hot fluid away from the surface and to the heat exchanger, but it is also to provide uniform heat removal over the entire cooling surface of all of the parts throughout the load being quenched. The batch quench system in Fig. 57 illustrates a system where axial (vertical) quenchant flow occurs throughout a load of round bars lying horizontally in a basket (Ref 22). In this case, the bottom surfaces of the bars experience greater agitation than the top surfaces. Cracks form on the upper surfaces because of the nonuniform heat loss. Agitation produces greater heat loss at the bottom, creating a large thermal gradient between the top and the bottom surfaces.

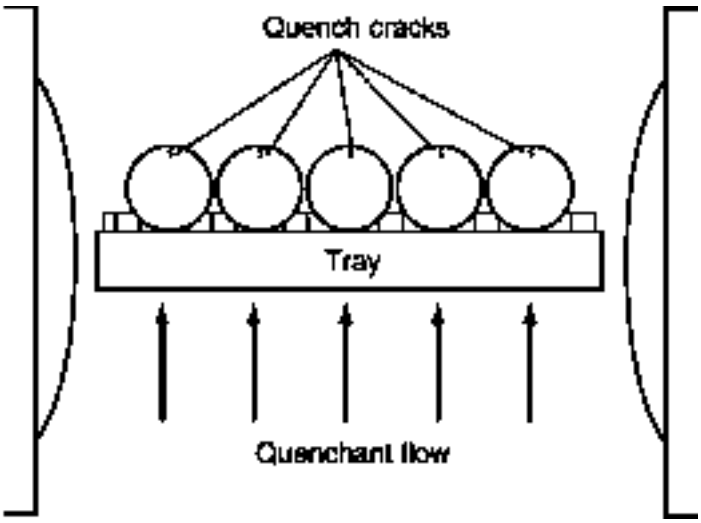


Fig. 57 Harmful effects of impeded vertical quenchant flow through the load of a batch quench system

If a submerged spray manifold is used to facilitate more uniform heat removal, the following design guidelines are recommended:

- The total surface of the part should experience uniform quenchant impingement.
- The largest holes possible (2.3 mm, or 0.09 in., minimum) should be used.
- The manifold face should be at least 13 mm (0.5 in.) from the surface of the parts being quenched.
- Repeated removal of hot quenchant and vapor should be possible.

Excessive distortion was also obtained with an agitation system, illustrated in Fig. 58, when the quenchant flow was either in the same direction relative to the direction of part immersion or in the opposite direction (Ref 35).

The solution to this problem was to minimize the quenchant flow to that required for adequate heat transfer during the quench and to provide agitation by mechanically moving the part up and down in the quenchant. Identifying sources of nonuniform fluid flow during quenching continues to be an important tool for optimizing distortion control and minimizing quench cracking.

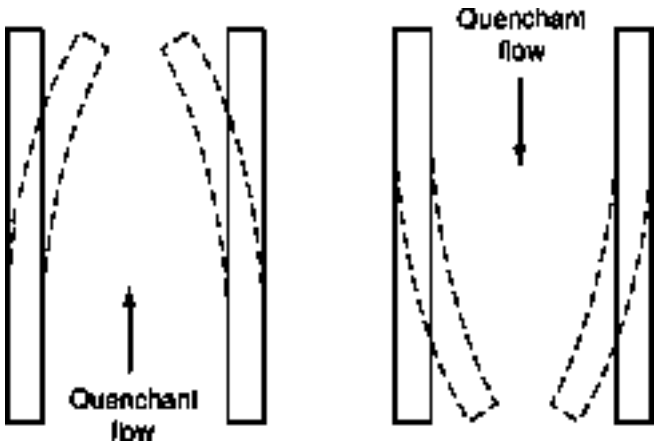


Fig. 58 Effect of quenchant flow direction on distortion

Nonuniform thermal gradients during quenching are also related to interfacial wetting kinematics, which is of particular interest with vaporizable liquid quenchants, including water, oil, and aqueous polymer solutions (Ref 36). Most liquid vaporizable quenchants exhibit boiling temperatures between 100 and 300 °C (210 and 570 °F) at atmospheric pressure. When parts are quenched in these fluids, surface wetting is usually time-dependent, which influences the cooling process and the achievable hardness.

Another major source of nonuniform quenching is foaming and contamination. Contaminants include sludge, carbon, and other insoluble materials. It includes water in oil, oil in water, and aqueous polymer quenchants. Foaming and contamination leads to soft spotting, increased distortion, and cracking.

Quench Distortion and Cracking

In the previous discussion, two quenchant-related phenomenon have been discussed: quench cracking and quench distortion. Although quench cracking can be eliminated, quench distortion cannot. Instead, the issue is distortion control, not elimination. Both quenching-related distortion control and quench cracking are discussed here.

Quench Distortion. One form of distortion that may occur on quenching is defined as shape distortion, which is dimensional variation that occurs when the parts are hot and the stress of their own weight leads to dimensional variation, such as bending, warpage, and twisting (Ref 34). Shape distortion is attributable to component support and loading, as discussed previously.

A second form of distortion is size distortion, which includes dimensional changes observable as elongation, shrinkage, thickening, and thinning. Size distortion is due to the volumetric variation that accompanies each of the transformational phases formed on quenching (Ref 34). Size distortion can be further classified as one-dimensional, two-dimensional, and three-dimensional. One-dimensional distortion is exemplified by the change in length when rods or wire are quenched and is primarily due to transformational distortion. Two-dimensional distortion is also primarily due to transformational distortion, and it is exemplified by size variation in two dimensions, such as would occur with sheet or plate stock on quenching. With bulk material that can undergo dimensional variation in three dimensions, the problem is due to both thermal and transformational distortion and is much more complicated (Ref 34). It has been shown that one of the primary contributors to quench distortion is uneven (nonuniform) cooling.

Quench Cracking. One major source of steel cracking is excessive cooling rates during the quench (quench severity). This is illustrated in Fig. 54 (Ref 27). Note that the crack passes straight from the surface to the core. Excessive cooling rates (high quench severity) produce greater thermal stresses in addition to greater transformation stresses. Steel cracking during transformation to martensite is largely due to the volumetric

increase that accompanies martensite formation. If the total residual stresses in the part are sufficiently high, quench cracking occurs.

Although quench distortion and cracking are most often due to nonuniform cooling, materials selection can be an important factor. The following materials selection rules are helpful:

- The compositional tolerances should be checked to ensure that the alloy is within specification.
- It is often better to choose a lower carbon content, because the higher carbon content often causes higher susceptibility for distortion and cracking.
- If possible, it is better to choose a combination of high-alloy steel and very slow cooling. As a matter of course, the selection of high-alloy steels markedly raises the material cost.

Cracking tendency decreases as the start of the martensitic transformation temperature (M_s) increases (Ref 37). Kunitake and Susigawa developed a relationship to interrelate the combined effect of both carbon content and elemental composition on the cracking propensity of steel. This was designated as the carbon equivalent (CE); the equation for calculating this value is provided in Fig. 59, where it is seen that quench cracks are prevalent at CE values above 0.525 (Ref 37).

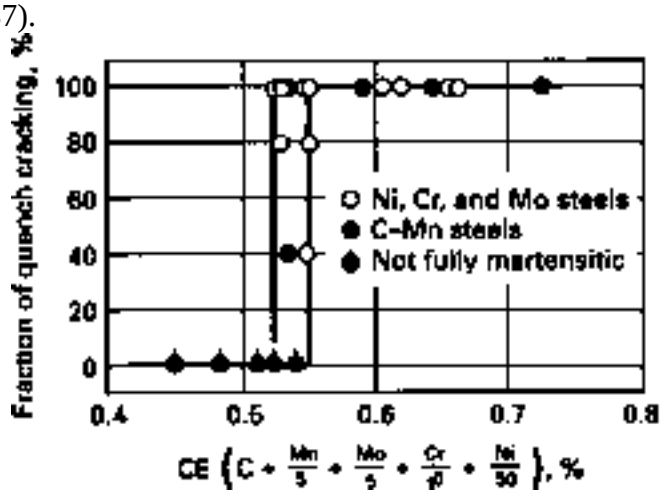


Fig. 59 Effect of M_s temperature and carbon equivalent (CE) on quench cracking of selected steels

Stresses that lead to cracking are thermal and often nonuniform cooling from T_A and M_s and transformational stress formed between M_s and martensite finish temperature (M_f). Nonuniform thermal stresses occur when nonuniform surface cooling occurs. In this case, there is a deferred contraction in the slower cooling areas, creating a pull stress resulting in pull cracking (Ref 36). This is illustrated in Fig. 60(a) (Ref 36), where rapid but nonuniform surface cooling to point A has occurred, at which time it affects adjacent point B, which is cooling more slowly, causing a deferred contraction at point B′—both of which occur while the martensitic transformation occurs at the M_s point. As a result, point A, which cooled rapidly, undergoes compression, and point B′, which cooled more slowly, undergoes pulling. This means that pull stress is a tensile stress, and if this stress becomes large enough, cracking occurs in areas where cooling is delayed.

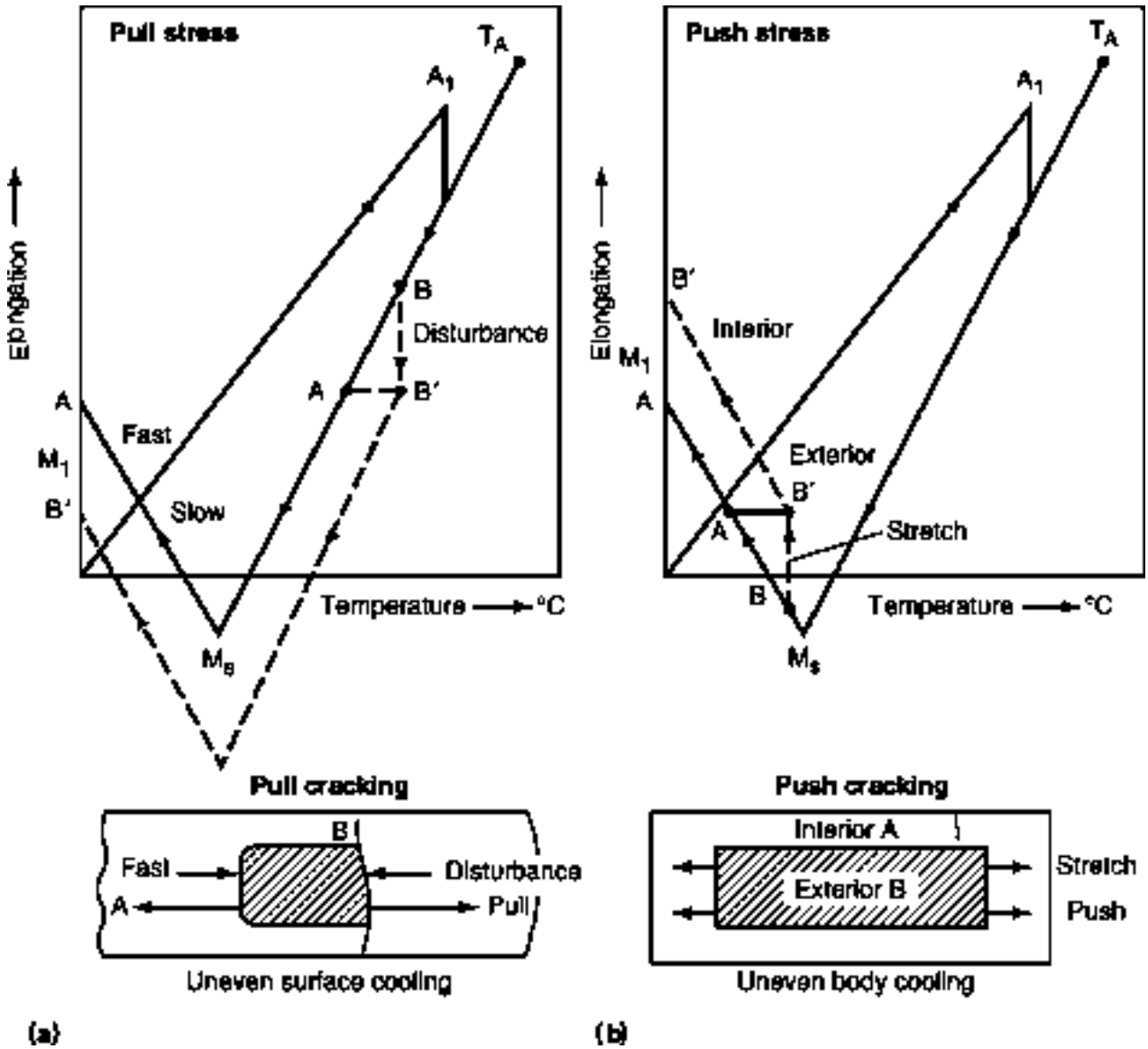


Fig. 60 Schematics of the mechanisms that generate (a) pull stress and pull cracking and (b) push stress and push cracking

When there is nonuniform cooling within the part between the M_s and M_f, there is a stretching or elongation in areas where the cooling is slow, which acts as a push stress, leading to push cracking (Ref 38). This is illustrated by Fig. 60(b) (Ref 36), where point A, which cooled rapidly, stretches point B (which cooled more slowly) to point B′ below the M_s temperature, which is a state that undergoes the expansion due to martensitic transformation. As a result, the rapidly cooled point A (exterior) undergoes pulling, and the slowly cooled point B′ (interior) undergoes compression. This results in interior point B′ causing pushing and cracking at point A, which occurs in areas where quenching occurs rapidly.

Therefore, pull cracking, which occurs with nonuniform surface cooling between T_A and M_s, and push cracking, which occurs with nonuniform cooling within the part between M_s and M_f, are the opposite of each other, although the cracking event takes place for both between the M_s and M_f. Figure 61 provides illustrations of both push and pull cracking (Ref 36).

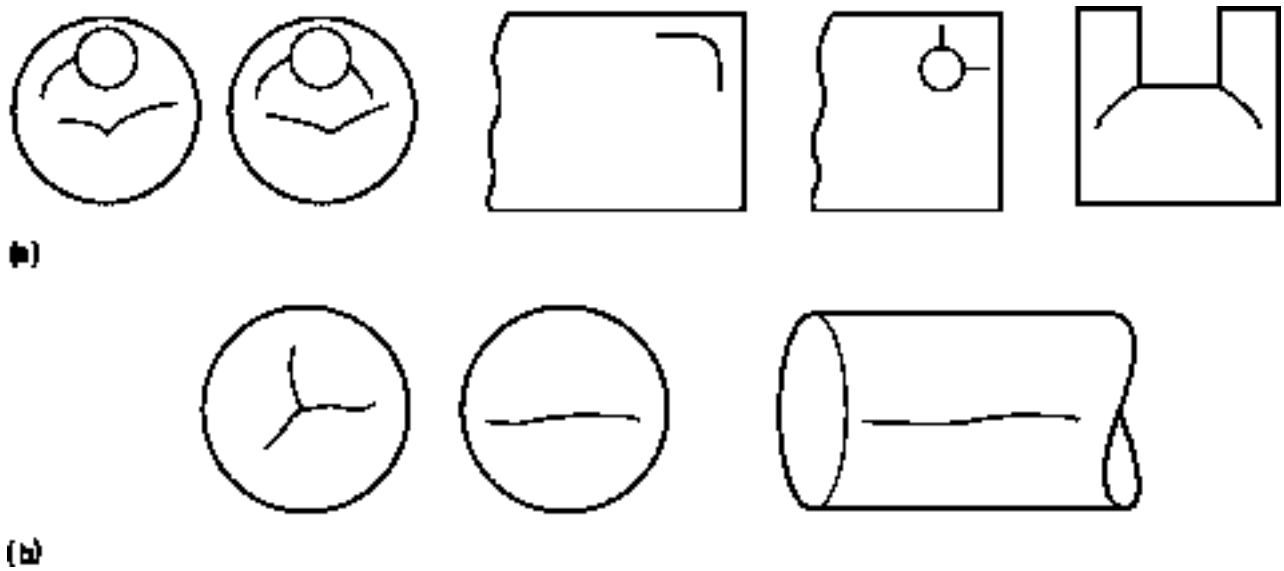


Fig. 61 Two forms of quench cracking. (a) Pull cracking (b) Push cracking

For cracking to occur, the presence of a stress raiser (points of stress concentration) is typically necessary. Two types of stress raisers may be found. One is a geometric notch, which includes cutting tool marks, sharp corner angles, and areas with rapid section thickness changes. The other type of stress raiser commonly found is a notch in the material, which may include intergranular effects, carbide segregation, and aggregates of impurities (Ref 36).

Although it would be desirable to eliminate all stress raisers to facilitate quench processing without any potential for cracking, this is not possible, in most cases. Instead, methods of quenching to minimize potential cracking often must be developed. Polyakov provided immersion recommendations for parts with complex shapes, where cracking potential is minimum and maximum, and these are summarized in Fig. 62 (Ref 39). The Polyakov rules for quenching such parts include (Ref 39):

- Conditions for quench cracking are most favorable if parts are immersed into the quenchant, so the perimeter of the stress raiser (P_{SC}) simultaneously touches the liquid along the entire length.
- In the case where the perimeter of the stress raiser slowly touches the quenchant at the moment of immersion in individual regions, quench cracking is reduced substantially.

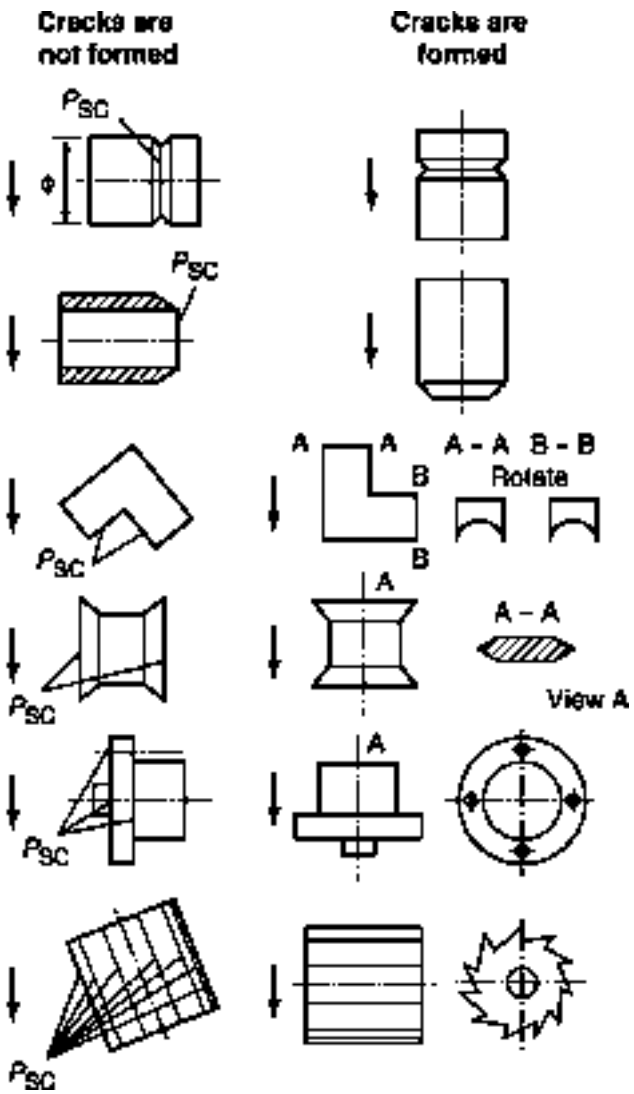


Fig. 62 Polyakov rules for immersion of cylindrical parts into a quenchant. The dashed horizontal lines indicate the level of the liquid; the vertical arrows indicate the direction of immersion of the part. P_{SC} indicates the perimeter of the stress concentrator.

Retained Austenite

The austenitic structure that remains in steel after quenching is called retained austenite. The amount of retained austenite is dependent on the M_s and M_f temperatures. To obtain complete transformation, steel must be quenched into the martensitic transformation region for sufficient time for austenite to transform to martensite. If the M_f temperature is below ambient temperature, the transformation to martensite is incomplete, and some austenite is retained in the as-quenched structure. In these cases, to obtain complete transformation, the steel must be quenched directly into a refrigerant to a temperature below the M_f. However, if steel is first quenched to ambient temperature and then cryogenically cooled, additional austenite is transformed to martensite, but there is still a small fraction of retained austenite (Ref 30).

Increasing carbon content of steel increases the potential for retained austenite on quenching, as illustrated in Fig. 63 (Ref 30). This is because the M_s temperature decreases with increasing carbon content. The effect of carbon content on the M_s temperature can be calculated using the Steven and Haynes equation for steels containing up to 0.5% C (Ref 40):

$$M_s(°C) = 561 - 474C - 33Mn - 17Ni - 17Cr - 21Mo$$

where carbon, manganese, nickel, chromium, and molybdenum are concentrations of these elements in percent. For carbon concentrations greater than 0.5%, corrections must be made using Fig. 64 (Ref 30).

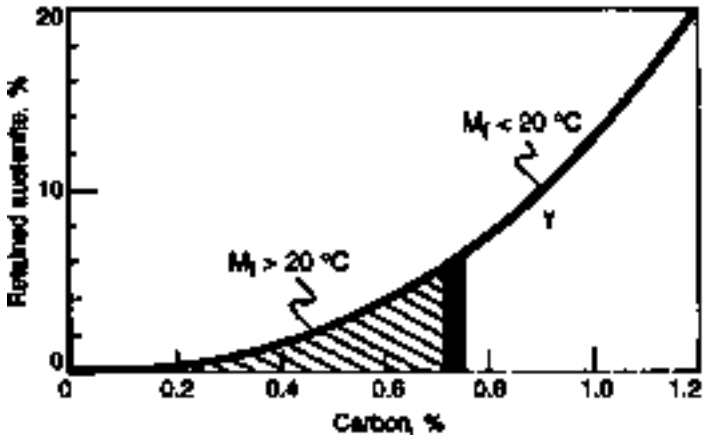


Fig. 63 Influence of carbon content on the formation of retained austenite. Source: Ref 30

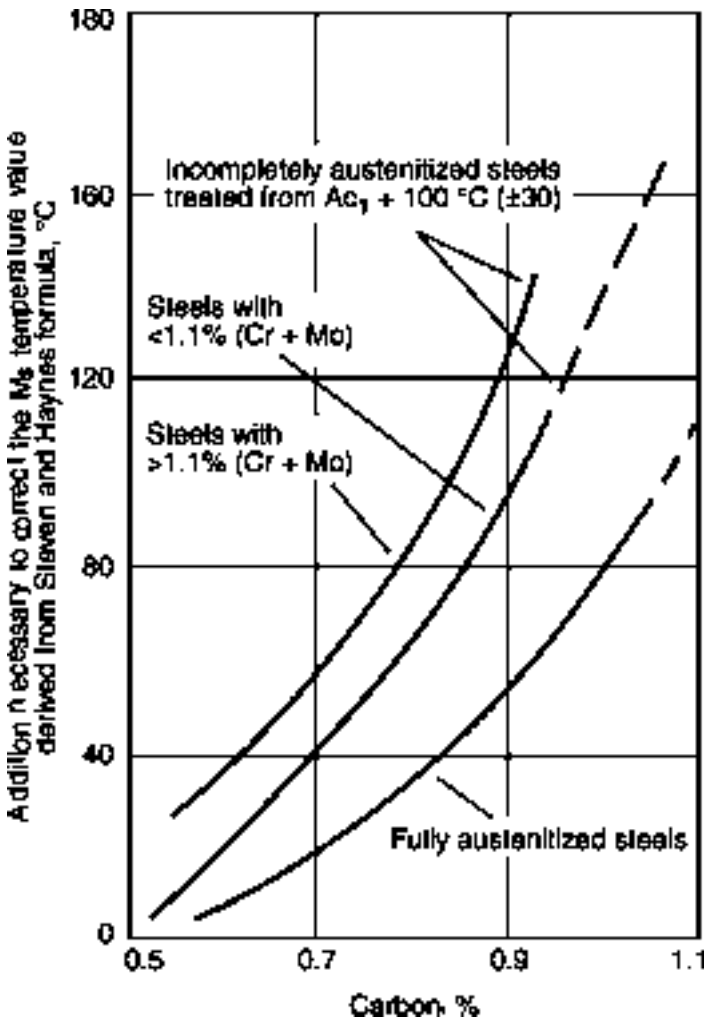


Fig. 64 Correction curves for use with the Steven and Haynes equation, with Parrish modifications indicated by dashed lines (see Ref 30). When the carbon content is less than 0.9%, an 830 °C (1530 °F) soak for greater than 2 h should produce a fully austenitic structure. Source: Ref 30

The M_f temperature is typically approximately 215 °C (385 °F) below the M_s (Ref 40). Typically, when the M_f temperature is less than the quenchant temperature, the transformation to martensite is incomplete. The volume of untransformed austenite (V_γ) is related to the M_s and the quenchant temperature (T_q) by the Koistinen and Marburger equation (Ref 41):

$$V_\gamma = \exp\left[-1.10 \times 10^{-2}(M_s - T_q)\right]$$

These equations are used to estimate the impact of steel chemistry and quenching conditions on the amount of retained austenite that may be expected.

Dimensional changes may occur slowly or fast and are due to the volume composition of the transformation products formed on quenching. One of the most important, with respect to residual stress variation, distortion, and cracking, is the formation and transformation of retained austenite. For example, the data in Table 12 indicate the slow conversion of retained austenite to martensite, which was still occurring days after the original quenching process for the two steels shown (Ref 18, 19). This is particularly a problem when dimensional control and stability is one of the primary goals of heat treatment. Therefore, microstructural determination is an essential component of any distortion-control process.

Table 12 Dimensional variation in hardened high-carbon steel with time at ambient temperature

Steel type	Tempering temperature		Hardness,	Change in length, % × 10^3			
	°C	°F	HRC	After 7 days	After 30 days	After 90 days	After 365 days
1.1% C tool steel (790 °C, or 1450 °F, quench)	None		66	-9.0	-18.0	-27.0	-40.0
	120	250	65	-0.2	-0.6	-1.1	-1.9
	205	400	63	0.0	-0.2	-0.3	-0.7
	260	500	61.5	0.0	-0.2	-0.3	-0.3
1% C/Cr (840 °C, or 1540 °F, quench)	None		64	-1.0	-4.2	-8.2	-11.0
	120	250	65	0.3	0.5	0.7	0.6
	205	400	62	0.0	-0.1	-0.1	-0.1
	260	500	60	0.0	-0.1	-0.1	-0.1

Figure 65 illustrates the microstructure of a nickel-chromium steel containing different amounts of retained austenite (Ref 30). The retained austenite resides in the interlath boundaries of martensite. The volume of retained austenite is dependent on the prior austenite grain size and the amount of austenite retained, which is dependent on the carbon content, alloying element concentration, and the quenching temperature.

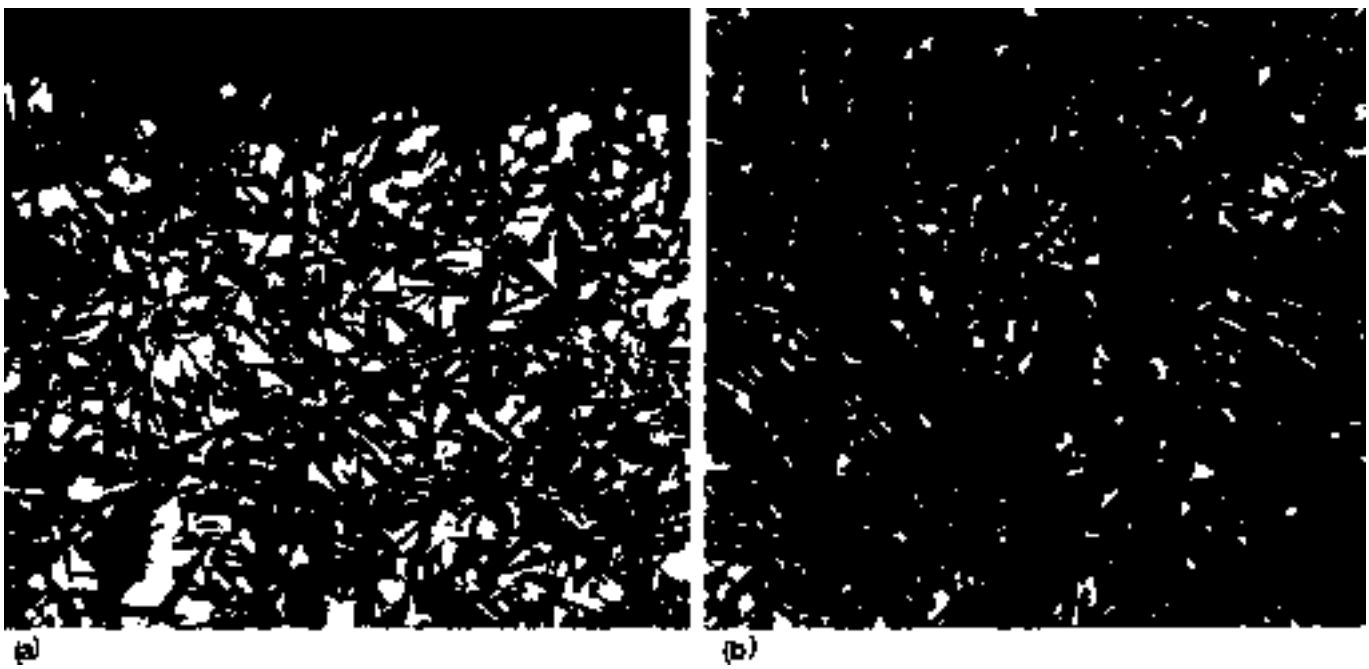


Fig. 65 Retained austenite (white) and martensite in the surfaces of carburized and hardened nickel-chromium steel testpieces. (a) Approximately 40% retained austenite. (b) Approximately 15% retained austenite. Both 550×. Source: Ref 30

The effect of retained austenite on the properties of steel can be summarized generally as (Ref 30):

- Retained austenite reduces hardness, which is proportional to the amount of retained austenite present.
- Tensile strength and yield strength decrease with increasing amounts of retained austenite.

- Because retained austenite relates to untransformed martensite, there is a reduction of maximum attainable surface compressive stresses that would have been attained if transformation to martensite was complete.
- Strength and compressive stresses are reduced by the presence of retained austenite, thus reducing fatigue resistance.
- Wear resistance decreases with increasing retained austenite.

The amount of retained austenite may be reduced by refrigeration to facilitate the transformation to martensite. Alternatively, surface-working methods, such as shot peening or surface rolling, may be used to effectively work harden the surface, resulting in higher compressive stresses.

Quenching Methods

Part design, materials selection, quenchant selection, and so on are important factors for suppressing quench distortion and cracking of steel parts. In addition, several methods for minimizing distortion and eliminating cracking may also be used, for example, interrupted quenching and press quenching.
Interrupted Quenching Techniques. Interrupted quenching refers to the rapid cooling of the steel parts from the austenitizing temperature to a temperature where it is held for a specified period of time, followed by slow cooling. There are several different types of interrupted quenching: austempering, marquenching (martempering), time quenching, and so on. The temperature at which the quenching is interrupted, the length of time the steel is held at temperature, and the rate of cooling can vary, depending on the type of steel and workpiece thickness.
Marquenching (martempering) is a term used to describe an interrupted quenching from the austenitizing temperature of steels. The purpose is to delay the cooling just above the martensitic transformation temperature until temperature equalization is achieved throughout the steel part. This minimizes the distortion, cracking, and residual stress. The term *martempering* is somewhat misleading, and the process is better described as marquenching. Figures 66(a) and (b) (Ref 42) show the significant difference between conventional quenching and marquenching. The marquenching of steel consists of:

- Quenching from the austenitizing temperature into a hot quenching medium (hot oil, molten salt, molten metal, or a fluidized particle bed) at a temperature usually above the martensite range (M_s point)
- Submerging in the quenching medium until the temperature throughout the steel workpiece is substantially uniform
- Cooling (usually in air) at a moderate rate to prevent large differences in temperature between the outside and the core of the workpiece

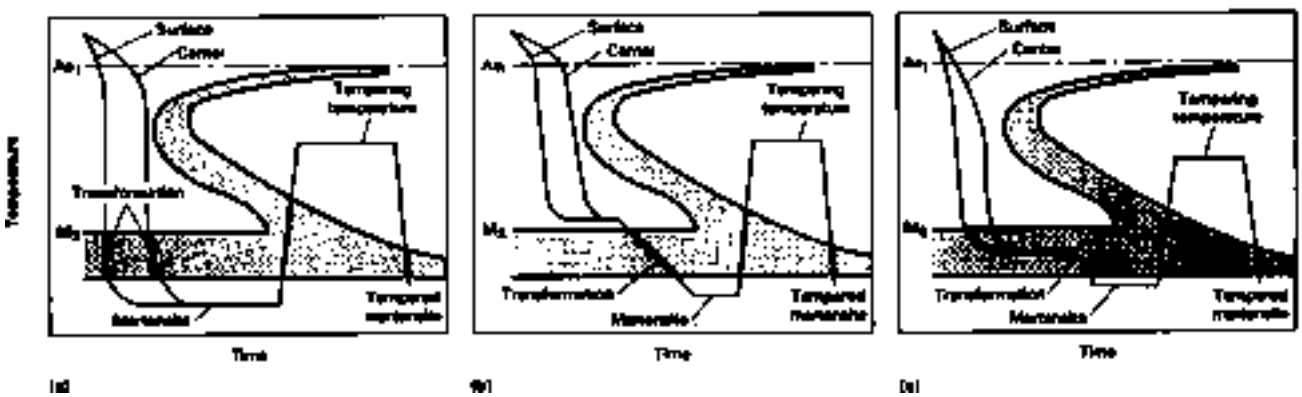


Fig. 66 Comparison of cooling curves as a workpiece cools into and through the martensite transformation range for a conventional quenching and tempering process and for interrupted quenching processes. (a) Conventional quenching and tempering. (b) Marquenching. (c) Modified marquenching. Ae_1, austenite start transformation under equilibrium conditions

The formation of martensite occurs fairly uniformly throughout the workpiece during the cooling to room temperature, thereby avoiding excessive amounts of residual stress. Straightening or forming is also easily

accomplished on removal from the marquenching bath while the part is still hot. The marquenching can be accomplished in a variety of baths, including hot oil, molten salt, molten metal, or a fluidized particle bed. Cooling from the marquenching bath to room temperature is usually conducted in still air. Deeper hardening steels are susceptible to cracking while martensite forms, if the cooling rate is too rapid. Marquenched parts are tempered in the same manner as conventional quenched parts. The time lapse before tempering is not as critical, because the stress is greatly reduced.

The advantage of marquenching lies in the reduced thermal gradient between surface and core as the part is quenched to the isothermal temperature and then is air cooled to room temperature. Residual stresses developed during marquenching are lower than those developed during conventional quenching, because the greatest thermal variations occur while the steel is in the relatively plastic austenitic condition, and because the final transformation and thermal changes occur throughout the part at approximately the same time. Marquenching also reduces susceptibility to cracking.

Modified marquenching differs from standard marquenching only in that the temperature of the quenching bath is below the M_s point (Fig. 66c) (Ref 42). The lower temperature increases the severity of quenching. This is important for steels of lower hardenability that require faster cooling in order to harden to sufficient depth. Therefore, modified marquenching is applicable to a greater range of steel compositions than is the standard marquenching.

Austempering consists of rapidly cooling the steel part from the austenitizing temperature to a temperature above that of martensite formation, holding at a constant temperature to allow isothermal transformation, followed by air cooling. The steel parts must be cooled fast enough so that no transformation of austenite occurs during cooling and then held at bath temperature long enough to ensure complete transformation of austenite to bainite. Molten salt baths are usually the most practical for austempering applications. Parts can usually be produced with less dimensional change by austempering than by conventional quenching and tempering. In addition, austempering can decrease the likelihood of cracking and can improve ductility, notch toughness, and wear resistance. However, austempering is applicable to a limited steel and parts size. Important considerations for the selection of steel parts for austempering are:

- The location for the nose of the time-temperature-transformation (TTT) curve and the speed of the quenching
- The time required for complete transformation of austenite to bainite at the austempering temperature
- The location of the M_s point
- Maximum thickness of section that can be austempered to a fully bainite structure

Carbon steels of lower carbon content are restricted to a lesser thickness. For 1080 steel, the maximum section thickness is approximately 5 mm (0.2 in.). Low-alloy steels are usually restricted to approximately 10 mm ($\frac{3}{8}$ in.) or thinner sections, while more hardenable steels can be austempered in sections up to 25 mm (1 in.) thick. Nevertheless, sections of carbon steel thicker than 5 mm (0.2 in.) are regularly austempered in production, when some pearlite is permissible in the microstructure (Ref 43).

Time quenching is one of the interrupted quenching methods and is used when the cooling rate of the part being quenched needs to be abruptly changed during the cooling cycle. The usual practice is to lower the temperature of the part by quenching in a medium with high heat-removal characteristics (for example, water) until the part has cooled below the nose of the TTT curve and then to transfer the part to a second medium (for example, oil, air, an inert gas), so that it cools more slowly through the martensite formation range. Time quenching is most often used to minimize distortion, cracking, and dimensional changes.

Restraint Quenching Techniques. Restraint quenching refers to quenching under restraint of distortion of steel, for example, quenching by using restraint fixtures, press quenching, cold die quenching, plague quenching, and so on. There are many round, flat, or cylindrical parts that distort to an unacceptable degree by conventional immersion quenching. Under such conditions, it is necessary to resort to special techniques. However, the equipment for those special techniques is expensive, and production rates are slow. Consequently, the resulting cost of heat treatment is relatively high. Therefore, use of these techniques should be considered only when minimal distortion is mandatory. Press quenching is closely related to intense quenching, which is discussed elsewhere in this article.

Restraint fixturing is costly and is used primarily for highly specialized applications. Representative examples are the quenching of rocket and missile casings or other large components with thin wall sections.

Press Quenching and Plug Quenching. Probably the most widely used special technique is press quenching. To minimize distortion caused by the quenching cycle, press and plug quenching dies must be made to provide the proper quenchant flow and hold critical dimensions of the part being quenched. In quenching, the die or plague contacts the heated part, and the pressure of the press restrains the part mechanically. This occurs before quenching begins, while the part is hot and plastic. The machine and dies then force the quenching medium into contact with the part in a controlled manner.

Cold Die Quenching. Thin disks, long thin rods, and other delicate parts that distort excessively by conventional immersion quenching can often be quenched between cold dies with no distortion. Cold die quenching is limited to parts with a large surface area and small mass, such as washers, rods of small diameter, thin blades, and so on. Large, thin thrust washers have to be flat after quench hardening, but considerable distortions are developed as a result of blanking and machining stresses. To ensure the required flatness, the washers are squeezed between a pair of water-cooled dies immediately after leaving the furnace.

Other quenching methods include selective quenching, spray quenching, and intensive quenching.

Selective quenching is used when it is desirable for certain areas of a part to be relatively unaffected by the quenching medium. This can be accomplished by insulating an area to be more slowly cooled, so the quenchant contacts only those areas of the part that are to be rapidly cooled. A clay-coating method may be employed for selective quenching, as in, for example, the water quenching of Japanese swords. The distortion and microstructure of Japanese swords is controlled by the thickness distribution of the clay layer. A thick clay layer (>0.1 mm, or 0.004 in.) suppresses the cooling rate by its insulating effect, and a thin clay layer increases the cooling rate during water quenching by its cooling accelerating effect (Fig. 67). Selective quenching is often effective for suppressing excessive distortion or cracking.

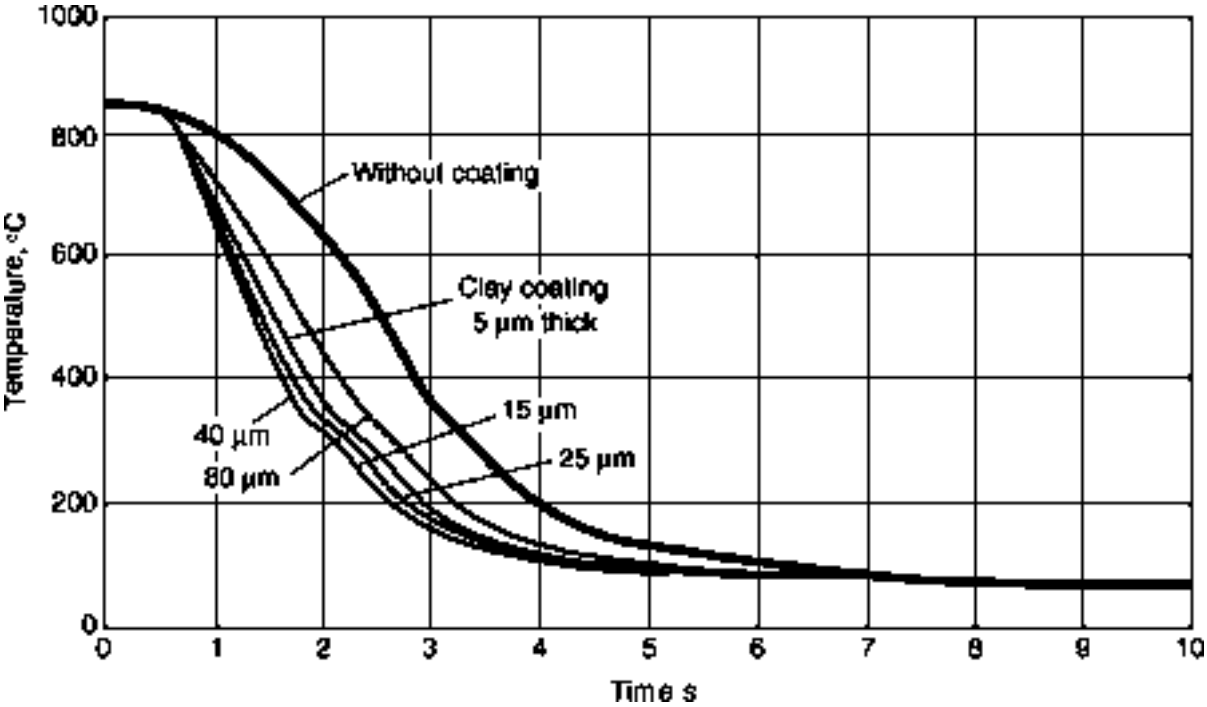


Fig. 67 Effect of clay coating on cooling curves of steel test specimen quenched into still water at 30 °C (86 °F). Test specimen, JIS S45C steel cylinder (10 mm, or 0.40 in., diam × 30 mm, or 1.2 in., long)

Spray quenching involves directing high-pressure streams of liquid onto the surfaces of steel parts. Spray quenching is often useful for minimizing distortion and cracking, because it can realize a uniform quenching by selecting optimal spray condition and the sprayed surfaces where higher cooling rates are desired. The cooling rate can be faster, because the quenchant droplets formed by the high-intensity spray impact the part surface and remove heat very effectively.

Fog quenching uses a fine fog or mist of liquid droplets in a gas carrier as the cooling agent. Although similar to spray quenching, fog quenching produces lower cooling rates because of the relatively low liquid content of the stream.

Intensive quenching (Ref 44) involves forced-convection heat exchange both during high-temperature cooling and also during the transformation of austenite into martensite. Intensive quenching promotes temperature uniformity during cooling through the cross section of the part. Cooling intensity should be sufficient to promote maximum surface compressive stresses. This occurs when the Biot number ≥18. The second criterion of intensive quenching is that the intensive cooling be stopped at the moment that maximum surface compressive stresses are formed.

References cited in this section

18. W.T. Cook, A Review of Selected Steel-Related Factors Controlling Distortion in Heat-Treatable Steels, *Heat Treat. Met.,* Vol 26 (No. 2), p 27–36

19. Y. Toshioka, Heat Treatment Deformation of Steel Products, *Mater. Sci. Technol.,* Vol 1, Oct 1985, p 1883–1892

22. R.F. Kern, Thinking Through to Successful Heat Treatment, *Met. Eng. Q.,* Vol 11 (No. 1), 1971, p 1–4

23. R. Kern, Distortion and Cracking, Part II: Distortion from Quenching, *Heat Treat.,* March 1985, p 41–45

27. R.R. Blackwood, L.M. Jarvis, D.G. Hoffman, and G.E. Totten, Conditions Leading to Quench Cracking other than Severity of Quench, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 575–585

28. F. Legat, "Why Does Steel Crack during Quenching," *Kovine Zlitine Technol.,* Vol 32 (No. 3–4), 1998, p 273–276

30. G. Parrish, *Carburizing: Microstructure and Properties,* ASM International, 1999

31. J.R. Davis, Ed., *ASM Materials Engineering Dictionary,* ASM International, 1992, p 407

32. P.F. Stratton, N. Saxena, and R. Jain, Requirements for Gas Quenching Systems, *Heat Treat. Met.,* Vol 24 (No. 3), 1997, p 60–63

33. H.M. Tensi, G.E. Totten, and G.M. Webster, Proposal to Monitor Agitation of Production Quench Tanks, *Heat Treating: Including the 1997 International Induction Heat Treating Symposium, Proc. of the 17th Conf.,* D.L. Milam, D.A. Poteet, G.D. Pfaffmann, V. Rudnev, A. Muehlbauer, and W.B. Albert, Ed., ASM International, 1997, p 423–441

34. S. Owaku, Quench Distortion of Steel Parts, *Netsu Shori (J. Jpn. Soc. Heat Treat.),* Vol 32 (No. 4), 1992, p 198–202

35. R.T. Von Bergen, The Effects of Quenchant Media Selection on the Distortion of Engineered Steel Parts, *Quenching and Distortion Control,* G.E. Totten, Ed., ASM International, 1992, p 275–282

36. H.M. Tensi, A. Stich, and G.E. Totten, Fundamentals of Quenching, *Met. Heat Treat.,* Mar/Apr 1995, p 20–28

37. T. Kunitake and S. Susigawa, *Sumitomo Search,* May 1971, p 16–25

38. V.D. Kalner and S.A. Yurasov, Internal Oxidation during Carburizing, *Met. Sci. Heat Treat. (USSR),* Vol 12 (No. 6), June 1970, p 451–454

39. A.A. Polyakov, Quenching Properties of Parts Having Stress Concentrators, *Met. Sci. Heat Treat.,* Vol 37 (No.7–8), 1995, p 324-325

40. A.G. Haynes, Interrelation of Isothermal and Continuous-Cooling Heat Treatments of Low-Alloy Steels and Their Practical Significance, *Heat Treat. Met.,* Special Report 95, The Iron and Steel Institute, 1966, p 13–23

41. D.P. Koistinen and R.E. Marburger, A General Equation Prescribing the Extent of the Austenite-Martensite Transformation in Pure Iron-Carbon Alloys and Plain Carbon Steels, *Acta. Metall.,* Vol 7, 1959, p 59–60

42. H. Webster and W.J. Laird, Jr., Martempering of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 137

43. J.R. Keough, W.J. Laird, Jr., and A.D. Godding, Austempering of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 152–163

44. N.I. Kobasko, M.A. Arpnov, G.E. Totten, and A.V. Sverdlin, Method and Equipment for Implementation of Intensive Quenching, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 616–621

Steel Transformation Products and Properties

Internal oxidation, also called subscale formation, is the formation of corrosion product particles below the surface of the metal. Internal oxidation has a deleterious effect on hardness, bending fatigue strength, residual stresses, and wear resistance.

Internal oxidation occurs preferentially with certain alloy constituents, primarily manganese, chromium, and silicon, by inward diffusion of oxygen, sulfur, and other gases to the site of oxidation during carburizing using an endothermic atmosphere at 900 to 950 °C (1650 to 1740 °F). An endothermic atmosphere (endogas) is produced by the controlled combustion of a hydrocarbon gas, such as natural gas and propane, with air. The oxygen required for internal oxidation is derived primarily from carbon dioxide and water vapor, the by-products of the combustion process. Figure 68 shows the oxidation potential of various alloying elements in an endothermic gas (Ref 30).

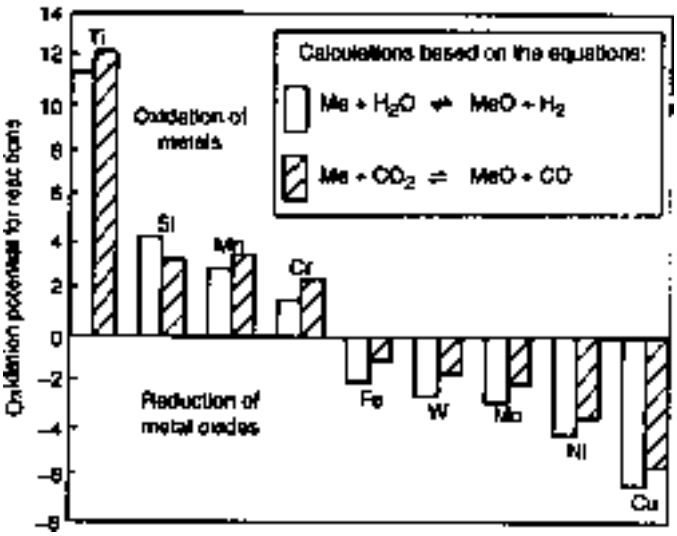


Fig. 68 Oxidation potential of alloying elements and iron in steel heated in endothermic gas with an average composition of 40% H_2, 20% CO, 1.5% CH_4, 0.5% CO_2, 0.28% H_2O (dewpoint, 10 °C, or 50 °F), and 37.72% N_2. Source: Ref 30

It is essentially impossible to eliminate internal oxidation when carburizing using an endothermic atmosphere. The presence of approximately 0.25% Si in carburizing grades of steel far exceeds the amount necessary to produce internal oxidation. In addition, the presence of silicon also reduces the threshold concentrations of manganese and chromium that produce internal oxidation. However, the sensitivity to internal oxidation may be decreased by the presence of alloying elements with atomic numbers greater than iron, such as nickel and molybdenum.

The oxygen content and depth of oxygen penetration increase with increasing oxygen potential (concentration) at a particular temperature. However, the depth of penetration increases with both temperature and time, as shown in Fig. 69 (Ref 30). Because the oxygen potential decreases with increasing carbon potential, the amount of internal oxidation is less at higher-carbon-potential carburizing processes, depending on the carburizing time.

Fig. 69 Internal oxidation of a nickel-chromium steel carburized in a laboratory furnace, showing both grain-boundary oxides and oxide precipitates within grains. 402×. Source: Ref 30

Metallographically, there are two common forms of internal oxidation observed. The first type appears as globular particles of approximately 0.5 μm in diameter from nearest to the surface to a depth of approximately 8 μm. Although these oxides may occur along the surface, they usually are found in the grain and subgrain boundaries and possibly within the grain. If the oxidation occurs on the surface, the grains are typically 0.5 to 1 μm, although they may be as large as 2 to 4 μm (Ref 38).

The second type of internal oxides occurs at greater depths, typically 5 to 25 μm and appear as a dark phase at prior austenite grain boundaries, as illustrated in Fig. 70 (Ref 30, 45). Thus, surface grain size also influences the amount of internal oxidation. As the grain size increases, the potential for internal oxidation within the grains increases (Ref 30).

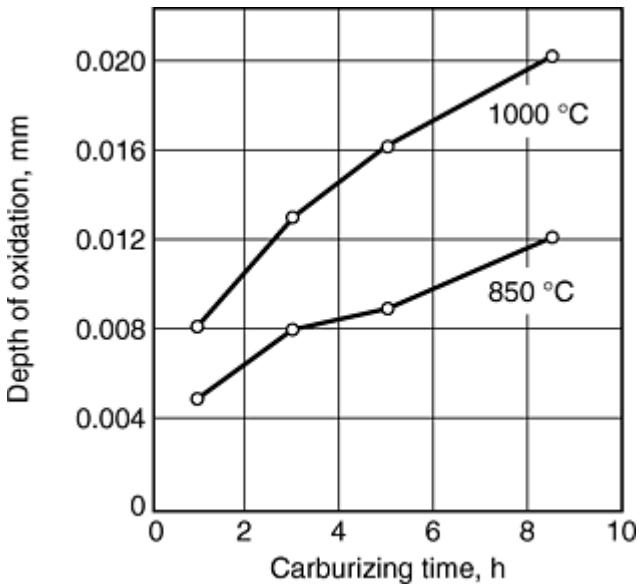


Fig. 70 Depth of oxidized zones as a function of carburizing time at different carburizing temperatures for SAE 1015. Source: Ref 30

Internal oxidation is also affected by the surface condition of the material being carburized, because organic impurities, such as lubricants, can contaminate the furnace atmosphere, and the presence of oxides due to scale or corrosion affect the formation and depth of penetration of internal oxidation. Although surface oxides may be removed by electropolishing, electrochemical machining, honing, grinding, shot blasting, or peening, these methods must be selected with care, because they may induce such negative effects as increased tensile stresses, risk of inducing decarburization during subsequent reheating processes, or these processes may not allow precise control.

Table 13 provides the maximum acceptable depth of internal oxidation relative to the case depth for carburized gears. (See also International Organization for Standardization (ISO) 6336-5.2; the equivalent grades are ML, MQ, and ME.) Note that grade 1 has no specification value. Also, recovery by shot peening is acceptable, with agreement of the customer.

Table 13 American National Standards Institute/American Gear Manufacturers Association (ANSI/AGMA) 2001-C95 standard values for acceptable depth of internal oxidation for carburized gears relative to case depth (tooth size)

Grade 2		**Grade 3**		**Case depth**	
mm	**in.**	**mm**	**in.**	**mm**	**in.**
0.0178	0.0007	0.0127	0.0005	<0.75	<0.030
0.0254	0.0010	0.0203	0.0008	0.75–1.50	0.030–0.059
0.0381	0.0015	0.0203	0.0008	1.50–2.25	0.059–0.089
0.0508	0.0020	0.0254	0.0010	2.25–3.00	0.089–0.118
0.0610	0.0024	0.0305	0.0012	>3.00	>0.118

Alloy Depletion. Another source of nonhomogeneous steel composition is alloy depletion. As with nonmetallic inclusions, alloy depletion leads to greater stresses and cracking, as illustrated in Fig. 71 (Ref 27). Examples of alloys particularly susceptible to alloy depletion are AISI 4100, 4300, and 8600 series.

Fig. 71 Micrograph of AISI 4140 steel as quenched and tempered; microstructure is tempered martensite, where cracking is promoted by alloy depletion. 91×; 2% nital etch. Source: Ref 27

High-temperature transformation products (**HTTP**) are products that form on the surface of a carburized part within the same area where internal oxidation occurs. High-temperature transformation products are nonmartensitic microstructures that include pearlite, upper and lower bainite, or mixtures of two or more of these products. A lean-alloy steel or large cross section favors pearlite at the surface, whereas higher-alloy steel or thinner cross section favors the formation of bainitic microstructures when quenched. However, if the

cooling rate is sufficiently fast or if sufficient quantities of reducing elements, such as nickel and molybdenum, are within the oxide-forming region, it is possible that no HTTP is formed (Ref 30, 46).
The formation of internal oxidation products effectively results in localized depletion of the elements that are preferentially oxidized, such as manganese and chromium. When HTTP are present, they often extend deeper than the internal oxidation, especially with leaner steel grades. The occurrence of HTTP results in lower hardness, reduction of surface compressive stresses, reduction of bending fatigue, and reduction of wear resistance.
The effects of HTTP can be controlled by (Ref 30):

- Using a steel grade with sufficient alloy content in the depleted layer to give martensite on quenching
- Reducing the amount of HTTP in the surface layer by a late increase of carbon potential or a late addition of ammonia to the furnace chamber
- Using a faster quench to suppress HTTP formation; however, beware of possible increased distortion
- Removing HTTP by grinding, grit blasting, and/or shot peening

Decarburization is defined as “a loss of carbon atoms from the surface of a workpiece, thereby producing a surface with a lower carbon gradient than at some short distance below the surface” (Ref 30). This is important, because the strength of steel is dependent on carbon content (carbides) within the steel structure. If surface carbon is greater than 0.6%, the surface hardness should be acceptable. If surface carbon is approximately 0.6% or less, all the main properties are adversely affected; for example, bending fatigue could be reduced by 50%.
The depth of decarburization is defined as “the thickness of the layer in which the structure differs significantly from that of the core.” Complete decarburization results in a completely ferritic microstructure within the surface layer. The depth of decarburization is the thickness of this layer. In the zone of partial decarburization, there is a gradual increase of carbon content within the ferritic layer to the core. Decarburization is observed metallographically. Figure 72(a) illustrates complete decarburization on the surface of 1018 steel. Figure 72(b) illustrates partial decarburization (Ref 30).

(a)

(b)

Fig. 72 Micrographs showing different degrees of decarburization. (a) Total decarburization caused by a furnace leak during gas carburizing of AISI 1018 steel. 500×; 1% nital etch. (b) Partially decarburized specimen. 190×. Source: Ref 30

Carbon may also be oxidized under the same conditions leading to metallic oxidation (scaling). Therefore, although decarburization may occur independently of scaling, they typically occur simultaneously. During carbon oxidation, gaseous by-products (CO and CO_2) are formed. Usually, the most important factor controlling the rate of decarburization is the diffusion rate of carbon within the metal that must replace the carbon lost due to volatilization.

Although decarburization may occur at temperatures above 700 °C (1290 °F), decarburization is typically most severe when steel is heated to temperatures above 910°C (1670 °F) (Ref 47). The depth of decarburization actually depends on both temperature and time, as shown in Fig. 73 (Ref 30).

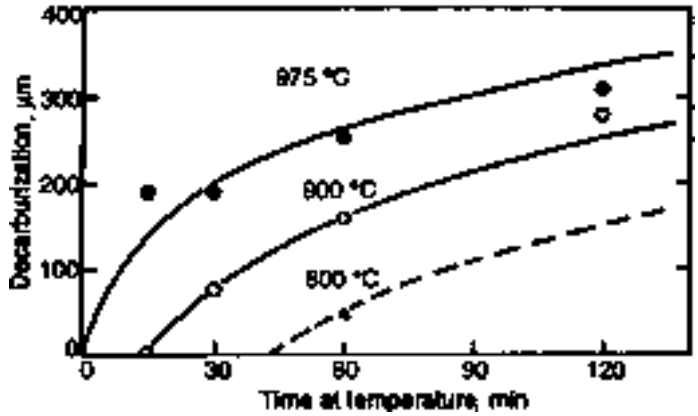


Fig. 73 Depth of decarburization of a cold-worked steel in a fluidized bed in air. Source: Ref 30

Exothermic atmosphere (exogas) is used as a protective atmosphere to prevent decarburization, scaling, and other undesirable surface reactions. To achieve the best results, the dewpoint must be minimized. This is accomplished by drying the gas. The relationship between dewpoint and carbon potential (concentration) is shown in Fig. 74 (Ref 30).

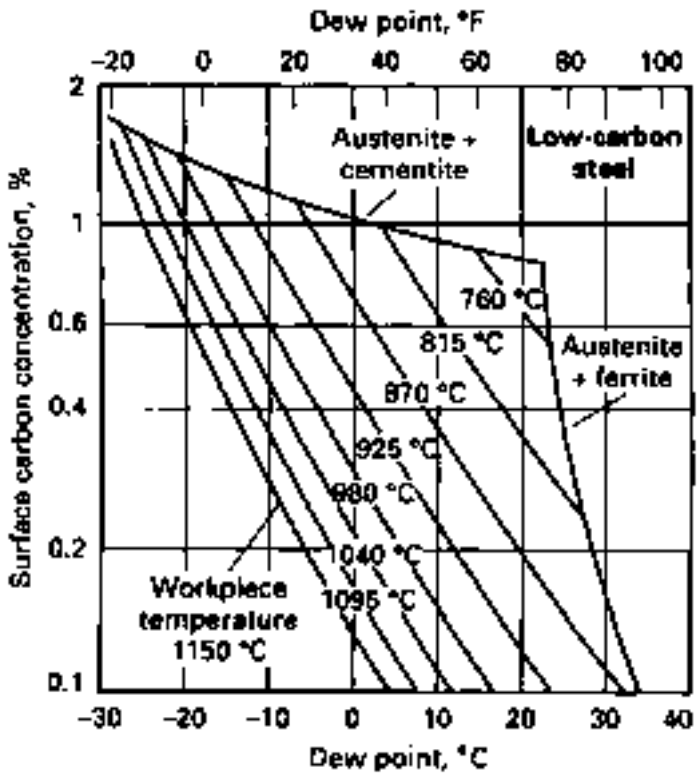


Fig. 74 Variation of carbon potential with dewpoint for an endothermic-based atmosphere containing 20% CO and 40% H_2 in contact with plain carbon steel at various steel temperatures. Source: Ref 30

The effects of decarburization may be corrected, depending on the degree of decarburization. Shot peening may be used for shallow decarburization or partial decarburization. For deeper total decarburization or partial decarburization, a process called restoration carburizing may be used, but it is usually accompanied by some additional distortion.

A specification for decarburization is ANSI/AGMA 2001-C95. There are no specification limits for grade 1. For grades 2 and 3, no partial decarburization is apparent in the outer 0.13 mm (0.005 in.), except in unground

roots. Another standard for decarburization is ISO 6336-5.2. For MQ and ME grades, the reduction of surface hardness due to decarburization in the outer 0.1 mm (0.004 in.) should not exceed 2 HRC on the test bar.

Carbides. When the carbon content is greater than the eutectoid carbon content, free carbides may form when the steel is heated above the austenite formation temperature (Ac_1) or when they are rejected by austenite during cooling and before the austenite-pearlite, austenite-bainite, or austenite-martensite transformations occur. Depending on the steel composition and processing conditions, free carbides may appear metallographically as networks, spheroids or large chunky particles, or as surface films (Ref 30). Different geometric models for carbide formation are shown schematically in Fig. 75 (Ref 48).

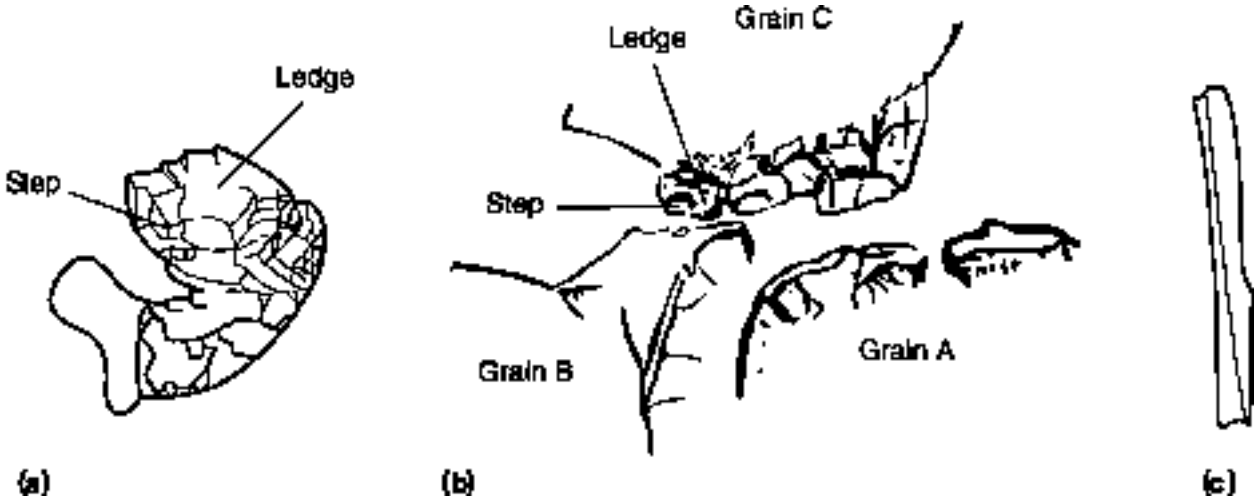


Fig. 75 Geometric models of carbides formed during case hardening. (a) Massive carbide grain, 4000×. (b) Film carbide, 2000×. (c) Intergranular carbide, 4000×. Source: Ref 30

So-called "massive carbides," as illustrated in Fig. 76, may be formed under isothermal conditions at austenite grain junctions. Under these conditions, carbides are formed in three directions by step and ledge growth mechanisms, where ledge heights are much greater than step heights, as illustrated in Fig. 75(a).

Fig. 76 Micrograph of 4% Ni-C-Cr carburized steel showing massive carbides produced during carburizing with surface carbon above Ac_{cm} carbon. Source: Ref 30

After diffusion to austenite grain boundaries below the Ac_{cm} temperature, excess carbon is deposited to form a film of small, thin platelets, which are illustrated in Fig. 77 (Ref 30). Although the ledges and steps are of

heights similar to those producing massive carbides (Fig. 75(b)), they differ by the temperatures at which they are formed. These carbides may appear as isolated grain-boundary particles forming a row with a film length that is equivalent to the prior austenite grain boundary. This film carbide is a continuous carbide network and is illustrated in Fig. 78 (Ref 30). A fine grain size may reduce the amount of carbon deposited at the grain boundaries. Lean-alloy grades with high manganese contents may be more prone to developing network carbides.

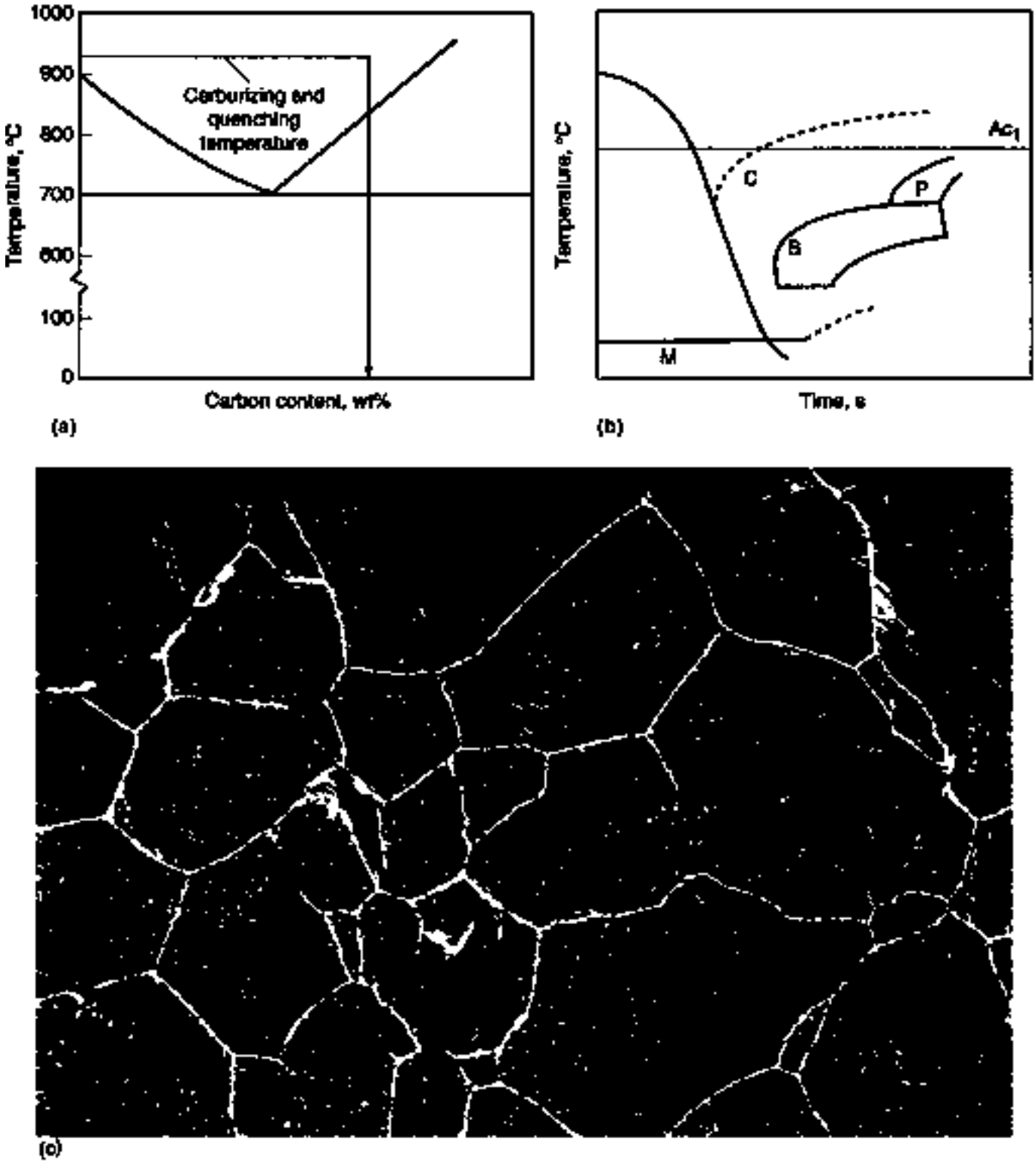


Fig. 77 Direct quenching from carburizing temperature. (a) Phase diagram schematic. (b) Continuous cooling transformation curve for a high-carbon surface. (c) Micrograph of direct quenched 3% Ni-Cr carburized steel. 280×. Source: Ref 30

Fig. 78 Micrograph of 4% Ni steel showing carbides formed after slow cool from carburizing temperature. 178×. Source: Ref 30

Intergranular carbides, which precipitate on planes within the grains, exist in a cylindrical pattern and a smooth spiraling surface, as shown in Fig. 75(c).

Network and dispersed carbides, while harder than martensite, typically exist in concentrations <10% to exhibit any significant effect on macrohardness. Surfaces containing network carbides exhibit tensile macroresidual stresses (Ref 49). However, spheroidal carbides exhibit compressive macroresidual stresses (Ref 49).

Continuous network carbides reduce bending fatigue strength, whereas partial carbide networks do not exhibit any significant effect (Ref 50). Even under controlled carburizing conditions, carbide buildup can occur at external edges and corners. For example, corner carbides can reduce bending fatigue strength, depending on the concentration and type of carbide formed. Interestingly, the effect can be eliminated by rounding the corners and edges before carburizing (Ref 30). Increasing network carbide concentrations improved high-stress/low-cycle rolling-contact fatigue (Ref 50).

Globular carbide dispersions, such as those illustrated in Fig. 79, are formed by slowly heating a steel surface through the Ac_1 to the ferrite-to-austenite tranformation completion (Ac_3) temperatures in the presence of a carburizing atmosphere (Ref 30). Film carbides or flake carbides, illustrated in Fig. 80 (Ref 30), exist as a continuous or discontinuous film on the surface of a carburized part. In gas carburizing, this carbide structure forms due to cooling from the carburizing temperature in an atmosphere with a high carbon potential.

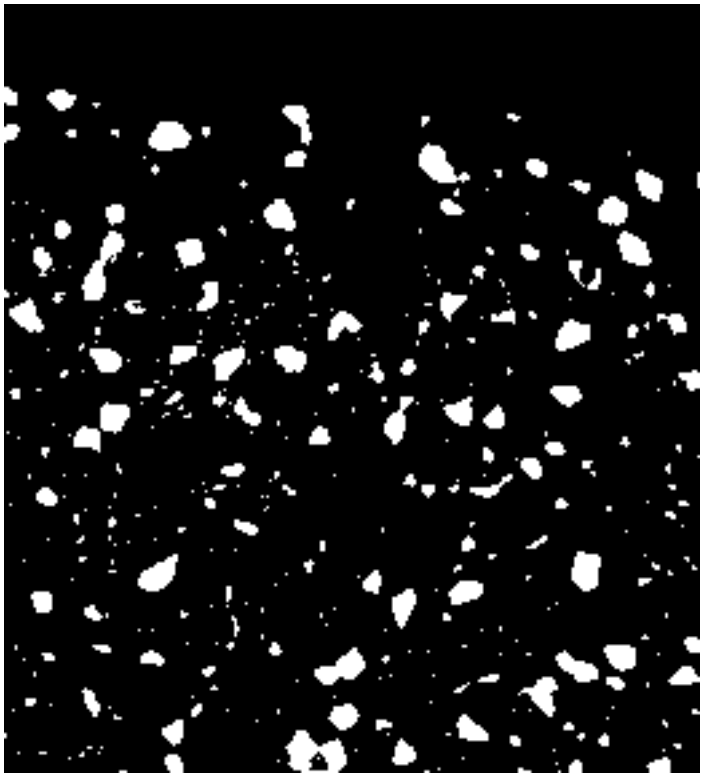


Fig. 79 Globular carbides at the surface of a carburized 1% Cr-Mo steel (reheat quenched). 836×. Source: Ref 30

Fig. 80 Surface film carbide (1% Cr-Mo steel). 874×. Source: Ref 30

Because carbides are hard, high surface-carbide concentrations should exhibit hardness values significantly higher than noncarbide-containing surfaces. The term *supercarburizing* is a carburizing process conducted at 930 °C (1710 °F) in a high carbon potential with subsequent subcritical annealing to a temperature just below the Ac_1 temperature; the process is repeated and finally quenched and tempered (Ref 51). This process results in surface carbon contents of 1.8 to 3% and greater than 25% carbides. Supercarburized parts exhibit up to 15%

better bending fatigue properties. Coarse carbide structure may result in significantly reduced contact fatigue (Ref 30). Wear resistance generally increases with increasing carbide concentration.

Carbide particles may be unavoidably formed during reheat-quenched carburized alloy steels. These dispersed, fine particles generally do not produce detrimental effects and may even be unavoidable in reheat-quenched carburized alloy steels. However, massive carbides, which may be formed at grain boundaries in the outer 0.05 mm (0.002 in.) at corners and edges, may be detrimental. Network carbides, which are formed during posthardening processes, can be detrimental to properties.

References cited in this section

27. R.R. Blackwood, L.M. Jarvis, D.G. Hoffman, and G.E. Totten, Conditions Leading to Quench Cracking other than Severity of Quench, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 575–585

30. G. Parrish, *Carburizing: Microstructure and Properties,* ASM International, 1999

38. V.D. Kalner and S.A. Yurasov, Internal Oxidation during Carburizing, *Met. Sci. Heat Treat. (USSR),* Vol 12 (No. 6), June 1970, p 451–454

45. I.Y. Arkhipov, V.A. Batyreva, and M.S. Polotskii, Internal Oxidation of the Case on Carburized Alloy Steels, *Met. Sci. Heat Treat. (USSR),* Vol 14 (No. 6), 1972, p 508–512

46. W.E. Dowling, W.T. Donlon, W.B. Copple, and C.V. Darragh, Fatigue Behavior of Two Carburized Low-Alloy Steels, *Proc. of Second International Conference on Carburizing and Nitriding with Atmospheres,* ASM International, 1995, p 55–60

47. B. Liscic, Heat Treatment of Steel, *Steel Heat Treatment Handbook,* G.E. Totten and M.A.H. Howes, Ed., Marcel Dekker, 1997, p 527–662

48. R. Ruxanda and E. Florian, Microscopic Observations of the Carbide/Matrix Interphasic Interface in a Low-Alloyed Hypercarburized Steel, *Proc. of Second International Conference on Carburizing and Nitriding with Atmospheres,* ASM International, 1995, p 117–121

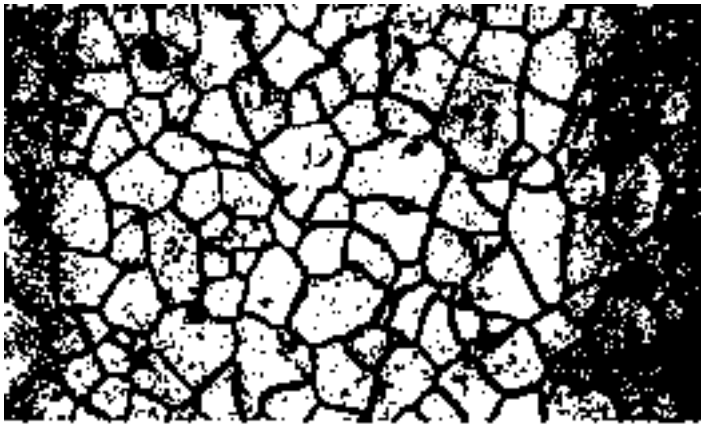
49. J. Wang, Z. Qin, and J. Zhou, Formation and Properties of Carburized Case with Spheroidal Carbides, *5th International Congress on the Heat Treatment of Materials,* (Budapest), International Federation for Heat Treating and Surface Engineering (Scientific Society for Mechanical Engineering), 1986, p 1212–1219

50. J.G. Wang, Z. Feng, and J.C. Zhou, Effect of the Morphology of Carbides in the Carburizing Case on the Wear Resistance and the Contact Fatigue of Gears, *J. Mat. Eng. (China),* Vol 6, Dec 1989, p 35–40

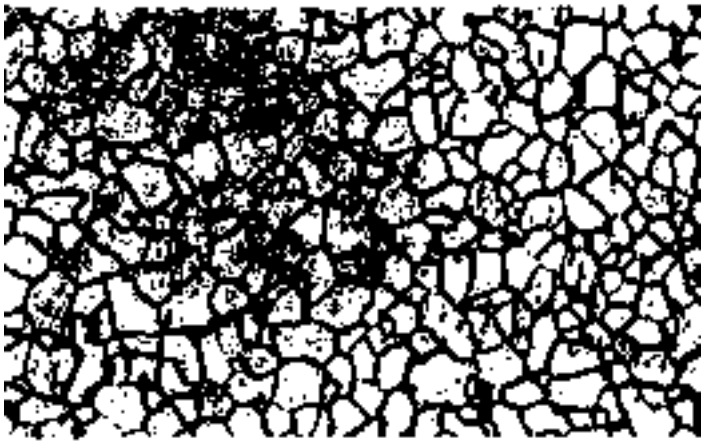
51. R. Kern, Supercarburizing, *Heat Treat.,* Oct 1986, p 36–38

Influential Microstructural Features

Grain Size. The mechanical properties and toughness of steel, both in the case and core if the steel is carburized, are not only affected by carbon and alloying element concentration and microstructure but also by the grain size. Typically, finer grain sizes are preferred. The ASTM grain size characterization comparison is illustrated in Fig. 81 (Ref 30).

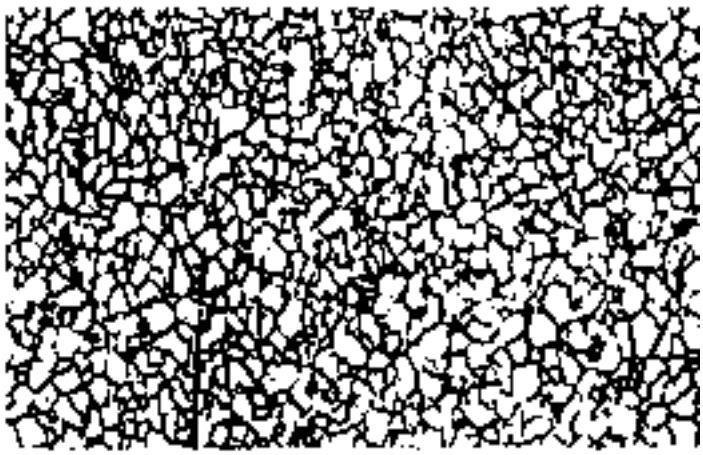


ASTM No. 6



ASTM No. 7

ASTM No. 8



ASTM No. 9

Fig. 81 Comparison of ASTM 6 to 9 grain size microstructures. 100×; nital etch. Source: Ref 30

Figure 82 illustrates the effect of increasing temperature on grain size (Ref 47). The following are reasons for grain growth when the temperature was increased from 850 to 1250 °C (1560 to 2280 °F) (Ref 52):

- Grain resorption occurs at 50 to 100 °C (90 to 180 °F) above Ac_3
- New boundaries and grains are formed at 250 to 300°C (450 to 540 °F) above Ac_3
- Although boundary migration occurs at all temperatures, it only affects grain growth above 1100 °C (2010 °F).

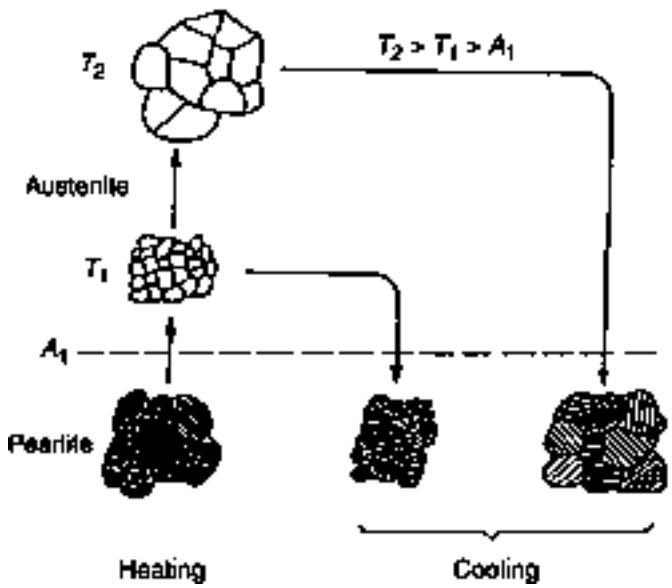


Fig. 82 Schematic illustration of the effect of austenitizing temperature on grain size of a eutectoid steel after normalizing. Source: Ref 47

If the fractured surface of a steel tool or machine component is coarse grained, this often indicates the use of excessively high austenitizing temperatures during heat treatment, or that an incorrect heat treatment for a given steel grade was used. This is a problem, because many properties of steel are affected by grain size.
Grain growth is affected by steel composition. Steels containing alloying elements such as aluminum nitride (AlN), which inhibits grain growth, are called fine-grained steels. Fine-grained steels do not exhibit grain growth, except at relatively high austenitizing temperatures when grain growth occurs rapidly, as illustrated in Fig. 83 (Ref 30). Coarse-grained steels contain silicon, and typically, the grain size increases slowly with increasing temperature, as illustrated in Fig. 83 (Ref 30).

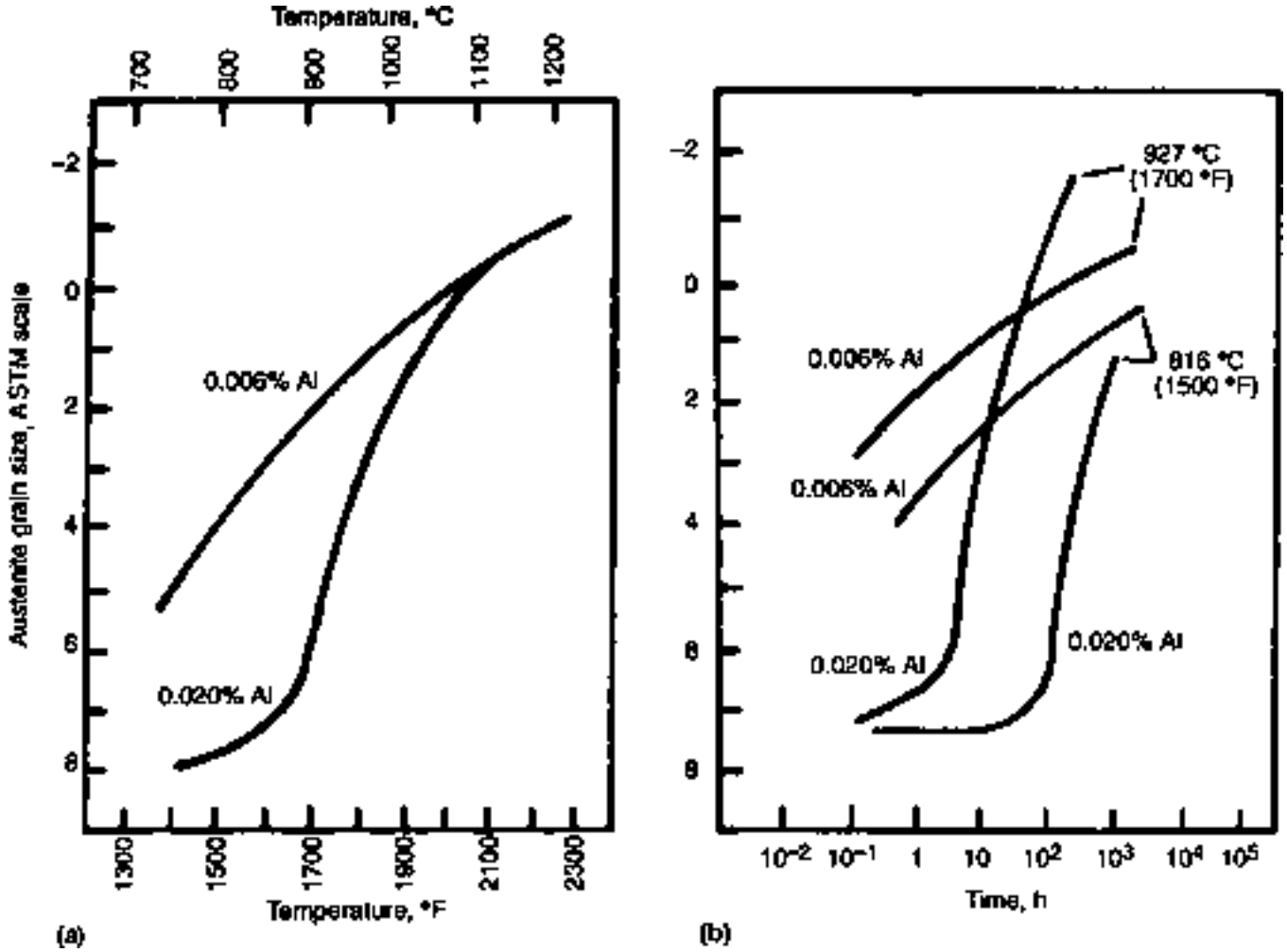


Fig. 83 Relationship between austenitization processing parameters and grain size for a grain-refined and non-grain-refined AISI 1060 steel. (a) Effect of austenitization temperature and 2 h soaking time. (b) Effect of austenitizing time. Source: Ref 30

Grain size may be reduced (grain refinement) by normalizing, which involves heating the steel to its austenitizing temperature, holding for 10 to 20 min, and then cooling in air. Typically, normalizing produces a uniform and smaller grain size.

A fine grain size may exhibit an adverse influence on the case and the core hardenability of lean-alloy carburizing steels. However, a fine grain size is generally preferred. Coarse-grained steels may exhibit greater distortion during heat treatment (Ref 30). An as-quenched coarse austenite grain produces large martensite plates and large austenite volumes. This is important, because large martensite plates are more prone to microcracks than are small plates. These structural effects reduce fatigue and impact strengths.

Microcracks refer to cracks that develop during the formation of plate martensite. Microcracking is most often observed when the martensite plates are long and thin. Microcracking does occur in lath martensite. Microcracks are formed when the strains generated at the tip of a growing martensite plate are sufficient to produce cracking within the plate or boundary between growing plates (Ref 30). Therefore, microcracks run across the plate or alongside of the plate. They have also been observed at prior austenite grain boundaries. They are confined to the high-carbon regions of the case. Large, high-carbon martensite plates are more likely to have microcracks. Microcracks in nickel-chromium steel are shown in Fig. 84 (Ref 30).

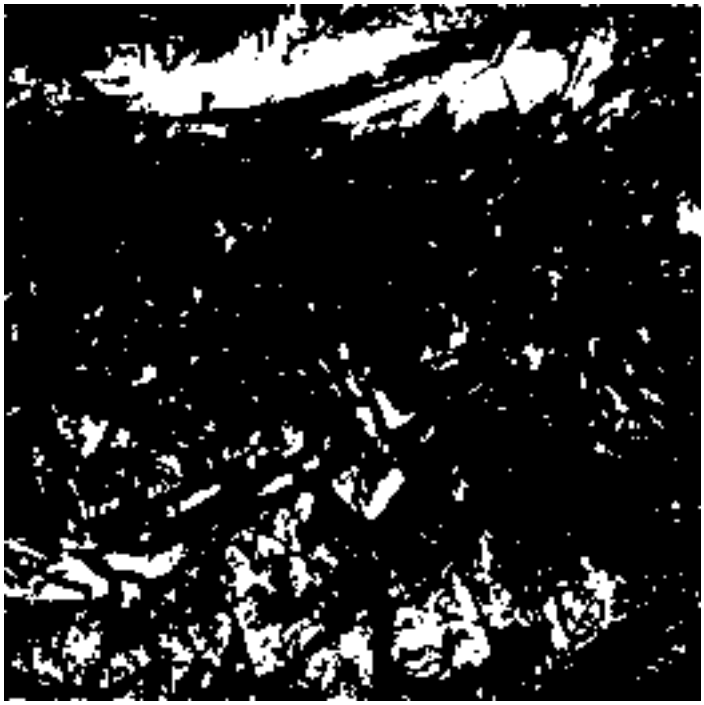


Fig. 84 Microcracking in a nickel-chromium steel that also exhibits microsegregation. 910×. Source: Ref 30

The propensity for microcracking increases as the carbon content of the steel increases above 0.6%. Microcracking potential increases with the amount of carbon in solution in the martensite, which increases as the temperature increases from Ac_1 up to 920 °C (1690 °F). Direct quenching retains more carbon in solution in the martensite; therefore, quenching method and temperature are important. Plain-carbon and very lean-alloy grades of steel are more prone to microcracking during case hardening. However, 8620 and 52100 are known to be susceptible to microcracking (Ref 30). The frequency and size of microcracks increased with grain size (Ref 53). Microcracking has been reported to be aggravated by the presence of hydrogen (Ref 54), and tempering, even at temperatures as low as 180 °C (360 °F) for 20 min, has been reported to exhibit a healing effect (Ref 55).

Although the presence of microcracks may potentially cause various problems, there are few studies reported in this area. It is possible that fatigue strength is reduced by as much as 20% (Ref 30). Similarly, there are few standards. There are no ANSI/AGMA specifications for grades 1 and 2. For the grade 3 quality, ten microcracks in a 0.06 mm^2 (0.0001 $in.^2$) field at 400× is specified.

Microsegregation always occurs in processes such as casting. The amount of microsegregation in steel castings is affected by the alloy content of the steel, including tramp elements. The propensity for microsegregation is alloy-dependent. Silicon and manganese affect the microsegregation of molybdenum, and manganese affects the microsegregation of sulfur. In addition to microstructural analysis, elemental analysis may also provide an invaluable insight into potential for inclusion formation. For example, sulfide inclusions may be obtained if insufficient manganese is present. Generally, the manganese content should be approximately five times the sulfur content to convert sulfur to manganese sulfide. The order of microsegregation susceptibility (from most to least) is sulfur, niobium, phosphorus, tin, molybdenum, chromium, silicon, manganese, and nickel (Ref 30). Figure 85 illustrates dendritic microsegregation of a fractured gear tooth (Ref 30).

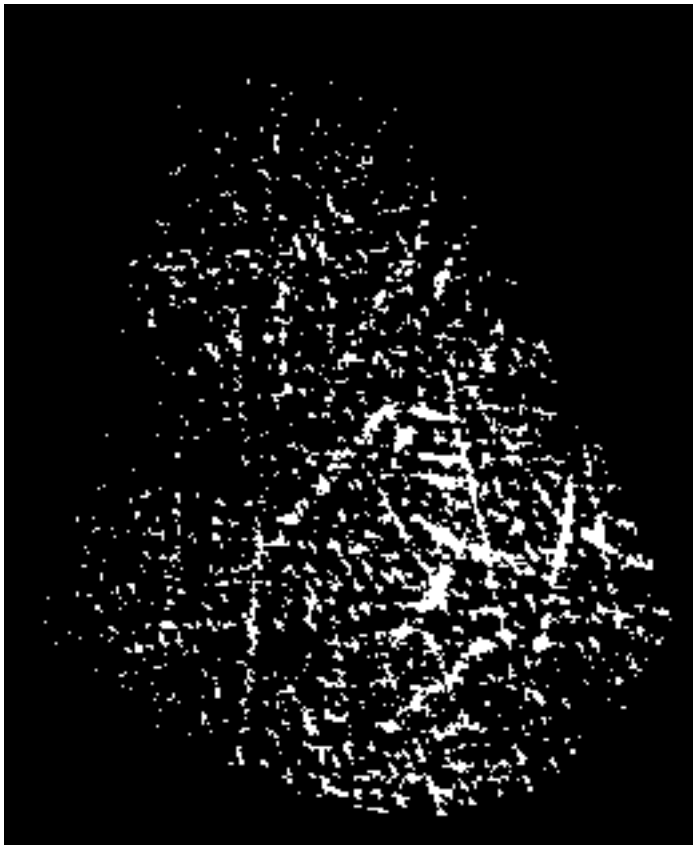


Fig. 85 Dendritic microsegregation in a fractured gear tooth. 2×. Source: Ref 30

Steel cracking may result from chemical segregation, which may be evident in the form of banding, as shown in Fig. 86 (Ref 27) Microstructural analysis showed that bands of tempered martensite were associated with bands of tempered martensite and bainite. The bainite appeared to originate from carbon and manganese segregation within the material, which would lead to increased internal stresses. Both surface and subsurface cracking was observed. Figure 87 shows subsurface cracking, which appeared to be intergranular (Ref 27).

Fig. 86 Micrograph of AISI 4140 steel as quenched and tempered, showing bands of tempered martensite and tempered martensite/bainite. 50×; 2% nital etch. Source: Ref 27

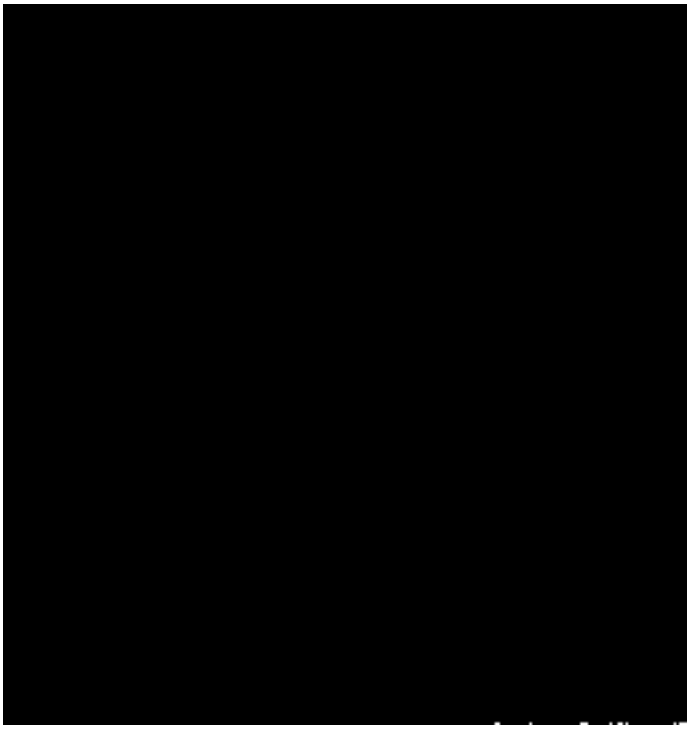


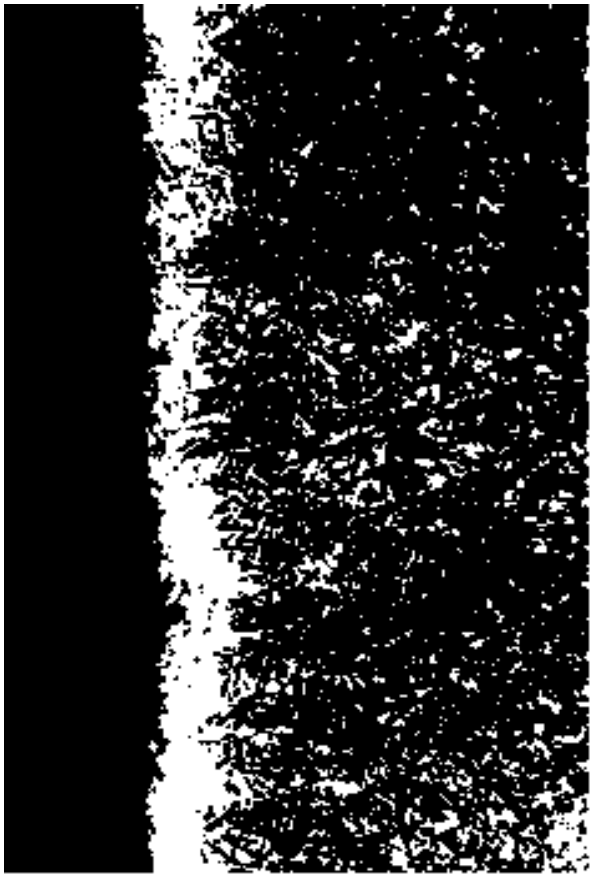
Fig. 87 Micrograph of AISI 4140 steel as quenched and tempered. Microstructure shows banded martensite and tempered martensite/bainite. Subsurface cracking is illustrated. 100×; 2% nital etch. Source: Ref 27

In a similar case, AISI 1144, a resulfurized steel, pins with subsequent cracking at the pin tips were accompanied with soft spotting. (The pins were through-hardened prior to induction hardening of the pin tip.) Microstructural analysis showed that the cracking and soft spotting condition was due to stringer inclusions, with bands of gross chemical segregation significantly greater than normally observed with this grade of steel. The stringers act as stress concentration sites for crack initiation in the presence of quenching stresses.

Mechanical and thermal treatments may substantially affect the distribution and directionality of microsegregation, and, in some cases, it may be reduced significantly. However, because of the long soaking times required to completely homogenize the steel, it is generally recommended that nothing be done.

Microsegregation may lead to machining problems, and severe cracking in the more alloyed carburizing grades of steel may occur during slow cooling from the carburizing temperature. In carburized and quenched steels, microsegregation may lead to unsatisfactory zones of martensite and austenite. Austenite-martensite or martensite-bainite banding probably reduces fatigue resistance, especially if bainite is present. Microsegregation may also influence the growth and distortion of a part during heat treatment. Microsegregation, along with nonmetallic inclusions, influences the transverse toughness and ductility of wrought steels.

Effects of Grinding. A properly designed case-hardening process should produce surface microstructures containing essentially fine, tempered martensite with minimal part-to-part and batch-to-batch variation. However, grinding may be performed to achieve the desired dimensions and surface finish. It is also performed to remove undesired metallurgical products resulting from heat treatment, such as carbide films and internal oxidation. However, the grinding process can unacceptably heat a surface, leading to grinding burns and even cracking. A grinding burn occurs when the part becomes sufficiently hot, as a result of the grinding action, to cause discoloration or to exhibit a microstructural change due to tempering or hardening. The temperature distribution and associated microstructure products during grinding are illustrated in Fig. 88 (Ref 30).



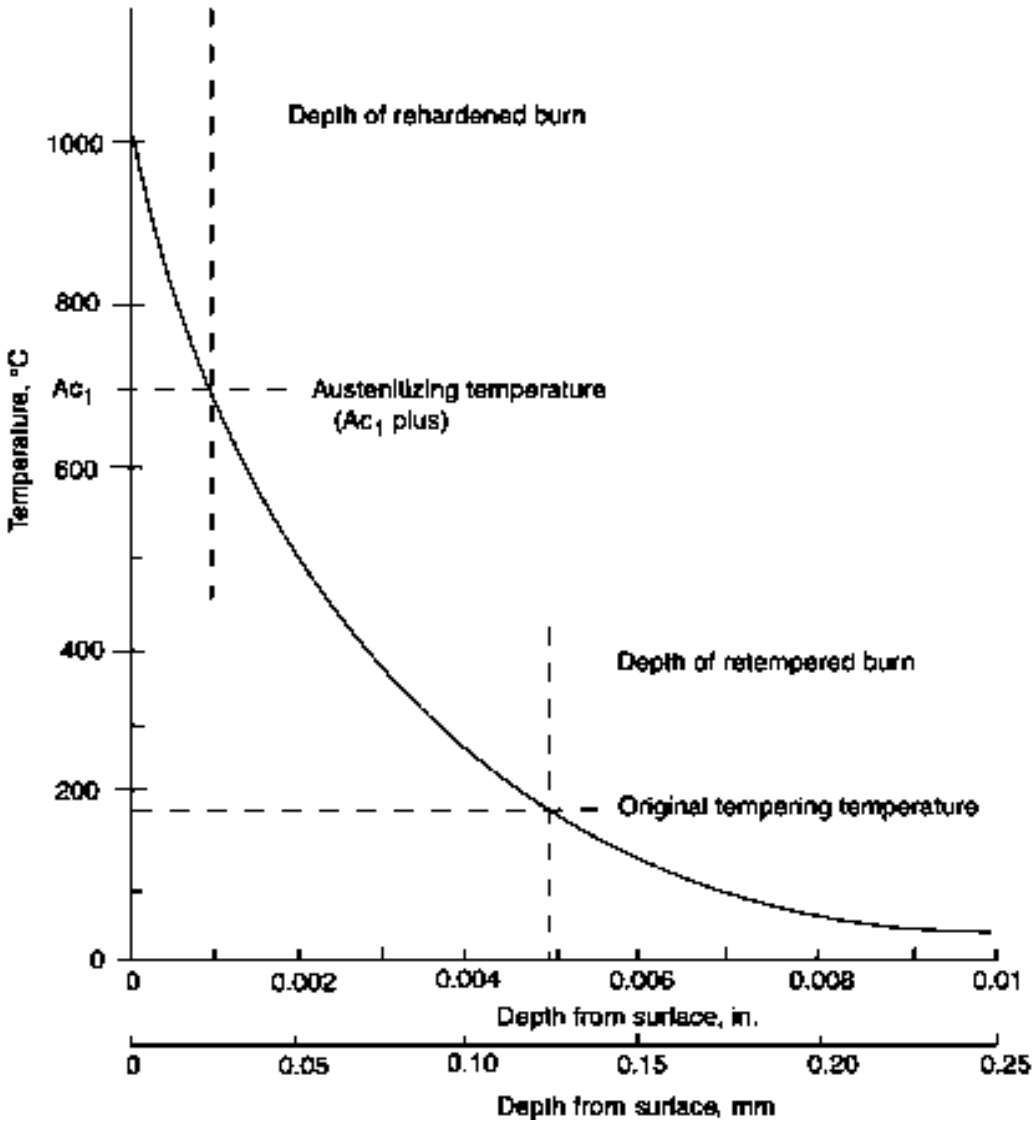

Depth of rehardened burn
Austenitizing temperature
(Ac_1 plus)
Depth of retempered burn
Original tempering temperature
Temperature, °C
1000
800
Ac_1
600
400
200
0
0
0.002
0.004
0.006
0.008
0.01
Depth from surface, in.
0
0.05
0.10
0.15
0.20
0.25
Depth from surface, mm

Fig. 88 Temperature distribution within a ground surface, as indicated by microstructural modifications. The Ac temperature indicated is for slow heating; at high heating rates, such as encountered in grinding, the Ac temperatures are elevated. 500×. Source: Ref 30

There are different types of grinding burns. One type results in localized tempering and is called overtemper burning. This occurs when heat is not transferred from the grinding surface sufficiently to prevent induced tempering of the martensitic surface. This results in a loss of hardness on the burned surface and possibly a loss of compressive stresses.

If the surface temperature during grinding exceeds the Ac_3, a thin layer of austenite is formed that is cooled by the mass of the part to form a light-etching martensite. This is called a rehardening burn. Because of the thermal process, the rehardened material is surrounded by overtempered steel and reduced hardness, as illustrated in Fig. 88 (Ref 30).

Grinding cracks, such as those illustrated in Fig. 89 and 90 are often accompanied by grinding burns (Ref 30). Grinding cracks typically form perpendicular to the grinding direction and penetrate at right angles to the surface. It is possible for the crack to penetrate substantially deeper than the depth of the burn.

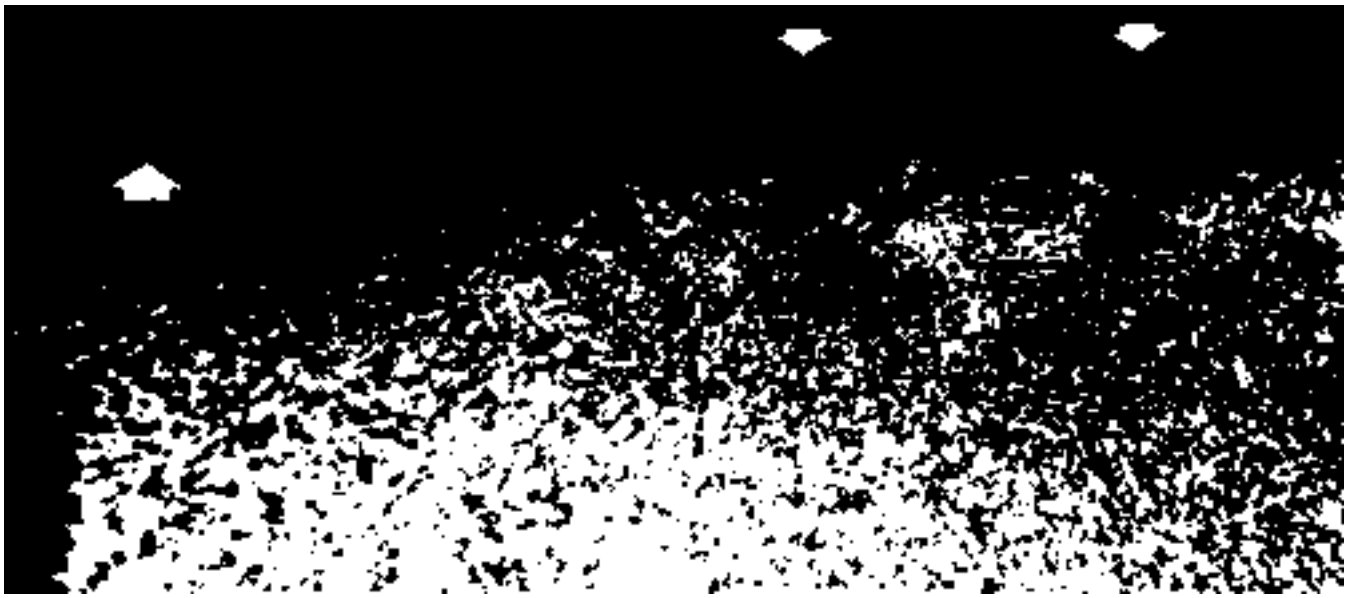


Fig. 89 Micrograph of grinding cracks in case-hardened 8620 steel showing small cracks (see small arrows) that passed through the hardened case to the core, and the burned layer on the surface (dark band with arrow at the left) that resulted in grinding burns. (Note: Nital and acidic ferric chloride are suitable etchants for grinding burns.) Source: Ref 30

Fig. 90 Grinding cracks on the flanks of a small spur gear wheel. Source: Ref 30

When grinding, the grinding wheel grade, wheel speed, and depth of cut are important process considerations. Generally, it is recommended to locate the high spots and then proceed with light cuts. If wet grinding is done, ensure that coolant feed is adequate and correctly directed. The part should be inspected for grinding burns, cracks, and steps in gear tooth fillets. Ground surfaces vary from good quality (with compressive residual stresses) to poor quality (with burns or even cracks). Bending fatigue strength is improved or seriously impaired, depending on the quality of the ground surface. Because grinding improves the contact accuracy and surface smoothness, the quality of this process affects lubrication and contact fatigue. Tempering and peening may be used to improve the quality of ground surfaces.

The following example illustrates the potential impact of grinding cracks. An AISI G-3500 gray iron camshaft was induction hardened to a depth of 1.5 to 3.8 mm (0.060 to 0.150 in.) and a lobe hardness of 45 HRC. Cracking was observed by magnetic-particle inspection of the cam lobe after subsequent machining. Elemental analysis showed that the carbon content of the steel was higher than the specification value (3.56% C versus 3.10 to 3.45% C). The surface hardness was 52 HRC versus a specification value of 45 HRC. The hardness was 94 HRB. Microstructural evidence suggested quenching from an excessively high austenitizing temperature, which contributed to the formation of excessive amounts of retained austenite at the core lobe surface and within the induction-hardened case. Grinding stresses, apparently caused during machining of the cam lobe surface, apparently caused localized martensitic transformation of the retained austenite. Cracking that occurred due to the martensitic conversion is shown in Fig. 91 (Ref 27).

Fig. 91 Micrograph of induction-hardened AISI G-3500 gray iron illustrating crack propagation into the hardened case. 88×; 3% nital etch. Source: Ref 27

Effects of Shot Peening. Shot peening improves surface hardness and induces compressive stresses. This is essentially a cold working process that plastically deforms the surface, producing increased surface compressive stresses. Figure 92 illustrates the formation of a case-hardened surface by shot peening (Ref 30).

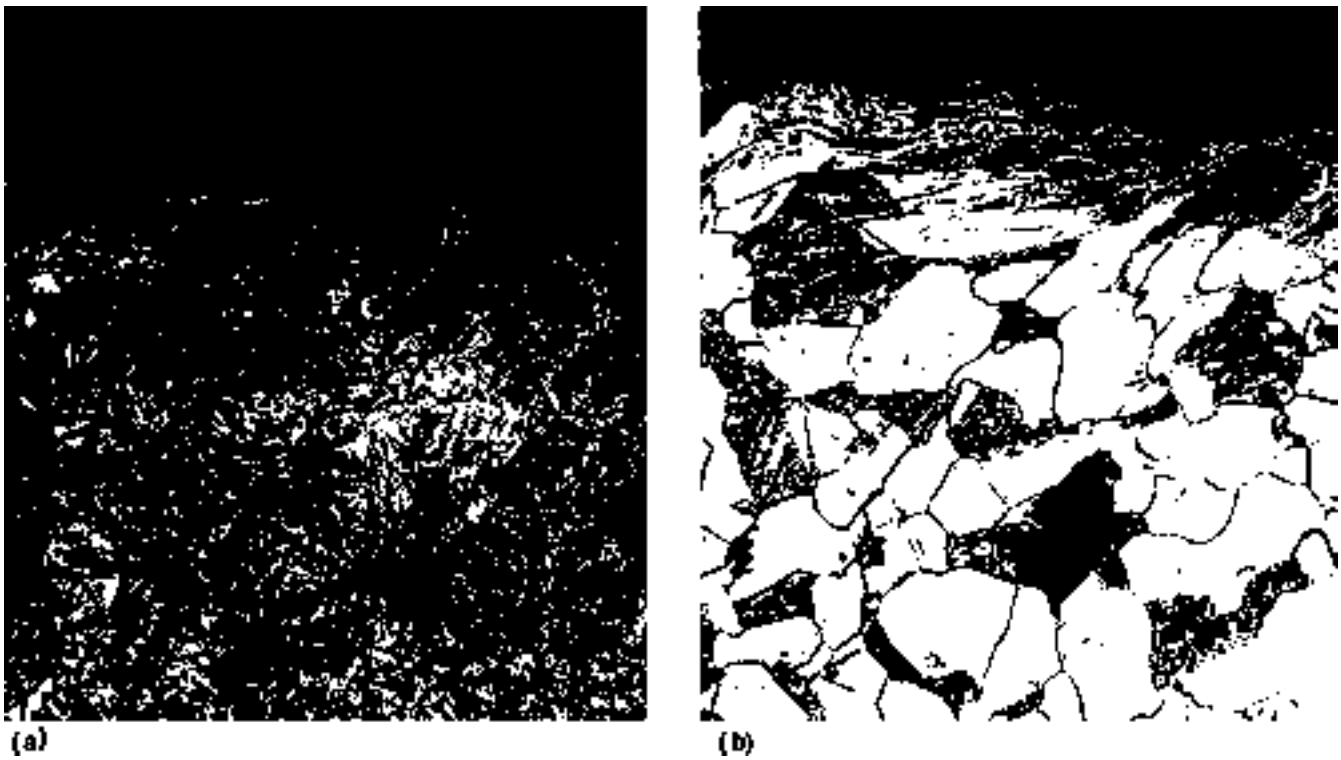


Fig. 92 Effect of hardening by plastic deformation. (a) Case-hardened surface. (b) Non-case-hardened surface. Both 243×. Source: Ref 30

Shot peening may also be used to remove directional machining marks. By masking, it is possible to localize this process to affect only the most critically stressed areas. The benefits may be lost if the initial surface is rough, and ragged surfaces may lead to folded-in defects.

Prior to conducting the peening process, the type of shot and the appropriate process parameters for the material and hardness must be determined. This may be done experimentally using an Almen strip. It is important that the shot are round, and the process must be precisely controlled to ensure complete coverage of the desired area.

Fatigue cracks generally initiate at the surface under bending conditions. Shot peening induces surface compressive residual stresses. Therefore, the possibility of fatigue crack initiation at the surface is reduced. A 20% improvement in fatigue limit or high-cycle fatigue life and contact-fatigue resistance has been quoted (Ref 30).

References cited in this section

27. R.R. Blackwood, L.M. Jarvis, D.G. Hoffman, and G.E. Totten, Conditions Leading to Quench Cracking other than Severity of Quench, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 575–585

30. G. Parrish, *Carburizing: Microstructure and Properties,* ASM International, 1999

47. B. Liscic, Heat Treatment of Steel, *Steel Heat Treatment Handbook,* G.E. Totten and M.A.H. Howes, Ed., Marcel Dekker, 1997, p 527–662

52. M.I. Vinograd, I.Y. Ul'Yanina, and G.A. Faivilevich, Mechanism of Austenite Grain Growth in Structural Steel, *Met. Sci. Heat Treat. (USSR),* Vol 17 (No. 1–2), 1975

53. R.P. Brobst and G. Krauss, The Effect of Austenite Grain Size on Microcracking in Martensite of an Fe-1.2C Alloy, *Metall. Trans.,* Vol 5, Feb 1974, p 457–462

54. J. Hewitt, The Study of Hydrogen in Low Alloy Steels by Internal Friction Techniques, *Hydrogen in Steel,* Special Report 73, The Iron and Steel Institute, 1962, p 83–89

55. T.A. Balliet and G. Krauss, The Effect of the First and Second Stages of Tempering of Microcracking in Martensite of an Fe-1.2C Alloy, *Metall. Trans. A,* Vol 7, Jan 1976, p 81–86

References

1. S. Mocarski, Carburizing and Its Control, Part I: Basic Considerations, *Ind. Heat.,* Vol 41 (No. 5), 1974, p 58–70

2. A. Bavaro, Heat Treatments and Deformation, *Trait. Therm.,* Vol 240, 1990, p 37–41

3. F. Legat, Why Does Steel Crack during Quenching, *Kovine Zlitine Technol.,* Vol 32 (No. 3–4), 1998, p 273–276

4. R.W. Bohl, Difficulties and Imperfections Associated with Heat Treated Steel, Lesson 13, Heat Treatment of Steel, Materials Engineering Institute Course 10, ASM International, 1978

5. K.E. Thelning, *Steel and Its Heat Treatment,* Butterworths, London, U.K., 1985

6. H. Berns, “Distortion and Crack Formation by Heat Treatment of Tools,” *Radex Rundsch.,* Vol 1, 1989, p 40–57

7. G. Beck, Internal Stresses Associated with Heat Treatment of Alloys, *Mém. Etud. Sci. Rev. Métall.,* June 1985, p 269–282

8. T. Ericsson and B. Hildenwall, Thermal and Transformational Stresses, *Sagamore Army Mater. Res. Conf., 28th Proc.,* 1981, p 19–38

9. C. Jensen, Dimensional Changes during Heat Treatment of Hot Work Steel, *Ind. Heat.,* March 1993, p 33–38

10. P. Mayr, “Dimensional Alteration of Parts due to Heat Treatment,” *Residual Stresses in Science and Technology,* Vol 1, Garmisch-Partenkirschen, FRG, 1986, DGM Informationsgesellschaft mbH, Adenauerallee 21, D-6370, Oberursel, FRG 1987, p 57–77

11. R.W. Edmonson, Dimensional Changes in Steel during Heat Treatment, *Met. Treat.,* Vol 20 (No 6), 1969, p 3–19

12. M. Wisti and M. Hingwe, Tempering of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 121–136

13. Y. Toshioka, Heat Treatment Deformation of Steel Products, *Mater. Sci. Technol.,* Vol 1, 1985, p 883–892

14. T. Inoue, K. Haraguchi, and S. Kimura, Analysis of Stresses due to Quenching and Tempering of Steel, *Trans. Iron Steel Inst. Jpn.,* Vol 18, 1978, p 11–15

15. G.E. Totten and M.A.H. Howes, Distortion of Heat Treated Components, *Steel Heat Treatment Handbook,* G.E. Totten and M.A.H. Howes, Ed., Marcel Dekker Inc., 1997, p 292

16. D.A. Canonico, Stress-Relief in Heat Treating of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 33–34

17. C.E. Bates, G.E. Totten, and R.L. Brennan, Quenching of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 67–120

18. W.T. Cook, A Review of Selected Steel-Related Factors Controlling Distortion in Heat-Treatable Steels, *Heat Treat. Met.,* Vol 26 (No. 2), p 27–36

19. Y. Toshioka, Heat Treatment Deformation of Steel Products, *Mater. Sci. Technol.,* Vol 1, Oct 1985, p 1883–1892

20. A. Thuvander and A. Melander, Calculation of Distortion during Quenching of a Low Carbon Steel, *First ASM Heat Treatment and Surface Engineering Conference in Europe (Proc.),* Trans Tech Publications, Part 2, 1992, p 767–782

21. P.C. Clarke, Close-Tolerance Heat Treatment of Gears, *Heat Treat. Met.,* Vol 25 (No. 3), 1998, p 61–64

22. R.F. Kern, Thinking Through to Successful Heat Treatment, *Met. Eng. Q.,* Vol 11 (No. 1), 1971, p 1–4

23. R. Kern, Distortion and Cracking, Part II: Distortion from Quenching, *Heat Treat.,* March 1985, p 41–45

24. R. Kern, Distortion and Cracking, Part III: How to Control Cracking, *Heat Treat.,* April 1985, p 38–42

25. H.-J. Chen and Z.W. Jiang, Microstructure Improvement and Low-Temperature Quenching of Dies Made of Cr12-Type Steel, *Jinshu Rechuli,* No. 8, 1992, p 39–41

26. H.-H. Shao, Analysis of the Causes of Cracking of a 12% Cr Steel Cold Die during Heat Treatment, *Jinshu Rechuli,* No. 11, 1995, p 43

27. R.R. Blackwood, L.M. Jarvis, D.G. Hoffman, and G.E. Totten, Conditions Leading to Quench Cracking other than Severity of Quench, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 575–585

28. F. Legat, “Why Does Steel Crack during Quenching,” *Kovine Zlitine Technol.,* Vol 32 (No. 3–4), 1998, p 273–276

29. X. Cheng and S. He, Analysis of Quenching Cracks in Machine-Tool Pistons under Supersonic Frequency Induction Hardening, *Heat Treat. Met. (China),* No. 4, 1991, p 51–52

30. G. Parrish, *Carburizing: Microstructure and Properties,* ASM International, 1999

31. J.R. Davis, Ed., *ASM Materials Engineering Dictionary,* ASM International, 1992, p 407

32. P.F. Stratton, N. Saxena, and R. Jain, Requirements for Gas Quenching Systems, *Heat Treat. Met.,* Vol 24 (No. 3), 1997, p 60–63

33. H.M. Tensi, G.E. Totten, and G.M. Webster, Proposal to Monitor Agitation of Production Quench Tanks, *Heat Treating: Including the 1997 International Induction Heat Treating Symposium, Proc. of the 17th Conf.,* D.L. Milam, D.A. Poteet, G.D. Pfaffmann, V. Rudnev, A. Muehlbauer, and W.B. Albert, Ed., ASM International, 1997, p 423–441

34. S. Owaku, Quench Distortion of Steel Parts, *Netsu Shori (J. Jpn. Soc. Heat Treat.),* Vol 32 (No. 4), 1992, p 198–202

35. R.T. Von Bergen, The Effects of Quenchant Media Selection on the Distortion of Engineered Steel Parts, *Quenching and Distortion Control,* G.E. Totten, Ed., ASM International, 1992, p 275–282

36. H.M. Tensi, A. Stich, and G.E. Totten, Fundamentals of Quenching, *Met. Heat Treat.,* Mar/Apr 1995, p 20–28

37. T. Kunitake and S. Susigawa, *Sumitomo Search,* May 1971, p 16–25

38. V.D. Kalner and S.A. Yurasov, Internal Oxidation during Carburizing, *Met. Sci. Heat Treat. (USSR),* Vol 12 (No. 6), June 1970, p 451–454

39. A.A. Polyakov, Quenching Properties of Parts Having Stress Concentrators, *Met. Sci. Heat Treat.,* Vol 37 (No.7–8), 1995, p 324-325

40. A.G. Haynes, Interrelation of Isothermal and Continuous-Cooling Heat Treatments of Low-Alloy Steels and Their Practical Significance, *Heat Treat. Met.,* Special Report 95, The Iron and Steel Institute, 1966, p 13–23

41. D.P. Koistinen and R.E. Marburger, A General Equation Prescribing the Extent of the Austenite-Martensite Transformation in Pure Iron-Carbon Alloys and Plain Carbon Steels, *Acta. Metall.,* Vol 7, 1959, p 59–60

42. H. Webster and W.J. Laird, Jr., Martempering of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 137

43. J.R. Keough, W.J. Laird, Jr., and A.D. Godding, Austempering of Steel, *Heat Treating,* Vol 4, *ASM Handbook,* ASM International, 1991, p 152–163

44. N.I. Kobasko, M.A. Arpnov, G.E. Totten, and A.V. Sverdlin, Method and Equipment for Implementation of Intensive Quenching, *Heat Treating, Including the Liu Dai Memorial Symposium, Proc. of the 18th Conf.,* R.A. Wallis and H.W. Walton, Ed., ASM International, 1998, p 616–621

45. I.Y. Arkhipov, V.A. Batyreva, and M.S. Polotskii, Internal Oxidation of the Case on Carburized Alloy Steels, *Met. Sci. Heat Treat. (USSR),* Vol 14 (No. 6), 1972, p 508–512

46. W.E. Dowling, W.T. Donlon, W.B. Copple, and C.V. Darragh, Fatigue Behavior of Two Carburized Low-Alloy Steels, *Proc. of Second International Conference on Carburizing and Nitriding with Atmospheres,* ASM International, 1995, p 55–60

47. B. Liscic, Heat Treatment of Steel, *Steel Heat Treatment Handbook,* G.E. Totten and M.A.H. Howes, Ed., Marcel Dekker, 1997, p 527–662

48. R. Ruxanda and E. Florian, Microscopic Observations of the Carbide/Matrix Interphasic Interface in a Low-Alloyed Hypercarburized Steel, *Proc. of Second International Conference on Carburizing and Nitriding with Atmospheres,* ASM International, 1995, p 117–121

49. J. Wang, Z. Qin, and J. Zhou, Formation and Properties of Carburized Case with Spheroidal Carbides, *5th International Congress on the Heat Treatment of Materials,* (Budapest), International Federation for Heat Treating and Surface Engineering (Scientific Society for Mechanical Engineering), 1986, p 1212–1219

50. J.G. Wang, Z. Feng, and J.C. Zhou, Effect of the Morphology of Carbides in the Carburizing Case on the Wear Resistance and the Contact Fatigue of Gears, *J. Mat. Eng. (China),* Vol 6, Dec 1989, p 35–40

51. R. Kern, Supercarburizing, *Heat Treat.,* Oct 1986, p 36–38

52. M.I. Vinograd, I.Y. Ul'Yanina, and G.A. Faivilevich, Mechanism of Austenite Grain Growth in Structural Steel, *Met. Sci. Heat Treat. (USSR),* Vol 17 (No. 1–2), 1975

53. R.P. Brobst and G. Krauss, The Effect of Austenite Grain Size on Microcracking in Martensite of an Fe-1.2C Alloy, *Metall. Trans.,* Vol 5, Feb 1974, p 457–462

54. J. Hewitt, The Study of Hydrogen in Low Alloy Steels by Internal Friction Techniques, *Hydrogen in Steel,* Special Report 73, The Iron and Steel Institute, 1962, p 83–89

55. T.A. Balliet and G. Krauss, The Effect of the First and Second Stages of Tempering of Microcracking in Martensite of an Fe-1.2C Alloy, *Metall. Trans. A,* Vol 7, Jan 1976, p 81–86

Failure Analysis and Life Assessment of Structural Components and Equipment

Introduction

LIFE ASSESSMENT of structural components is used to avoid catastrophic failures and to maintain safe and reliable functioning of equipment. The articles in the “Structural Life Assessment Methods” Section in this Volume are written to provide an overview of the prevalent life assessment methodologies for structural components. Because the failure analyst is often asked questions concerning remaining life, fitness-for-service, inspection intervals, and reliability of structural components and equipment, it is necessary that the failure analyst be aware of life assessment methodologies to address the questions and concerns of the industry he or she serves. Life assessment method advances and changes in technologies for structural components and equipment will require the investigator to adapt to the need of the industry. Furthermore, the failure investigator role has expanded from providing accurate identification of life-limiting failure mechanisms and degradation phenomena to also providing the time for degradation or damage, and crack growth rate to be used in life assessment estimates. Thus, the failure investigator's input is essential for meaningful life assessment of structural components. This article provides an overview of the structural design process, the failure analysis process, the failure investigator's role, and how failure analysis of structural components integrates into determination of remaining life, fitness-for-service, and other life assessment concerns. The topics discussed in this article include:

- Industry perspectives on failure and life assessment of components
- Structural design philosophies
- Life-limiting factors
- The role of the failure analyst in life assessment
- The role of nondestructive inspection
- Fatigue life assessment
- Elevated-temperature life assessment
- Fitness-for-service life assessment
- Probabilistic and deterministic approaches

Industry Perspectives on Failure and Life Assessment of Components

As noted previously, life assessment of structural components is a means to avoid catastrophic failures and to maintain safe and reliable functioning of equipment. Catastrophic failures of structural components occur rather infrequently, but when they do, they take a heavy toll on human lives in addition to the cost of repairs, replacement power, and litigation costs. In 1982, the National Bureau of Standards commissioned a study to determine the direct and indirect cost of fracture in the United States. It was estimated that $120 billion are spent annually to cover direct costs and costs associated with fracture-related accidents (Ref 1). The estimates took into account the necessity of overdesigning to prevent failure, added inspections, repairs, and replacement of degraded materials. Needless to say, the costs and stakes for failure are high. In addition, the cost savings are great if failure can be mitigated or prevented.

For the failure investigator, a failure is often defined as the rupture, fracture, or cracking of a structural member. The industrial definition of failure is often quite different from the textbook definition. A component, in practice, is deemed to have failed when it can no longer perform its intended function safely, reliably, and economically. Any one of these criteria can constitute failure. For example, a steam turbine blade whose tip has eroded affects turbine efficiency and hence affects the economics of operation adversely. The blade should therefore be replaced even though it can continue to operate. Component failures are thus defined in terms of “functional” rather than “structural” failures. Replacement of parts can be based on economic considerations, reliability, and material properties. In the discipline of life assessment, equipment and structures are evaluated

to determine if they are suitable, reliable, and economical for continual service. If deemed unreliable, the equipment or structure may be repaired, refurbished, or replaced.

In any component the failure criteria need to be defined and established. Failure does not always involve fracture or rupture. Progressive damage of structure and components under operating conditions leads to exhaustion of life, thus leading to failure. Damage may be defined as a "progressive and cumulative change acting to degrade the structural performance of the load-bearing component or components which make up the plant" (Ref 2). Life may be defined as the "period during which a component can perform its intended function safely, reliably, and economically" (Ref 3). With some modifications, the definitions used by Viswanathan and Dooley for fossil-fuel power steam plant components (Ref 3) can be used to define failure and life of components; that is, component life is expended when:

- Design life has elapsed.
- Calculations predict life exhaustion.
- Service time has reached some arbitrarily chosen fraction of calculated or experimental failure life.
- Previous failure statistics indicate high probability of failure.
- Frequency of repair renders continued operation uneconomical.
- Nondestructive inspection reveals cracking.
- Surface degradation from corrosion, including coating degradation, is excessive.
- Grain-boundary attack and/or pitting by oxidation/hot corrosion, is excessive.
- Foreign object damage is severe.
- Destructive sampling and testing indicate life exhaustion.
- Excessive deformation has occurred due to creep, causing distortion and unfavorable changes in clearances.
- Sudden and complete fracture occurs.

References cited in this section

1. J.J. Duga et al., *The Economic Effects of Fracture in the United States*, Report to the National Bureau of Standards, Battelle Columbus Laboratories, March 1983

2. L.F. Coffin, Damage Evaluation and Life Prediction for High-Temperature Gas Turbine Materials, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials*, EPRI AP-4477, Project 2382-3, Electric Power Research Institute, April 1986, p 1.1–1.17

3. R. Viswanathan and R.B. Dooley, Creep Life Assessment Techniques for Fossil Power Plant Boiler Pressure Parts, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials*, EPRI AP-4477, Project 2382-3, Electric Power Research Institute, April 1986, p 2.1–2.28

Structural Design Philosophies

Historic Failures. It is often stated that history repeats itself. Yet, when it comes to structural components and equipment, structural designers, original equipment manufacturers (OEMs), and users do not want a repeat of history. The consequences and costs of fractured, cracked, corroded, and malfunctioned equipment are unwanted. Through the years, history has demonstrated that failures occur; history has also shown that the engineering communities have responded to prevent failure from occurring again. Table 1 (Ref 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15) identifies some of the historic structural failures that have occurred in the 20th century. These historic failures as well as other failures have revolutionized design philosophies, inspection techniques and practices, material development, and material processing and controls and have redefined the criteria for failure. Furthermore, the pursuit of understanding how and why these failures occurred have resulted in the development of structural-integrity programs, enhanced analytical modeling and prediction techniques, accurate life assessment methods, and a fortified commitment to avoid the recurrence of these failures through improved designs. The examples cited in Table 1 were serious and often tragic failures that had a great impact on structural designs and life assessment developments. However, not all failures or malfunctions of equipment is as pivotal in history as those mentioned in Table 1. Yet, it is emphasized that any failure, no matter how seemingly insignificant, should be investigated and the findings used to improve the design and increase the life and reliability of that component or equipment.

Table 1 Historic failures and their impact on life assessment concerns

Failure	Year	Reason for failure	Life assessment developments
Titanic (Ref 4)	1912	Ship hits iceberg and watertight compartments rupture.	Improvement in steel grades Safety procedures established for lifeboats Warning systems established for icebergs
Molasses Tank Failures (Ref 5)	1919, 1973	Brittle fracture of the tank as a result of poor ductility and higher loads	Design codes for storage tanks developed Consideration given to causes for brittle fracture
Tacoma Bridge Failure (Ref 6, 7	1940	Aerodynamic instability and failure caused by wind vortices and bridge design	Sophisticated analytical models developed for resonance Bridge design changed to account for aerodynamic conditions
World War II Liberty ships (Ref 8)	1942–1952	1289 of the 4694 warships suffered brittle fracture or structure failure at the welded steel joints.	Selection of increased toughness material Improved fabrication practices Development of fracture mechanics
Liquefied natural gas (LNG) storage tank (Ref 9)	1944	Failure and explosion of an LNG pressure vessel due to a possible welding defect and improperly heat treated material resulting in subsequent fatigue crack growth	Selection and development of materials with improved toughness at the service temperature of -160 °C (-250 °F)
Comet aircraft failures (Ref 10)	1950s	Fatigue crack initiation in pressurized skins due to high gross stresses and stress concentration effects from geometric features	Development of the fatigue “safe-life” approach Evaluation of the effects of geometry and notches on fatigue behavior Evaluation of the effects of stiffeners on stress distribution Establishment of aircraft structural integrity program (ASIP) in 1958
F-111 Aircraft No. 94 wing pivot fitting (Ref 11)	1969	Fatigue failure due to material defect in high-strength steel	Improved inspection techniques Change from fatigue “safe-life” to damage-tolerant design philosophy Development of materials with improved toughness
Seam-welded high-energy piping failures (Ref 12)	1986–2000	Cavitation and creep voids in welds resulting in catastrophic high-energy rupture	Development of elevated-temperature life assessment techniques for cavitation and creep failure
Aloha Incident, Boeing 737 (Ref 13)	1988	Accelerated corrosion and multiple fatigue crack-initiation sites in riveted fuselage skin	Improved aircraft maintenance and inspection procedures Life assessment methods developed for multiple-site damage (MSD)
Sioux City Incident (Ref 14)	1989	Hard alpha case present in titanium fan disk resulted in fatigue crack initiation and catastrophic failure.	Increased process controls on processing of titanium ingots

Failure	Year	Reason for failure	Life assessment developments
			Development of probabilistic design approach and analytical life assessment using dedicated computer programs for titanium disks
Earthquakes in Kobe City, Japan, and Northridge, California (Ref 15)	1994, 1995	Failure occurred in I-beams and columns due to joint configuration and welding practices that resulted in low ductility of the steel.	Development of earthquake resistant structures Improved joint designs and welding practices for structural steels Improved controls on steel manufacture

Overview of the Design Process. Because of failures similar to those in Table 1, predicting performance and assessing the remaining life with greater confidence becomes increasingly important as costs for manufacturers and operators need to be reduced. Furthermore, the cost of failure is progressively greater as systems become more complex, downtime costs increase, and liability for failure increases. A brief discussion follows on the design process because it is important for failure investigators and life assessment engineers to understand some of the design issues. Each structure has unique design requirements, but all structures are designed using some basic design principles. Figure 1 illustrates the relationship among the design phase, testing, systematic failure analysis, and life assessment of components.

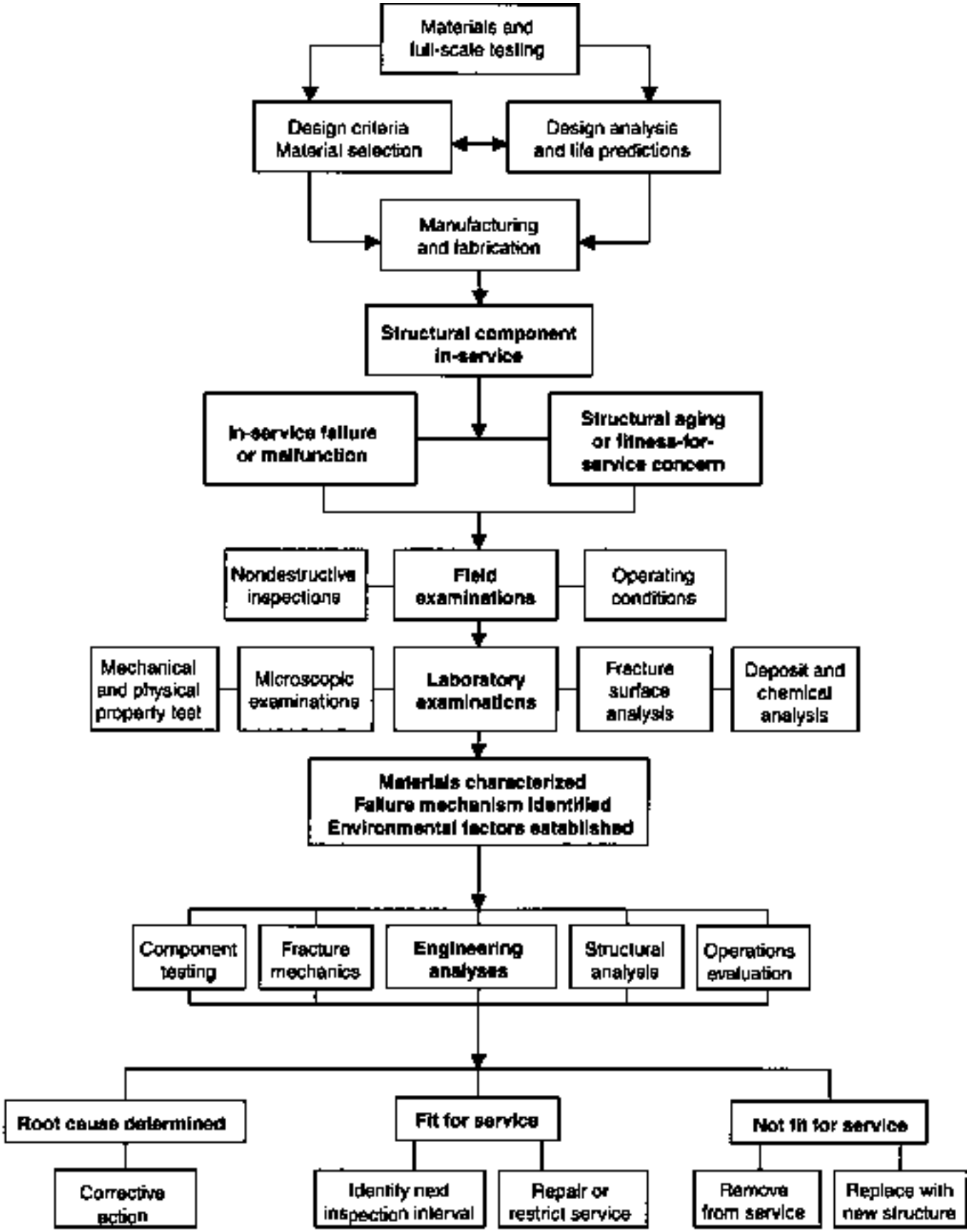


Fig. 1 Flow diagram showing the relationship between the design phase and the investigative tasks for in-service failure, structural aging, and fitness-for-service of structural components

One alternative for avoiding failures used in the past was to overdesign and to operate overconservatively. The economic penalties for both are increasingly significant; however, the economic penalties for failures are significant as well. It is

necessary, then, to pay more attention to predicting and ensuring performance. Predicting and ensuring performance is fundamentally a part of the design process for buildings, power plants, aircraft, refineries, and ships (Ref 16).

For any given design, the mission and the intended use are established. Predicting the performance and design life of a component depends on defining what life or performance is required for a given duration while the component operates generally in combinations of mechanical and chemical environments. Defining performance may involve defining end points such as: acceptable length of propagating cracks, maximum depth of propagating pits, acceptable remaining thickness of corroding pipes, maximum number of fatigue cycles or extent of cumulative damage, maximum number of plugged tubes, maximum number of failed circuits, maximum leakage, or appearance of a maximum area or number of rust spots. Defining such end points is a critical part of predicting life since prediction defines when these end points will be reached and therefore when “failure” occurs.

Defining failure is also related to what is meant by the “design life.” For example, for the aerospace industry, an airplane may be designed for 8000 flight hours and analyzed for two lifetimes or 16,000 flight hours. For the power industry, the design life of components is sometimes taken as 40 years. This means that the equipment is expected to perform satisfactorily at its rated output for 40 years. This is not to say that some maintenance is not necessary. However, to assert to a customer that a component has a 40 year design life, it is necessary to develop bases for such a claim. Such bases are usually provided by analyses and by accelerated testing in the laboratory and with prototype and model testing.

As part of the life assessment process, it is important to understand how a structural component—whether a pressure vessel, shaft, or structural member—is designed in order to understand how it may fail and to perform meaningful life assessment. For example, the first step in the design of any pressure vessel is to select the proper design code based on its intended use. For example, a pressure vessel may be a power or heating boiler, a nuclear reactor chamber, a chemical process chamber, a hydrostatic test chamber used to test underwater equipment, or a pressure vessel for human occupancy. Once the intended use is identified, the appropriate design code can be selected. For example, pressure vessels use codes provided by many organizations and certifying agencies, such as the American Society of Mechanical Engineers (ASME), the American Bureau of Shipping (ABS), and European agencies have pressure vessel design codes. Strict adherence to these codes for the design, fabrication, testing, and quality control and assurance allows the finished pressure vessel to be certified by the appropriate authorizing agency (Ref 17).

One of the first incentives to develop a pressure vessel code occurred after the Boston Molasses tank incident in 1919 when the tank failed by overstress, consequently releasing more than 2 million gallons of molasses and resulting in the loss of life and property (Ref 5). Even after that catastrophic failure and understanding the nature of the failure, another molasses tank failure occurred in New Jersey in 1973. Figure 2 shows the destruction caused by the molasses tank incident. These molasses tank incidences demonstrate how important it is to prevent failures, and it underscores that good designs consider the operating conditions and limitations of materials of construction.

Fig. 2 Failed molasses tank, which fractured suddenly in New Jersey in March 1973. This catastrophic and sudden brittle fracture resulted in the release of the molasses in the tank similar to the Boston Molasses tank disaster in 1919.

The next step in the design process is to identify the design parameters, such as configuration, design pressure, and so forth. Table 2 presents an example of a design parameter list applicable for a chemical process chamber (Ref 17). These design parameters are the same parameters considered when conducting a pressure vessel failure investigation and life assessment.

Table 2 Pressure vessel design parameters

Required design code[(a)]	Penetration and location requirements
Basic chamber configuration. (Cylindrical or spherical; flat, spherical, or elliptical end details; etc.)	Contents and/or process within the pressure vessel
Internal volume capacity	Estimated operational pressure and temperature cycle history (number of cycles at what pressures and temperatures over the vessel's lifetime)
Minimum inside diameter	Piping, external and internal attachment requirements
Minimum inside length	Test chamber surroundings (enclosed in building or exposed to elements)
Chamber orientation (for cylindrical chambers, longitudinal axis vertical or horizontal)	Test chamber physical geographical location
Support configuration (saddle supports, bottom cylindrical skirt, legs, etc.)	Vessel special material requirements (vessel material other than carbon steel, internal cladding, etc.)
Maximum internal operating pressure	Vessel protection requirements (painted surfaces, stainless steel overlays at seals, cathodic protection, etc.)
Maximum external operating pressure (vacuum, etc.)	Fabrication requirements
Design operating temperature range	Material selection

(a) ASME Boiler and Pressure Vessel Code, Section VII, Div. 1, 2, or 3; ABS; other

Given the design parameters, the proper material(s) is selected for the structural component. Safety and economy are often the governing factors when selecting a material for pressure vessels. The material is selected based on its mechanical, corrosion, creep, toughness, and thermal properties as applicable. If necessary, the appropriate weld material is selected based on the chosen base material. Material is assigned an allowable stress value based on its ultimate and yield strengths and operating temperature range. This allowable stress value is then used in design equations or compared to results obtained from detailed analyses.

The design process then proceeds with the determination of the sizes and/or thickness of the various components. The design process is completed with the creation of the engineering and fabrication drawings. These drawings should include the dimensional information, but also specify materials, weld identification, weld procedures, and required weld inspections. Other helpful information to include on the drawings is basic parameters such as design pressure, design temperature range, design code, and other information deemed necessary for the particular structural component.

Structural Design Approaches. The criteria of failure are determined by the strength of materials, fracture toughness, creep resistance, fatigue behavior, and the corrosion resistance of materials. These are briefly discussed in this section.

Strength of Materials. In the strength-of-materials design approach one typically has a specific structural geometry (assumed to be defect free) for which the load-carrying capacity must be determined. To accomplish this, a calculation is first made to determine the relation between the load and the maximum stress that exists in the structure. The maximum stress so determined is then compared with the strength of the material. An acceptable design is achieved when the maximum stress is less than the strength of the material, suitably reduced by a factor of safety.

It can be assumed that failure will not occur unless σ_{max} exceeds the yield strength of the material, σ_Y. To ensure this, a factor of safety (S) can be introduced to account for material variability and/or unanticipated greater service loading. The strength-of-materials approach is a good approach for materials with no defects and simple structures. Figure 3 shows the strength-of-materials approach and the engineering design regime based on a factor of safety.

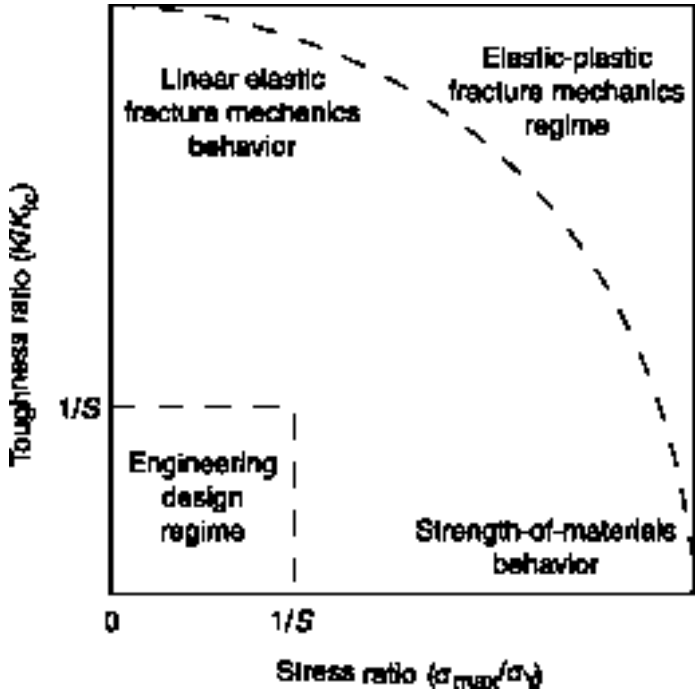


Fig. 3 A general plot of the ratios of the toughness and stress showing the relationship between linear elastic fracture mechanics and strength of materials as it relates to fracture and structural integrity (Ref 18)

Linear Elastic Fracture Mechanics. Brittle fractures, similar to those mentioned in Table 1, are avoided using a linear elastic fracture mechanics design approach. This approach considers that the structure, instead of being defect-free, contains a crack (Ref 18). The governing structural mechanics parameter when a crack is present, at least in the linear approach, is an entity called the stress-intensity factor. This parameter, which is conventionally given the symbol *K*, can be determined from a mathematical analysis similar to that used to obtain the stresses in an uncracked component. For a relatively small crack in a simple structure, an analysis of the flawed structure beam would give to a reasonable approximation:

$$K = 1.12 s_{\max} \sqrt{p\,a} \qquad \text{(Eq 1)}$$

where α is the depth of the bracket and σ_{max} is the stress that would occur at the crack location in the absence of the crack.
The basic relation in fracture mechanics is one that equates *K* to a critical value. This critical value is often taken as a property of the material called the plane-strain fracture toughness, conventionally denoted as K_{Ic}. When equality is achieved between *K* and K_{Ic}, the crack is presumed to grow in an uncontrollable manner. Hence, the structure can be designed to be safe from fracture by ensuring that *K* is less than K_{Ic}. The Liberty warship fractures are a classic example of a structural failure caused by K_{Ic} exceeding *K* and uncontrolled crack growth.
The essential difference from the strength of materials approach is that the fracture mechanics approach explicitly introduces a new physical parameter: the size of a (real or postulated) cracklike flaw. In fracture mechanics the size of a crack is the dominant structural parameter. It is the specification of this parameter that distinguishes fracture mechanics from conventional failure analyses.
The generalization of the basis for engineering structural-integrity assessments that fracture mechanics provides is portrayed in terms of the failure boundary shown in Fig. 3. Clearly, fracture mechanics considerations do not eliminate the traditional approach. Structures using reasonably tough materials (high K_{Ic}) and having only small cracks (low *K*) will lie in the strength of materials regime. Conversely, if the material is brittle (low K_{Ic}) and strong (high σ_Y), the presence of even a small crack is likely to trigger fracture. The fracture mechanics assessment is then the crucial one.
The special circumstances that would be called into play in the upper right-hand corner of Fig. 3 are worth noting. In this regime, a cracked structure would experience large-scale plastic deformation prior to crack extension. Additional information is provided in the article “Failure Assessment Diagrams” in this Volume and in Ref 19.
Damage Tolerance Approach. Life assessment of aircraft and power-plant equipment stems largely from the development of the damage-tolerance philosophy based on fracture mechanics. Damage tolerance is the philosophy used for maintaining the structural safety of commercial transport, military aircraft, structures, and pressure vessels. The use of fracture mechanics and damage tolerance has evolved into the design program for structures that are damage tolerant, that is, designed to operate with manufacturing and in-service-induced defects (Ref 20).

Damage-tolerance evaluation has been interpreted in the past as a means to allow continued safe operation in the presence of known cracking. This interpretation is incorrect. No regulations allow the strength of the structure to be knowingly degraded below ultimate strength (1.5 × limit). The damage-tolerance evaluation is merely a means of providing an inspection program for a structure that is not expected to crack under normal circumstances, but may crack in service due to inadvertent circumstances. If cracks are found in primary structure, they must be repaired. The only allowable exception is through an engineering evaluation, which must show that the strength of the structure will never be degraded below ultimate strength operations or in-service conditions.

After many major fatigue failures in the 1950s on both military and commercial aircraft, the most notable of which were the DeHavilland Comet failures in early 1954, the U.S. Air Force (USAF) initiated the Aircraft Structural Integrity Program (ASIP) in 1958 (Ref 10). The fatigue methodology adopted in the ASIP was the reliability approach, which became known as the "safe-life" method. This safe-life approach, used in the development of USAF aircraft in the 1960s, involved analysis and testing to four times the anticipated service life. On the commercial scene, another philosophy, "fail safety," was introduced in the early 1960s, and a choice between safe-life and fail-safe methods was allowed by commercial airworthiness requirements. However, it was found that the safe-life method did not prevent fatigue cracking within the service life, even though the aircraft were tested to four lifetimes to support one service life (i.e., scatter factor of 4). One notable example is the F-111 aircraft 94, which crashed in 1969 (Ref 11). The F-111 aircraft had a safe-life of 4000 flight hours. However, a material defect caused the F-111 aircraft, which used high-strength steel (ultimate tensile strength of 1655 to 1793 MPa, or 240 to 260 ksi, toughness of about 66 MPa $\sqrt{m}$, or 60 ksi $\sqrt{in}$) for the wing box (Fig. 4). The defect was not observed during inspection, and a fatigue crack initiated and grew for only about 0.38 mm (0.015 in.). The aircraft was flown for 107 flights safely, at which time catastrophic failure occurred, causing the destruction of the aircraft.

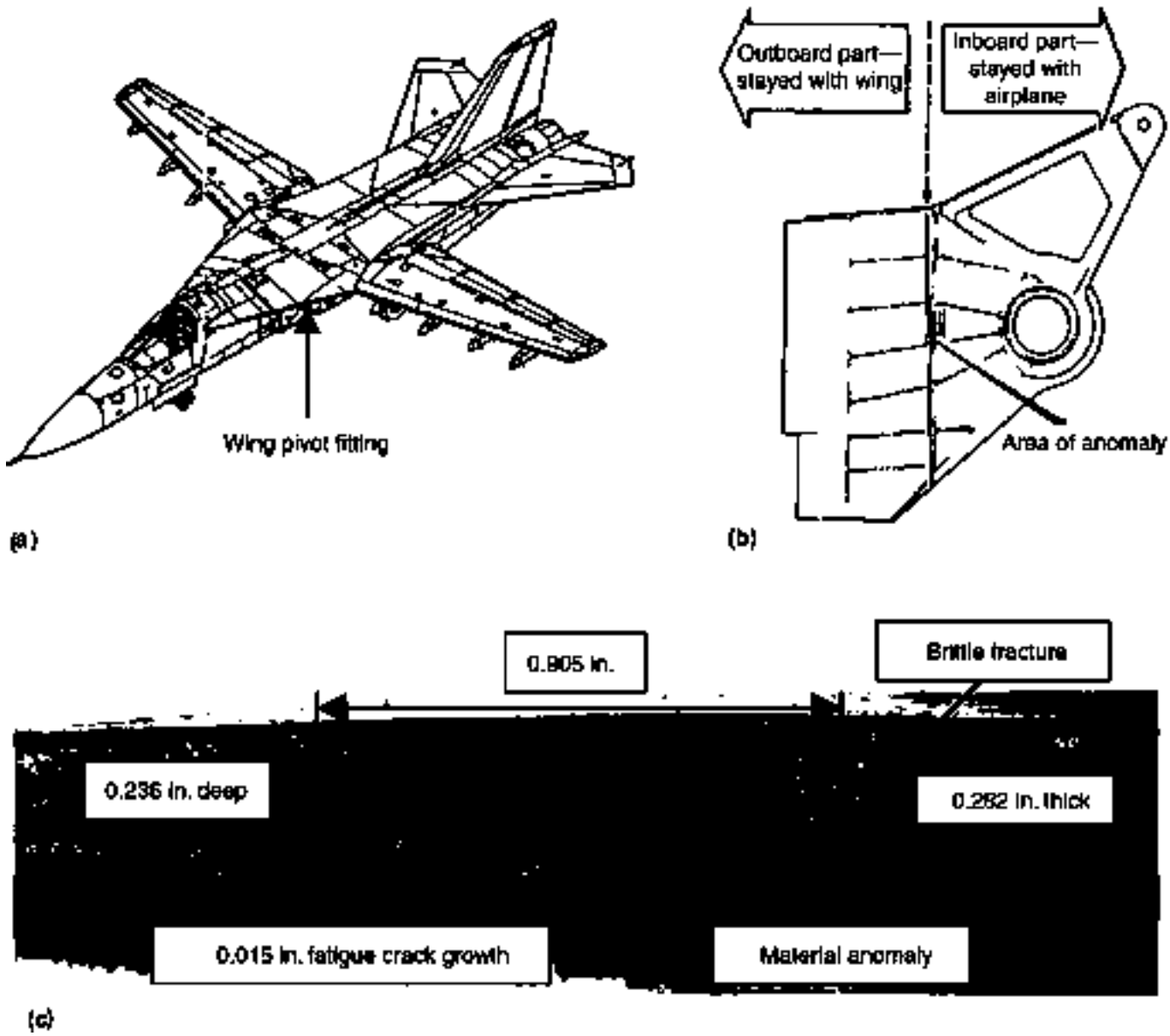


Fig. 4 Fatigue cracking in an aircraft wing fitting for the F-111 Aircraft 94 that crashed in 1969. (a) and (b) Location of the left wing-pivot box fitting. The 22 mm (0.91 in.) material defect was not observed during inspection, and a fatigue crack initiated and grew for only about 0.38 mm (0.015 in.) before unstable brittle fracture occurred.

The leak-before-break design approach is prevalent in pressure-containing equipment such as pressure vessels and piping used in the nuclear and fossil-fuel power-generation plants, refineries, and chemical plants. Failure analysis and life assessment of pressure-containing systems is essential. Although leak-before-break failures are not catastrophic, they are costly and can affect plant operations. Therefore, analyses are often performed to predict when the next internal or external inspection should be performed. Typical life-limiting mechanisms include stress-corrosion cracking, fatigue, and thermal fatigue. Welded structures that could initiate a crack are often susceptible to these mechanisms.

The leak-before-break concept generally refers to a pressure-contaminant system failure in which a part-through wall crack extends to become a through-wall crack, thus allowing fluid to escape. If no further crack extension occurs, then the loss of the fluid is detected and no further crack growth occurs. Alternatively, when a through-wall crack propagates along the wall, a catastrophic event can occur (Ref 18).

Power-plant piping materials that are ductile, such as stainless steel and nickel-base alloys, often leak before break. Figure 5 shows small-bore, socket-welded piping that will initiate fatigue cracks at either the toe of the weld or the root of the weld. These ductile socket-welded pipes leak before catastrophic failure occurs (Ref 21).

Fig. 5 Stainless steel piping such as small-bore piping is designed to leak before break. A fatigue crack either initiates at the toe or the root of the weld. (a) Typical socket fitting with a fillet weld. (b) Micrograph of a cross section through a socket-welded joint showing fatigue crack that initiated from the weld root and extended through the weld. (c) Micrograph of axial weld toe crack showing a fatigue-induced crack that extended through the pipe wall

Elevated-Temperature Concerns. For elevated-temperature equipment and structures subjected to steady-state or cyclic stresses, the principal design considerations are creep control, oxidation prevention through the use of oxidation-resistant materials or coatings, and selection of materials that have good stress-rupture and creep properties. The criteria for failure is (1) to not go below a minimum stress-rupture strength for a given operating stress and temperature and (2) to not operate above a certain temperature that alters the microstructure or oxidizes the material.

Corrosion Allowances. Designs are configured such that the operating or loading stresses can be minimized for safe operations. It is necessary to consider the effects that an environment will have on the material; it is just as important as considering structural loads on a component. It is important that environments are known and controlled in such a way that corrosion is minimized on all surfaces. This means that designs consider effects of crevices, galvanic couples, flows, stresses, and temperatures to ensure that all the surfaces of materials will be minimally degraded within the design life (Ref 16).

A common design approach for pressure vessels and tanks to deal with corrosion is to provide a “corrosion allowance,” which takes the form of additional thickness based on available information on rates of general corrosion over the design life. For example, a carbon steel vessel designed for 25 years of service in sulfuric acid at a corrosion rate of 5 mils/yr (0.005 in./yr) would have a corrosion allowance of 125 mils. However, such allowances cannot deal with stress-corrosion cracking, pitting, intergranular cracking, or effects of long-range cells. Use of a corrosion allowance can be disastrously misleading since its use suggests that all corrosion problems have been solved. It should be pointed out that exceeding the corrosion allowance does not necessarily mean the vessel would fail or is unsuitable for service. It is an indication that the vessel should be evaluated for continued service.

References cited in this section

4. H.P. Leighly, Jr., B.L. Bramfitt, and S.J. Lawrence, RMS Titanic: A Metallurgical Problem, *Pract. Failure Anal.,* Vol 1 (No. 2), April 2001, p 10–13, 33–37

5. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 230

6. S. Ross, Tacoma Narrows, 1940, *Construction Disasters,* 1st ed., McGraw-Hill, 1984, p 216–239

7. Y. Billah and R. Scanlan, Resonance, Tacoma Narrows Bridge Failure and Undergraduate Physics, *J. Physics,* Vol 59 (No. 2), Feb 1991, p 118–123

8. E.R. Parker, *Brittle Behavior of Engineering Structures,* Wiley, 1957

9. S. Atallah, U.S. History's Worst LNG Disaster, *Firehouse,* Jan 1979, p 29

10. T. Swift, Damage Tolerance Certification of Commercial Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 566–576

11. V.A. Hindas, *F-111 Design Experience: Use of High Strength Steels,* AIAA 2nd Aircraft Design and Operations Meeting

12. R. Viswanathan and J. Stringer, Failure Mechanisms of High Temperature Components in Power Plants, *Trans. ASME,* Vol 122, July 2000

13. W.R. Hendricks, The Aloha Incident: A New Era for Aging Aircraft, *Structural Integrity of Aging Aircraft,* S.N. Altussi, et al., Ed., Springer-Verlag, 1991, p 153

14. N1819U, NTSB-AAR-90-06, Aircraft Accident Report, United Airlines, Inc., Nov 1990

15. *Interim Guidelines: Evaluation, Repair, Modification and Design of Steel Moment Frames,* FEMA-267, Federal Emergency Management Agency

16. R.W. Staehl, Combining Design and Corrosion for Predicting Life, *Proc. Conf. Life Prediction of Corrodible Structures,* Vol 1, NACE International, 1994, p 138

17. D.J. Benac and L.J. Goland, Failures of Pressure Vessels, Lesson 15, *ASM Practical Failure Analysis Course,* to be published

18. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

19. P.F. Timmins, Operating Stress Maps for Failure Control, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 460

20. M.P. Kaplan and T.A. Wolff, Life Extension and Damage Tolerance of Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* p 557–565

21. D.J. Benac and D.N. Hopkins, Investigation of Fatigue-Induced Socket-Welded Joint Failures for Small-Bore Piping Used in Power Plants, *Pract. Failure Anal.,* Vol 1 (No. 2), April 2001, p 71–82

Life-Limiting Factors

To assess the life of a structure, one first must understand the factors that limit life (Ref 20). It is the purpose of the following discussion to briefly discuss some of the aspects that control the durability and remaining life of structures. Each structure or structural system has unique parameters of life expectancy. For example, the life of an aircraft is dependent on a number of variables, such as flight profiles, usage rate, external environment, material selection, and

geometry. For fossil-fuel plant equipment that is exposed to elevated temperatures, life-limiting factors involve pressure, temperature, environment, fatigue cycles, and time at temperature. The following life-limiting factors are common to most structures and should be considered in a failure analysis and a life assessment:

- Material defects
- Fabrication practices
- Stress, stress concentration, and stress intensity
- Temperature
- Thermal and mechanical fatigue cycles
- Corrosion concerns
- Improper maintenance

Material Defects. With the advent of strict quality-assurance controls and nondestructive inspection, systems, manufacturing and material defects are not as common as in early years. Yet, they have and will be a potential concern in structural components. To determine if a defect is a cause for failure, it is important that the component or structure undergoes a comprehensive failure investigation to determine if such is the case. Defects greatly affect remaining life because the time for crack initiation has been reduced.

Defects are either material defects or mechanical defects. The material defect may manifest itself because of a metallurgical anomaly. This anomaly or metallurgical defect may occur as a result of faulty furnace control during heat treating, improper plating that may result in unwanted hydrogen in high-strength steel, processing problems (forging laps, shuts, undesirable grain structure, etc.), welding defects, and so forth. For example, a material defect, shown in Fig. 4, caused the historic failure of an F-111 aircraft in 1969. Another example of a preexisting material defect is failures of a titanium disk on United Flight 232, which failed near Sioux City, Iowa, in July 1989 (Ref 14). The Sioux City mishap was due to a manufacturing problem in a titanium fan disk, which had the presence of hard alpha case. As a result of the Sioux City incident, a probabilistic life assessment approach was taken to evaluating titanium disks (Ref 22). It can be seen that these defects occur as a result of both manufacturing assembly and material processing.

A mechanical defect is often induced after the manufacturing process when the structure is in service. A mechanical defect such as a dent, gouge, impact, or deformation of the material can reduce component life. The result often is yielding or cracking of the structural component. If the mechanical defect is located at a high-stress location, the reduction in life can be significant and possibly catastrophic. For example, Fig. 6 shows damage at a fastener hole in a wing skin spar mechanically induced during the drilling of the hole. The failure analyst was able to identify the location of the damage and extent of fatigue crack growth, which was used to establish inspection intervals (Ref 23)

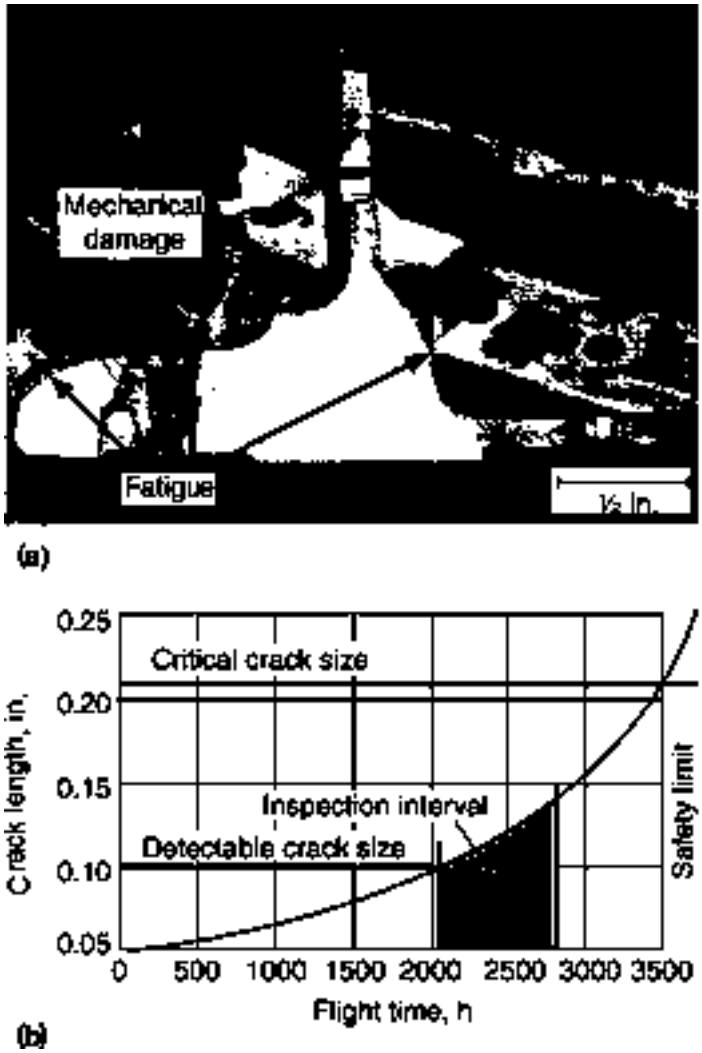


Fig. 6 Fractographic evaluation for life assessment purposes. (a) Wing spar that had mechanically induced damage at a fastener hole (indicted by an arrow). (b) Plot showing how fracture information is used to establish initial and recurring inspection requirements

Fabrication practices such as welding can affect the lifetime of a part. For example, in the fabrication of pressure vessels or pressure piping unfavorable results from welding, such as brittle cracking in the heat-affected zone (HAZ), often result from the use of steels containing excessive amounts of residual elements that increase hardenability and consequently their susceptibility to cracking. Therefore, complete control of composition is of the utmost importance when welding is involved. In producing welds without embrittlement, it is necessary to follow the proper welding procedure. Selection of the proper filler metal is also important.

Sometimes imperfections are undetected and become evident during operations. For example, small cracks were not detected in a carbon steel ASTM A 105 steam drum nozzle that was arc gouged. Consequently, during a cold start-up, a 46 cm (18 in.) crack occurred. Figure 7 shows the cracked nozzle. Prior to the cold start-up, the steam drum had been in service for five years with no cracking problems. Investigation of the cracking problem determined that crack initiation occurred from the root of an arc-gouged notch. A metallurgical cross section through the arc-gouged notch revealed microcracks and a very hard layer (54 HRC) that were not removed after arc gouging the nozzle. The carbon steel base metal was 72 to 88 HRB. The crack extended during the cold start-up because of preexisting microcracks and an as-forged nozzle that had poor fracture toughness (Ref 17).

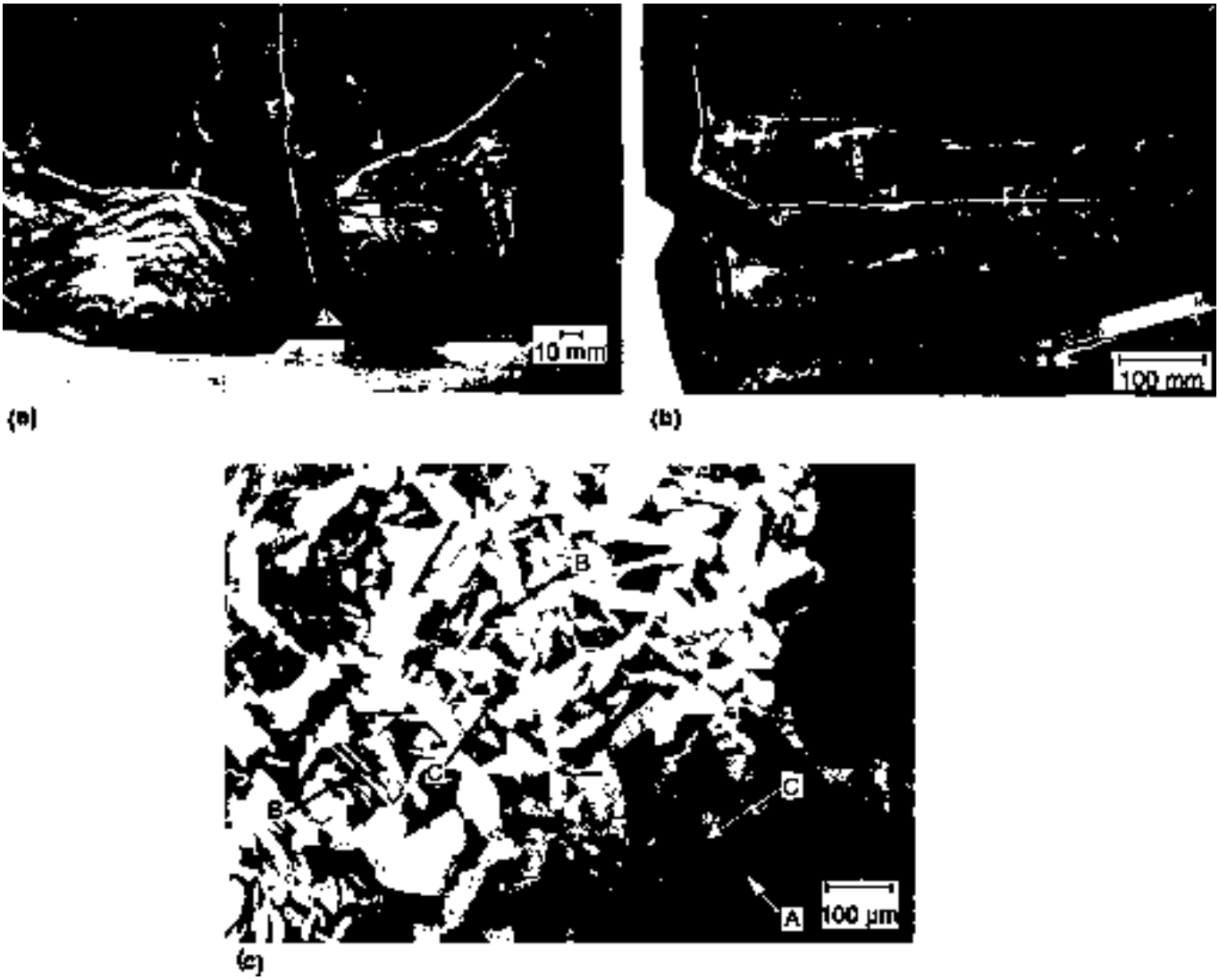


Fig. 7 Effect of welding on the life of a carbon steel structure. (a) and (b) 46 cm (18 in.) long crack found in a carbon steel as-forged nozzle that was arc gouged. Failure occurred after five years in service during cold start-up procedure. (c) Micrograph showing a hardened layer (A) and the as-forged (acicular) microstructure (B). The black square shapes (C) are diamond pyramid hardness indentations.

Operating and Nonoperating Environmental Effects. Probably the single most significant element affecting structural service life and structural integrity is the actual usage or operating condition. Very early in the design and development phase, a design usage is identified. The design usage for an aircraft system considers design capability, operating environments, intended mission goals, types of weapons, and power-plant capabilities. Based on all considerations, the design usage is an attempt to truly identify the total environment as to how a structural component or system will be utilized in its expected service life. At that time, it is probably adequate in that it represents the intentions for how the structure is going to be tested and eventually operated. However, the actual usage may be different. An example of this is unanticipated cracking that occurred in an aircraft structure at one structural control point location because the operational spectrum was 11 times more severe than the actual design fatigue spectrum (Ref 24).

The influence and degradation of material properties and therefore the reduction of life cycle is often caused by the strong influence of the operating and nonoperating environments. Life-limiting factors occur both while the structure or equipment is in service and when not in use. For example, while the aircraft is operational, the dynamic factors (fatigue, flutter, and vibration) are the responsible parameters for determining the life aspects of selected components, but while the aircraft is sitting on the ground, corrosion behavior is normally the dominant driver. In a similar manner, process equipment subjected to elevated temperatures can suffer reduced the creep life, but also detrimental is equipment that is left stagnant with a corrosive medium in a pressure vessel. For pressure vessels and piping equipment, actual usage or operating conditions can even change from company to company or plant to plant. For example, the operation of a land-based turbine can vary from plant-to-plant within the same company.

Stress, Stress Concentration, and Stress Intensity. In the earlier design days the principal design consideration was stress, that is, the load to cause failure at a specific wall thickness or cross-sectional area. The criterion for failure could be yielding resulting in plastic deformation or failure or fracture due to exceeding the ultimate tensile strength of the material. The stress on structural members is still critical, yet, with the increased understanding of crack growth and failure mechanisms, stress-concentration effects and the stress intensity at a crack tip must be considered.

High-stress regions in a structural component are typically located around geometric details, such as a hole, fillet radius, or notch. These local high-stress regions are described in terms of the remote or far-field stress of a component by multiplying the remote stress by a stress-concentration factor. Stress-concentration factors for various geometric details are derived from the theory of elasticity and/or obtained experimentally. Stress-concentration curves are available for different design geometries (Ref 25). These curves are useful when investigating a design and doing a life assessment.

Stress intensity is the controlling factor in subcritical crack growth propagation rates and the identification of the critical crack size for the onset of rapid overload (critical fracture). Stress and crack geometry are parameters that determine the stress intensity. The stress-intensity factor describes the stress distribution at a crack tip. For a small edge crack in a thin flat plate subject to a uniform tensile stress, the stress-intensity factor can be determined from:

$$K_1 = b\,s\,\sqrt{p\,a} \qquad \text{(Eq 2)}$$

where K_I is the stress-intensity factor, σ is the applied stress, and a is the crack length. The subscript on the stress-intensity factor refers to the mode of crack loading. Equation 2 is simplified; often each geometry and crack will have a correction factor, β, to account for the crack shape, loading direction, and geometry. The failure investigator often provides the size and ratio of the crack length (c) to the crack depth (a), which is used to identify the correction factor, β. For example, an edge crack normal to the free edge that is loaded in tension would have a β factor of 1.12 (Eq 3). For each crack and structure geometry, the approximate stress-intensity factor should be determined by (Ref 18):

$$K_1 = 1.12\,s\,\sqrt{p\,a} \qquad \text{(Eq 3)}$$

Temperature Effects. Temperature affects structures because it can reduce the stress to failure, cause the material to be more brittle, or oxidize the surface. One of the first questions the failure investigator asks is: what was the temperature when the failure occurred, or what is the operating temperature? If the answer to this question is unknown, then the investigator deduces the temperature based on physical evidence and material properties.

Life of a metal component at elevated temperature, when subjected to either steady or fluctuating stress, is limited. In contrast, at ambient temperatures and in the absence of a corrosive environment, the life of a component in steady-state load conditions may be unlimited, provided operating loads do not exceed the yield strength of the specific metal and wear is not serious. Stress produced at elevated temperature produces a condition of continuous strain called creep. By definition, creep is deformation as a function of time at constant load or stress. Creep, after a period of time, may terminate in stress-rupture fracture, also known as creep rupture.

Most brittle fractures of steels occur because structural components are operating at temperatures below the ductile-brittle transition temperature or the nil-ductility transition temperature. Consequently, the structure is unable to maintain subcritical crack growth and failure occurs in a catastrophic manner. The ductile-to-brittle transition temperature can be above room temperature, putting at risk structures that operate below room temperature. This occurs in tanks such as molasses tanks, pressure vessels, ships such as the Liberty warships, and high-strength alloy steel structures.

Metallic structures oxidize when subjected to elevated temperatures. For example, carbon steel boiler tube used above 540 °C (1000 °F) is prone to oxidation. Increase of oxidation resistance can occur with the addition of alloying elements such as chromium and molybdenum. A boiler tube for a 9Cr-1Mo material can have good oxidation resistance to 650 °C (1200 °F). Oxidation and high-temperature corrosion is controlled through the use of protective coatings. Elevated-temperature life assessment concerns are discussed in detail in the article “Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing” in this Volume and in Ref 26.

Thermal and Mechanical Fatigue Cycles. Fatigue fractures result from cyclic stressing, which progressively propagates a crack or cracks until the remaining section is no longer able to support the applied load. For example, pressure vessels and pressure piping are subject to high static stresses arising from the pressure of contained liquids or gases, to stresses resulting from misalignment of components, and to residual stresses induced during welding. The cyclical component may be added mechanically—by vibration of associated equipment, pulsation from a compressor. It may also be added thermally, resulting in thermal fatigue if the component is cycled through a temperature range in service. Fatigue cracks nucleate from a stress riser such as a discontinuity or a notch, which produces triaxial stressing in the material. The stress riser may be macroscopic in size, such as a notch or discontinuity, or microscopic and not visible.

Mechanical conditions are not the only sources of cyclic loading, which can contribute to fatigue failure. Transient thermal gradients within a vessel can induce plastic strains; if these thermal gradients are applied repeatedly, the resulting cyclic strain can induce failure. This process is known as thermal fatigue. Thermal fatigue is often considered a low-cycle failure mechanism that occurs due to operating conditions. Thermal fatigue can be defined as the gradual deterioration and eventual cracking of a material by alternating heating and cooling during which free thermal expansion is partially or

fully constrained. The constraint of a part or material does not allow free expansion to occur. For example, piping attachments that are mechanically constrained will develop thermal fatigue cracks when thermally cycled.

The cyclic thermally induced stresses associated with frequent start-ups and shutdowns are frequently more severe than stresses of steady-state operations, and designers cannot always predict the use cycles of equipment that owner-operators impose. Equipment sometimes is operated only during peak-demand periods; this imposes many start-ups and shutdowns, with attendant thermal stresses on pressure vessels and piping systems. When cracking occurs, it is in areas least able to accommodate severe cyclic loading; such areas are likely to be rigid or massive sections under conditions of high constraint or thermal stress.

Corrosion Concerns. In most cases, corrosion itself does not lead to structural failure, but it is the most expensive aspect of maintaining structural integrity. The effect of corrosion needs to be addressed by denoting the types of corrosion and where they can occur, inspection techniques to find the corrosion in situ, and repair or replacement methods to remove the offending material. There are many corrosion mechanisms such as stress-corrosion cracking, hydrogen embrittlement, and general corrosion that reduce the life expectancy of components. These life-limiting corrosion mechanisms need to be identified and considered in performing life assessments. These mechanisms are discussed in greater detail in this Volume and are addressed in the literature.

Improper Maintenance. The manufacturer of structures, equipments, and components often requires special maintenance as identified in handbooks, structural repair manuals, and technical orders (Ref 20). It is imperative that these manuals be rigidly adhered to. It is easy to devise methods that may be shorter and easier than those discussed in the manual, but the manufacturer selected that particular procedure for sound reasons. It is dangerous to deviate from these methods, for the deviation itself may place undue stress or cause latent damage to components apart from those being examined. This high stress or latent damage can lead to premature failure of components, and that may result in undesirable consequences. Proper maintenance and inspection of tanks, aircraft, and pressure vessels are mandatory. Extreme care during maintenance and the inspection following maintenance must be taken to ensure that proper procedures and techniques are rigorously followed.

References cited in this section

14. N1819U, NTSB-AAR-90-06, Aircraft Accident Report, United Airlines, Inc., Nov 1990

17. D.J. Benac and L.J. Goland, Failures of Pressure Vessels, Lesson 15, *ASM Practical Failure Analysis Course,* to be published

18. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

20. M.P. Kaplan and T.A. Wolff, Life Extension and Damage Tolerance of Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* p 557–565

22. H.R. Millwater, S.H. Fitch, Y.-T. Wu, D.S. Riha, M.P. Enright, G.R. Leverant, R.C. McClung, C.J. Kuhlman, G.G. Chell, and Y. Lee, “A Probabilistically-Based Damage Tolerance Analysis Computer Program for Hard Alpha Anomalies in Titanium Rotors,” Paper 2000-GT-0421, IGTI, ASME Turbo Expo, 2000

23. “Aircraft Structural Integrity and Failure Analysis of Structural Components,” Southwest Research Brochures

24. D.J. Benac and J.P. Pendley, “Integration of Design, Failure Analysis and Fleet Management to Prevent Aircraft Structural Failures,” Asian Aircraft Conference (Singapore), 1991

25. W.D. Pilkey et al., *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, 1997

26. A.F. Liu, High-Temperature Life Assessment, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 520–526

Role of the Failure Investigator

The typical role of the failure investigator is that of a materials failure analyst looking at fracture surfaces or examining the microstructure for degradation, microstructural changes, or unusual crack morphologies. With the advent of life assessment methodologies the investigator's responsibilities have expanded to be an integral part of life assessment for structures. The flow diagram in Fig. 1 shows the tasks the investigator undertakes for evaluation of in-service failures and

assessing the remaining life of aged structures. The failure analyst must be a part of the life assessment process to determine:

- The location of origin
- The life-limiting failure mechanism
- The rate of propagation or time for degradation
- The actual service or operating conditions (stress, temperature, environment)
- The validation of life assessment models

These responsibilities are discussed in the following sections.

Determining the Origin of the Failure. The first and foremost question the investigator must answer is where the crack initiated or where did the degradation start? In other words, where is the origin of the failure or location of degradation? A trained, patient, investigating engineer can identify the origin site through examination of the topographical and microscopic footprints. If the footprints point back to the origin, the investigator may subsequently find a dent, a pit, intergranular attack, a material defect, or a crack initiating from a corner, radius, or notch. If the origin is not identified, it may be difficult to identify the initiation event and hence the root cause. Furthermore, predicting the life of a structure must take into account the initiation event because some life assessment methodologies consider the time and mechanism for initiation. If the initiating event is not accurately determined, then the life assessment results will not accurately assess the part or structure. For example, Fig. 8 shows a defect that was found in the radius of a gate valve stem (Ref 27). In the gate valve stem, the failure analyst determined that a preexisting forging defect was present at the radius. Because other gate valves were in service with similar gate valve stems, a question was raised by the power plant whether this and other gate valve systems could be used or if a costly replacement effort was needed. The investigator showed that no crack extension had occurred from the forging defect. Thus, with a periodic inspection of the gate valve stem, it was concluded that the risk of using the gate valves was low and immediate replacement of the stems was not necessary.

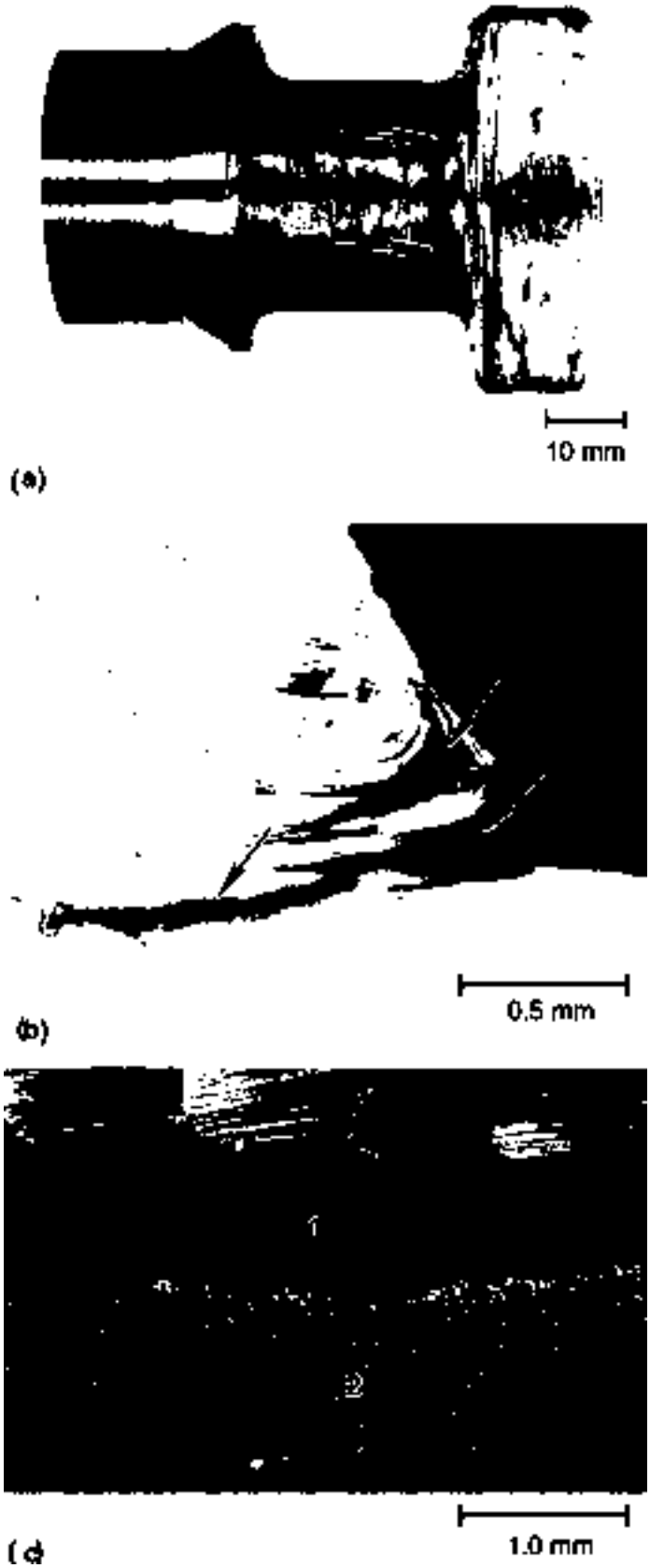


Fig. 8 Failure evaluation of a cracked gate valve stem determined that no subcritical crack growth had occurred from the forging defect. Therefore, similar gate valve stems could remain in service with a periodic inspection. (a) Cracked gate valve stem. 1×. (b) Forging defect in the radius. 50×. (c) Scanning electron fractograph of the defect (1) and lab fracture (2). 15×. No subcritical crack growth had occurred from the forging defect.

Often failure occurs not at one location, but at multiple locations. This complicates and compounds the life assessment approach. This applies to components such as rotating shafts, turbine blades, and pressure vessels to name a few. It also applies to complex structures such as bridges, aircraft, and buildings. For example, the effect of fatigue crack growth on the life of aircraft has been much studied in recent years. Since the accident involving Aloha Airlines Flight 243 in April 1988 (Ref 13), the interest regarding multiple-site damage (MSD) has increased significantly. Additionally, much analytical and test work has been accomplished to study the effect of many small cracks and their linking together. Figure 9 is an example of multiple fatigue crack initiation that occurred in a lower wing skin because of abusive machining marks. Fractographic evaluation identified multiple crack fronts that linked together and formed one crack front.

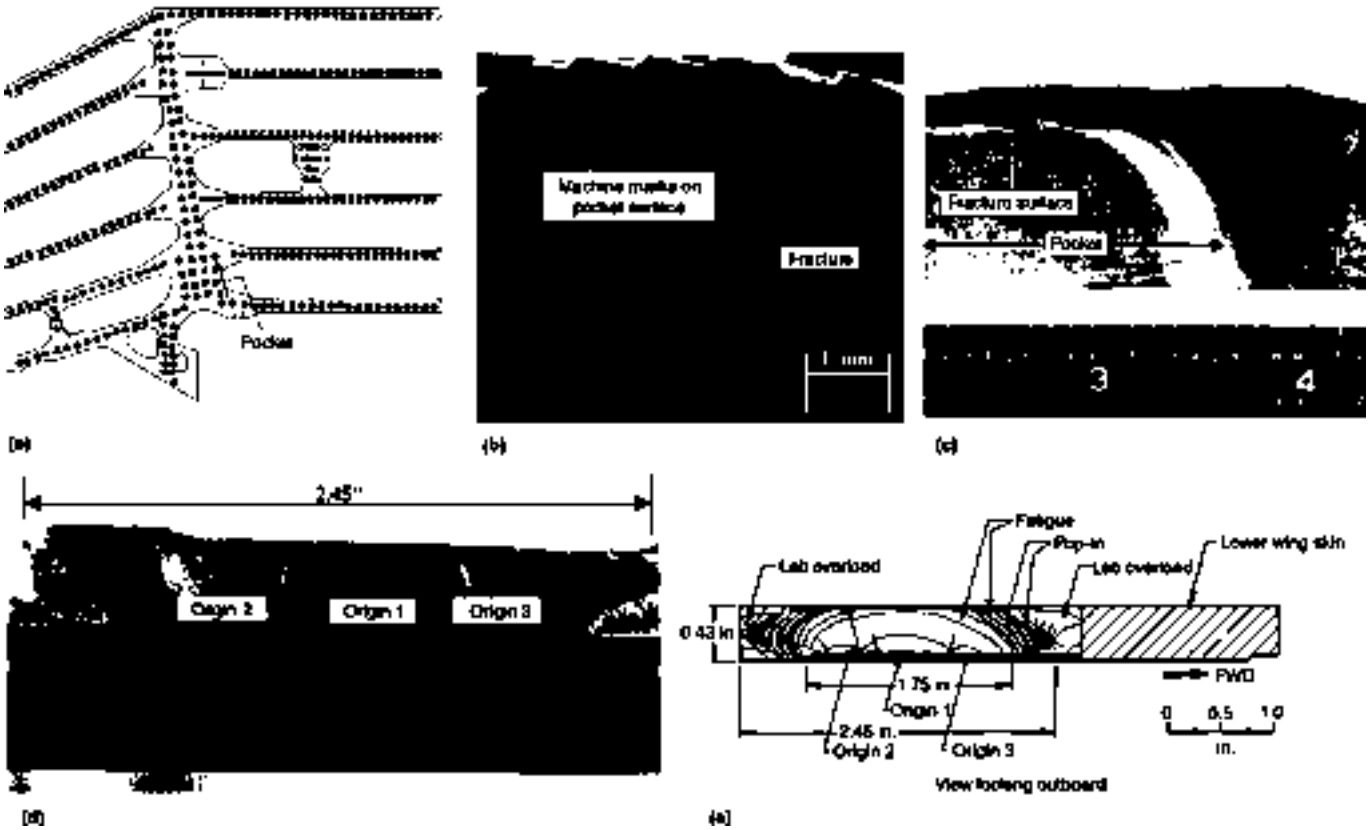


Fig. 9 Fractographic evaluation of multiple-site fatigue damage in a lower wing skin. The evaluation determined that the damage was caused by abusive machining marks that accelerated the fatigue crack growth rate. (a) Location of cracks in the a lower wing skin pocket. (b) Machine marks on the pocket surface. (c) and (d) Fracture surface. (e) Schematic of multiple origins, pop-ins, and 62 cm (2.45 in.) long total crack length

Determining the Life-Limiting Mechanism. When conducting a failure investigation, it is important to identify the failure mechanism—how the component or structure failed, degraded, or malfunctioned. To perform a meaningful life assessment it is likewise important to know the failure mechanism or degradation process. The principal failure mechanisms are discussed in this Volume. Some more common failure mechanisms are brittle fracture, creep, low- or high-cycle fatigue, thermal fatigue, and corrosion. Embrittlement phenomena can occur, and metallurgical instabilities can also detrimentally affect the life of components; for example, carbide coarsening, σ-phase formation, temper embrittlement, and so forth can facilitate rapid brittle fracture at low temperatures during transient conditions. Combinations of these conditions can, however, cause failures if the environment modifies them.

As a rule, the failure mechanism is established by examination of the physical features observed in the microstructure, fracture surface, or changes in geometry. Each life-limiting failure mechanism has characteristic fingerprints that a trained failure investigator can identify and correlate to a particular failure mechanism. These physical features are then correlated to the operating conditions to determine the driver for the failure and the time frame that failure initiated and propagated.

The failure analyst must be given the opportunity to investigate how something malfunctioned or failed and to determine how the usage or operating conditions affected the microstructure. The oversight of investigating the failure of a component can result in improper understanding of how something failed. What may appear as the obvious cause or potential cause for failure may not be the root cause, but rather a result. This in the long run is not only dangerous, but

also very costly. For this reason, closer examination of broken, malfunctioned, or aged components is often needed. A very simple axiom should be followed: “Things are not always as they appear to be.” Therefore, inaccurate determination of how something is assumed to have failed or malfunctioned will result in inaccurate life assessments.

Determining the Rate of Crack Growth or Time for Degradation. Once the origin of failure has been located or the life-limiting mechanism has been established, another question related to life assessment is raised: what was the time for the degradation phenomenon, crack initiation, or crack propagation? The failure analyst role is critical because often the investigator can establish the time for the life-limiting mechanism to occur based on fracture surface features and microstructures. Although this is not always the case, many time-dependent failure mechanisms, such as fatigue, creep, and stress-corrosion cracking, demonstrate time, load, and environment fingerprints that aid the investigator in estimating how fast a particular crack or degradation mechanism was acting. These features, such as fatigue striations, creep voids, and corrosion species, can be used to date times of degradation.

Once the origin, life-limiting failure mechanism, and rate of crack growth or time for degradation have been determined, the failure analyst may have enough information to perform a life assessment. The following case history illustrates how the fractographic fatigue features are used to determine when a crack initiated and how fast it may propagate and to predict when to inspect, replace, or continue to use a valve (Ref 27).

Cracking occurred in two 25 mm (1 in.) solenoid valve bodies reported to be 1100 kg (2500 lb), schedule 160 valves manufactured from type 316 stainless steel. The valves were in service for about ten years and were opened periodically about once a week. Pressurized hot water at about 316 °C (600 °F) flowed through the open valve. The closed valve was at near-ambient temperature.

Circumferential cracks initiated at the bottom of the bore in which the seat was located Fig. 10a. The cracking extended about 100° to 110° around the base of the bore. Fractographic examination using a stereomicroscope (10 to 50×) and a scanning electron microscope determined the crack depth to be shallow—about a maximum depth of 1 mm (0.04 in.) measured along the inlet port. The fractographic evaluation determined the failure mechanism was fatigue. The transgranular cracks and striations in Fig. 10(c) are characteristic of fatigue. The fatigue striations were coarse and on the order of 5×10^{-3} mm/cycle (2×10^{-5} in./cycle). The reason for cracking was the thermal stresses that were generated when the valve was cycled from ambient to 316 °C (600 °F) and back to ambient.

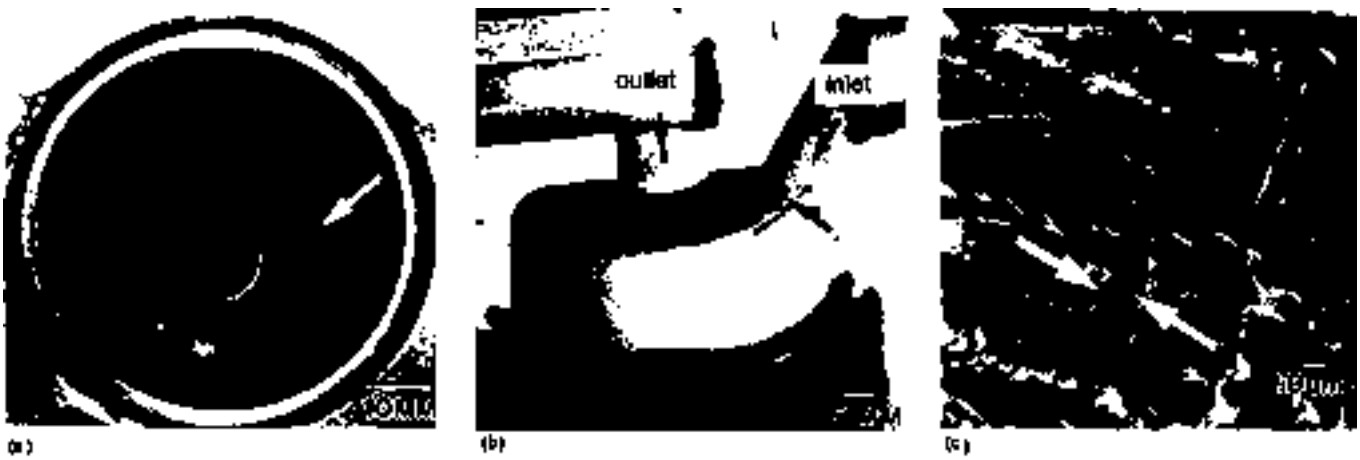


Fig. 10 Remaining life estimated on a failed solenoid valve by determining the total crack depth and rate of crack propagation. This information was correlated to the operating condition of the valve and the valve geometry, after which it was determined how much remaining life was left in the similar valve bodies. (a) Crack indicated by dye-penetrant inspection in bore of valve. 1×. (b) Crack relative to the inlet and outlet ports. 1.25×. (c) Fatigue striation found on the fracture surface. 7000×

During operation, one of the valves (valve A) was opened approximately three times per week, while the other valve (valve B) was opened approximately once per week. Thus, over the approximately 450 weeks of operation, the two valves may have experienced on the order of 1350 and 450 cycles, respectively. Based on the fracture surface striations and an average crack growth rate of 5×10^{-3} mm/cycle (2×10^{-5} in./cycle), the two cracks in valve body B, which were 0.4 and 0.3 mm (0.016 and 0.012 in.) deep, would require 800 and 600 cycles, respectively, to grow to their depths. It is evident from this simple calculation that the computed growth cycles are of the same order as the estimated number of large thermal excursions experienced by the valve body. Furthermore, we can conclude that the cracks initiated very early in the service history of the valves.

It is also possible to use the fracture data approach to estimate the remaining life of the valve bodies. First, however, the end of the useful life of the valve body must be defined. A breach between the bore and the outer-pressure boundary would cause the valve to leak to the external environment and certainly end useful life. Approximately 22.2 mm (0.875 in.) of additional crack growth would be required for the cracks to reach the outer surface of the body and create such a leak. A breach across the seat producing a leak path between the inlet and outlet ports, although less catastrophic than a leak to the outer surface, could also be considered to end the useful life of the valves. The shortest distance between the bore and the outlet port in these valve bodies was 8.0 mm (0.313 in.). Although, no cracking was present directly over the outlet port, if the conservative assumption is made that a crack initiates in this location on the very next cycle, 15,560 cycles would be required at a growth rate of 5×10^{-3} mm/cycle (2×10^{-5} in./cycle) before the crack reached the outlet port. Thus, a conservative estimate of the minimum remaining life for these valve bodies would be 15,000 cycles to produce a leak across the seat from the inlet port to the outlet port. Growth of the existing cracks through to the outer surface would require almost three times as many cycles, that is, 45,000 cycles. Thus, it was concluded that significant life remains in these valve bodies although cracks were detected during the inspection process.

Determining the Operation History. The case history on the valve bodies demonstrates how the failure investigator can correlate physical features to the operational history. This is important because the component may often be operating in a different manner than the intended mode of operation. Thus, the investigator, using his investigative tools and examining the component, can sometimes determine, for example, that a component such as a turbine was operating much hotter than it was intended, or the stress condition on an aircraft structural member was higher than the design stress. If a life assessment were conducted using the expected operating temperature or intended design stress and not the actual operating temperature or stress, then the life assessment would inaccurately determine the remaining life. Therefore, the investigator needs to make a determination of how the physical features correlate to the intended design usage and the design usage to the actual operating conditions.

Validation of Life Assessment Models. Another role the investigator serves is for validation of life assessment and methodologies models. Life assessment model validation often occurs using material coupon tests or full-scale testing of components. These tests control essential parameters of stress, environment, and cyclic conditions. These same parameters are input into a life assessment model, which estimates a component life. The validity of the model can be demonstrated by making macroscopic and microscopic measurements of the crack lengths at predetermined time intervals and at the end of the test. Figure 11 shows the comparison between predicted crack growth and measured crack growth.

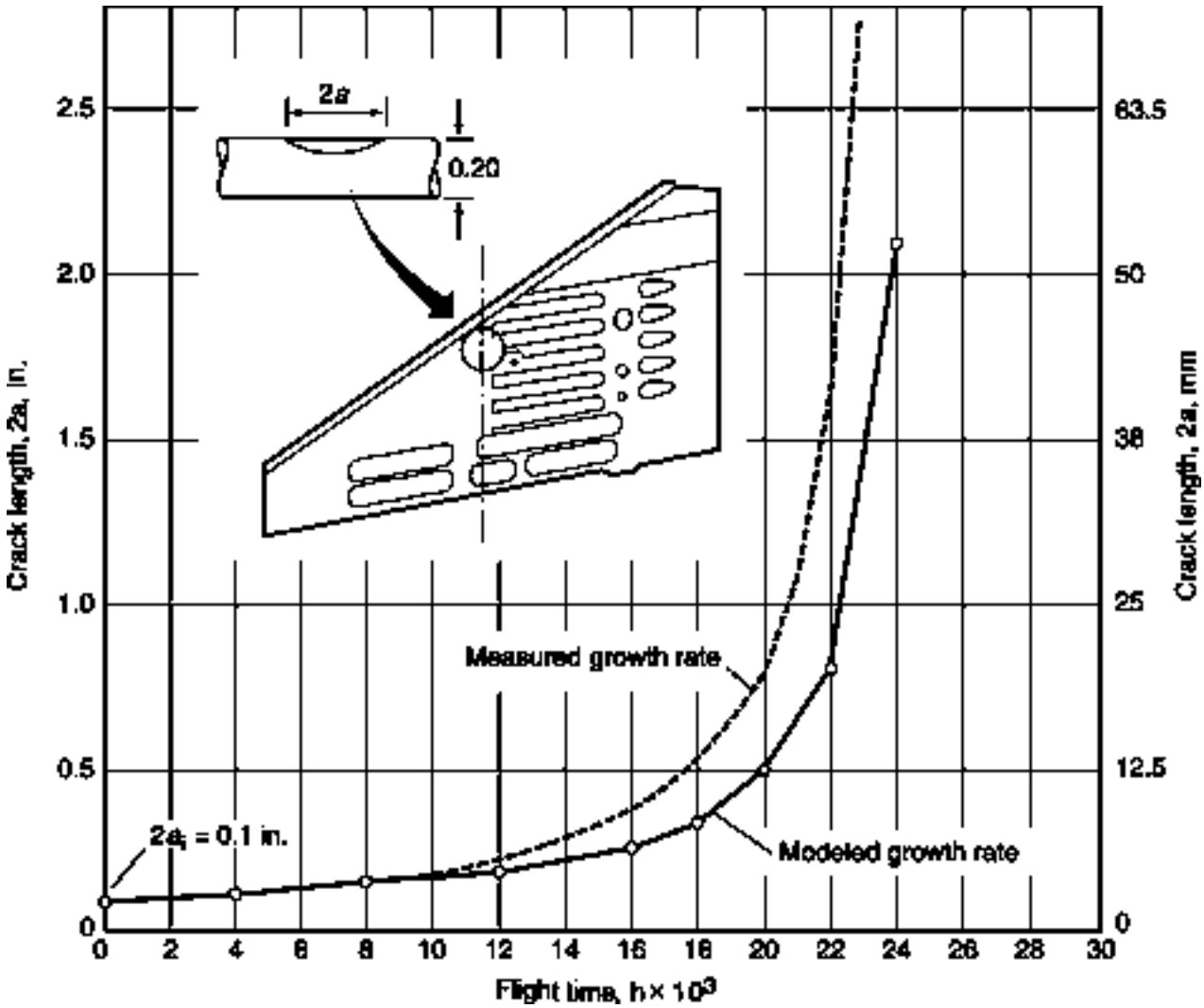


Fig. 11 The measured crack growth rate (crack length versus time) determined by optical measurements or fractographic evaluation used to validate life prediction estimates. In this example, for an aircraft wing, the predicted crack growth and the actual crack growth based on measured crack growth rates for a surface flaw are compared to validate the model.

References cited in this section

13. W.R. Hendricks, The Aloha Incident: A New Era for Aging Aircraft, *Structural Integrity of Aging Aircraft,* S.N. Altussi, et al., Ed., Springer-Verlag, 1991, p 153

27. D.J. Benac and R.A. Page, Integrating Design, Maintenance, and Failure Analysis to Increase Structural Valve Integrity, *Pract. Failure Anal.,* Vol 1 (No. 3), June 2001, p 31–43

Role of Nondestructive Inspection

Nondestructive inspection methodologies were and are still used for quality control of structures and components that are fabricated or manufactured. Even more so, however, the nondestructive inspection is being performed on structures and equipment in-service. The in-service nondestructive inspection techniques are used to quantify the size of a defect or flaw. This information is used in the life assessment model. The failure analyst often makes recommendations on what method should be used to determine the condition of the component. Therefore, the failure analyst as well as the life assessment engineer should understand the nondestructive methods and their limitations.

Nondestructive testing (NDT) and nondestructive evaluation (NDE) involve the use of noninvasive measurement techniques to gain information about defects in and various properties of materials, components, and structures, information that is needed to determine their ability to perform their intended function and prevent failure. With the

increasing performance demands imposed by today's highly competitive economic climate, it is often necessary to extend the service life or expectancy. Thus, it is critical to understand how safe and reliable that equipment or structure might be over this additional time of service. Nondestructive evaluation is used throughout the life cycle and includes the inspection of new, aging, and in-service structures and equipment.

Many measurement techniques are employed in NDE. Those most widely used are visual, liquid penetrant, magnetic particle, eddy-current, ultrasonic, and radiographic testing. These techniques are discussed in detail in the article "Nondestructive Evaluation and Life Assessment" in this Volume and in *Nondestructive Evaluation and Quality Control*, Vol 17 of the *ASM Handbook.*

Nondestructive evaluation plays a large role in damage-tolerant approach employed by the aerospace and structures communities. It is assumed that the service-induced damage in a part will increase in time. For the case of fatigue, the measure of damage is crack size. However, the same concepts would apply to high-temperature creep, corrosion, wear, or some other degradation mechanism. In the damage-tolerant approach, the use of inspection allows a reduction of this conservatism of the safe-life approach. Figure 12 is an illustration of the damage-tolerant approach to life assessment. It is recognized that flaws will be present, and NDE techniques are expected to remove those components from service that contain a flaw whose size is greater than the inspectable flaw. The expected life of the remaining components will now be greater. As indicated, the process could be repeated additional times, extending the life of remaining components.

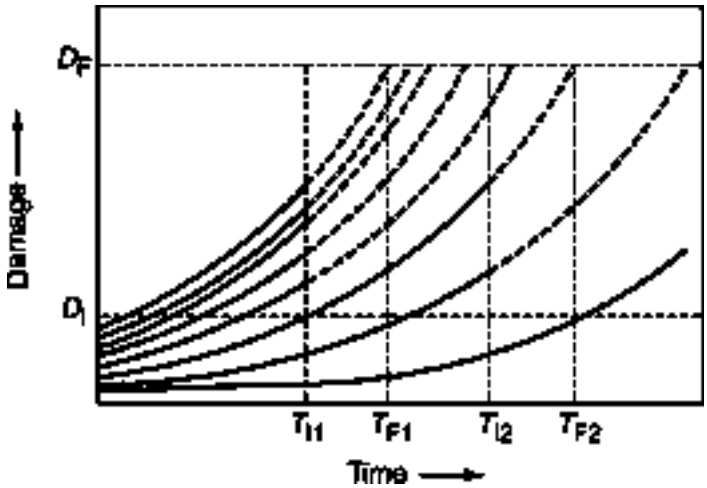


Fig. 12 Damage-tolerance approach to life assessment. Curves indicate progression of damage from different nucleating conditions, with broken segments representing regimes in which a perfect NDE technique would remove the component from service.

In addition to damage-tolerance requirements, nondestructive inspection of plant equipment, such as pressure vessels, tanks, and piping, is performed to assess fitness for service when exposed to corrosive media found in refineries, paper mills, and chemical plants.

Specific Life Assessment Methodologies

Fatigue/Damage-Tolerance Life Assessment. The lessons learned from the historic aircraft failures, such as the Comet, F-111 and the Aloha incident, have created the need to understand how to evaluate the life of structures subjected to fatigue loading. Life assessment of aircraft and also power-plant equipment subjected to fatigue loading largely stems from the development of the damage-tolerance philosophy based on fracture mechanics. Damage tolerance is the philosophy used for maintaining the structural safety of commercial transport, military aircraft, structures, and pressure vessels.

The individuals tasked with the determination of the component life depend on the knowledge and results of the many who can supply the necessary information. Life assessment of fatigue-loaded structures requires considering some of the following issues:

- Identification of fatigue critical structures
- Mission or profile of fatigue loading history
- Part geometries
- Damage-tolerant materials behavior
- Determination of inspection intervals
- Linear elastic and plastic fracture mechanics considerations
- Constant and variable loading on a structure

- The retardation effects of fatigue crack growth rate
- Corrosion effects for crack initiation and propagation
- Effects of stress concentration and notches
- The size, shape, and number of fatigue crack fronts
- Validation of prediction models using test data and field failures

For more detailed discussions on these fatigue-related topics, see the article “Fatigue Life Assessment” in this Volume and Ref 28.

The failure investigator's role in the fatigue life assessment is so critical because he or she often is the one who can provide information on whether multiple cracks occurred, how fast the crack was growing, and was any environment present to cause cracking or even reduce the crack rate due to corrosion buildup in the crack. Using investigative tools, the investigator can often identify the crack length and depth, which is used in estimating β factors.

The following case history demonstrates how the investigator contributes to the fatigue life assessment process by correlating the load profile to the rate of crack propagation (Ref 29). During testing, a thrust reverser aluminum fitting with a swaged bearing cracked, Fig. 13. The thrust reverser typically experiences a maximum load condition when deployed and a ground idle condition. Fractographic evaluation determined that fatigue crack initiation started in the bore of the hole. Further evaluation, using the scanning electron microscope, identified a fatigue pattern of narrow and wide striation spacing Fig. 13d that correlated to the load profile (Fig. 13e). Fatigue-striation measurements were made to estimate the total number of fatigue cycles for crack extension. When the cycles for crack extension were correlated to the total number of test cycles and predicted life cycle, it was concluded that an atypical stress condition was present that caused premature crack initiation and propagation.

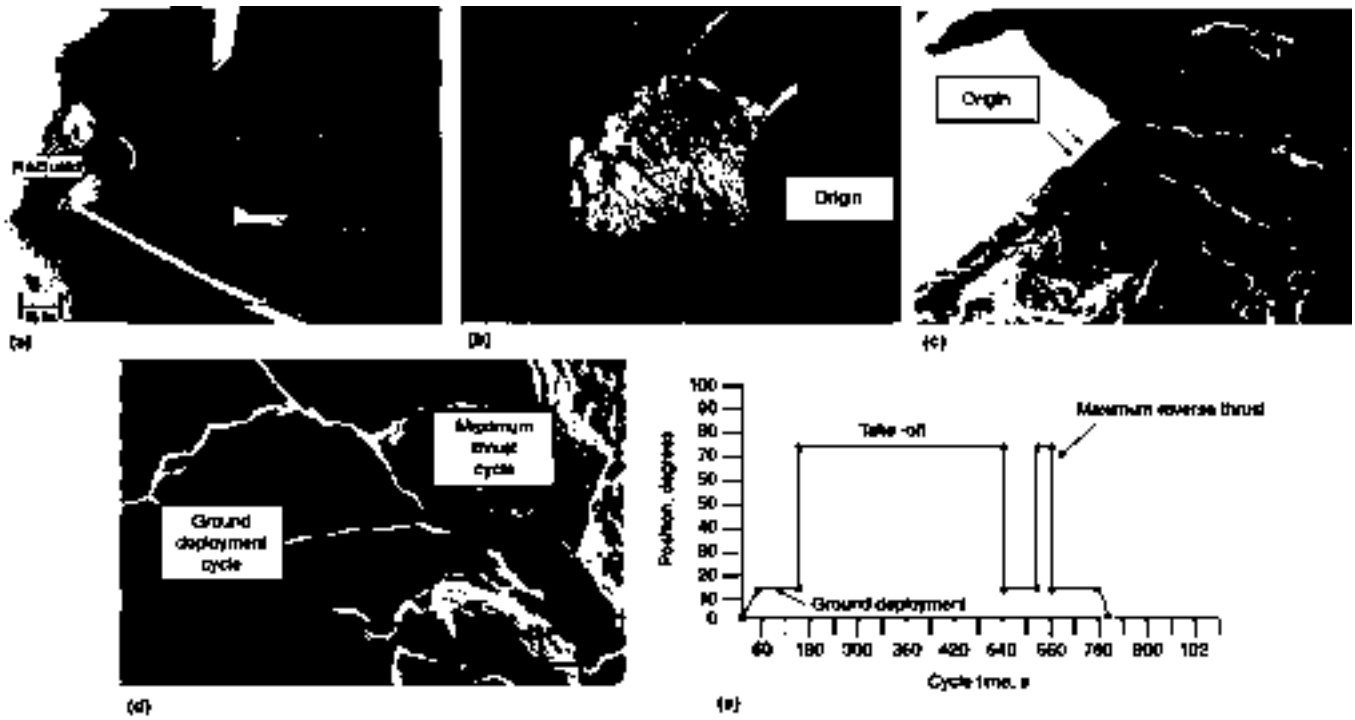


Fig. 13 Correlation of fatigue-striation patterns with the load history to estimate when a crack started in relation to the actual or predicted life of a part. (a) Cracked fitting. (b) and (c) Crack origin location. (d) Two distinct fatigue-striation patterns that were correlated to the load profile shown in (e)

Elevated-Temperature Life Assessment. Besides fatigue loading, elevated-temperature exposure is one of the leading causes of reduced life of structure components. Gas turbine blades, steam lines, heater, boiler, and superheater tubes are subjected to elevated temperatures that can cause degradation, deformation, bulging, cracking, or bursting. Thus, it is possible for the life expectancy to be reduced. For example, combustion turbine hot section rotating blades are fabricated from nickel-base superalloys and operate at high temperatures and aggressive environments. Both metallurgical and mechanical property degradation of the blade material occurs during service, which can limit the useful service life of blades. It is important to assess the condition of blades periodically and remaining life estimated to provide guidelines for replacement or reconditioning of the blades. Similarly, steel piping and tubing subjected to elevated temperature and stress suffer a reduction in life because of unplanned temperature excursions or higher operating stresses.

Therefore, it is often necessary to conduct blade, piping and tube life assessments. The advancements in elevated-temperature life assessment have made it possible to estimate remaining life in critical components used for power generation. Traditionally, microstructure-based and hardness-based techniques have been used to assess the condition of blades and tubes. In the past, they have been used mainly to estimate the "equivalent temperatures" of the tubes, which are then utilized in conjunction with standard rupture data to calculate the fractional creep life consumed. In recent years, analytical models correlating physical features such as cavities and scale are used to predict life. Relationships have also been established between the degree of creep cavitation, as measured in replicas, and the creep life fraction consumed. Furthermore, verification of the equivalent temperature in piping and tubing are based on scale thickness. Typical life assessment techniques for elevated-temperature exposure include:

- Life-fraction rule
- Parameter-based assessments
- Thermal mechanical fatigue
- Coating evaluations
- Hardness testing
- Microstructural evaluations
- Creep cavitation damage assessment
- Oxide-scale-based life prediction
- High-temperature crack growth methods

Greater detail of the elevated-temperature life assessment methods is discussed in the article "Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing" in this Volume and in Ref 26.

The role of the investigator is vital in determining the life-limiting elevated-temperature failure mechanisms such as creep, thermal fatigue, or embrittling phenomenon such as carbide coarsening. For example, in refineries and power plants, superheaters and heater tubes require remaining life assessments. Examination of the microstructure and correlating that to operating temperature and stress-rupture data, the life assessment engineer can estimate the remaining life on a component.

Elevated-temperature life assessment involves performing in situ replication of the microstructure, scale analysis, and sectioning of tube to perform stress-rupture testing. This remaining life analysis allows possible life extension of superheater or heater tubes past the previously considered retirement age. Accelerated stress-rupture tests allows for calculation of the actual operating stress and temperature. In addition, the maximum heater operating temperature can be calculated to obtain a safe run life until the desired tube replacement at a planned turnaround. Figure 14 shows a stress-rupture curve used for evaluation of turbine blades and heater tubes. Often the life-fraction rule is used to determine the remaining life (see the article "Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing" in this Volume).

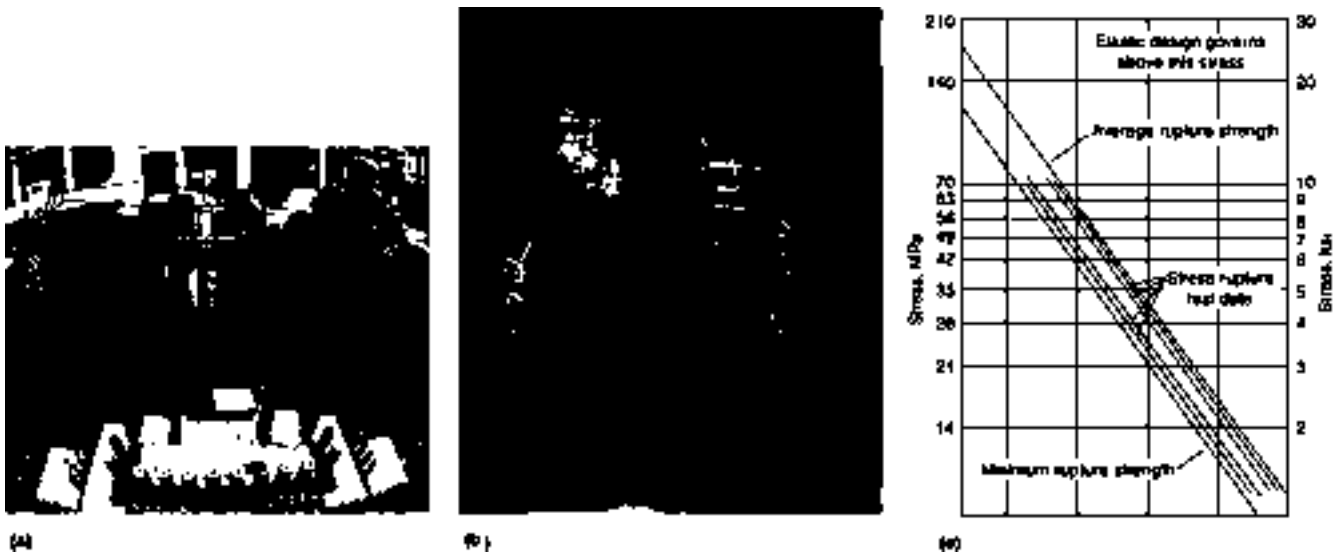
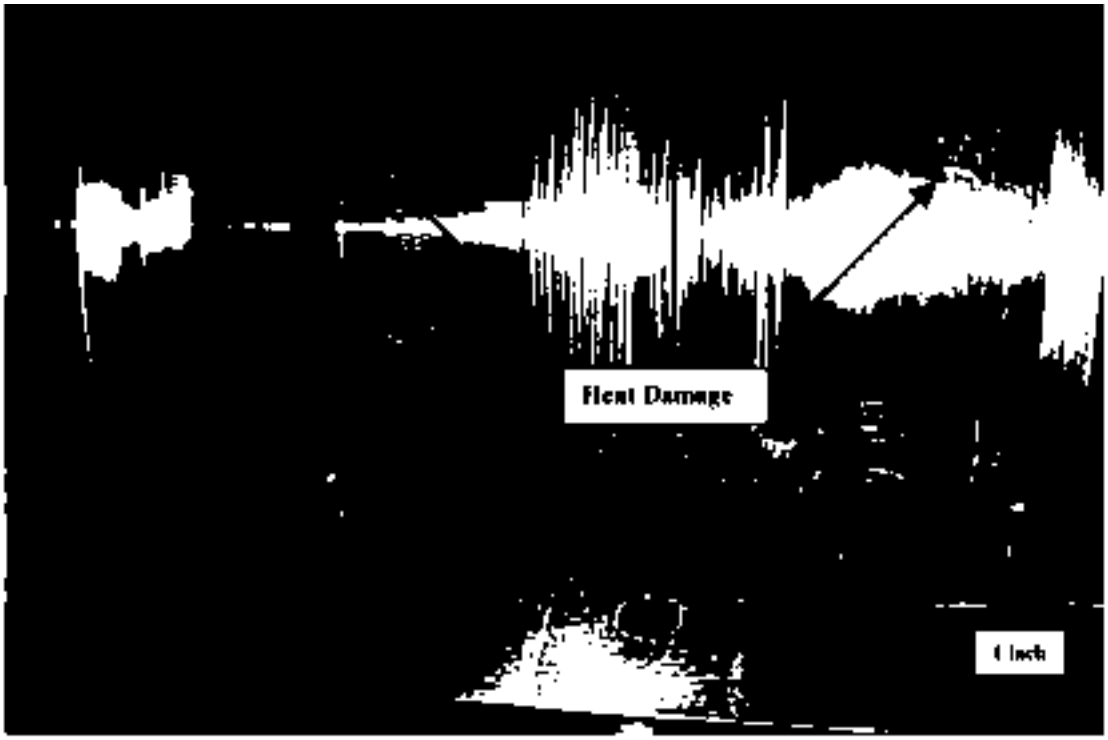


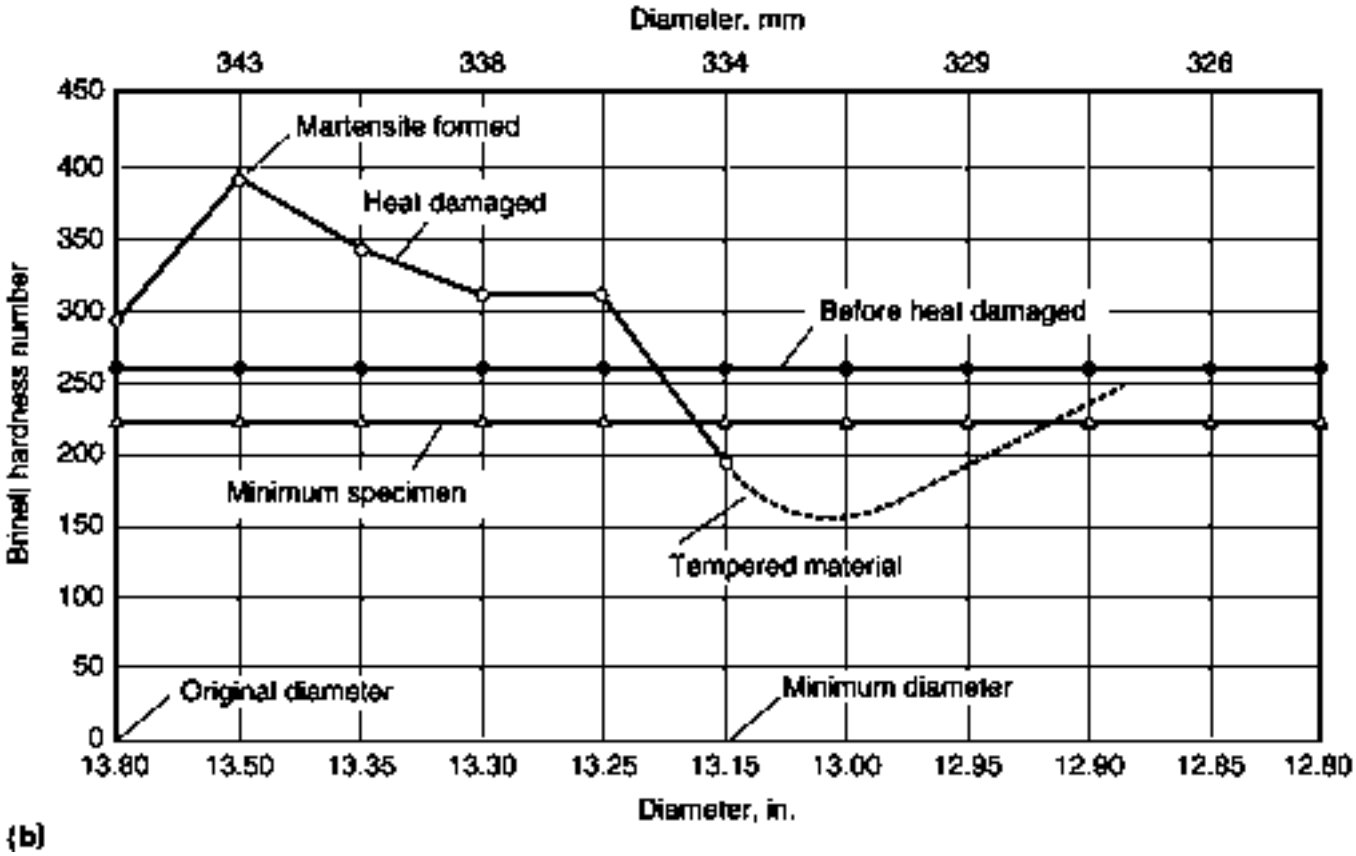
Fig. 14 Replication and microstructural evaluation combined with stress-rupture testing as performed on turbine blades and heater tubes to estimate the remaining life. (a) Land-based turbine. (b) Heater tubes. (c) Typical stress-rupture curve for 9Cr-1Mo material showing that the stress-rupture test results were still above the minimum rupture strength. Thus, the tubes were considered usable until the next inspection and testing.

Another type of elevated-temperature life assessment is evaluating structures and components that have been heat damaged by fire, lack of lubricant, or excessive temperature exposure. This often requires hardness testing, microstructural evaluations, or heat-damage discoloration assessment. API 579 (American Petroleum Institute) provides some guidance on evaluating structures exposed to fire damage (Ref 30).

During operation of rotating equipment that requires lubrication to reduce heating and the coefficient of friction, overheating can occur if the lubricant is lost or degraded. Therefore, a life assessment and reliability question is raised whether the equipment is fit for service. As an example, a rotating journal in a power-generator plant had lost lubricating oil, resulting in the overheating of the journal (Fig. 15). The loss of lubricant caused the temperatures to be hot enough to heat the journal above the critical transformation temperature. Consequently, a very brittle hard layer of martensite formed. Hardness testing showed that the surface could be ground to remove the martensite layer and still be above the minimum diameter of the journal, but hardness testing also showed a slight drop in hardness below the minimum diameter, indicating tempered material. A torsional fatigue analysis that considered load events and reduction in strength was performed and showed that the as-ground journal was acceptable and fit for service.



(a)



(b)

Fig. 15 Life assessment of rotating equipment damaged by exposure to excessive temperatures or loss of lubricant. (a) Journal that overheated because of the loss of lubricating oil. (b) Hardness test results of the heat-damaged region indicated the formation of a hard martensite layer and a tempered structure. After removal of the martensite, fatigue analysis verified that the journal could be put back in service.

Fitness for Service. When conducting investigations and determining how a component or structure failed or may fail, the users often want to determine if other equipment is fit for service (FFS). Many analytical methodologies are employed

and being developed to determine how fit a structure or equipment is for continued service. Historically, the ASME and API design codes for new pressure vessels do not address the fact that equipment degrades while in service and that structural deficiencies due to degradation from the original fabrication may be found during subsequent investigations. Due to this concern in the refinery and chemical industries, API 579, "Fitness-for-Service," was developed by a joint technical community to ensure structural integrity of pressurized equipment (Ref 30). Results of an FFS assessment can be used to make run-repair-replace decisions to help ensure that pressurized equipment containing detected flaws can continue to operate safely. One such analytical tool developed and mentioned in API 579 is the failure assessment diagram (FAD), which is discussed in the following paragraphs. The failure analyst may or may not be directly involved in structural life assessments using these analytical tools.

During life assessment, often there is a need for optimized modeling of fracture critical structural components and connections using elastic-plastic fracture mechanics. Such applications can require sophisticated analytical techniques that require time and/or resources beyond those available to the designer or analyst. One of the first engineering tools to address this dilemma was the Welding Institute Dimensionless CTOD Design Curve and was included in the first edition of the British Standards Institution (BSI) fitness-for-purpose guidance PD 6493 (Ref 31). The engineering tool receiving attention currently is the FAD. This approach has been used, primarily, in the electric power industry both in Great Britain as the R6 criteria (Ref 32) and the United States as both the failure assessment diagram (Ref 33) and the deformation plasticity failure assessment diagram (DPFAD) (Ref 34). Both the R6 and DPFAD approaches utilize the *J*-integral criteria for fracture driving force and resistance (i.e., toughness).

The United Kingdom Control Electricity Geometry Board (CEGB) first proposed the failure analysis diagram shown in Fig. 16(a). The approach addressed postyield fracture using an interpolation formula between the two cases of linear elastic fracture and plastic collapse. If a point describing the state of a component or structure falls below the R6 curve then the structure is considered safe. A point falling on or above the R6 curve is considered failure. Figure 16(b) shows a typical FAD curve for steels (Ref 35). K_r is the fracture ratio, which is the ratio of the crack driving force (including residual stress) to the material toughness. L_r or S_r is the collapse ratio, which is the ratio of the applied stress at design load to the applied stress at plastic collapse. For more details, see the article "Failure Assessment Diagrams" in this Volume.

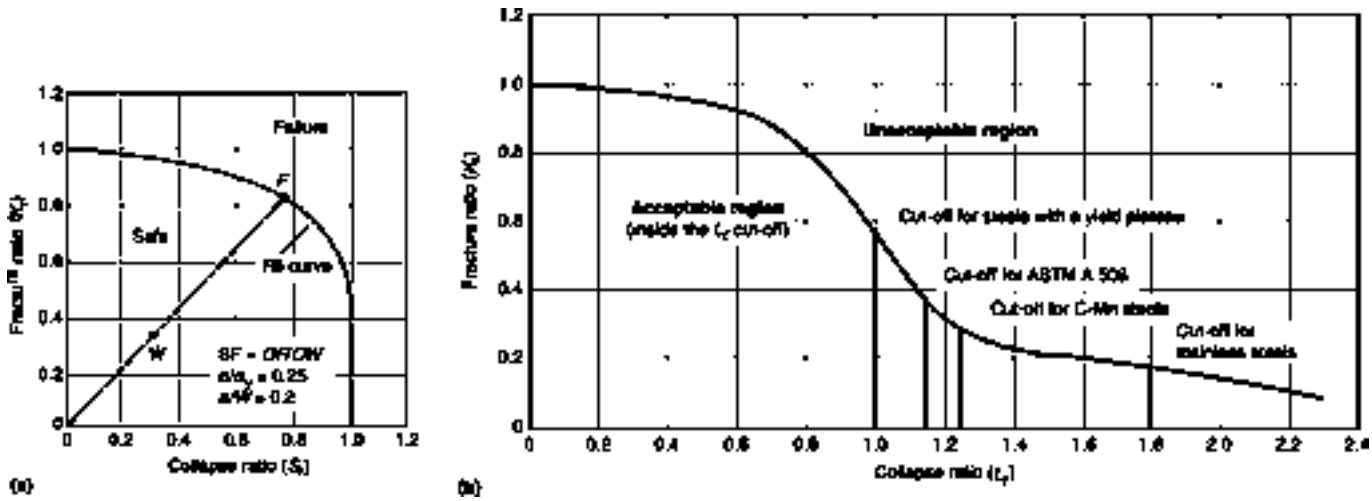


Fig. 16 Failure assessment diagrams used to evaluate the elastic-plastic behavior of structures to estimate safe operation. (a) The CEGB R6 curve. If a point describing the state of a component or structure falls below the R6 curve then the structure is considered safe. A point falling on or above the R6 curve is considered failure. (b) The failure assessment diagram from API 579 for steel (Ref 35). K_f is the fracture ratio, which is the ratio of the crack driving force (including residual stress) to the material toughness. L_r is the collapse ratio, which is the ratio of the applied stress at design load to the applied stress at plastic collapse. Similar to the CEGB R6 curve, points beneath the curve are considered safe and points on or above the curve are considered unacceptable.

Probabilistic Analyses. Life assessment often involves statistical calculations to quantify the probability of failure. This probability of failure is used to make decisions on the risk of continuing to operate a piece of equipment.

Application of probabilistic analysis methods for life assessment of structural components is becoming more prevalent as:

- Failure models continue to be enhanced for structural life prediction.
- Engineers increasingly realize that significant uncertainties and variations in loads, material properties, geometrical, and other parameters occur in real world structures.
- The need increases to quantify the probability of failure of structural components and to design for appropriate levels of safety.
- Computational resources become less of an impediment through enhancements in computational algorithms and computer efficiency.
- Engineers recognize that factor-of-safety approaches may not give the desired reliability or may lead to overdesigned structures with significant penalties.
- The need for a rational consistent methodology on which to base important decisions such as to inspect, replace, or repair a structure is realized.

The modern era of probabilistic structural design and analysis started after the Second World War. In 1947, a paper entitled, "The Safety of Structures," appeared in the *Transactions of the American Society of Civil Engineers.* This historical paper, written by A.M. Freudenthal, suggested that rational methods of developing safety factors for engineering structures should give due consideration to observed statistical distributions of the design factors.
An illustration of the probabilistic approach is shown in Fig. 17(a). In this methodology, uncertainties are explicitly modeled with probability density functions and the probability of failure is quantified. Several solutions methods exist to compute the probability of failure. One such method, Monte Carlo sampling, is shown schematically in Fig. 17(b). A more detailed discussion is found in the article "Analysis Methods for Probabilistic Life Assessment" in this Volume.

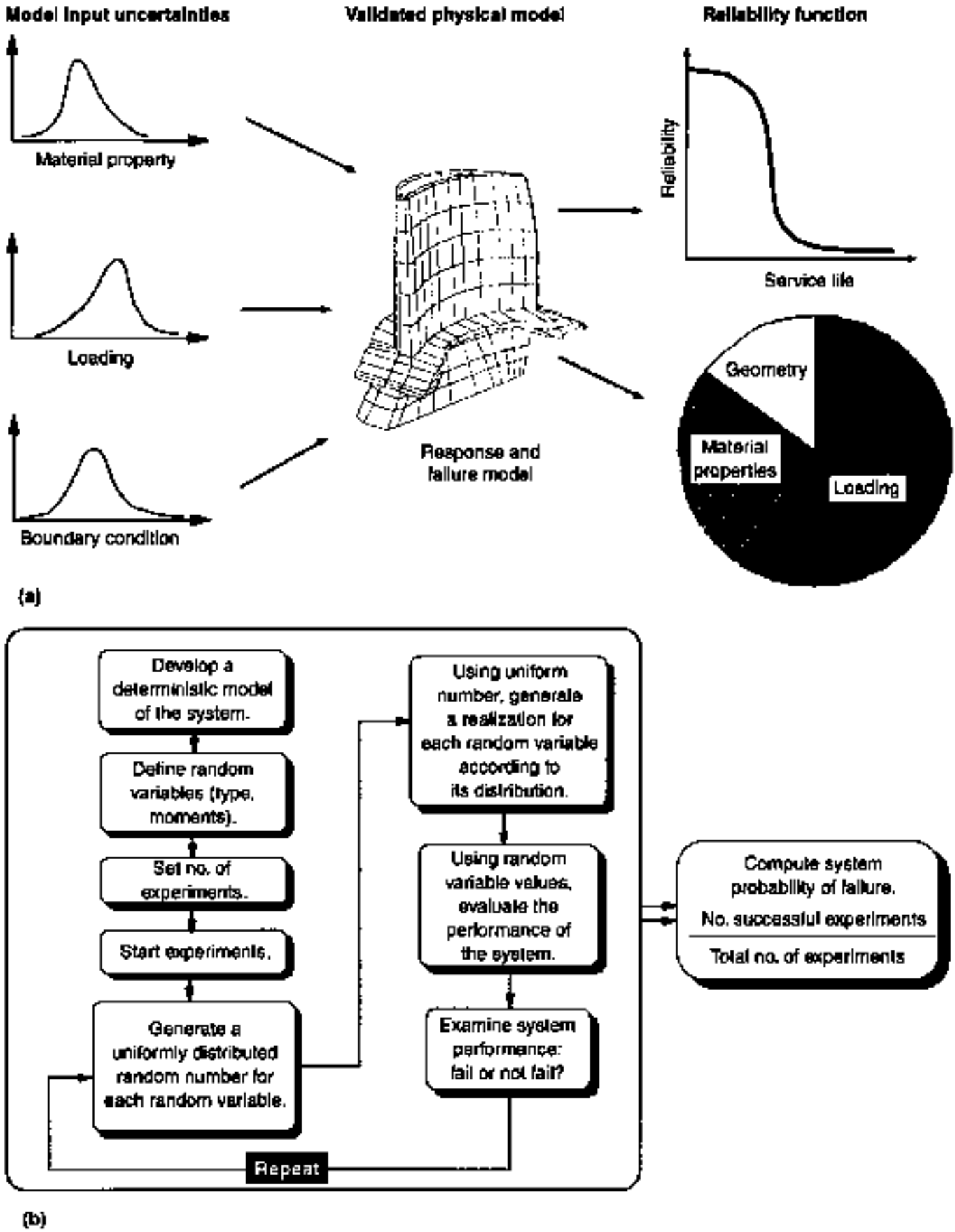


Fig. 17 (a) The probabilistic approach and (b) the Monte Carlo flow diagram used for a turbine blade

References cited in this section

26. A.F. Liu, High-Temperature Life Assessment, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 520–526

28. N.E. Dowling, Estimating Fatigue Life, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 250–262

29. D.J. Benac, "Interpreting Fatigue Fracture Features for Life Assessment of Aircraft Components," presented at ASM Materials Solutions 2000 (St. Louis, MO)

30. "Fitness-For-Service," Section 11, API Recommended Practice 579, American Petroleum Institute

31. "Guidance on Some Methods for the Derivation of Acceptance Levels for Defects in Fusion Welded Joints," PD 6493, 1st ed., British Standards Institution, London, March 1980

32. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, England, May 1986

33. V. Kumar, M.D. German, and C.F. Shih, "An Engineering Approach for Elastic-Plastic Fracture Analysis," EPRI NP-1931, Project 1237-1, General Electric Company, R&D Center, July 1981

34. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, *Fracture Resistance Curves and Engineering Applications,* Vol II, *Elastic-Plastic Fracture: Second Symposium,* C.F. Shah, and J.P. Gudas, Ed., STP 803, American Society for Testing and Materials, 1983, p II-206 to II-238

35. "Fitness-For-Service," Section 11, API Recommended Practice 579, American Petroleum Institute, 2000, p 9–61

Conclusions

Because the costs and stakes are high when failure occurs and because equipment and structures are required to last longer with higher-performance demands, life assessment of structures and equipment is necessary. For meaningful and accurate life assessment of components, a multidisciplined approach is needed that considers design parameters, life-limiting factors, and the correct understanding of the life-limiting mechanism for fatigue life or elevated-temperature life assessment.

The failure analyst provides engineering insights for life assessment of components such as the location of origin, the life-limiting failure mechanism, the rate of propagation or time for degradation, the actual service or operating conditions, and validation of the life assessment model. If the failure analyst is not involved in these tasks or an incorrect determination is made related to these failure and degradation mechanisms, then the life assessment will prove of little to no value.

The understanding of failures has revolutionized design philosophies, nondestructive inspection techniques and practices, material development, material processing and controls; redefined the criteria for failure; created structural-integrity programs; enhanced analytical modeling and prediction techniques; and resulted in probabilistic approach to understanding failure. Furthermore, industries and the engineering community have fortified a commitment to avoid the recurrence of failures through improved life assessment techniques. No matter how large or small a failed structural component or equipment, understanding how and why the component or structure failed or degraded is accomplished through a thorough failure investigation and accurate life assessment.

Acknowledgment

This article is a result of over twenty-five years of working alongside many individuals who have demonstrated by example the value of in-depth failure investigations and the commitment to understand the mysteries of failure and how to prevent recurrence. The writer appreciates the authors who have contributed to the life assessment section of this Volume and whose technical input were used to add breath to this article.

Furthermore, worthy of mention are the numerous original equipment manufacturers and users of structural components and equipments that have designated, operated, and maintained safe structures and equipments.

References

1. J.J. Duga et al., *The Economic Effects of Fracture in the United States,* Report to the National Bureau of Standards, Battelle Columbus Laboratories, March 1983

2. L.F. Coffin, Damage Evaluation and Life Prediction for High-Temperature Gas Turbine Materials, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials,* EPRI AP-4477, Project 2382-3, Electric Power Research Institute, April 1986, p 1.1–1.17

3. R. Viswanathan and R.B. Dooley, Creep Life Assessment Techniques for Fossil Power Plant Boiler Pressure Parts, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials,* EPRI AP-4477, Project 2382-3, Electric Power Research Institute, April 1986, p 2.1–2.28

4. H.P. Leighly, Jr., B.L. Bramfitt, and S.J. Lawrence, RMS Titanic: A Metallurgical Problem, *Pract. Failure Anal.,* Vol 1 (No. 2), April 2001, p 10–13, 33–37

5. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* John Wiley & Sons, 1976, p 230

6. S. Ross, Tacoma Narrows, 1940, *Construction Disasters,* 1st ed., McGraw-Hill, 1984, p 216–239

7. Y. Billah and R. Scanlan, Resonance, Tacoma Narrows Bridge Failure and Undergraduate Physics, *J. Physics,* Vol 59 (No. 2), Feb 1991, p 118–123

8. E.R. Parker, *Brittle Behavior of Engineering Structures,* Wiley, 1957

9. S. Atallah, U.S. History's Worst LNG Disaster, *Firehouse,* Jan 1979, p 29

10. T. Swift, Damage Tolerance Certification of Commercial Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 566–576

11. V.A. Hindas, *F-111 Design Experience: Use of High Strength Steels,* AIAA 2nd Aircraft Design and Operations Meeting

12. R. Viswanathan and J. Stringer, Failure Mechanisms of High Temperature Components in Power Plants, *Trans. ASME,* Vol 122, July 2000

13. W.R. Hendricks, The Aloha Incident: A New Era for Aging Aircraft, *Structural Integrity of Aging Aircraft,* S.N. Altussi, et al., Ed., Springer-Verlag, 1991, p 153

14. N1819U, NTSB-AAR-90-06, Aircraft Accident Report, United Airlines, Inc., Nov 1990

15. *Interim Guidelines: Evaluation, Repair, Modification and Design of Steel Moment Frames,* FEMA-267, Federal Emergency Management Agency

16. R.W. Staehl, Combining Design and Corrosion for Predicting Life, *Proc. Conf. Life Prediction of Corrodible Structures,* Vol 1, NACE International, 1994, p 138

17. D.J. Benac and L.J. Goland, Failures of Pressure Vessels, Lesson 15, *ASM Practical Failure Analysis Course,* to be published

18. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

19. P.F. Timmins, Operating Stress Maps for Failure Control, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 460

20. M.P. Kaplan and T.A. Wolff, Life Extension and Damage Tolerance of Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* p 557–565

21. D.J. Benac and D.N. Hopkins, Investigation of Fatigue-Induced Socket-Welded Joint Failures for Small-Bore Piping Used in Power Plants, *Pract. Failure Anal.,* Vol 1 (No. 2), April 2001, p 71–82

22. H.R. Millwater, S.H. Fitch, Y.-T. Wu, D.S. Riha, M.P. Enright, G.R. Leverant, R.C. McClung, C.J. Kuhlman, G.G. Chell, and Y. Lee, "A Probabilistically-Based Damage Tolerance Analysis Computer Program for Hard Alpha Anomalies in Titanium Rotors," Paper 2000-GT-0421, IGTI, ASME Turbo Expo, 2000

23. "Aircraft Structural Integrity and Failure Analysis of Structural Components," Southwest Research Brochures

24. D.J. Benac and J.P. Pendley, "Integration of Design, Failure Analysis and Fleet Management to Prevent Aircraft Structural Failures," Asian Aircraft Conference (Singapore), 1991

25. W.D. Pilkey et al., *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, 1997

26. A.F. Liu, High-Temperature Life Assessment, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 520–526

27. D.J. Benac and R.A. Page, Integrating Design, Maintenance, and Failure Analysis to Increase Structural Valve Integrity, *Pract. Failure Anal.,* Vol 1 (No. 3), June 2001, p 31–43

28. N.E. Dowling, Estimating Fatigue Life, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 250–262

29. D.J. Benac, "Interpreting Fatigue Fracture Features for Life Assessment of Aircraft Components," presented at ASM Materials Solutions 2000 (St. Louis, MO)

30. "Fitness-For-Service," Section 11, API Recommended Practice 579, American Petroleum Institute

31. "Guidance on Some Methods for the Derivation of Acceptance Levels for Defects in Fusion Welded Joints," PD 6493, 1st ed., British Standards Institution, London, March 1980

32. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, England, May 1986

33. V. Kumar, M.D. German, and C.F. Shih, "An Engineering Approach for Elastic-Plastic Fracture Analysis," EPRI NP-1931, Project 1237-1, General Electric Company, R&D Center, July 1981

34. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, *Fracture Resistance Curves and Engineering Applications,* Vol II, *Elastic-Plastic Fracture: Second Symposium,* C.F. Shah, and J.P. Gudas, Ed., STP 803, American Society for Testing and Materials, 1983, p II-206 to II-238

35. "Fitness-For-Service," Section 11, API Recommended Practice 579, American Petroleum Institute, 2000, p 9–61

Failure Assessment Diagrams

Harold S. Reemsnyder, Fracture Technology Associates

Introduction

OPTIMIZED MODELING of fracture-critical structural components and connections requires the application of elastic-plastic fracture mechanics. Such applications, however, can require sophisticated analytical techniques that require time and/or resources beyond those available to the designer or analyst. One of the first engineering tools to address this dilemma was The Welding Institute dimensionless crack tip opening displacement (CTOD), design curve (Ref 1), which was included in the first edition of the British Standards Institution (BSI) fitness-for-purpose guidance PD 6493 (Ref 2). The engineering tool receiving attention currently is the failure assessment diagram (FAD). This approach has been used primarily in the electric power industry both in Great Britain as the R6 criteria (Ref 3) and in the United States as both the FAD (Ref 4) and the deformation plasticity failure assessment diagram (DPFAD) (Ref 5). Both the R6 and DPFAD approaches use the *J*-integral criteria for fracture-driving force and resistance (i.e., toughness). The second edition of the BSI PD 6493 (Ref 6) presents failure assessments in the form of FADs using CTOD to characterize both crack-driving force and fracture toughness. The British Standards Institution has issued an updated and expanded version of PD 6493 as BS 7910 (Ref 7). In addition, the FAD is the recommended approach for the damage-tolerance analysis of ship structures (Ref 8). The background and derivations of the FAD were the subject of a theme issue of *The International Journal of Pressure Vessels and Piping* (Ref 9). Applications of, and experience with, the FAD concept were discussed in a seminar held by the Institution of Mechanical Engineers (Ref 10).

References cited in this section

1. F.M. Burdekin and M.G. Dawes, Practical Uses of Linear Elastic and Yielding Fracture Mechanics with Particular Reference to Pressure Vessels, *Practical Applications of Fracture Mechanics to Pressure Vessel Technology,* Institution of Mechanical Engineers, London, 1971, p 28–37

2. "Guidance on Some Methods for the Derivation of Acceptance Levels for Defects in Fusion Welded Joints," PD 6493:1980, 1st ed., British Standards Institution, London, March 1980

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

4. V. Kumar, M.D. German, and C.F. Shih, "An Engineering Approach for Elastic-Plastic Fracture Analysis," EPRI NP–1931, Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, July 1981

5. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, *Elastic-Plastic Fracture: Second Symposium, Vol II, Fracture Resistance Curves and Engineering Applications,* STP 803, C.F. Shah and J.P. Gudas, Ed., ASTM, 1983, p II-206 to II-238

6. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," PD 6493:1991, 2nd ed., British Standards Institution, London, August 1991

7. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," BS 7910:1998, British Standards Institution, London, 1998

8. "Guide to Damage Tolerance Analysis of Marine Structures," SSC-409, Ship Structures Committee, Washington, DC, 2000

9. Theme Issue: The Revisions to the Structural Integrity Assessment Method CEGB/R6, *Int. J. Pressure Vessels Piping*, Vol 32, 1988

10. "20 Years of R6," S 458, Institution of Mechanical Engineers, London, Nov 1996

Origin and Description of the Failure Assessment Diagram

The United Kingdom Central Electricity Generating Board (CEGB) first proposed a failure assessment diagram (Fig. 1) based on the two-criteria approach of Dowling and Townley (Ref 11). The CEGB approach (Ref 3, 12, 13) addressed post-yield fracture by an interpolation formula between two limiting cases: linear elastic fracture and plastic collapse. The interpolation formula, known as the *failure assessment* or *R6 curve* (Fig. 1), developed from the Dugdale solution for a cracked, infinite plate, is:

$$K_r = \frac{S_r}{\sqrt{\frac{8}{p^2}.\ln\left[\sec\left(0,5.p.S_r\right)\right]}} \qquad \text{(Eq 1)}$$

The right-hand side of Eq 1 may be viewed as the plastic correction to the small-scale yielding prediction. In Eq 1, the fracture ratio K_r is:

$$K_r = \frac{K_1}{K_{mat}} \qquad \text{(Eq 2)}$$

and the collapse ratio S_r is:

$$S_r = \frac{s}{s_c} \qquad \text{(Eq 3)}$$

In Eq 2, K_I is the stress-intensity factor (a function of nominal stress σ, crack size *a*, and geometry) at fracture of the component, and K_{mat} is the linear elastic fracture toughness of the component. (K_{mat}, the linear elastic fracture toughness, is the value of K_I at fracture and, for a given material condition, is a function of thickness and temperature. The lower bound to K_{mat}, for a given temperature, is the plane-strain fracture toughness K_{Ic}.) In Eq 3, σ is the applied (remote) stress in the component at fracture and σ_c is the applied (remote) stress at plastic collapse of the cracked component.

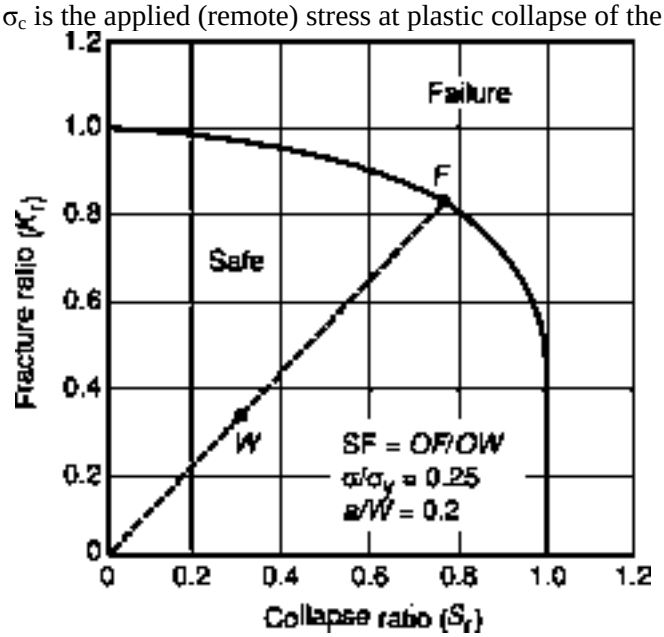


Fig. 1 Failure assessment diagram (R6 curve) as proposed by the United Kingdom Central Electricity Generating Board. SF, safety factor; σ, applied stress; σ_y, yield strength; a, crack length; W, panel width

If a point describing the state of a component or structure (e.g., point *W*) falls below the R6 curve (Fig. 1), the structure is considered to be *safe.* A point falling on or above the R6 curve represents *failure.* If a ray were constructed from the origin *O* through point *W* to the R6 curve (i.e, point F), the safety factor on load is the length of the ray from the origin to the intersection of the ray and the R6 curve divided by the length of the ray from the origin to point *W* (e.g., *OF*/*OW*). One application of the R6 curve is shown in Fig. 1 for a center-cracked panel with the crack size 20% of the panel width *W* and the applied stress σ 25% of the yield stress σ_y. In Fig. 1, the safety factor on load is *OF*/*OW* = 2.5.

The FAD as generally used today is, essentially, a graphical model, shown schematically in Fig. 2, that reflects the stress-strain curve of the material. The FAD consists of two elements: the failure assessment curve (FAC) and the failure assessment point (FAP). The collapse ratio, S_r or L_r, is the ratio of the applied stress at design load to the applied stress at plastic collapse. (S_r is a function of flow stress, σ_f, which is the average of the yield and tensile strengths, respectively, σ_y and σ_u; L_r is a function of yield stress.) K_r, the fracture ratio, is the ratio of the crack-driving force (including residual stress when applicable) to the material toughness (expressed as K_I, *J*-integral, or CTOD, δ). The FAC defines the critical combination of service loads, material stress-strain properties, and geometry of the cracked member at which failure might be expected. The FAP defines the state of a member containing a flaw of given size under specific service loads. The factor of safety (on load) against failure for a given FAP (in the absence of residual stresses) is the ratio of the line segments *OB*/*OA* (Fig. 2). In Fig. 2, it is observed that, in the presence of residual stresses, the factor of safety is reduced (i.e., *O′B′*/*O′A′* is less than *OB*/*OA*).

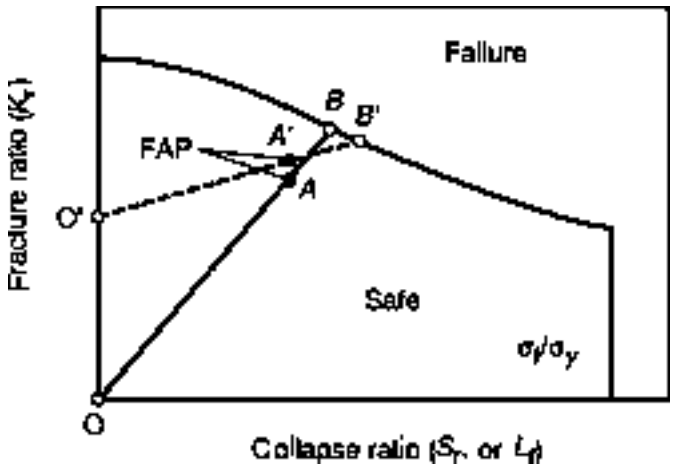


Fig. 2 Example of a typical failure assessment diagram. FAP, failure assessment point; FAC, failure assessment curve; σ_f, flow strength, average of yield and tensile strengths; σ_y, yield strength; solid line, without residual stress; dashed line, with residual stress

Current applications of this form of the FAD are: CEGB R6, Revision 3 (Ref 3), BS 7910 (Ref 7), EPRI/GE model (Ref 4), deformation plasticity FAD (Ref 5), ASME Section XI Code Case (DPFAD) for ferritic piping (Ref 14), and damage-tolerance analysis of ship structures (Ref 8).

References cited in this section

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, “Assessment of the Integrity of Structures Containing Defects,” R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

4. V. Kumar, M.D. German, and C.F. Shih, “An Engineering Approach for Elastic-Plastic Fracture Analysis,” EPRI NP–1931, Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, July 1981

5. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, *Elastic-Plastic Fracture: Second Symposium, Vol II, Fracture Resistance Curves and Engineering Applications,* STP 803, C.F. Shah and J.P. Gudas, Ed., ASTM, 1983, p II-206 to II-238

7. “Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures,” BS 7910:1998, British Standards Institution, London, 1998

8. “Guide to Damage Tolerance Analysis of Marine Structures,” SSC-409, Ship Structures Committee, Washington, DC, 2000

11. A.R. Dowling and C.H.A. Townley, The Effect of Defects on Structural Failure: A Two-Criteria Approach, *Int. J. Pressure Vessels Piping,* Vol 3, 1975, p 77–107

12. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

13. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, Assessment of the Integrity of Structures Containing Defects, *Int. J. Pressure Vessels Piping*, Vol 32, theme issue, 1988, p 3–104

14. J. Bloom, Appendix J—The Deformation Plasticity Failure Assessment Diagram (DPFAD) Approach to Evaluation of Flaws in Ferritic Piping, presented at the *European Symposium on Elastic-Plastic Fracture Mechanics* (Frieburg, W. Germany) Oct 1989

Current R6 Failure Assessment Diagrams

Chell (Ref 15) and Bloom (Ref 16) pointed out that K_r (Eq 1) could be interpreted as a ratio of *J*-integrals:

$$K_r = \sqrt{\frac{J_e}{J}} = f(S_r) \qquad \text{(Eq 4)}$$

where J_e, the elastic component of total *J*-integral *J*, is:

$$J_e = \frac{K_1^2}{E}.(1-\nu^2) \qquad \text{(Eq 5)}$$

where *E* is Young's modulus and ν is Poisson's ratio.

The current version of the R6 FAD (Ref 3, 12, 13) expresses K_r as a ratio of the linear elastic stress-intensity factor associated with the stress and the flaw to the *J*-integral toughness. Further, the current R6 presents three options for the FAC and three analysis categories for the FAP.

Option 1 FAC is recommended for materials with a low initial work-hardening rate, or as an initial screening test, or when details of the material stress-strain curve are not known. Option 1 is not recommended for materials with discontinuous stress-strain curves. The Option 1 FAC is an empirical fit to option 2 FACs for a variety of steels (including an elastic-perfectly plastic material) but biased toward a lower bound (Ref 9).

Option 2 FAC is recommended for materials with a high initial work-hardening rate, (e.g., strain-aging mild steels), or for materials with a discontinuous yield point, or when the complete material true stress-strain curve is known. The Option 2 FAC was developed from expressions for *J*-integrals from the EPRI/GE handbook (Ref 4) reformulated to use actual true stress-strain curves. Also, approximations erring on the conservative side were introduced to make the formulae geometrically independent, thus obviating the need for fully plastic power-law solutions for each geometry (Ref 9).

Option 3 FAC is the most sophisticated of the three options and requires a *J*-integral *R*-curve for the subject material. (An *R*-curve is a plot of the crack driving force (K_I, *J*, or CTOD δ) versus stable crack extension under monotonic, increasing load. Stable crack extension ceases when the loading rate goes to zero.) Option 3 acknowledges the presence of stable crack extension and a concomitant increase in fracture toughness. The construction of the Option 3 FAC is identical to that of the EPRI/GE FAD (Ref 4) and Bloom's DPFAD (Ref 5). Also, Appendix 8, R6, Revision 3, presents an FAC for C-Mn steels. The Appendix 8 FAC is an empirical lower bound to R6 Option 1 FACs for a variety of C-Mn steel plates and weld metal, finite element *J*-analyses (R6 Option 3), and tests on center-cracked panels and compact tension specimens (Ref 9). In Options 1, 2, and 3, the collapse ratio is expressed as L_r, while Appendix 8 uses S_r to represent the same.

Analysis Categories. Analysis Category 1 acknowledges elastic-plastic interaction of applied and residual stresses in the computation of K_I. Category 2 consists of two FAPs—one using a_0 and one using $a_0 + \Delta a_g$ where Δa_g is the maximum stable crack extension permitted in a *J*-integral fracture toughness test. Category 3 consists of several FAPs computed using various amounts of stable crack extension Δa as in the EPRI/GE FAD (Ref 4) and the DPFAD (Ref 5). The plastic-collapse load formulae used to compute L_r for the FAP are reviewed in Ref 9.

References cited in this section

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, “Assessment of the Integrity of Structures Containing Defects,” R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

4. V. Kumar, M.D. German, and C.F. Shih, “An Engineering Approach for Elastic-Plastic Fracture Analysis,” EPRI NP–1931, Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, July 1981

5. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, *Elastic-Plastic Fracture: Second Symposium, Vol II, Fracture Resistance Curves and Engineering Applications,* STP 803, C.F. Shah and J.P. Gudas, Ed., ASTM, 1983, p II-206 to II-238

9. Theme Issue: The Revisions to the Structural Integrity Assessment Method CEGB/R6, *Int. J. Pressure Vessels Piping,* Vol 32, 1988

12. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

13. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, Assessment of the Integrity of Structures Containing Defects, *Int. J. Pressure Vessels Piping,* Vol 32, theme issue, 1988, p 3–104

15. G.G. Chell, A Procedure for Incorporating Thermal and Residual Stresses into the Concept of a Failure Assessment Diagram, *Elastic-Plastic Fracture,* STP 668, ASTM, 1979, p 581–605

16. J.M. Bloom, Prediction of Ductile Tearing Using a Proposed Strain-Hardening Failure Assessment Diagram, *Int. J. Fract.,* Vol 16, 1980, p R73–R77

Current Fracture Criteria of BS 7910

BS 7910 (Ref 7), the successor to PD 6493 (Ref 6), incorporates a three-tiered (or 3-level) FAD criterion using CTOD as the characterization of both crack-driving force and fracture toughness.

The tiered approach to the FAD treats the problem with increasing analytical sophistication and data requirements and decreasing conservatism. The first tier or level (the least sophisticated and the most conservative) requires only an elastic fracture mechanics analysis and uses single-value estimates of fracture toughness (e.g., δ_c, δ_u, or δ_m) (Table 1).

Table 1 Characterizations of elastic-plastic fracture toughness

Final event	**Location on toughness-temperature curve**	**Significant stable crack extension**	***J*-integral toughness, J_{mat}**	**CTOD toughness, δ_{mat}**
Unstable fracture or onset of pop-in	Lower shelf	No	J_c	δ_c
Unstable fracture or onset of pop-in	Transition	Yes	J_u	δ_u
First attainment of maximum load plateau	Upper shelf	Yes	J_m	δ_m

J_{mat}, general *J*-integral fracture toughness of material; CTOD, crack-tip opening displacement; δ_{mat}, general CTOD toughness of material; J_c, *J*-integral fracture toughness, no significant stable crack extension, unstable fracture; δ_c, CTOD fracture toughness, no significant stable crack extension, unstable fracture; J_u, *J*-integral fracture toughness, significant stable crack extension, unstable fracture; δ_u, CTOD fracture toughness, significant stable crack extension, unstable fracture; J_m, *J*-integral fracture toughness, significant stable crack extension, plastic collapse; δ_m, CTOD fracture toughness, significant stable crack extension, plastic collapse

On the other hand, the highest tier (the most sophisticated and least conservative) requires an elastic-plastic fracture mechanics analysis and uses the tearing resistance of the material (*J* or CTOD *R*-curve). In addition, the highest tier requires the expression of the true stress-strain curve of the material. The procedure is to evaluate a given situation at the first tier, and if the FAP falls below the FAC, safe performance is concluded. If, however, the FAP falls on or above the FAC, then the investigator goes to the second tier analysis and so on. If an unsafe situation is predicted at the highest tier, a redesign or selection of a tougher material is required.

Level 1 FAC (Fig. 3), a recasting of the dimensionless CTOD design curve (Ref 1) (based on wide-plate fracture tests), may be thought of as a screening method and includes a safety factor of approximately 2 on flaw size in the construction

of the FAC. For the FAP (not shown in Fig. 3), K_r is the square root of the ratio of the applied elastic CTOD δ_e to the material CTOD toughness δ_{mat} or

$$K_r = \sqrt{\frac{d_e}{d_{\max}}} \qquad \text{(Eq 6)}$$

The elastic CTOD δ_e, for steels (including stainless steels) and aluminum alloys, is determined from the elastic stress-intensity factor K_I concomitant with σ_{max}, the sum of the applied and residual stresses:

$$d_e = \frac{K_1^2}{s_y.E} \text{ for } \frac{s_{\max}}{s_y} \le 0,5 \qquad \text{(Eq 7)}$$

or

$$d_e = \frac{K_1^2}{s_y.E}\left(\frac{s_y}{s_{\max}}\right)\left(\frac{s_{\max}}{s_y} - 0,25\right) \text{ for } \frac{s_{\max}}{s_y} \le 0,5 \qquad \text{(Eq 8)}$$

No elastic-plastic interaction of applied and residual stresses is acknowledged in the computation of K_r (i.e., the validity of superposition is assumed). The material CTOD toughness δ_{mat} is the fracture-test-estimated value δ_c, δ_u, or δ_m (Ref 17, 18).

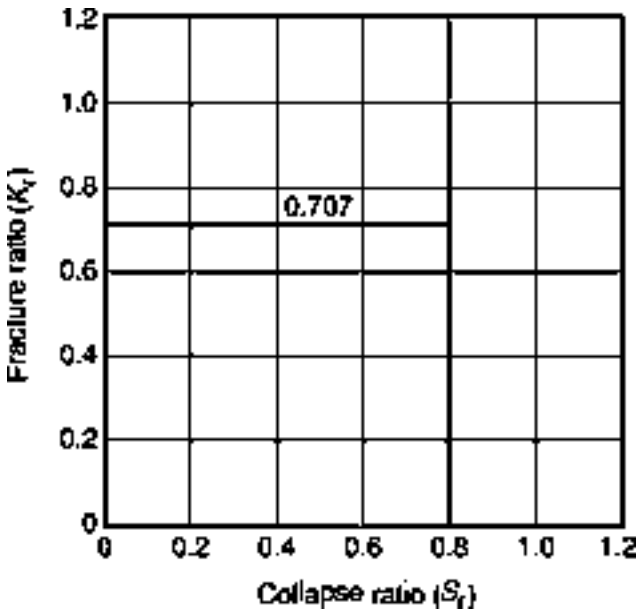


Fig. 3 Level 1 failure assessment diagram (BS 7910)

S_r is the ratio of the effective net-section stress σ_{ref} to the flow stress:

$$S_r = \frac{s_{\text{ref}}}{s_f} \qquad \text{(Eq 9)}$$

(The effective net-section stress σ_{ref} is the net-section stress computed in such a way that S_r is equivalent to the ratio of the applied (remote) stress to the applied (remote) stress at plastic collapse [Ref 7, 19, 20].)

Level 2a FAC (Fig. 4) is meant for low work-hardening materials and is identical to the original form of the R6 curve (Eq 1) based on Dugdale's strip-yield model and is expressed as an inverse function of the log-secant of L_r.

$$K_r = \frac{L_r \cdot \sigma_y / \sigma_f}{\sqrt{\frac{8}{\pi^2} \cdot \ln\left[\sec\left(\frac{\pi \cdot \sigma_y}{2 \cdot \sigma_f} \cdot L_r\right)\right]}} \qquad \text{(Eq 10)}$$

L_r in Eq 10 is the ratio of the effective net-section stress σ_{ref} to the yield stress:

$$L_r = \frac{s_{\text{ref}}}{s_y} \qquad \text{(Eq 11)}$$

The expression for the coordinate K_r of the FAP (not shown in Fig. 4) includes a constraint factor X and acknowledges the inelastic stress redistribution due to the interaction of the applied and residual stresses through the plasticity correction factor ρ (in the manner of the present R6 approach). K_r is expressed as:

$$K_r = \sqrt{\frac{\delta_e}{\delta_{max}}} + \rho \qquad \text{(Eq 12)}$$

where

$$\delta_e = \frac{K_I^2}{X \cdot \sigma_y \cdot E'} \qquad \text{(Eq 13)}$$

For plane stress and low work-hardening materials, X and E' are, respectively, 1 and E, while for plane strain and low work-hardening materials, X and E' are, respectively, 2 and $E/(1 - \nu^2)$. X is affected by work hardening and can be determined from either elastic-plastic analyses or fracture toughness tests:

$$X = \frac{J_{mat}}{\sigma_y \cdot \delta_{total}} \qquad \text{(Eq 14)}$$

The plasticity correction factor ρ to accommodate residual stresses is determined by:

$$\begin{aligned} \rho &= \rho_1 && \text{for} \quad L_r \le 0.8 \\ \rho &= 4 \cdot \rho_1 \cdot (1.05 - L_r) && \text{for} \quad 0.8 < L_r < 1.05 \\ \rho &= 0 && \text{for} \quad L_r \ge 1.05 \end{aligned} \qquad \text{(Eq 15)}$$

where

$$\begin{aligned} \rho_1 &= 0 && \text{for} \quad \chi \le 0 \\ \rho_1 &= 0.1 \cdot \chi^{0.714} - 0.007 \cdot \chi^2 - 3.x10^{-5} \cdot \chi^3 && \text{for} \quad 0 < \chi < 5.2 \\ \rho_1 &= 0.25 && \text{for} \quad \chi \ge 5.2 \end{aligned} \qquad \text{(Eq 16)}$$

In Eq 16, χ is:

$$\chi = \frac{K_I^{residual}}{K_I^{applied}} \cdot L_r \qquad \text{(Eq 17)}$$

where $K_I^{applied}$ and $K_I^{residual}$ are the linear elastic stress-intensity factors for, respectively, the applied and residual stresses. Equations 15 and 16 are applicable when χ is less than or equal to 4. When χ exceeds 4, a more detailed procedure for the computation of ρ is presented in Ref 7.

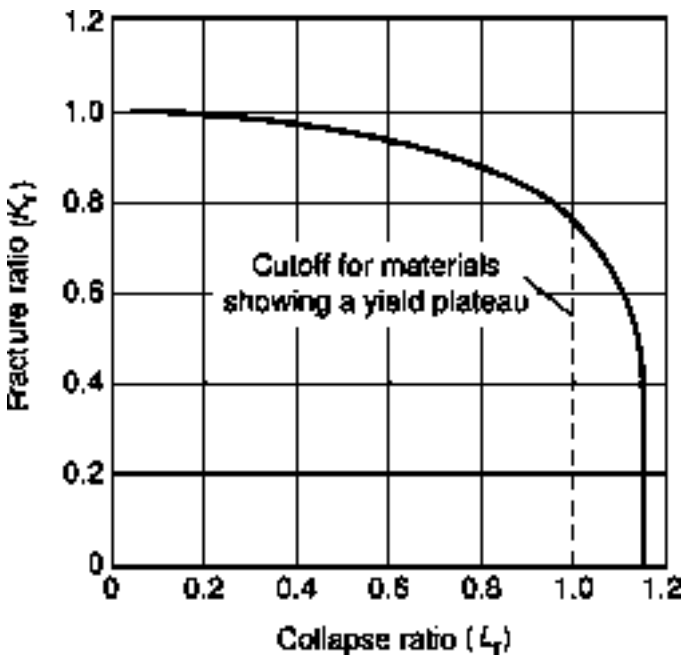


Fig. 4 Level 2a failure assessment diagram (BS 7910)

There is *no* safety factor inherent to the Level 2a FAC, and the latter should be considered critical in the assessment. It is believed that Level 2a is the preferred method for most cases. However, the Level 2a model of the FAC is inaccurate for high work-hardening materials.

Levels 2b and 2c (Fig. 5) generally are used in the assessment of high work-hardening materials and/or stable tearing where the Level 2a approach would prove too restrictive.

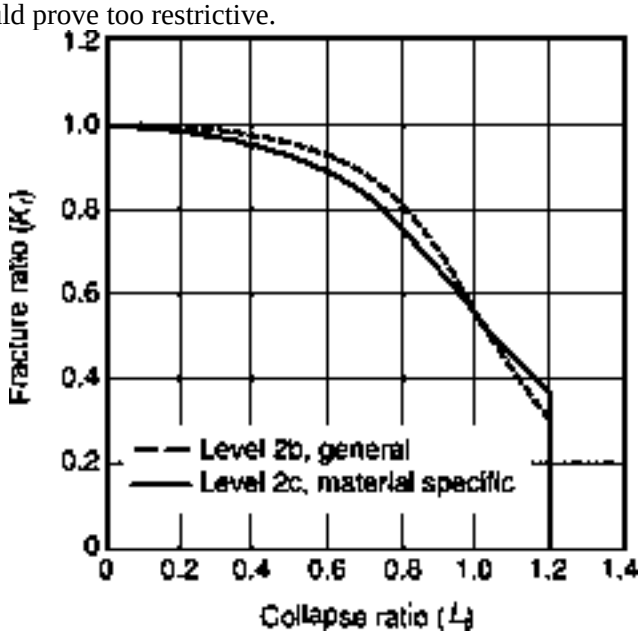


Fig. 5 Level 2b and level 2c failure assessment diagram (BS 7910)

Level 2b (Fig. 5) is a general FAC similar to Option 1 of the current R6 method (Ref 3) and is expressed as:

$$K_r = \left(1 - 0.14 \cdot L_r^2\right) \cdot \left[0.3 + 0.7 \cdot \exp\left(-0.65 \cdot L_r^6\right)\right] \text{ for } L_r \le L_r^{max} \qquad \text{(Eq 18)}$$

or

$$K_r = 0 \text{ for } L_r > L_r^{max} \qquad \text{(Eq 19)}$$

where L_r is defined by Eq 11 and L_r^{max} is defined as:

$$L_r^{max} = \frac{\sigma_f}{\sigma_y} \qquad \text{(Eq 20)}$$

Level 2b may be used for all cases but is recommended (Ref 7) for high work-hardening materials (i.e., when $\sigma_f > 1.2\ \sigma_y$).

Level 2c (Fig. 5), a material-specific FAC similar to Option 2 of the current R6 (Ref 3), is suitable for base and weld metal of all types and provides more accurate results than either Level 2a or 2b but requires the stress-strain curve for the material examined. Level 2c FAC is expressed as:

$$K_r = \frac{1}{\sqrt{\dfrac{E \cdot \varepsilon}{L_r \cdot \sigma_y} + \dfrac{L_r^3 \cdot \sigma_y}{2 \cdot E \cdot \varepsilon}}} \text{ for } L_r \le L_r^{max} \qquad \text{(Eq 21)}$$

and

$$K_r = 0 \text{ for } L_r > L_r^{max} \qquad \text{(Eq 22)}$$

where L_r is expressed by Eq 11, L_r^{max} is expressed by Eq 20, and ε is the true strain for a true stress of $L_r\ \sigma_y$ from the true stress-strain curve of the material.

The coordinate L_r (Fig 5) is the ratio of the net-section stress to the yield stress σ_y. Inelastic stress redistribution due to the interaction of the applied and residual stresses is handled as in Level 2b.

K_r for the general FAC is a function only of L_r, and for the material specific FAC, is a function of L_r, yield strength σ_y, and the true stress-strain curve. K_r for the FAP is identical in form to that of Level 2a in Eq 12. Residual stresses are accommodated in Levels 2b and 2c as in Level 2a (i.e., Eq 15, 16, and 17). There is no safety factor inherent to either Level 2b or 2c FAC, and they should be considered critical in the assessment.

It should be noted that, as in Levels 1 and 2a, only the elastic stress-intensity factor K_I for the flawed member is needed. Level 2c FACs for two materials with identical yield and tensile strengths, respectively, σ_y and σ_u (Fig. 6), are compared with the Level 2b (general FAC) in Fig. 7. Both stress-strain curves were developed from tension tests on an HSLA 60 steel (API 2Y grade 60T, a plate steel used in offshore welded construction) base metal (showing a yield plateau in Fig. 6) and the deposited weld metal (showing a roundhouse in Fig. 6) in the same steel. Below an L_r of 0.6, there is little difference between the Level 2b and Level 2c FACs, (Fig. 7). However, for L_r greater than 0.6, the effect of stress-strain-curve shape is marked. Note that in Fig. 6 and 7, the base metal and deposited weld metal are represented by, respectively, the solid and dashed curves.

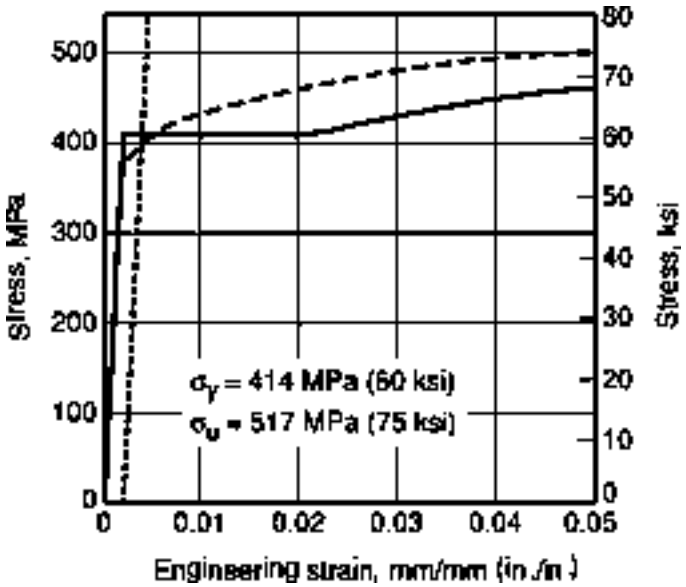


Fig. 6 Engineering stress-strain curve for HSLA 60 (API 2Y grade 60T) plate steel. σ_y, yield strength; σ_u, tensile strength

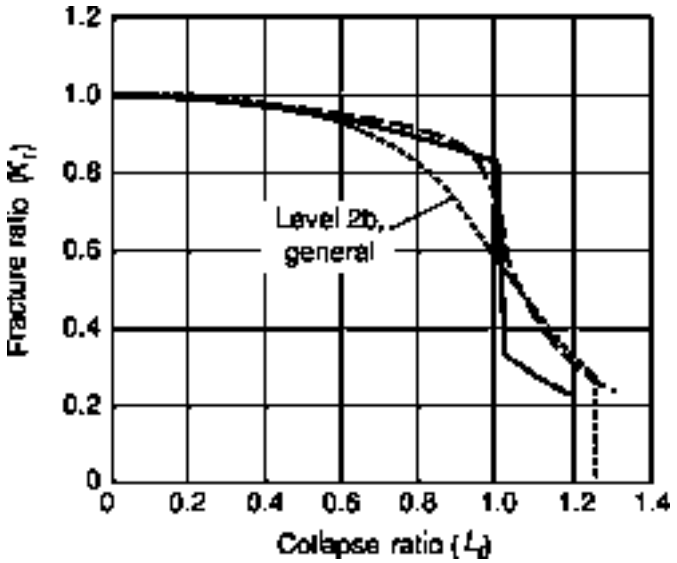


Fig. 7 Level 2b and 2c failure assessment diagram for HSLA 60 (API 2Y grade 60T) plate steel

Level 3 FADs include a ductile tearing assessment. Levels 3a and 3b apply ductile tearing assessments to, respectively, the FAD of Level 2a or 2b, and to the FAD of Level 2c. Level 3c uses the *J*-integral in a tearing assessment similar to R6 Option 3 (Ref 3) or the EPRI/GE handbook (Ref 4). Such an analysis requires the *J*-integral solutions for any geometry. The complete solutions for many geometries are listed in the EPRI/GE handbook (Ref 4), subsequent EPRI/GE reports (Ref 21, 22, 23), and the textbook by Kanninen and Popelar (Ref 12). EPRI and Novetech jointly developed the three-volume *Ductile Fracture Handbook* (Ref 24) that assembled and synthesized the elastic-plastic fracture mechanics solutions for cracked cylinders developed in EPRI-sponsored research. The first volume of the *Ductile Fracture Handbook* presents solutions for circumferential through-wall cracks. The second volume presents solutions for circumferential part-through-wall cracks and axial through-wall cracks. The third volume presents solutions for axial part-

through-wall cracks and cracks in elbows, tees, and nozzles and flaw-evaluation procedures for piping and pressure vessels.

References cited in this section

1. F.M. Burdekin and M.G. Dawes, Practical Uses of Linear Elastic and Yielding Fracture Mechanics with Particular Reference to Pressure Vessels, *Practical Applications of Fracture Mechanics to Pressure Vessel Technology,* Institution of Mechanical Engineers, London, 1971, p 28–37

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

4. V. Kumar, M.D. German, and C.F. Shih, "An Engineering Approach for Elastic-Plastic Fracture Analysis," EPRI NP–1931, Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, July 1981

6. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," PD 6493:1991, 2nd ed., British Standards Institution, London, August 1991

7. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," BS 7910:1998, British Standards Institution, London, 1998

12. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

17. "Standard Test Method for Crack-Tip Opening Displacement (CTOD) Fracture Toughness Assessment," E 1290, *Annual Book of ASTM Standards,* ASTM, 2001

18. "Fracture Mechanics Toughness Tests, Part 1: Method for Determination of K_{Ic}, Critical CTOD and Critical J Values of Metallic Materials," BS 5748:1991, British Standards Institution, London, 1991

19. A.G. Miller, Review of Limit Loads of Structures Containing Defects, *Int. J. Pressure Vessels Piping,* Vol 32, theme issue, 1988, p 197–327

20. A.A. Willoughby and T.G. Davy, Plastic Collapse at Part Wall Flaws in Plates, *Fracture Mechanics: Perspectives and Directions,* STP 1020, R.P. Wei, Ed., ASTM, 1989, p 390–409

21. V. Kumar, W.W. Wilkening, W.R. Andrews, M.D. German, H.G. deLorenzi, and D.F. Mowbray, "Estimation Technique for the Prediction of Elastic-Plastic Fracture of Structural Components of Nuclear Systems," 5th and 6th Semiannual Report to EPRI, Contract RP1237-1, General Electric Company, Schenectady, NY, 01 Feb 1981–30 Jan 1982

22. V. Kumar, M.D. German, W.W. Wilkening, W.R. Andrews, H.G. deLorenzi, and D.F. Mowbray, "Advances in Elastic-Plastic Fracture Analysis," EPRI NP-1931, Research Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, August 1984

23. V. Kumar and M.D. German, "Further Developments in Elastic-Plastic Fracture Analysis of Through-Wall Flaws and Surface Flaws in Cylinders," Final Report 87-SED-017, Research Project 1237-5, General Electric Company, Research and Development Center, Schenectady, NY, June 1987

24. EPRI/Novetech, *Ductile Fracture Handbook,* Novetech Corporation, Gaithersburg, MD, Vol 1, 1989; Vol 2, 1990; Vol 3, 1991

Considerations in Use

The characterizations for the fracture ratio K_r of the FAD are shown in Table 2 for linear elastic fracture mechanics (LEFM) and elastic-plastic fracture mechanics (EPFM). It is vital to recognize that the FAD acknowledges elastic-plastic response both in the fracture event and plastic collapse. Also, only the linear elastic stress-intensity factor K_I for a specific geometry is required, thus using the existing compendia of K_I solutions and obviating the need for complex elastic-plastic solutions. However, application of the FAD does require the knowledge of the elastic-plastic fracture toughness (characterized as either *J*-integral or CTOD) for the material in the thickness of application.

Table 2 Fracture ratios, K_r

FAD element	LEFM	EPFM	
		J-integral	CTOD
FAC	1	$\sqrt{\frac{J_e}{J_e+J_p}}$	$\sqrt{\frac{\delta}{\delta_e}}$
FAP	$\frac{K_J}{K_{mat}}$	$\sqrt{\frac{J_e}{J_{mat}}}$	$\sqrt{\frac{\delta}{\delta_e}}$

FAD, failure assessment diagram; LEFM, linear elastic fracture mechanics; EPFM, elastic-plastic fracture mechanics; CTOD, crack-tip opening displacement; FAC, failure assessment curve; FAP, failure assessment point; J_e, elastic component of *J*-integral; J_p, plastic component of *J*-integral; δ_e, elastic component of crack-tip opening displacement; δ_p, plastic component of crack-tip opening displacement; K_I, mode I stress-intensity factor; K_{mat}, general linear elastic fracture toughness of material; J_{mat}, general *J*-integral fracture toughness of material; δ_{mat}, general CTOD toughness of material

Also, R6 (Ref 3) and BS 7910 (Ref 7) give guidance on reliability, partial safety factors, number of fracture toughness tests, selection of failure risk for Levels 2 and 3, and sensitivity analyses.

Limitations to be addressed in the future include the improvement of estimates of material toughness through:

- The evaluation of constraint
- Improved estimation equations for *J*-integral and CTOD from test results
- Use of shallow cracks more typical of service

Improvements in the construction of the FAD include:

- Combining the applied and residual stresses to acknowledge inelastic redistribution of residual stresses
- Consideration in mismatch between the mechanical properties of the base metal and deposited weld metal
- Ductile tearing (Level 3, PD 7910) beyond that allowed by the validity criteria of current fracture toughness test standards
- Consideration of constraint in the definition of the FAC and FAP

Caution. Many practitioners of fracture mechanics today determine the critical stress-crack-size $K_I^{critical}$ from the test estimations of elastic-plastic fracture toughness (i.e., J_{mat} or δ_{mat}) (Table 1) through:

$$K_I^{critical} = \sqrt{\frac{J_{mat} \cdot E}{(1-\nu^2)}} \qquad \text{(Eq 23)}$$

or

$$K_I^{critical} = \sqrt{\delta_{mat} \cdot E \cdot \sigma_y} \qquad \text{(Eq 24)}$$

Such an approach implies that $K_r = 1$ regardless of the value of L_r (Fig. 8). In Fig. 8, it may be seen that K_r, equivalent to $\sqrt{\delta_e/\delta_{mat}}$ (Table 2), can be significantly less than 1, depending on the value of L_r. The observed ratios of $\sqrt{\delta_e/\delta_{mat}}$ versus the test-estimated CTOD toughness for an HSLA 50 steel (API 2Y Grade 50T, a plate steel used in offshore welded construction) are shown in Fig. 9. With the observed range of values δ_e/δ_{mat} for the various δ_{mat} (i.e., δ_c, δ_u, and δ_m, defined in Table 1), $K_I^{critical}$ may be computed through a combination of Eq 12, 13, and 24 (Table 3).

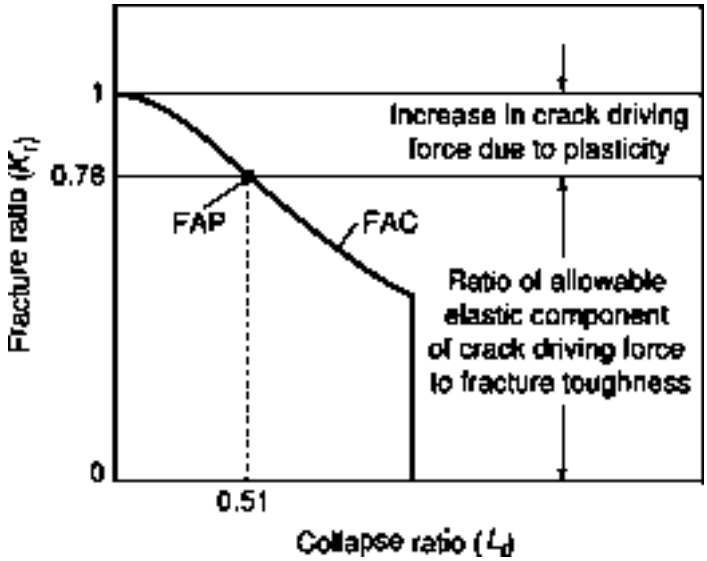


Fig. 8 Failure assessment diagram with K_r < 1, depending on the value of L_r. FAP, failure assessment point; FAC, failure assessment curve

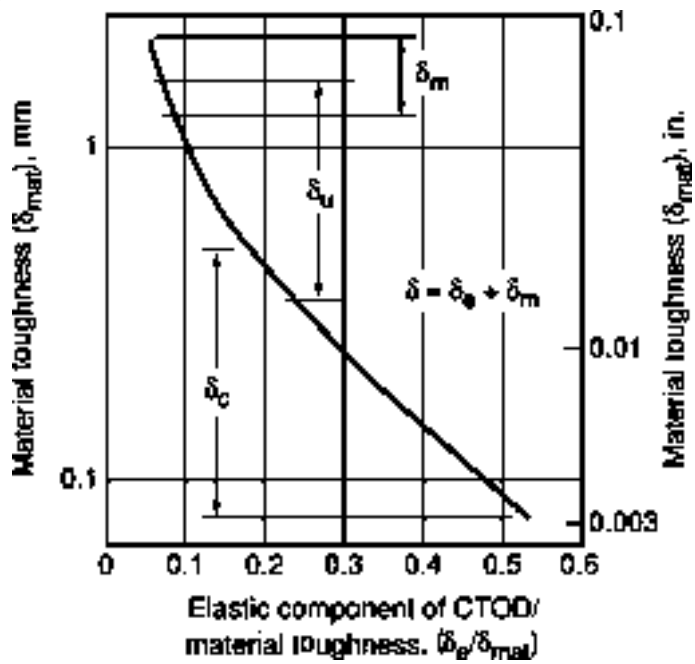


Fig. 9 Crack-tip opening displacement (CTOD) toughness, HSLA 50. δ, CTOD; δ_c, CTOD fracture toughness, no significant stable crack extension, unstable fracture; δ_e, elastic component of CTOD; δ_m, CTOD fracture toughness, significant stable crack extension, plastic collapse; δ_p, plastic component of CTOD; δ_u, CTOD fracture toughness, significant stable crack extension, unstable fracture

Table 3 $K_I^{critical}$ from CTOD toughness

δ_{mat} (Table 1)	δ_e/δ_{mat} (Fig. 9)	$K_I^{critical}/\sqrt{\delta_{mat}\cdot\sigma_y\cdot E}$	
		Plane stress	**Plane strain**
δ_c	0.19–0.54	0.44–0.74	0.64–1.09
δ_u	0.07–0.23	0.26–0.48	0.39–0.71
δ_m	0.05–0.08	0.22–0.28	0.33–0.42

$K_I^{critical}$, mode I stress-intensity factor estimated from J_{mat} or δ_{mat}; CTOD, crack-tip opening displacement; δ_{mat}, general CTOD toughness of material; δ_e, elastic component of crack-tip opening displacement; σ_y, yield strength; E, Young's

modulus; δ_c, CTOD fracture toughness, no significant stable crack extension, unstable fracture; δ_u, CTOD fracture toughness, significant stable crack extension, unstable fracture; δ_m, CTOD fracture toughness, significant stable crack extension, plastic collapse

From Eq 24, the ratio $K_I^{critical}/\sqrt{\delta_{mat}\sigma_y E}$ would always be unity. However, recognition of elastic-plastic response at fracture, demonstrated by the FAD (e.g., Fig. 8) shows that the ratio $K_I^{critical}/\sqrt{\delta_{mat}\sigma_y E}$ can be significantly less than unity (Table 3). Obviously, the use of Eq 24 to estimate the critical stress–crack-length combination can be very unconservative.

References cited in this section

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

7. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," BS 7910:1998, British Standards Institution, London, 1998

Failure Assessment Diagrams

Harold S. Reemsnyder, Fracture Technology Associates

Application Examples

Example 1: Application of FAD to determine the required CTOD toughness to develop the full potential for a transverse butt weld in a very wide 33 mm (1.3 in.) plate containing a surface flaw at the weld toe that penetrates the plate fully. Consider both the as-welded (AW) and postweld heat treated (PWHT) cases. The yield and tensile strengths are, respectively, σ_y = 414 MPa (60 ksi) and σ_u = 517 MPa (75 ksi). The applied stress σ_m is 331 MPA (48 ksi), or 80% of the yield strength. Assume that the residual stress σ_{resaw} transverse to the weld is equal to the yield stress. The residual stress after PWHT σ_{repwht} is 20% of that in the AW condition (Ref 7), or 82.8 MPA (12 ksi).

Hint: Estimate the fracture toughness when the FAP falls on the FAC for Levels 1 and 2, BS 7910 (Ref 7). Assume: (1) plane strain for Level 2, (2) the stress-strain curves for Level 2c are those shown in Fig. 6, and (3) the crack geometry is that of through-crack of length 2*a* in an infinitely wide plate of thickness *a*:

$$K_I = \sigma\sqrt{\pi \cdot a} \qquad \text{(Eq 25)}$$

and

$$\sigma_{ref} = \sigma_m = 331 \text{ MPa } (48 \text{ ksi}) \qquad \text{(Eq 26)}$$

E, ν, and σ_f are, respectively, 206,897 MPa (30,000 ksi); 0.3; and 466 MPa (67.5 ksi). From Eq 9 and 11, the collapse ratios, S_r and L_r, are, respectively, 0.711 and 0.800.

Level 1, BS 7910. The maximum stresses for the AW and PWHT conditions are:

$$\sigma_{maxaw} = \sigma_m + \sigma_{resaw} = 745 \text{ MPa } (108 \text{ ksi}) \qquad \text{(Eq 27)}$$

$$\sigma_{maxpwht} = \sigma_m + \sigma_{respwht} = 414 \text{ MPa } (60 \text{ ksi}) \qquad \text{(Eq 28)}$$

and, from Eq 25, with $\sigma = \sigma_{maxaw}$ and $\sigma = \sigma_{maxpwht}$:

$$K_I^{maxaw} = 239 \text{ MPa}\sqrt{\text{m}} \left(217.5 \text{ ksi}\sqrt{\text{in.}}\right) \qquad \text{(Eq 29)}$$

$$K_I^{maxpwht} = 133 \text{ MPa}\sqrt{\text{m}} \left(121.1 \text{ ksi}\sqrt{\text{in.}}\right) \qquad \text{(Eq 30)}$$

From Fig. 3, K_r = 0.707, and from Eq 8, with $\sigma_{max}/\sigma_y > 0.5$:

$$\delta_I^{aw} = 0.323 \text{ mm } (0.0127 \text{ in.})$$
$$\delta_I^{pwht} = 0.156 \text{ mm } (0.00613 \text{ in.}) \qquad \text{(Eq 31)}$$

From Eq 6, for Level 1, the optimum fracture toughnesses are:

$$\delta_{max}^{aw} = 0.643 \text{ mm } (0.0253 \text{ in.})$$
$$\delta_{max}^{pwht} = 0.310 \text{ mm } (0.01222 \text{ in.}) \quad \text{(Eq 32)}$$

Level 2a, BS 7910. The maximum stress-intensity factors, from Eq 25, for the AW and PWHT conditions are:

$$K_I^{aw} = K_I^m + K_I^{resaw} = \sigma_m \cdot \sqrt{\pi \cdot a} + \sigma_{resaw} \cdot \sqrt{\pi \cdot a} \quad \text{(Eq 33)}$$

$$K_I^{pwht} = K_I^m + K_I^{respwht}$$
$$= \sigma_m \cdot \sqrt{\pi \cdot a} + \sigma_{respwht} \cdot \sqrt{\pi \cdot a} \quad \text{(Eq 34)}$$

and, from Eq 25, with $\sigma = \sigma_{maxaw}$ and $\sigma = \sigma_{maxwht}$:

$$K_I^{aw} = 239 \text{ MPa}\sqrt{\text{m}} \left(217.5 \text{ ksi}\sqrt{\text{in.}}\right) \quad \text{(Eq 35)}$$

$$K_I^{pwht} = 133 \text{ MPa}\sqrt{\text{m}} \left(121.1 \text{ ksi}\sqrt{\text{in.}}\right) \quad \text{(Eq 36)}$$

From Eq 10, $K_r = 0.870$, and from Eq 17:

$$\chi^{aw} = 1.00\chi^{pwht} = 0.200 \quad \text{(Eq 37)}$$

From Eq 16, for $0 < \chi < 5.2$:

$$\rho_1^{aw} = 0.0930 \qquad \rho_1^{pwht} = 0.0314 \quad \text{(Eq 38)}$$

From Eq 15, for $L_r < 0.8$ (i.e., $\rho = \rho_1$):

$$\rho^{aw} = 0.0930\rho^{pwht} = 0.0314 \quad \text{(Eq 39)}$$

From Eq 13, for plane strain with $X = 2$ and $E' = E/(1 - \nu^2)$:

$$\delta_I^{aw} = 0.305 \text{ mm } (0.0120 \text{ in.})$$
$$\delta_I^{pwht} = 0.0945 \text{ mm } (0.00372 \text{ in.}) \quad \text{(Eq 40)}$$

For Level 2a, from:

$$\delta_{max} = \frac{\delta_I}{(K_r - \rho)^2} \quad \text{(Eq 41)}$$

the optimum plane-strain fracture toughnesses are:

$$\delta_{max}^{aw} = 0.508 \text{ mm } (0.0200 \text{ in.})$$
$$\delta_{max}^{pwht} = 0.134 \text{ mm } (0.00529 \text{ in.}) \quad \text{(Eq 42)}$$

Level 2b, BS 7910. The maximum stress-intensity factors, χ, ρ, and ρ_1, for the AW and PWHT conditions, are identical to those for Level 2a above. From Eq 18, $K_r = 0.811$. From Eq 13, for plane strain with $X = 2$ and $E' = E/(1 - \nu^2)$:

$$\delta_I^{aw} = 0.305 \text{ mm } (0.0120 \text{ in.})$$
$$\delta_I^{pwht} = 0.0945 \text{ mm } (0.00372 \text{ in.}) \quad \text{(Eq 43)}$$

For Level 2b, from Eq 41, the optimum plane-strain fracture toughnesses are:

$$\delta_{max}^{aw} = 0.594 \text{ mm } (0.0234 \text{ in.})$$
$$\delta_{max}^{pwht} = 0.155 \text{ mm } (0.00612 \text{ in.}) \quad \text{(Eq 44)}$$

Level 2c, BS 7910. The maximum stress-intensity factors, χ, ρ, and ρ_1, for the AW and PWHT conditions, are identical to those for Levels 2a and 2b above. The strain ε is computed from $\varepsilon = \sigma_m/E = 0.00160$ (Fig. 6). From Eq 21, $K_r = 0.870$. From Eq 13, for plane strain with $X = 2$ and $E' = E/(1 - \nu^2)$:

$$\delta_I^{aw} = 0.305 \text{ mm } (0.0120 \text{ in.})$$
$$\delta_I^{pwht} = 0.0945 \text{ mm } (0.00372 \text{ in.}) \quad \text{(Eq 45)}$$

For Level 2c, from Eq 41 the optimum plane-strain fracture toughnesses are:

$$\delta_{max}^{aw} = 0.505 \text{ mm } (0.0199 \text{ in.})$$
$$\delta_{max}^{pwht} = 0.134 \text{ mm } (0.00528 \text{ in.}) \quad \text{(Eq 46)}$$

The fracture toughnesses estimated from Levels 1, 2a, 2b, and 2c, BS 7910, are compared in Table 4. The toughnesses for Level 2c are slightly lower than those for Level 2b, which reflects the slight difference between the FAC for Level 2b and those of Level 2c (Fig. 7). The toughnesses for Levels 2a and 2c are practically identical. However, this is fortuitous because the values of K_r, for L_r = 0.8, are practically identical (Fig. 4, 7). If the stress-strain curves for the material demonstrated a greater strain-hardening effect than in Fig. 6, the fracture toughnesses for Level 2c would be greater than those of Level 2a. The fracture toughnesses required from the Level 1 analysis are 1.2 to 2.2 times those of the more accurate analysis of Level 2c, which reflects the conservative nature of Level 1. The decrease in required fracture toughness when PWHT is applied reflects only the reduction in residual stress that reduces the crack-driving force. The reader should remember that an improper PWHT could reduce the fracture toughness of the material.

Table 4 Optimum CTOD toughness mm (in.), at indicated level

Case	1	2a	2b	2c
AW	0.64 (0.025)	0.51 (0.020)	0.58 (0.023)	0.51 (0.020)
PWHT	0.31 (0.012)	0.13 (0.0053)	0.16 (0.0061)	0.13 (0.0053)

CTOD, crack-tip opening displacement; AW, as welded; PWHT, postweld heat treatment

Example 2: Application of FAD. The structural-integrity analysis of a heavy rolled W14 × 730 H-section (Fig. 10) used in high-rise building construction is presented in Ref. 25. Normally, such sections are used as columns in high-rise buildings, but in some cases, these sections are used as tension chords of trusses in such construction. The material fracture toughness, from tests on compact tension specimens taken from the flange-web core was linked to the estimation of the critical crack size in the section through an elastic-plastic structural-integrity analysis. The assessment methodology was that of the FAD. The fracture model for the section was assumed to be a surface flaw in one flange joined at the flange-web intersection to an edge flaw in the web (Fig. 10). The steel grade of the section was A 572 Grade 50, yield and tensile strengths of, respectively, 345 and 552 MPa (50 and 80 ksi). Crack-tip opening displacement toughness values δ_m (upper shelf, Table 1) of 0.25 to 0.76 mm (0.01 to 0.03 in.) estimated from the fracture toughness tests were used in the analysis. Both the presence and absence of rolling residual stresses were considered. (Maximum tensile residual stresses on the order of 69 MPa (10 ksi) were measured in the flange-web core.) Critical surface-flaw depths were estimated with and without residual stresses for a range of applied stresses S of 69 to 345 MPa (10 to 50 ksi) and for surface flaw aspect ratios a/c (Fig. 10) varying from 0.1 to 0.6. (Applied stress S is the applied load divided by the area of the uncracked cross section.) The fracture model for the cracked section was assumed to be a surface flaw in one flange joined at the web-flange intersection to an edge flaw in the web. Elastic continuity was established by matching the crack-mouth opening displacements (CMOD) of the two crack geometries at their junction. The crack depth a (Fig. 10) was varied from 19 to 83 mm (0.75 to 3.25 in.).

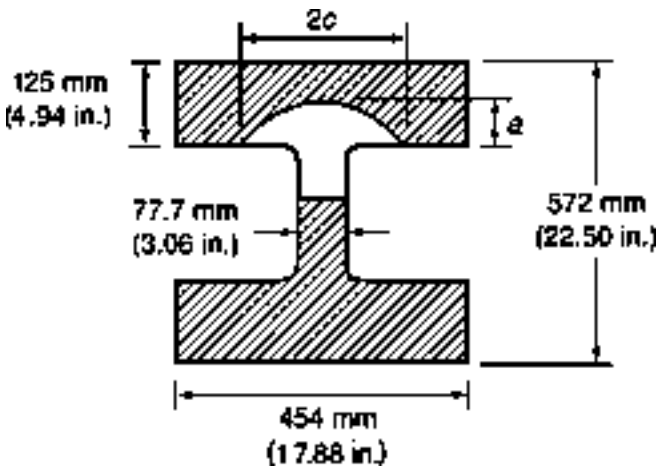


Fig. 10 W14 × 730 column H-section with assumed crack shape, *a*, depth of surface crack; 2*c*, length of surface crack

Three FAD approaches were used in the analyses:

- Level 2, PD 6493 (Ref 2), now Level 2a BS 7910 (Ref 7)
- Level 3 general, PD 6493, now Level 2b, BS 7910
- Appendix 8, R6, Revision 3, for C-Mn steels (Ref 3)

The results of the R6, Appendix 8, FAD analysis were the most conservative of the three approaches and are presented herein. The FAC for C-Mn steels from Appendix 8, R6, Revision 3 (Ref 3) is:

$$K_r = \frac{1 - 0.1 \cdot S_r^2 + 0.1 \cdot S_r^4}{1 + 3 \cdot S_r^4} \text{ for } S_r \le 1 \qquad \text{(Eq 47)}$$

and

$$K_r = 0 \text{ for } S_r > 1 \qquad \text{(Eq 48)}$$

where the fundamental definition of the collapse ratio S_r was used:

$$S_r = \frac{\text{applied load on cracked section}}{\text{limit load of cracked section}} \qquad \text{(Eq 49)}$$

Solutions for an array of FAPs, (e.g., Fig. 11) were performed and resulted in an FAP surface for each aspect ratio *a*/*c* with nodes located at the intersections *S*, *a* (Fig. 11). In Fig. 11, two examples of FAP loci for one aspect ratio, *a*/*c* = 0.5, are shown: (1) *S* = 138 MPa (20 ksi), varying *a* from 19 to 82.6 mm (0.75 to 3.25 in), and (2) *a* = 38.1 mm (1.50 in), varying *S* from 69 to 345 MPa (10 to 50 ksi). The critical combination of *S* and *a* for each aspect ratio *a*/*c* is estimated from the intersection of the FAP surface with the FAC (Fig. 11). The results of this process are shown in Fig. 12 for one aspect ratio, *a*/*c* = 0.5, and the lower bound to the test-estimated upper-shelf CTOD toughness, δ_m = 0.25 mm (0.010 in). Conclusions drawn for these analyses were (Ref 25):

- Typical surface flaws in flanges might be expected to have aspect ratios *a*/*c* on the order of 0.3 or greater. For such aspect ratios and values of applied stress typical of service, 138 to 207 MPa (20 to 30 ksi), the FAD analysis indicated that the minimum critical crack depth *a* would be on the order of 38 to 51 mm (1.5 to 2 inches) (about 30 to 40% of the flange depth) when the upper shelf CTOD toughness δ_m was 0.25 mm (0.01 in).
- The critical flange crack depth would be increased by about 30 to 50% when the upper shelf CTOD toughness δ_m was increased from 0.25 mm (0.01 in.) to 0.76 mm (0.030 in.) (the upper bound from the fracture toughness tests).
- For an aspect ratio *a*/*c* of 0.5, CTOD toughness δ_m of 0.25 mm (0.01 in.), and applied stresses *S* from 138 to 207 MPa (20 to 30 ksi), the elimination of rolling residual stresses would increase the critical crack depth *a* by about 15 to 34%. (The rolling residual stresses would be reduced significantly by a thermal treatment, e.g., normalizing.)

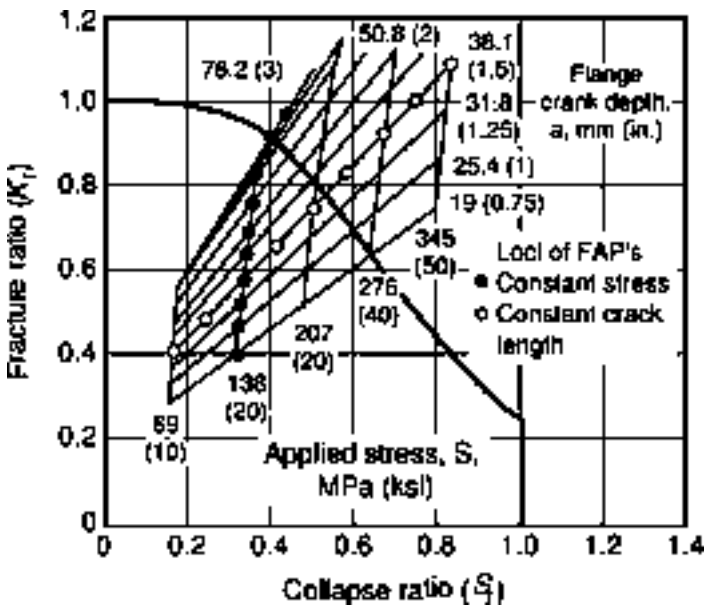


Fig. 11 Failure assessment diagram (R6, Appendix 8) for WF14 × 730 H-section, *a*/*c* = 0.5, with residual stresses. Collapse ratio is a function of flow stress σ_f; FAP, failure assessment point

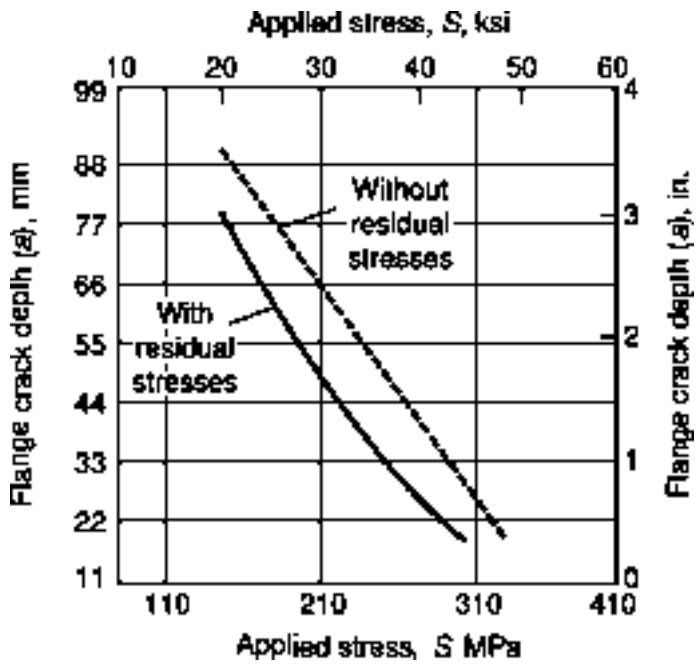


Fig. 12 Critical crack depth vs. applied stress for WF14 × 730 H-section, *a/c* = 0.5, δ_m = 0.25 mm (0.010 in.)

Essentially, the analysis of this example was a sensitivity analysis (see Ref 3 and 7) to establish the effects on the structural integrity of the cracked H-section by variations in the significant parameters (e.g., applied stress *S*, crack depth *a*, surface-crack aspect ratio *a*/*c*, fracture toughness δ_m, and the presence or absence of residual stresses).

References cited in this section

2. "Guidance on Some Methods for the Derivation of Acceptance Levels for Defects in Fusion Welded Joints," PD 6493:1980, 1st ed., British Standards Institution, London, March 1980

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, "Assessment of the Integrity of Structures Containing Defects," R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

7. "Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures," BS 7910:1998, British Standards Institution, London, 1998

25. H.S. Reemsnyder, Structural Integrity Analysis of a Heavy Rolled Structural Section, presented at *IIW Conference on Fitness for Purpose of Welded Structures,* International Institute of Welding (Key Biscayne, FL), Oct 1991

Failure Assessment Diagrams

Harold S. Reemsnyder, Fracture Technology Associates

References

1. F.M. Burdekin and M.G. Dawes, Practical Uses of Linear Elastic and Yielding Fracture Mechanics with Particular Reference to Pressure Vessels, *Practical Applications of Fracture Mechanics to Pressure Vessel Technology,* Institution of Mechanical Engineers, London, 1971, p 28–37

2. "Guidance on Some Methods for the Derivation of Acceptance Levels for Defects in Fusion Welded Joints," PD 6493:1980, 1st ed., British Standards Institution, London, March 1980

3. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, “Assessment of the Integrity of Structures Containing Defects,” R/H/R6, Revision 3, Central Electricity Generating Board, Leatherhead, Surrey, May 1986

4. V. Kumar, M.D. German, and C.F. Shih, “An Engineering Approach for Elastic-Plastic Fracture Analysis,” EPRI NP–1931, Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, July 1981

5. J.M. Bloom, Validation of a Deformation Plasticity Failure Assessment Diagram Approach to Flaw Evaluation, *Elastic-Plastic Fracture: Second Symposium, Vol II, Fracture Resistance Curves and Engineering Applications,* STP 803, C.F. Shah and J.P. Gudas, Ed., ASTM, 1983, p II-206 to II-238

6. “Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures,” PD 6493:1991, 2nd ed., British Standards Institution, London, August 1991

7. “Guidance on Methods for Assessing the Acceptability of Flaws in Fusion Welded Structures,” BS 7910:1998, British Standards Institution, London, 1998

8. “Guide to Damage Tolerance Analysis of Marine Structures,” SSC-409, Ship Structures Committee, Washington, DC, 2000

9. Theme Issue: The Revisions to the Structural Integrity Assessment Method CEGB/R6, *Int. J. Pressure Vessels Piping,* Vol 32, 1988

10. “20 Years of R6,” S 458, Institution of Mechanical Engineers, London, Nov 1996

11. A.R. Dowling and C.H.A. Townley, The Effect of Defects on Structural Failure: A Two-Criteria Approach, *Int. J. Pressure Vessels Piping,* Vol 3, 1975, p 77–107

12. M.F. Kanninen and C.H. Popelar, *Advanced Fracture Mechanics,* Oxford University Press, 1985

13. I. Milne, R.A. Ainsworth, A.R. Dowling, and A.T. Stewart, Assessment of the Integrity of Structures Containing Defects, *Int. J. Pressure Vessels Piping,* Vol 32, theme issue, 1988, p 3–104

14. J. Bloom, Appendix J—The Deformation Plasticity Failure Assessment Diagram (DPFAD) Approach to Evaluation of Flaws in Ferritic Piping, presented at the *European Symposium on Elastic-Plastic Fracture Mechanics* (Frieburg, W. Germany) Oct 1989

15. G.G. Chell, A Procedure for Incorporating Thermal and Residual Stresses into the Concept of a Failure Assessment Diagram, *Elastic-Plastic Fracture,* STP 668, ASTM, 1979, p 581–605

16. J.M. Bloom, Prediction of Ductile Tearing Using a Proposed Strain-Hardening Failure Assessment Diagram, *Int. J. Fract.,* Vol 16, 1980, p R73–R77

17. “Standard Test Method for Crack-Tip Opening Displacement (CTOD) Fracture Toughness Assessment,” E 1290, *Annual Book of ASTM Standards,* ASTM, 2001

18. “Fracture Mechanics Toughness Tests, Part 1: Method for Determination of K_{Ic}, Critical CTOD and Critical J Values of Metallic Materials,” BS 5748:1991, British Standards Institution, London, 1991

19. A.G. Miller, Review of Limit Loads of Structures Containing Defects, *Int. J. Pressure Vessels Piping,* Vol 32, theme issue, 1988, p 197–327

20. A.A. Willoughby and T.G. Davy, Plastic Collapse at Part Wall Flaws in Plates, *Fracture Mechanics: Perspectives and Directions,* STP 1020, R.P. Wei, Ed., ASTM, 1989, p 390–409

21. V. Kumar, W.W. Wilkening, W.R. Andrews, M.D. German, H.G. deLorenzi, and D.F. Mowbray, “Estimation Technique for the Prediction of Elastic-Plastic Fracture of Structural Components of Nuclear Systems,” 5th and

6th Semiannual Report to EPRI, Contract RP1237-1, General Electric Company, Schenectady, NY, 01 Feb 1981–30 Jan 1982

22. V. Kumar, M.D. German, W.W. Wilkening, W.R. Andrews, H.G. deLorenzi, and D.F. Mowbray, "Advances in Elastic-Plastic Fracture Analysis," EPRI NP-1931, Research Project 1237-1, General Electric Company, Research and Development Center, Schenectady, NY, August 1984

23. V. Kumar and M.D. German, "Further Developments in Elastic-Plastic Fracture Analysis of Through-Wall Flaws and Surface Flaws in Cylinders," Final Report 87-SED-017, Research Project 1237-5, General Electric Company, Research and Development Center, Schenectady, NY, June 1987

24. EPRI/Novetech, *Ductile Fracture Handbook*, Novetech Corporation, Gaithersburg, MD, Vol 1, 1989; Vol 2, 1990; Vol 3, 1991

25. H.S. Reemsnyder, Structural Integrity Analysis of a Heavy Rolled Structural Section, presented at *IIW Conference on Fitness for Purpose of Welded Structures*, International Institute of Welding (Key Biscayne, FL), Oct 1991

Analysis Methods for Probabilistic Life Assessment

Harry R. Millwater, Jr., The University of Texas at San Antonio; Paul H. Wirsching, The University of Arizona

Introduction

APPLICATION of probabilistic analysis methods for life assessment of structural components is becoming more prevalent because:

- Failure models continue to be enhanced for structural life prediction.
- Engineers realize that significant uncertainties and variations in loads, material properties, geometrical, and other parameters occur in real-world structures.
- A need exists to quantify the probability of failure of structural components to design for appropriate levels of safety.
- Computational resources are becoming less of an impediment through enhancements in computational algorithms and computer efficiency.
- Factor of safety approaches may not give the desired reliability or may lead to overdesigned structures with significant penalties.
- A need exists for a rational consistent methodology on which to base important decisions such as to inspect, replace, or repair a structure.

The field of probabilistic analysis is far too large to cover in a single article. Rather, this article provides an outline of the issues for performing a probabilistic life assessment. The article is organized in terms of an introduction, description of the elements of probabilistic analysis, evaluation of the probability of failure, advanced concepts, case histories, and available probabilistic analysis software. The section "Introduction to Probabilistic Analysis" briefly describes the historical background, definition of uncertainties, classifications used in probabilistic analysis, and application areas. The "Elements of Probabilistic Analysis" section describes the basic definition of random variables, some common distribution functions used in engineering, selection of a probability distribution, the failure model definition, and the definition of the probability of failure. The "Evaluation of Probability of Failure" section describes solution techniques for special cases as well as the general solution techniques: first-second-order reliability methods, the advanced mean value method, the response surface method, and Monte Carlo sampling. The "Advanced Concepts" section describes a brief introduction to importance sampling, time-variant reliability, system reliability, and risk analysis and defines some generally accepted target reliabilities. The "Case Histories" section describes two application problems for which probabilistic analysis is an essential element. The "Probabilistic Analysis Software" section provides an overview of some of the commercially available software programs for performing probabilistic analysis.

Harry R. Millwater, Jr., The University of Texas at San Antonio; Paul H. Wirsching, The University of Arizona

Introduction to Probabilistic Analysis

Historical Background

The modern era of probabilistic structural design and analysis started after World War II. In 1947, a paper titled "Safety of Structures" by A.M. Freudenthal appeared in the *Transactions of the American Society of Civil Engineers* (Ref 1). This historical paper suggested that rational methods of developing safety factors for engineering structures should give due consideration to observed statistical distributions of the design factors. It was not until the 1960s that rapid growth of academic interest in structural reliability theory occurred, stimulated in part by the publication of another paper by Freudenthal et al. (Ref 2).

In light of the practical difficulties in employing a probabilistic-statistical approach to design criteria development, C.A. Cornell (Ref 3) suggested the use of a second-moment format (called mean value first-order second-moment analysis, MVFOSM) and introduced the concept of a safety index. The safety index was the probabilistic analog of the factor of safety, widely employed to account for uncertainties in the design process.

But the MVFOSM was not an accurate predictor of failure probability. A mechanics design problem in which there are several random design factors presents a difficult mathematical challenge when the goal is to estimate the failure probability, particularly small probabilities typical of structural life assessments. A sequence of contributions made between 1974 and 1990 (Ref 4, 5, 6, 7, 8, 9) have produced very efficient gradient-based numerical algorithms for estimating probabilities of failure. Efficient sampling-based methods, called "importance" sampling, have been developed for addressing more complicated problems (Ref 10, 11, 12, 13).

These advanced methods have been employed to develop probability-based design codes. Reliability goals and descriptions of the random design factors are translated into safety factors, thereby giving the safety check expressions the appearance of conventional factor of safety forms. The American Institute of Steel Construction steel design code using load and resistance factor design (LRFD) was developed by Galambos and Ravindra (Ref 14). The American Petroleum Institute (API) has adopted an LRFD format for design of offshore structures (Ref 15, 16).

Over the past decade, there has been considerable interest in predicting the system reliability of structural and mechanical systems and also in time-variant reliability. These problems are more challenging, and efficient robust solution methods are still being developed.

Uncertainties in Structural Parameters

Uncertainty in a structural parameter is observed when an experiment is repeated and different values are observed. In general, a *random sample* can be defined as $>X> = (X_1, X_2, \ldots, X_n)$, where X_i is the ith observation and n is the sample size. The engineer needs to make decisions in face of this uncertainty. Thus, the goal of analysis is to provide a mathematical synthesis of the uncertainty that is convenient for the decision-making process.

For a statistical summary, two indexes are useful. The *sample mean* is defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The mean is a measure of the central tendency of the distribution.

The sample standard deviation is defined as:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The sample standard deviation is a measure of the dispersion or spread of the data. Often, what the engineer seeks is a design value, for example, a value on the safe side of the distribution of yield strength that can be used to produce high-reliability designs.

The distribution of a variable can be described using a histogram, examples of which are shown in Fig. 1.

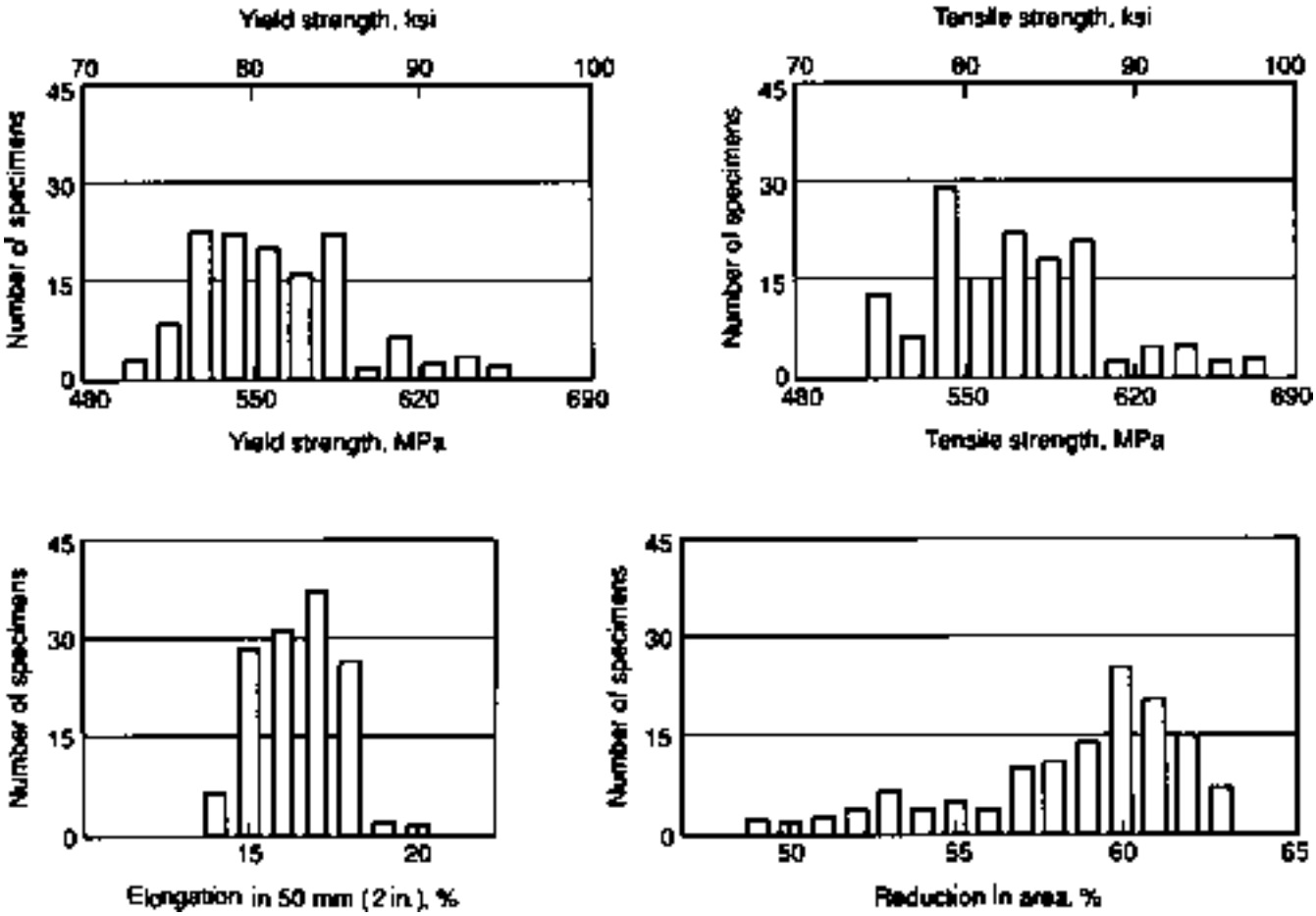


Fig. 1 Examples of histograms. These show distribution of tensile properties among different heats of grade 1117 cold-drawn steel bars. Tested were round bars, 25 mm (1 in.) in diameter, 25 heats from 2 mills.

The coefficient of variation is defined as:

$$C_X = \frac{s_X}{\bar{X}}$$

The coefficient of variation (COV) is widely used in mechanics and design to characterize the uncertainty in certain variables. It has been found that the COV is relatively insensitive to the mean value. Also, the COV is unitless. The values shown in Table 1 are representative of the COVs of structural parameters.

Table 1 Representative coefficient of variation (COV) values

Variable	COV
Yield strength of metals	0.07
Ultimate strength of metals	0.05
Fracture toughness of metals	0.15
Buckling strength of columns	0.15
Cycles to failure in fatigue	0.50
Design load (mechanical systems)	0.05–0.15
Design load (structural systems)	0.15–0.25

This is only an introduction to uncertainty. The representation of this uncertainty in a formal mathematical sense is described in the section “Basic Definition of Random Variables” in this article in terms of random variable distribution functions.

Classifications

Probabilistic analysis can be classified into several distinctive types. The analysis procedures to be applied are dependent on the classification.

Component versus System. The designation of a “component” or “system” probabilistic (or reliability) analysis depends on the number of failure equations required to describe the structural behavior. A system reliability analysis is required if a structure is composed of multiple components, has multiple failure modes, and/or has multiple failure locations, for example, multi-site damage. In each case, multiple failure equations are needed to describe the failure condition of the system.

System reliability is significantly more complex than component reliability analysis requiring specialized tools. This subject is too complex to be covered in detail in this article; a synopsis is provided in the section “System Reliability” in this article. However, a fundamental ingredient of a system reliability analysis is always a component analysis of the basic events.

Time Invariant versus Time Variant. While many reliability analyses involve time, often the problem can be formulated as being time independent or time invariant. For the simple strength (R)-stress (S) case, S could represent the largest stress during the service life and R the strength. The distribution (mean, standard deviation, and distribution family) of each is assumed to be constant, that is, not a function of time. This is an example of a time-invariant problem. For the time-variant case, the parameters (e.g., mean and standard deviation) of S and R are modeled as functions of time. A common example is that of a structural element whose mean strength, μ_R, decreases as a function of time as a result of a growing fatigue crack and/or corrosion. But S, which would now be treated as a random process, may also have time-variant parameters. It is possible, often in the simpler cases, to use analysis to convert a time-variant problem into a time-invariant problem, the latter of which is generally easier to solve.

This subject is too complex to be covered in detail in this article; a synopsis is provided in the section “Time-Variant Reliability” in this article.

Application Areas

The application of probabilistic methods has blossomed in the past two decades. This is evidenced by the number of journal papers describing the application of probabilistic analysis and design to structures. A number of application areas pertaining to structures are given in Ref 17. Some application areas are highlighted:

- Civil engineering has developed the LRFD methodology (Ref 14, 18), which uses probabilistic analysis to develop partial safety factors that engineers may use to design fabricated steel structures.
- The petrochemical industry developed API 579 (see the section “Example 2: Development of Recommended Practice API 579”), which develops partial safety factors for the fitness-for-service of petrochemical equipment.
- The National Aeronautics and Space Administration (NASA) funded a significant effort to develop and apply probabilistic methods to space hardware. The Probabilistic Structures Analysis Methods for Select Space Propulsion System Components program developed analytical techniques and probabilistic software and investigated applications to space shuttle main engine components over the period 1985 to 1995 (Ref 19, 20, 21). Development and application of probabilistic methods has continued at NASA.
- The sixth edition of the well-known machine design book, *Mechanical Engineering Design*, (Ref 22) contains a chapter on addressing uncertainty in machine design. Other examples are given in ref 23.
- Nuclear engineering has a long and rich history in applying probabilistic methods for probabilistic risk assessment of nuclear vessels.
- The United States Navy has been actively developing probability-based structural design criteria. Incidents in which modern cruisers have experienced structural failures in heavy seas have prompted the navy to rethink their decades-old basic design procedures.
- The United States Air Force has funded development of and applied the probability of fracture (PROF) computer code to quantify the probability of failure of aircraft structures with consideration of inspection, structural replacement, and repair (Ref 24).
- The commercial aircraft gas turbine engine industry is adopting a probabilistically-based damage tolerance analysis method for certifying new designs of titanium rotors (see the section “Case Histories: Examples of the Use of Probabilistic Analysis” and “Example 1: Probabilistic Damage Tolerance Analysis of Gas Turbine Rotors” in this article) (Ref 25, 26).

References cited in this section

1. A.M. Freundenthal, Safety of Structures, *Trans. ASCE*, Vol 112, 1947, p 125–180

2. A.M. Freudenthal, J.M. Garrelts, and M. Shinozuka, The Analysis of Structural Safety, *J. Struct. Div. (ASCE),* Vol 92 (No. ST2), Feb 1966

3. C.A. Cornell, A Probability-Based Structural Code, *J. Am. Concr. Inst.,* Vol 66 (No. 12), 1969

4. A.M. Hasofer and N.C. Lind, Exact and Invariant Second Moment Code Format, *J. Eng. Mech., (ASCE),* Vol 100 (No. EM1), 1974, p 111–121

5. R. Rackwitz and B. Fiessler, Structural Reliability under Combined Load Sequences, *J. Comput. Struct.,* Vol 9, 1978, p 489–494

6. L. Tvedt, "Two Second Order Approximations to the Failure Probability," Veritas Report, RDIV/20-004-83, Det Norske Veritas, Oslo, Norway, 1983

7. K. Breitung, Asymptotic Approximations for Multinormal Integrals, *J. Eng. Mech. (ASCE),* Vol 110, 1984, p 357–366

8. Y.-T. Wu and P.H. Wirsching, New Algorithm for Structural Reliability Estimation, *J. Eng. Mech.,* Vol 113 (No. 9), 1987, p 1319–1336

9. Y.-T. Wu, H.R. Millwater, and T.A. Cruse, Advanced Probabilistic Structural Analysis Method for Implicit Performance Functions, *AIAA J.,* Vol 28 (No. 9), 1990

10. A. Harbitz, An Efficient Sampling Method for Probability of Failure Calculations, *Struct. Saf.,* Vol 3, 1986, p 109–115

11. C.G. Bucher, Adaptive Sampling—An Iterative Fast Monte Carlo Procedure, *Struct. Saf.,* 1988, p 119–126

12. P. Bjerager, On Computational Methods for Structural Reliability Analysis, *Struct. Saf.,* Vol 9, 1990, p 79–96

13. Y.-T. Wu. Computational Methods for Efficient Structural Reliability and Reliability Sensitivity Analysis, *AIAA J.,* Vol 32, 1994, p 1717–1723

14. T.V. Galambos and M.K. Ravindra, Properties of Steel for Use in LRFD, *J. Struct. Eng. (ASCE),* Vol 104 (No. 9), 1978, p 1459–1468

15. F. Moses, "Implementation of a Reliability-Based API RP2A Format," final report, API PRAC 83-22, American Petroleum Institute, 1985

16. F. Moses, "Development of Preliminary Load and Resistance Design Document for Fixed Offshore Platforms," final report, API-PRAC 95-22, American Petroleum Institute, 1986

17. C. Sundararajan, Ed., *Probabilistic Structural Mechanics Handbook, Theory and Industrial Applications,* Chapman & Hall, 1995

18. L. Geschwindner, R. Disque, and R. Bjorhovde, *Load and Resistance Factor Design of Steel Structures,* Prentice Hall, 1994

19. "Probabilistic Structural Analysis Methods for Select Space Propulsion System Components (PSAM)," 1st–5th annual reports, NASA Contract NAS3-24389, Southwest Research Institute, 1985–1990

20. "Probabilistic Structural Analysis Methods for Select Space Propulsion System Components (PSAM), final report, NASA Contract NAS3-24389, Southwest Research Institute, 1995

21. K.R. Rajagopal, J. Keremes, H. Ho, and G. Orient, "A Probabilistic Approach to the Evaluation of Fatigue Damage in a Space Propulsion System Injector Element," 33rd Structures, Structural Dynamics, and Materials Conf., April 1992 (Dallas, TX), AIAA/ASME/ASCE/AHS/ASC

22. J.E. Shigley and C.R. Mischke, *Mechanical Engineering Design*, 6th ed., McGraw-Hill, 2000

23. S. Rao, *Reliability-Based Design*, McGraw-Hill, 1992

24. A.P. Berens, P.W. Hovey, and D.A. Skinn, "Risk Analysis for Aging Aircraft Fleets, Vol 1—Analysis," WL-TR-91-3066, Flight Dynamics Directorate, Wright-Patterson Air Force Base, OH 1991

25. G.R. Leverant, D.L. Littlefield, R.C. McClung, H.R. Millwater, and Y.-T. Wu, "A Probabilistic Approach to Aircraft Turbine Rotor Material Design," Paper 97-GT-22, IGTI Congress, 1997

26. G.R. Leverant, R.C. McClung, H.R. Millwater, and M.P. Enright, "A New Tool for Design and Certification of Aircraft Turbine Rotors," Paper GT-2002-30303, ASME Turbo Expo 2002, 3–6 June 2002 (Amsterdam, The Netherlands), American Society of Mechanical Engineers

Analysis Methods for Probabilistic Life Assessment

Harry R. Millwater, Jr., The University of Texas at San Antonio; Paul H. Wirsching, The University of Arizona

Elements of Probabilistic Analysis

A probabilistic analysis always consists of three elements:

- Failure model definition—generically called a "limit state" in probabilistic analysis
- Random variable definitions—variables in the limit state that are considered random
- Probabilistic computations

These elements are described in the next several sections.

Basic Definition of Random Variables

Some elements of probability theory are introduced here. Formally, a random variable is a function that assigns real numbers to an outcome of an experiment. For example, the yield strength of 1035 steel can be considered as a random variable. This random variable would be considered to be continuous, because any real number from zero to infinity would be a possibility. A discrete random variable is limited to the integers, for example, the number of trucks on a bridge at any time. Continuous random variables are more important in basic mechanics and design and are the only variables considered here.

Random variables are denoted as upper case. Let X be a generic random variable in this discussion. Specific values of the random variable are denoted as lower case, for example, x.

Probability can be considered as the numerical measure of the likelihood of occurrence of an event. The probability measure is bounded between 0 (impossible event) and 1 (certain event). In engineering practice, for example, the probability that a member of 1035 will have a yield strength less than 310 MPa (45 ksi) or, more generally, the probability of failure of a component during the intended service life is sought.

To help answer these questions, define the probability density function (PDF), $f_X(x)$, and the cumulative distribution function (CDF), $F_X(x)$ (see Fig. 2). Areas under the PDF define the probabilities as shown. The CDF defines probability of X less than x. The PDF and CDF are related as indicated.

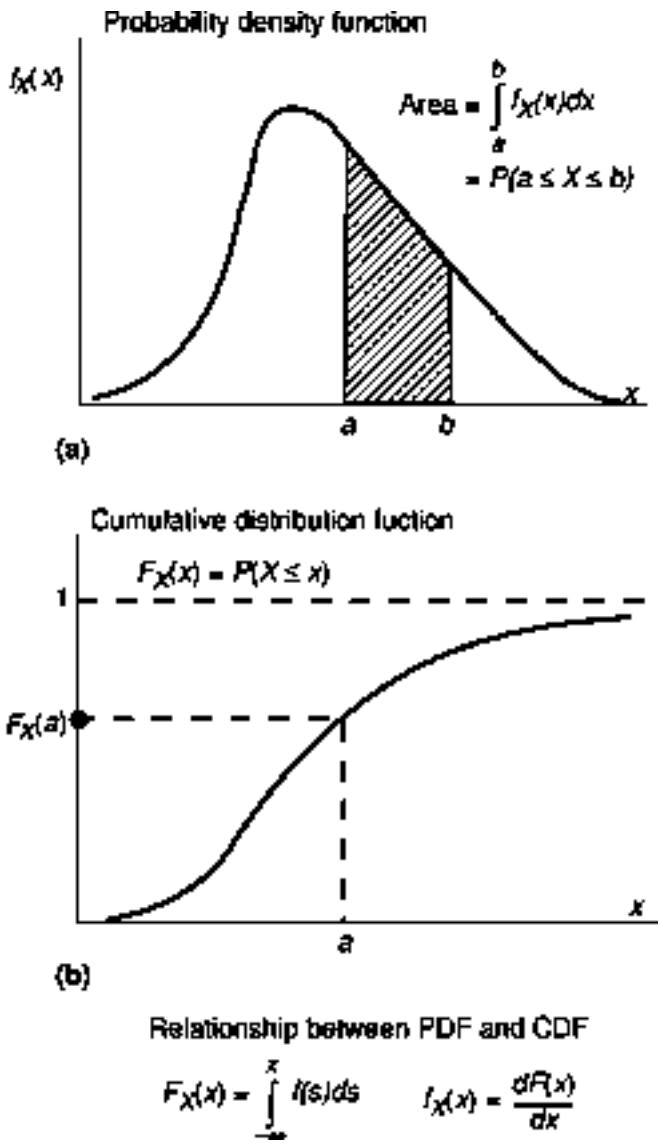


Fig. 2 Definitions of (a) probability density function (PDF) and (b) cumulative distribution function (CDF)

Analytical models that are used by engineers to describe the PDF and CDF are described in the next section.

Two important properties of the distribution are the mean, expected, or average value and the variance. The mean, expected, or average value is represented as:

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x) dx$$

This is the centroid of the PDF and an index of the central tendency of the distribution.

The variance is represented as:

$$V(x) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

The variance is the moment of inertia of the area of the PDF about a vertical axis through the centroid (mean). It is an index of the dispersion of the random variable.

When the sample size $n \to \infty$, the sample mean approaches the mean ($\overline{X} \to \mu_X$), the sample standard deviation approaches the standard deviation ($s_X \to \sigma_X$), and the histogram (vertical scale so that the area equals unity) approaches the PDF.

The coefficient of variation is defined as:

$$C_X = \frac{\sigma_X}{\mu_X}$$

Described in the following section are statistical models (standard mathematical forms) that are used to describe the PDF and CDF.

Common Engineering Distributions

There are a large number of probability distributions. Here, only the more common distributions for engineering problems are summarized. The equations for the PDF, CDF, mean, standard deviation, mode, and so on may be found in many references (e.g., Ref 27). The authors focus primarily on continuous distributions that are commonly used in engineering.
Mathematics. Any function that is strictly positive and for which the area is unity qualifies for a PDF, but there are standard functions, or distributions, that are commonly used in engineering practice. Some of these are the normal, lognormal, extreme value, and Weibull distributions. The PDF and CDF for these distributions are defined as follows.
Normal Distribution. The PDF of the random variable X having a normal distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad -\infty < X < \infty$$

where μ and σ are the distribution parameters. These are adjusted to fit the data. The two parameters are actually the mean and standard deviation of X.
To compute probabilities, it is necessary to integrate f_X, that is, derive the CDF. No closed-form solution exists. The approach used is to first define the variable Z as:

$$Z = \frac{X-\mu}{\sigma}$$

Z will have a normal distribution, with mean of 0 and standard deviation of 1. Numerical integration of this standard normal distribution produces tables and polynomial approximations that make evaluation of the CDF straightforward. (Ref 28).
The standard normal CDF of Z has its own notation, $\Phi(z)$, that is, $F_Z(z) = P(Z \le z) = \Phi(z)$.
Lognormal Distribution. Let X be a random variable. If $Y = \log X$ is normally distributed, then X is said to be lognormally distributed. To compute probabilities when X is lognormal:

$$P(X \le x) = \Phi\left[\frac{\ln(x/\tilde{X})}{\sigma_y}\right] \qquad 0 < X < \infty$$

where:

$$\sigma_y^2 = \ln\left(1 + C_X^2\right)$$

and where C_X is the COV of X, and $\tilde{X}$ is the median of X, calculated by:

$$\tilde{X} = \frac{\mu_X}{\sqrt{1+C_X^2}}$$

Extreme Value Distribution (EVD) (Type 1 EVD of Maxima, also Called Gumbel). The CDF of X having an EVD is

$$F_X(x) = \exp[-\exp(-\alpha(x - \beta))] \quad -\infty < X < \infty$$

where α and β are the distribution parameters. The mean and standard deviations in terms of the parameters are:

$$\mu_X = \beta + \frac{0.577}{\alpha}$$

$$\sigma_X = \frac{1.283}{\alpha}$$

Weibull Distribution. The CDF of X having a Weibull distribution is:

$$F_X(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \qquad 0 < X < \infty$$

where α and β are the shape and scale parameters, respectively. The mean and the COV are:

$$\mu_X = \beta\Gamma\left(\frac{1}{\alpha}+1\right)$$

$$C_X = \left[\frac{\Gamma\left(\frac{2}{\alpha}+1\right)}{\Gamma^2\left(\frac{1}{\alpha}+1\right)} - 1\right]^{1/2}$$

where Γ(·) is the gamma function, defined as:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$$

The parameters in terms of the mean and standard deviation (and COV) are difficult to determine. However, the approximation $\alpha = C_X^{-1.08}$ is a good one. Therefore, use:

$$\alpha = C_X^{-1.08}$$

$$\beta = \frac{\mu_X}{\Gamma\left(\frac{1}{\alpha}+1\right)}$$

Some General Properties. Probability distributions have certain general properties that may be of use in determining, or narrowing down, a distribution or distributions to use. Some properties are given subsequently. Four of the distributions described have been summarized previously; others can be found in any basic applied text on probability theory and statistics.

Normal

- Two parameters
- Symmetric
- Unbounded in both directions
- Sum of normal variables is normal
- Sum of a large number of variables approaches normal, independent of their distributions, as long as no single variable dominates (central limit theorem)
- May be truncated in tails, that is, truncated normal distribution
- *Applications:* widely used for geometric tolerances and material properties, if the coefficient of variation is relatively small, for example, yield and ultimate strength of metals

Lognormal

- Two parameters (sometimes three)
- Nonsymmetric, long right tail
- Bounded to the left (cannot be negative)
- Product of lognormal variables is lognormal
- Product of a large number of variables approaches lognormal, independent of their distributions, as long as no single variable dominates (central limit theorem)
- *Applications:* fatigue (cycles to failure), loading variables, material strengths; found to be a good default distribution for many engineering variables

Extreme value

- Two parameters
- Nonsymmetric
- Unbounded in both directions
- Maximum of a large sample of most common distributions (except lognormal)
- *Applications:* widely used to model the largest load or environment that a component or system experiences over its intended service life

Weibull

- Two parameters (often three)
- May be symmetric or nonsymmetric, depending on parameters
- Type of extreme value distribution
- Bounded to the left (cannot be negative)
- *Applications:* life of electronic components; material strengths (fracture toughness), time to wear out (bearings, pumps)

Exponential

- One parameter (sometimes two)
- Nonsymmetric
- Bounded to the left (cannot be negative)
- *Applications:* life and time between failures of electronic components and systems; fatigue (initial crack size)

Beta

- Two parameters and two bounds
- May be symmetric or nonsymmetric, depending on parameters
- Bounded on both sides
- *Applications:* material strengths (fracture), wear out (bearings, pumps)

Uniform

- The PDF is flat (even distribution of probability).
- Symmetric
- Bounded on both sides
- *Applications:* used to generate random numbers in Monte Carlo; used to model a random variable when nothing can be assumed with regard to its distribution

Rayleigh

- Special case of Weibull (shape parameter = 2)
- *Applications:* distribution of peaks in a random vibration that is narrow band, stationary, and Gaussian

Poisson

- Discrete random variable
- One parameter
- Bounded on both sides
- *Applications:* imperfections (number of defects/volume); occurrences of rare events such as earthquakes, accidents, fires, hurricanes, and so on

Selecting a Probability Distribution and Its Parameters

General Comments. One of the significant challenges to performing a meaningful probabilistic analysis is to determine the input distributions and parameters for a given random variable. The probability of failure results are dependent, in general, on the probability distribution selected, and it may be difficult to determine which distribution is the best choice.

As described previously, certain physical variables are known to be modeled well by the standard distributional forms. If a distribution can be assumed for a given variable, for example, the Weibull to model fracture toughness, then the distribution can be fit to data using point estimation. The method of moments is described subsequently. The performance of the more accurate, and complicated, maximum likelihood method was found by the second author of this document to be insignificantly better than the method of moments.

There are a number of statistical tests to quantify how well a distribution fits data, such as Chi-Square, Kolmogorov-Smirnov (Ref 29, 30) and Anderson-Darling (Ref 31), and software to apply these tests. These tests are not discussed here.

An expedient, subjective method (widely used, however, by engineers) is probability plotting, a graphical technique described as follows.

Probability Plotting. A probability plot provides a simple graphical but subjective method for seeing how well a given distribution fits the data. It also provides another method for estimating the parameters. A transformation of the cumulative distribution function of a given model is made to provide scales on rectangular paper, so that a CDF of that distribution is plotted as a straight line in this space. Then, the empirical distribution function (based on the data) is plotted on this paper, and if the points follow a straight line, one can conclude that the distribution is reasonable.

Examples of the transformation are given as follows.

Extreme Value. The CDF for the EVD is

$$F(x) = \exp\{-\exp[-\alpha(x - \beta)]\}$$

Taking the log twice of both sides and letting:

$$y = -\ln(-\ln(F(x)))$$

it follows that $y = \alpha x - \alpha\beta$. Thus, using this transformation, the EVD $F(x)$ plots as a straight line in x-y space.

Weibull. The CDF for the Weibull distribution is:

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$$

Taking the log twice of both sides and letting:

$$Y = \ln[-\ln(1 - F(x))] \quad X = \ln x$$

it follows that $Y = \alpha X - \alpha\ln\beta$. Thus, using this transformation, the Weibull $F(x)$ plots as a straight line in X-Y space.

Normal. The CDF of normal distribution cannot be computed analytically. The CDF is represented as:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Taking the inverse normal CDF of both sides and letting:

$$y = \Phi^{-1}[F(x)]$$

it follows that:

$$y = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$$

The standard normal CDF and its inverse is included in many common programs. Numerical approximations may be found in Ref 32.

Example Probability Plot. An experiment was performed to measure fracture toughness, K, of a brittle material. A random sample of $n = 5$ was obtained. It is assumed that the observations in this random sample are independent and from the same distribution. The tilde, ˜, indicates a random sample, that is, a vector, >K> = (7.8, 3.2, 1.2, 9.8, 5.7).

A Weibull probability plot is made of the fracture toughness data. The empirical distribution function (estimate of the CDF) is constructed as follows:

- The sample is ordered, smallest to largest. $K_{(i)}$ denotes the ith smallest value.
- The estimate of the distribution function corresponding to $K_{(i)}$ is:

$$F(K_{(i)}) = \frac{i - 0.5}{n}$$

- F_i and K_i are translated into Y_i and X_i using the equations:

$$Y_i = \ln[-\ln(1 - F(x_i))] \quad X_i = \ln K_i$$

- The points are plotted on rectangular paper. If the data follow a linear trend, it is suggested that the Weibull is a reasonable model. This decision is purely subjective.

The data analysis for this example is given in Table 2. *Y* versus *X* is shown in Fig. 3. A straight line can be fit through the points. The least-squares estimators of α and β are then:

$$\hat{\alpha} = \text{slope of } Y\text{-}X \text{ line}$$

$$\hat{\beta} = \exp\left(-\frac{(y \text{ intercept})}{\hat{\alpha}}\right)$$

Table 2 Data analysis for Weibull plot

Instance (*i*)	$K_{(i)}$	$F_i = i - 0.5/n$	Y_i	$\ln(K_i)$
1	1.2	0.1	-2.250	0.182
2	3.2	0.3	-1.031	1.163
3	5.7	0.5	-0.367	1.740
4	7.8	0.7	0.186	2.054
5	9.8	0.9	0.834	2.282

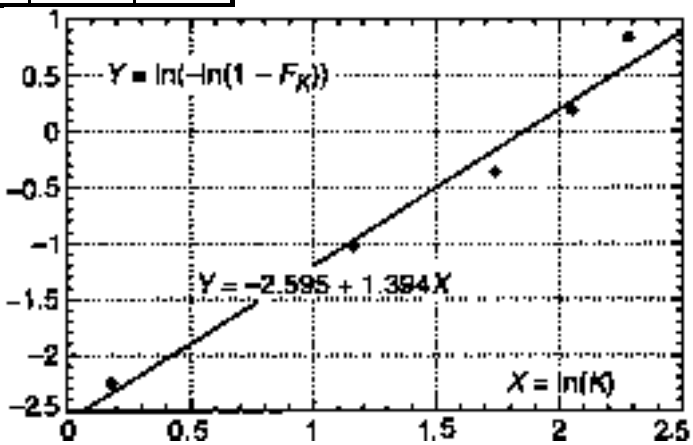


Fig. 3 Weibull probability plot

As shown in the figure, the slope is 1.394 and the *Y*-intercept is -2.595. Thus, the least-squares estimators are:

$$\hat{\alpha} = 1.394$$

$$\hat{\beta} = \exp\left[\frac{2.595}{1.394}\right] = 6.43$$

Usually, a linear regression analysis is done on the transformed data and the correlation coefficient used to indicate the quality of the match. The higher the correlation coefficient, the better the fit.

Method of Moments. The method of moments is demonstrated using fracture toughness data. To fit the Weibull distribution to the data, it is necessary to estimate the parameters α and β. This is done in three steps: (1) compute the sample mean $\overline{K}$ and sample standard deviation s_K, described in the section "Uncertainties in Structural Parameters" in this article. Here $\overline{K} = 5.54$, $s_K = 3.45$; (2) let the sample mean and standard deviation equal the population mean and standard deviation, μ_K and σ_K, respectively; (3) substitute into the equations that relate mean and standard deviation to the parameters and solve for α and β as:

$$\hat{\alpha} = \left(\frac{s_K}{\overline{K}}\right)^{-1.08}$$

$$\hat{\beta} = \frac{\overline{K}}{\Gamma\left(\frac{1}{\hat{\alpha}} + 1\right)}$$

Substituting the values for $\overline{K}$ and s_K, the estimates of μ_K and σ_K become:

$$\hat{\alpha} = 1.67$$

$$\hat{\beta} = 6.20$$

While the agreement with the probability plot is not excellent, it should be noted that the sample size is very small.

Failure Model (or Limit State)

The failure model, or limit state, is a statement of the physics of the problem of interest and defines the criteria for failure. In effect, the limit state divides the domain into “safe” and “failure” regions. The probability of failure is the amount of probability located in the failure region.

The limit state for probabilistic computations is formulated such that when evaluated mathematically, a value less than or equal to zero defines failure and a value greater than zero defines safety. The limit state is typically given the symbol “g” and formulated such that $g \leq 0$ is failure and $g > 0$ is safe.

The limit state depends on the analysis being considered and is purely a deterministic equation. Some common definitions of failure that could be formulated in a limit state are given in Table 3.

Table 3 Some common limit state definitions of failure

Yield strength ≤ stress
Clearance ≤ maximum displacement
Fracture toughness ≤ stress-intensity factor
$J_{material} \leq J_{integral}$
Critical crack size ≤ growing crack size
Material thickness ≤ corrosion depth
Cycles to failure ≤ cycles of interest
Creep strength ≤ Larson-Miller parameter

There is no unique definition required of the limit state. Equivalent representations, such that $g \leq 0$ defines failure and $g > 0$ defines safety, are sufficient. For example, if the limit state is such that the failure occurs when the stress (S) exceeds the material strength (R), the limit state may be represented equivalently in the following forms:

$$g = R - S$$

$$g = R/S - 1$$

$$g = \ln(R/S)$$

Probability of Failure

The probability of failure is determined by solving the multidimensional integral:

$$P_f = \int_{\Omega} f_{\underline{X}}(\underline{x})d\underline{x}$$

where >x> indicates a vector of random variables, $f_{>X>}$ is the joint PDF of the random variables, and Ω denotes the failure region, that is, the region of the random variable space where $g \leq 0$. In two dimensions, this integral is the volume of probability in the failure region.

This integral is difficult to solve numerically for several reasons. First, the joint PDF may not be known, especially if significant correlations exist among random variables. Often, however, the random variables are independent; then the joint PDF is simply the product of the individual density functions.

A second and more significant reason why this integral is difficult to solve is that it is often of high dimension. The dimension of the integral is the number of random variables, for example, a ten dimensional integral if there are ten random variables. This fact rules out numerical integration, except for low dimensions, for example, $N \leq 4$.

A third significant difficulty to solving this integral is that the integration domain, Ω, may be defined implicitly, say by a numerical program, perhaps a finite-element analysis.

As a result of these complications, numerical integration is rarely used. Other techniques that have been developed are discussed in the following two sections, “Evaluation of Probability of Failure” and “Advanced Concepts.”

References cited in this section

27. M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, 2nd ed., John Wiley & Sons, 1993

28. S.M. Shelby, Ed., *Standard Mathematical Tables,* 23rd ed., CRC Press, 1973

29. J. Benjamin and C. Cornell, *Probability, Statistics, and Decisions for Civil Engineers,* McGraw-Hill, 1970

30. A.H.-S. Ang and W. Tang, *Probability Concepts in Engineering Planning and Design,* Vol I, John Wiley & Sons, New York, 1975

31. M. Stephens, "Goodness of Fit with Special Reference to Tests for Exponentiality," Technical Report 262, Department of Statistics, Stanford University, Stanford, CA, 1977

32. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* Applied Mathematics Series, Vol 55, National Bureau of Standards, 1964

Analysis Methods for Probabilistic Life Assessment

Harry R. Millwater, Jr., The University of Texas at San Antonio; Paul H. Wirsching, The University of Arizona

Evaluation of Probability of Failure

The evaluation of the probability of failure is often challenging. Significant progress has been made in this area within the last thirty years. However, there is not one solution method that works in a practical sense for all problems, and it is useful to have a variety of computational techniques available. These techniques are presented subsequently.

Special cases affording a simple solution are presented first. These methods are functions of one random variable, strength minus stress (two random variables), the normal format, and the lognormal format. If the problem fits the domain of application of these methods, then these methods provide an efficient accurate solution.

The mean value first-order second-moment method is a convenient approximate technique to obtain an indicator, the safety index, of the reliability of the structure. However, its results may be significantly in error, and it provides only weak information on the probability of failure.

First- and second-order methods and the advanced mean value method that have been developed over the past few decades are very powerful for computing the probability of failure—particularly for small probabilities typically found in structural analysis. These methods are often attempted first, due to their efficiency; however, they are not guaranteed to work—there are certain mathematical requirements, such as a smooth and differentiable limit state.

The technique of Monte Carlo sampling is the most robust method and is guaranteed to work for any problem, if sufficient sampling is carried out. Also, formal error estimates of the solution are available, unlike the other methods. However, Monte Carlo sampling may require an excessive amount of limit state evaluations, particularly for small probability of failures, and, therefore, may not be practical.

Functions of One Random Variable

For problems where there is only one random variable, the probability of failure may be determined in a straightforward and efficient manner. Let X be a continuous random variable, and let random variable y be defined in terms of x through the functional relationship $y = H(x)$. The CDF of y, $G(Y)$ may be determined given the CDF of x, $F(x)$, and the function H. (H is a representation of the structural behavior, such as shear stress, cycles to crack initiation, etc.)

There is no requirement on the relationship between y and x, in general, and a one-dimensional integral is used to solve for the CDF of y (Ref 23, 30). However, often the relationship between y and x is single valued. In this case, the solution may be determined as follows. The CDF of y, $G(y)$, may be obtained as:

$$G(y) = P[Y \le y] = P[H(X) \le y] = P[X \le H^{-1}(y)] = F(H^{-1}(y))$$

Thus, the knowledge of the CDF of x, $F(x)$, and the inverse mapping $H^{-1}(\cdot)$ is sufficient to determine the CDF of y. In practice, $H^{-1}(\cdot)$ is often not known explicitly and is determined numerically, usually through a search method.

For example, consider a crack propagation fatigue analysis with the initial crack size, a_i, as the random variable. The problem of interest is to determine the probability of failure before a specific number of cycles, say N_0. For most problems, the cycles to failure given the initial crack size (and other essential parameters) must be solved numerically using a fatigue crack growth program, thus defining $N = H(a_i)$.

At the cycles of interest, N_0, the initial crack size that just causes failure is denoted a_i^*. This variable may be determined through trial and error or a search method, for example, bisection. The probability of failure at N_0 is the probability of the

initial crack size exceeding a_i^*, that is, $P[N \le N_0] = P[a_i \ge a_i^*] = 1 - F(a_i^*)$. This pro- cedure may be repeated for any number of values of N_0.

Strength Minus Stress (Two Random Variables)

The problem of estimating the probability of failure given a strength (or capacity), denoted *R*, and a stress (or load), denoted *S*, occurs frequently (shown schematically in Fig. 4). As long as the probability distributions of *R* and *S* are known (or can be estimated), this problem is one of two random variables, and the probability of failure may be determined in a straightforward and efficient manner.

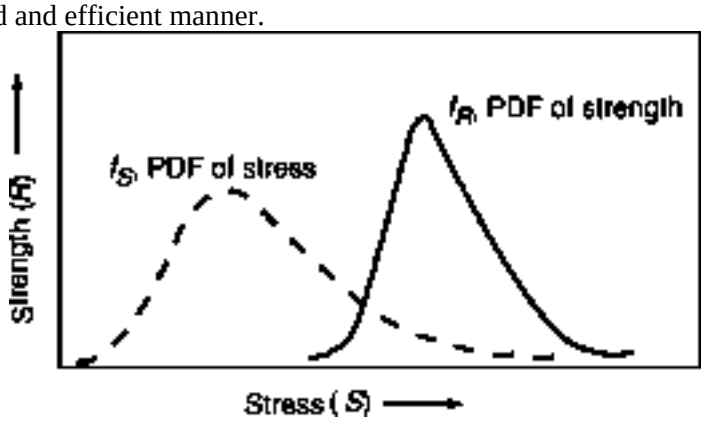


Fig. 4 Strength-stress diagram

The limit state is $g = R - S$, and the probability of failure can be determined by solving the two-dimensional integral:

$$P_f = \int_0^\infty \int_0^s f_S(s) f_R(r) dr ds$$

with *r* and *s* determined from the integration domain shown in Fig. 5, and f_S and f_R are the PDFs of the stress and strength, respectively.

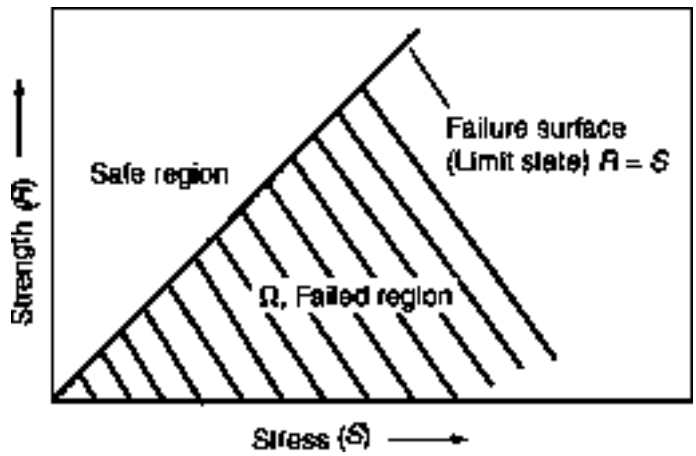


Fig. 5 Strength-stress integration domain

This integral may be reduced to a one-dimensional integral and is straightforward to compute numerically as:

$$P_f = \int_0^\infty f_S(s) F_R(s) ds$$

where F_R is the CDF of the strength.

Normal Format

If all the random variables, X_i, are normally distributed, and the limit state has a linear form:

$$Y = g(X) = A + \sum_{i=1}^{N} B_i X_i$$

where A and B_i are constants and N is the number of random variables, then the random variable Y will also have a normal distribution. The mean and standard deviation of Y are:

$$\mu_Y = A + \sum_{i=1}^{N} B_i \mu_i$$

$$\sigma_Y = \sqrt{\sum_{i=1}^{N} B_i^2 \sigma_i^2}$$

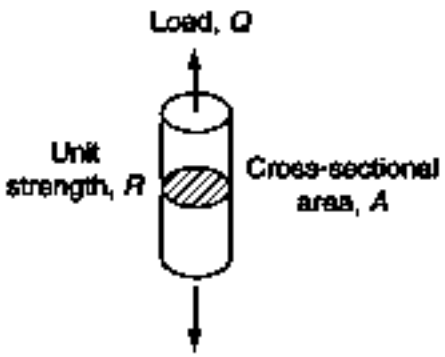
The failure condition is $Y \leq 0$, and the probability of failure is $P_f = P(Y \leq 0)$.
Because Y is normally distributed, the probability can be determined from a table of a standard normal distribution by transforming the variable Y linearly such that the resulting variable has a 0 mean and a standard deviation of 1:

$$Z = \frac{Y - \mu_Y}{\sigma_Y}$$

Then

$$P_f = P(Y \leq 0) = P\left(\frac{Z - \mu_Z}{\sigma_Z} \leq -\frac{\mu_Z}{\sigma_Z}\right)$$
$$= \Phi(-\beta)$$

where $\beta = -\mu_z/\sigma_z$ is called the safety index.
Thus, the probability of failure can be determined by reading the value of $-\mu_Z/\sigma_Z$ from the standard normal tables. An example is shown in Fig. 6.



Statistics of design variables

Both R and Q have normal distributions;
A is constant.
Mean and standard deviation in parenthases

$R \sim N(62, 6.2)$

$Q \sim N(100, 20)$

$A = 2.8$

Failure function (formulated so that failure corresponds to $g \leq 0$)

$g(\underset{\sim}{X}) = R - Q/A$

where $\underset{\sim}{X}$ denotes a vector or random variables

Mean and standard deviation of Z (using basic properties of the mean and variance)

$\mu_Z = \mu_R - \mu_Q/A = 26.286$

$\sigma_Z = \sqrt{\sigma_R^2 + \sigma_Q^2/(1/A)^2} = 9.458$

Safety index and probability of failure

Safety index

$$\beta = \mu_Z/\sigma_Z = \frac{26.286}{9.458} = 2.78$$

Probability of failure

$p_f = \Phi(-\beta) = 0.0027$

Fig. 6 Application of normal format

Lognormal Format

If all the random variables are lognormally distributed and the limit state has a multiplicative form:

$$g(\underset{\sim}{X}) = B\prod_{i=1}^{N} X_i^{a_i}$$

where B and all a_i are constants and N is the number of random variables, then $Y = \ln(g)$ is normally distributed:

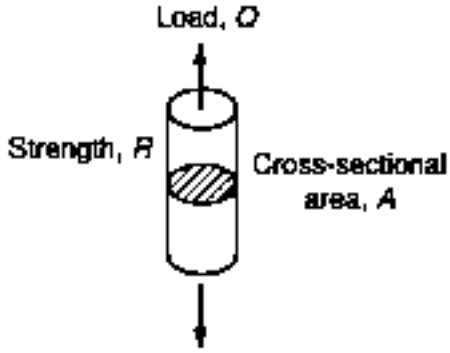
$$Y = \ln B + \sum_{i=1}^{N} a_i \ln(X_i)$$

Because X_i are lognormal, $\ln(X_i)$ are normal. Then, this format becomes identical to the normal format. The limit state is $g \leq 1$, or $Y \leq 0$, or $Z = (Y - \mu_Y)/\sigma_Y \leq 0$, and $P_f = \Phi(-\beta)$, where $\beta = -\mu_Y/\sigma_Y$, and

$$\mu_Y = \tilde{Y} = \ln \tilde{g} = \ln\left[B\prod_{i=1}^{N} \tilde{X}_i^{a_i}\right]$$

$$\sigma_Y = \left\{\ln\left[\prod_{i=1}^{N}\left(1 + C_i^2\right)^{a_i^2}\right]\right\}^{1/2}$$

$\tilde{X}$ denotes the median of X, and C_i denotes the coefficient of variation of X_i. An example is given in Fig. 7.



Statistics of design variables

Mean and standard deviation in parentheses. Both R and Q have lognormal distributions

$R \sim ln(62, 6.2)$

$Q \sim ln(100, 20)$

$A = 2.8$

Failure function (Formulated so that $Z = \ln g \leq 0$ is failure.)

$g(\underset{\sim}{X}) = RA/Q$

where $\underset{\sim}{X}$ denotes a vector or random variables

Compute median and COV of R and Q

$$C_R = \frac{\sigma_R}{\mu_R} = 0.10 \qquad \tilde{R} = \frac{\mu_R}{\sqrt{1 + C_R^2}} = 61.7$$

$$C_Q = \frac{\sigma_Q}{\mu_Q} = 0.20 \qquad \tilde{Q} = \frac{\sigma_Q}{\sqrt{1 + C_R^2}} = 98.1$$

$$\mu_Z = \ln \tilde{R}A/\tilde{Q} = 0.5659$$

$$\sigma_Z = \sqrt{\ln[(1 + C_R^2)(1 + C_Q^2)]} = 0.2217$$

Safety index and probability of failure

Safety index

$$\beta = \mu_Z/\sigma_Z = \frac{0.5659}{0.2217} = 2.55$$

Probability of failure

$$p_f = \Phi(-\beta) = 0.0054$$

This compares with a safety index of 2.78 using the normal format. The lognormal format could be described as a somewhat more conservative model.

Fig. 7 Application of lognormal format

Mean Value First-Order Second-Moment (MVFOSM)

The MVFOSM may be applied to perform reliability analysis in those general cases where the limit state, $g = (>X>)$, can be any function of the vector of design factors, $>X>$, of any number. Mean value first-order second-moment is based on only two parameters (the mean and standard deviation) of all random variables; no distributional information is used.
In general, a basic measure of reliability is the safety index, defined as:

$$\beta = \frac{\mu_g}{\sigma_g}$$

The safety index was defined previously in the discussion of the normal format, where it was used to compute the probability of failure. But more generally, β can be used as an index of reliability even in those cases where the requirements of the normal format are not met. An estimate for the probability of failure given β is $P_f = \Phi(-\beta)$.
Approximations to the mean and standard deviation of g are:

$$\mu_g \cong g(\underline{\mu})$$

$$\sigma_g \cong \sqrt{\sum_{i=1}^{N}\left(\frac{\partial g}{\partial X_i}\right)^2_{\underline{\mu}} \sigma_i^2}$$

where >μ> is a vector of mean values of >*X*>, σ_i, is the standard deviation of X_i, and *N* is the number of random variables. These approximations are developed by taking the expectation operator of the Taylor's series expansion of *g*(>*X*>). For smaller variances, higher-order terms are negligible. While these approximations depend on the degree of nonlinearity of the *g*-function, they are generally considered to be reasonable if all of the variables have coefficients of variation less than 15%.

However, while convenient to compute numerically, there are several problems with the MVFOSM method. For one, the probability of failure defined by the safety index β may not be accurate for nonnormal variables with moderate to large coefficients of variation. Second, the value of β depends on the particular formulation of the limit state. Different but mathematically equivalent formulations give different values for the safety index.

First-Second-Order Methods

In the general case of a single component, the limit state function, $g(\underline{X})$, is a function of *N* design factors, X_i, $i = 1, N$, each having a different statistical distribution. The general expression for the probability of failure, P_f, was presented in the section "Probability of Failure," but, as explained, its evaluation may be numerically difficult.

Monte Carlo sampling may not be practical, because of long solution times due to a computationally time-intensive limit state evaluation and a large number of samples required for small probability of failure. As mentioned previously, a second-moment formulation, MVFOSM, was convenient to use but was not an accurate predictor of failure probability. As a result, the first-order reliability method (FORM) and, subsequently, the second-order reliability method (SORM) (Ref 4, 5, 6, 7) were developed. The basis of a FORM/SORM analysis is the concept of the most probable point (MPP). The MPP is the most likely location on the limit state in random variable space. The location of the MPP is important, because linear approximations of the limit state about the MPP are often sufficient to closely estimate the probability of failure. The MPP is usually located using an optimization algorithm that requires gradients of the limit state with respect to the random variables. An example of the MPP concept is shown in Fig. 8.

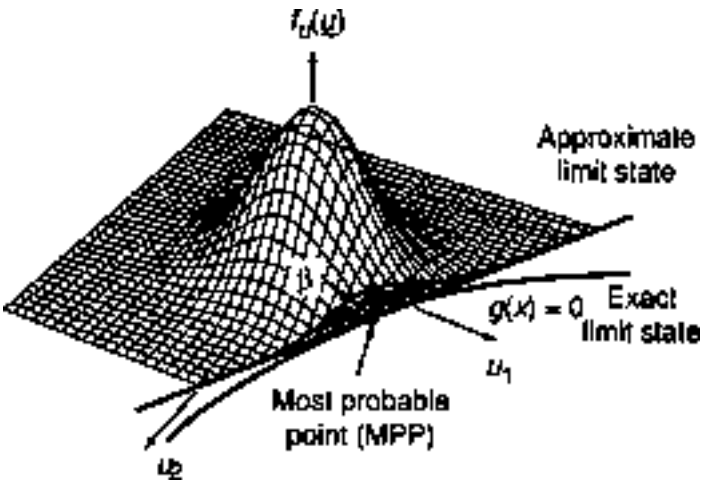


Fig. 8 Schematic of the most probable point concept

The FORM is summarized as follows (Ref 33):

1. Each random variable has a distribution defined by the CDF, F_{xi}. Transform all variables to standard normal variates, u_i, using the transformation $F_{Xi}(x) = \Phi(u_i)$. Thus,

$$x_i = F_{X_i}^{-1}\left[\Phi(u_i)\right]$$

2. Substituting each x_i into the limit state function *g*, *g* now becomes a function of the vector, >*u*>. Thus, *g*(>*u*>) can be thought of as a surface in *N*-dimensional *u*-space.

3. Determine the "generalized" safety index, β_g, as the minimum distance from origin of u-space out to the surface, $g(>u>)$. This can be accomplished with an optimization program.
4. An estimate of the probability of failure is $P_f = \Phi(-\beta_g)$.

If the surface is a plane (no curvature), then P_f is exact. If the surface has curvature, then P_f as computed by FORM is only approximate, with the error as a function of the nonlinearity. A numerical correction can be made to the FORM process to treat this nonlinearity and produce estimates of P_f that are reasonably accurate (typically less than 5% error). This algorithm is called a SORM (Ref 7, 34).
A FORM/SORM is most effective for small probabilities for which a large number of Monte Carlo samples would be required. However, because a FORM/SORM analysis typically requires gradients, the solution time is dependent on the number of random variables. In addition, the algorithm may not be able to locate the MPP, or there may be multiple MPPs. Thus, FORM/SORM may not be able to solve some problems.

Advanced Mean Value

Wu et al. (Ref 9) developed the advanced mean value (AMV) method specifically for problems with a time-intensive limit state, with the goal of obtaining the distribution of a response and estimates of small probabilities of failure with a minimum of limit state evaluations. Advanced mean value is similar to FORM/SORM in many concepts but often requires fewer limit state evaluations.
A brief summary of the process is described. Consider a response variable, $Y(>X>)$, where $>X>$ is a vector of random variables. The goal is to derive the distribution function, $F_Y(y)$. If the response variable is defined as the g-function, then a reliability problem can also be formulated; that is, the probability of failure is the distribution function of g evaluated at zero. The steps for the AMV method are:

1. Expand $Y(>X>)$ about the mean of $>X>$ (linear terms of a Taylor's series):

$$Y = g(X) = A + \sum_{i=1}^{N} B_i X_i$$

2. where N is the number of random variables, A is a constant, and:

$$B_i = \left.\frac{\partial Y}{\partial X_i}\right|_{\mu_i}$$

3. The B_i are typically evaluated numerically by performing perturbations of each variable about the mean. Note this requires $N + 1$ function evaluations of the limit state.
4. Use a reliability method, for example, FORM/SORM, to compute the safety index, β_i, and the corresponding MPP, $>X>$, at each of K selected points of Y. Typically, only approximately eight points are necessary to define F_Y. These points produce the mean value solution. This result is expected to be a crude approximation to F_Y, because of the linear Taylor series approximation.
5. Reevaluate Y at each MPP, that is, perform a function evaluation to get S_i; $i = 1, K$. The points (β_i, S_i) provide an improved estimate of F_Y.
6. For those problems where it is suspected that the AMV solution may have significant error, a linear expansion of Y can be constructed at each MPP and the process repeated. This is referred to as AMV+.

The authors have extensive experience with the AMV method and found it to be very efficient and accurate for well-behaved limit states.

Response Surface Method

Structural reliability analyses that require evaluation of a time-consuming limit state may not be practical to solve using standard Monte Carlo sampling, due to excessive computer time. This may occur if the evaluation of the limit state requires significant computation time and many evaluations, that is, samples are required to obtain the probability of failure. For example, if a fillet radius is a random variable and a finite-element analysis (FEA) is used to compute the

local stresses, then a FEA would be required for each Monte Carlo sample. Clearly, this would be prohibitive for most Monte Carlo-based analyses.

A simple approach is to approximate the relationship between the local stress and the fillet radius with a polynomial, usually second order, before performing the probabilistic analysis. Then, for each Monte Carlo sample, the polynomial is evaluated instead of performing a FEA. This significantly improves the solution time but adds a potential source of error—the quality of the polynomial approximation.

The technique of approximating a relationship between an output, for example, stress, and input parameters, for example, fillet radius, is termed the response surface method (RSM). This technology was originally developed for experimental design and analysis as a way to develop simple models relating inputs and outputs. It has since found use in probabilistic analysis.

Response surface method has a rich historical and technical background. There has been much work on determining the correct set of input values to select. The reader is referred to Ref 35, 36, 37, 38 for further details. Here, only the basic essentials for probabilistic analysis are covered.

Methodology. The RSM basic concepts are:

- Generate a series of output responses versus random variable inputs using design of experiments (DOE) techniques.
- Generate a polynomial approximation of the response with respect to the random variables, that is, a response surface.
- Evaluate the quality of the fit, and adjust if necessary.
- Perform probabilistic computations, typically using Monte Carlo sampling although other methods can be used, using the polynomial approximation as a substitute for, or as a component of, the limit state.

The procedure is shown schematically in Fig. 9.

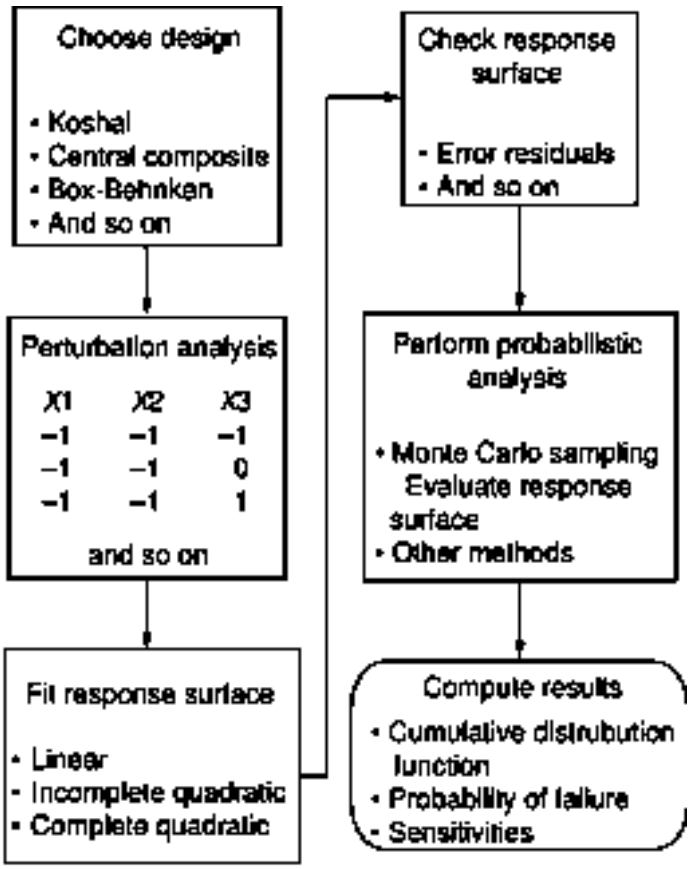


Fig. 9 Schematic of response surface method

The mathematical relationship of the response surface may be written:

$$r(\underline{\hat{X}}) = a_0 + \sum_{i=1}^{p} a_i (x_i - x_i^*) + \sum_{i=1}^{p} b_i (x_i - x_i^*)^2 + \sum_{i=1}^{p} \sum_{j=1}^{i-1} c_{ij} (x_i - x_i^*)(x_j - x_j^*)$$

where r represents the response, $>\hat{X}>$ represents those p random variables that affect the r, a_i are linear coefficients, b_i are the quadratic coefficients, c_{ij} are the mixed-terms coefficients, x_i represent the random variables, and x_i^* is an expansion point, typically the mean. Here, b_i and c_{ij} are zero for a linear function.

The quality of the polynomial approximation, r, may be estimated through statistical tests. Thus, a good fitting program gives the analyst feedback as to the quality of the fit. If the fit is poor, more solution points may need to be added, or the fit may be performed in terms of transforms of the basic variables, for example, log(x) or x raised to a power (Ref 39).

The choice of input values of the random variables to use to generate the response values to fit is a broad subject that has undergone considerable study and is generally called DOE. Certain statistical properties are desirable, such as a solution that is unbiased, that is, no variable is unduly emphasized over any other.

A number of experimental designs have been developed that give good statistical properties, and several try to minimize the number of required analyses (Ref 37). Common choices are Koshal (linear), factorial, fractional factorial, central composite, and Box-Behnken. A typical design, central composite, is shown in Fig. 10. The solid circles represent combinations of variables required as input. From the figure, it is seen that 15 analyses are needed for a central composite design with three random variables.

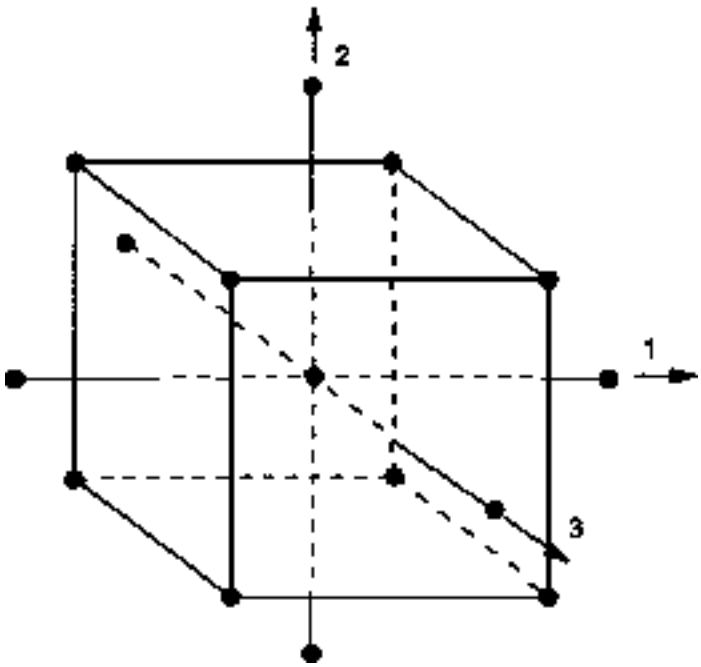


Fig. 10 Central composite design using three variables

The minimum number of runs needed to generate a response surface depends on the order of the polynomial and the number of random variables, as shown in Table 4. From the table, it is seen that the number of coefficients and, therefore, the minimum number of computer runs increases dramatically as the number of random variables becomes large.

Table 4 Minimum number of analyses required to fit polynomial

Order	***N* = 5**	***N* = 20**	***N* = 50**
First	6	21	51
Second (no mixed terms)	11	41	101
Full second (mixed terms)	21	231	1326

The designs define the relative combination of variables required but not the absolute values. For example, the design may require that random variables {x1, x2, x3} be set at {+1, +1, -1} levels, respectively, but the value of a “level” is not defined. A typical approach in probabilistic analysis is set to a level equal to one standard deviation. However, this value may not give accurate solutions for small probabilities of failure, because the failure region would most likely be far

outside of one standard deviation. Larger perturbations, for example, two or three standard deviations, may be more accurate. As yet, there are no standard guidelines on this.

Advantages and Disadvantages. The RSM is a technique that can be useful for certain probabilistic analysis problems. However, it has advantages and disadvantages, as summarized in Table 5.

Table 5 Advantages and disadvantages of the response surface method (RSM)

Advantages
Easy to do. Can be done with numerical codes or testing
Decouples probabilistic and deterministic analyses. Easy transfer of information between modelers
Probabilistic analysis (using Monte Carlo sampling) is straightforward and relatively fast
Disadvantages
Requires a large number of analyses for a large number of random variables (see Table 4)
May not be as efficient as other techniques
Sometimes not accurate for small probabilities of failure
Difficult to assess error in probability

The advantages are that it is straightforward to understand and perform and may save considerable computer time. The primary disadvantages are that a large number of analyses are required for problems with a large number of random variables that affect the response surface and that the probability of failure solution may not be accurate for small probabilities of failure, nor the error easily quantified.

Monte Carlo Sampling

Overview. Monte Carlo sampling is a method of statistical trials. The procedure is to generate realizations of random variable inputs (e.g., load, material properties, and geometries) based on their probability distributions, evaluate the model, and record the results (e.g., fail or no fail). This procedure is repeated many times with new randomly generated inputs. When a sufficient number of samples has been carried out, estimates of the probability of failure and the moments (e.g., mean and standard deviation) of the response may be computed (see Fig. 11).

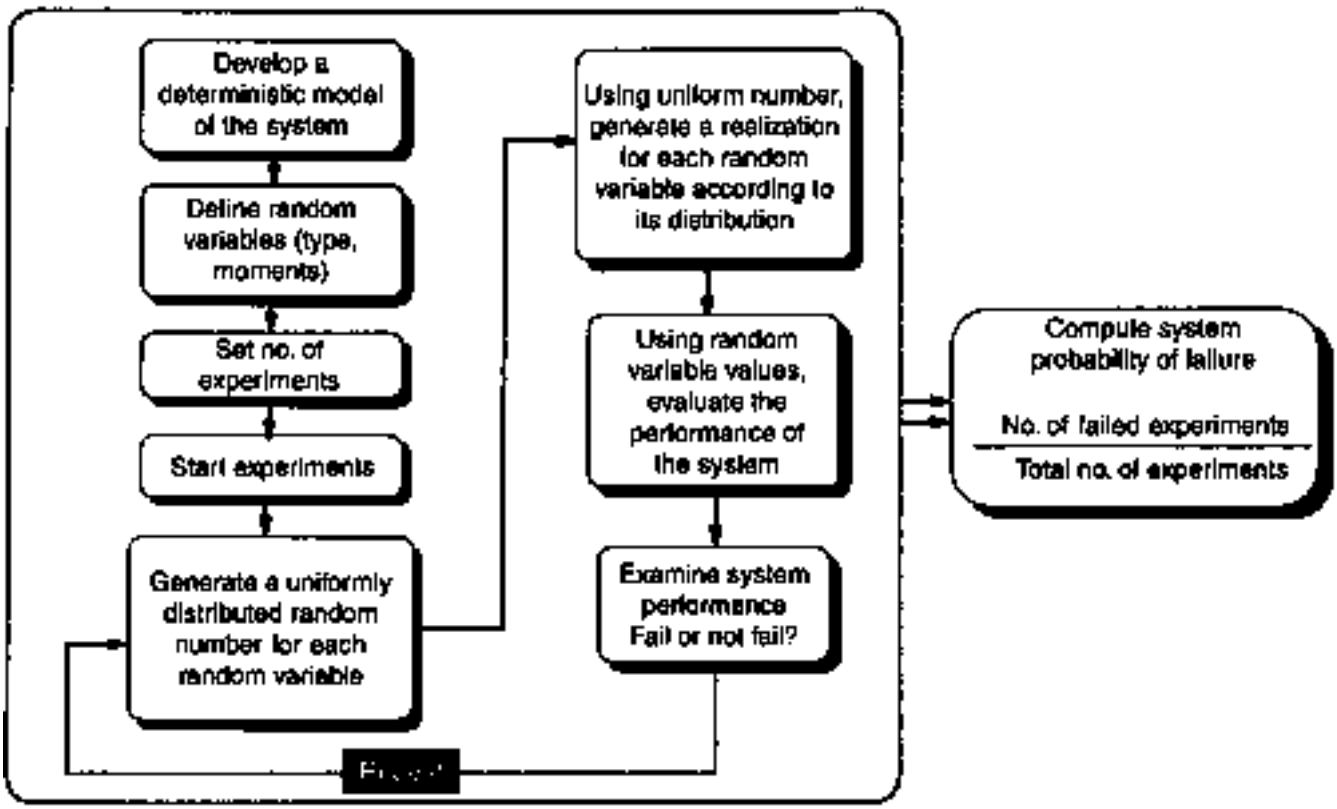


Fig. 11 Overview of Monte Carlo method

Monte Carlo is a powerful technique that is very robust and straightforward to use. It is guaranteed to converge to the correct probability solution in the limit as the number of samples, that is, statistical trials, increases to infinity. The

potential drawback is related to computational time, in that many realizations, and thus much computational time, may be required to obtain accurate probability of failure estimates for the low probabilities often found in structural life assessment.

Monte Carlo simulation has extremely broad applicability. It is used in the context of probabilistic analysis in many fields, such as engineering (structural), physics, materials science, biology, biomechanics, and finance and games, among others. It is also useful in a nonprobabilistic context for evaluating high-dimensional integrals.

The term “Monte Carlo” was coined in the 1940s during the Manhattan project, because of the similarity of statistical games of chance and because the city of Monte Carlo was a gambling center.

Generating Random Variables. Monte Carlo sampling requires repeated generation of random variable (RV) values based on their probability distributions. The generation of RV values is typically done using the inverse function method if the RVs are independent. More sophisticated methods are required if the RVs are correlated.

Values for independent RVs are generated as follows:

1. Generate a uniform random number between 0 and 1; this value becomes the probability of the RV.
2. Using this probability, invert the CDF to obtain the RV value that corresponds to this probability.

This procedure is shown mathematically as:

$$U^* = U[0, 1]$$

$$F(x^*) = U^*$$

$$x^* = F^{-1}(U^*)$$

where $U[0, 1]$ indicates a uniform distribution with limits 0 and 1, U^* is a random number between 0 and 1 generated from the uniform distribution, F denotes the random variable CDF, and x^* is the generated realization of random variable X.

The procedure is shown schematically in Fig. 12.

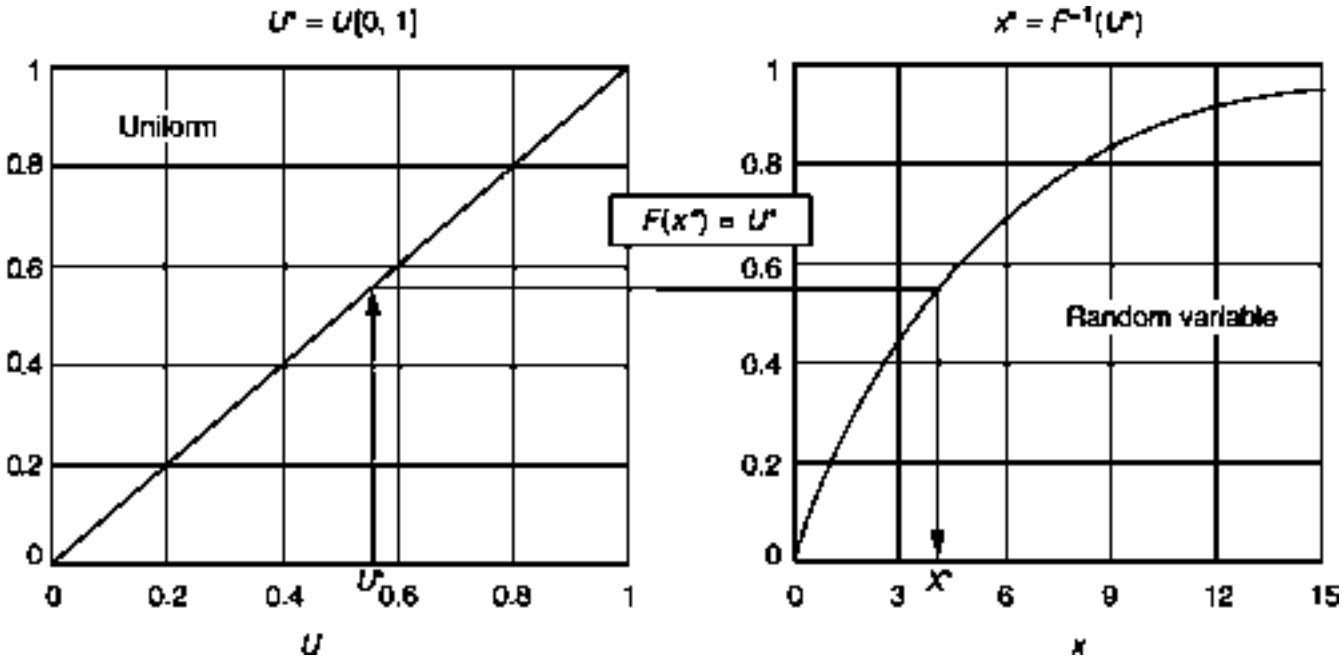


Fig. 12 Schematic of inverse function method

Figure 13 presents an example of generating realizations for an exponential distribution. The values for U^* are merely illustrative. In general, U^* values should be generated randomly, as described in the next section.

Example: exponential distribution
with $\lambda = 0.1$ (mean = 10)

$$F(x) = 1 - e^{-\lambda x}$$

$$x^* = F^{-1}(U^*) = -\frac{1}{\lambda}\ln(1 - U^*)$$

U^*	X^*
0.1	1.053
0.3	3.567
0.5	6.931
0.7	12.040
0.9	23.026

Fig. 13 Example: generating random variable values

Generating Uniform Numbers. There are computer algorithms that may be used to generate pseudorandom numbers on a computer. The numbers generated are not truly random, in that the sequence generated eventually repeats. However, the pseudorandom numbers are indistinguishable from truly random numbers for a good random number generator and for many applications that do not require an exceedingly large number of random variables, such that the period is exceeded. Most computer languages and spreadsheets have an intrinsic function for generating uniform random numbers. These methods are often sufficient but sometimes are not of good quality. Developing good random number algorithms is still a topic of current research. References 40– 42 discuss this topic in more detail.

A popular and typically reasonably good random number generator (depending on the choice of the numerical constants) is the linear congruential random number generator. Start with a seed, apply mathematical equations to generate a random value, and use this random value as a seed to generate the next random value. The equations are:

$$x_{i+1} = (ax_i + b)(\text{mod } m)$$

$$U_i = \frac{x_i}{m}$$

where x_i is the seed, that is, the last generated random number, a, b, and m are constants, mod is the modulo function, that is, the remainder, x_{i+1}, becomes the new seed, and U_i is the uniform random number. Values of the constants are critical to the performance of the random number generator. The period is $\leq m$, so m is typically chosen as large as possible. Reference 42 lists some values of typical constants and their properties.

Probability Estimates. In probabilistic life assessment, one is typically interested in determining the probability of failure, that is, $P[g \leq 0]$. The probability of $g \leq 0$ is estimated by the ratio of the number of failure points divided by the total number of samples, as:

$$P_f = \frac{n_f}{n}$$

where n is the total number of samples, n_f is the number of samples for which $g \leq 0$, and P_f is the estimate of the probability of failure. It is important to realize that P_f is merely an estimate of the true probability of failure that remains forever unknown. Thus, P_f is a random variable and changes if the simulation is rerun, for example, with a new seed. The variation in P_f, that is, its standard deviation, is an indication of the accuracy of the answer and can be determined from the binomial distribution as:

$$\sigma(P_f) = \sqrt{\frac{(1 - P_f)P_f}{n}}$$

Note that the standard deviation of the probability of failure reduces as the square root of the number of samples. Thus, for example, if the standard deviation is to be reduced by a factor of two, four times as many samples are required. The accuracy of the estimate of P_f is discussed in the section “Accuracy: Error and Confidence Estimation” in this article.

Statistical Analysis. Sometimes the probabilistic response of a structure is to be evaluated, not just the probability of failure. For example, the goal is to generate a histogram of the stress, or cycles to failure, and/or compute the mean, median, and standard deviation. The standard statistical procedures may be applied to the Monte Carlo sampling results.

Histograms. A histogram is a plot (or tabulation) of the number of data points versus selected intervals. A frequency diagram is a plot of the frequency of occurrence versus selected intervals generated by dividing the histogram by the total number of samples.

The histogram or frequency diagram can be easily generated from the Monte Carlo results by collecting the response values into “bins.” The bins are ranges of the response, typically equally spaced. The number of bins depends on the size of the data set. Rules of thumb for selecting the number of bins are:

$$M = \sqrt{n}$$

or

$$M = 1 + 3.3 \log_{10}(n)$$

where M is the number of bins and N is the number of samples. The bin boundaries are then set by dividing the range between the maximum and minimum values by the number of bins.
As an example (shown in Fig. 14), a set of 40 data points is divided into six bins.

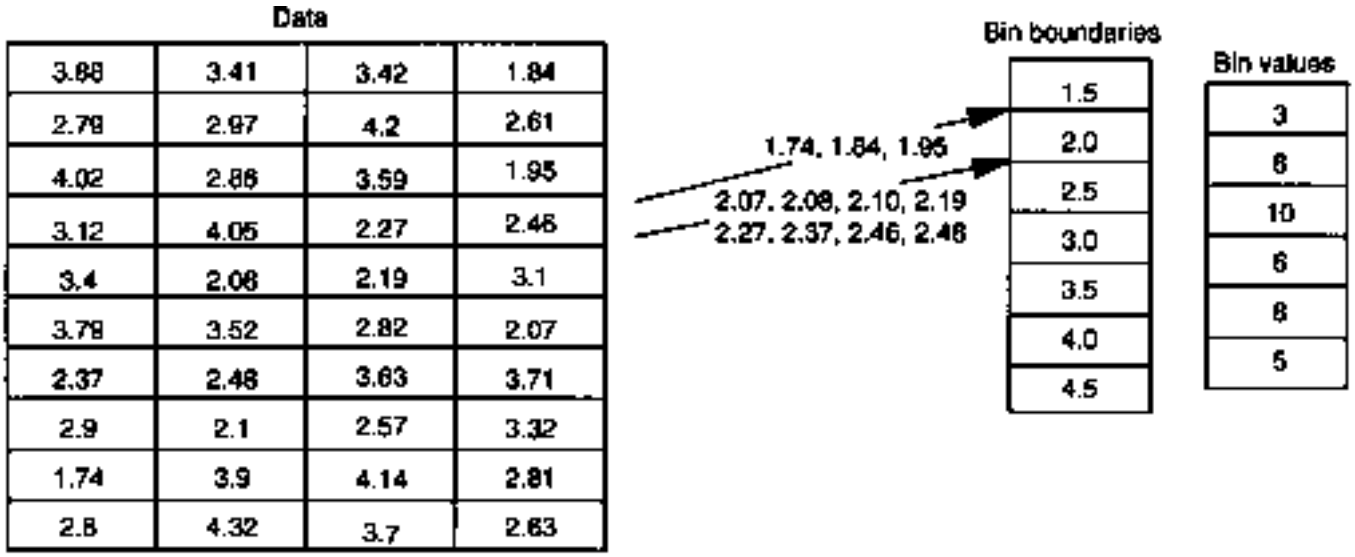


Fig. 14 Generating a histogram from Monte Carlo results

Figure 15 shows a frequency plot (and the resulting CDF, discussed in the section “Estimating CDF” in this article).

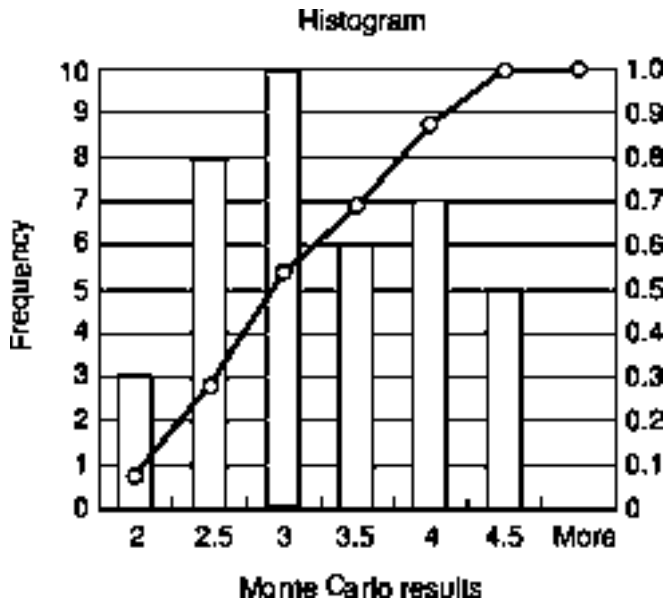


Fig. 15 Histogram example

Estimating Moments. The moments of the response may be obtained using standard statistical formulas. For example, the mean and standard deviation may be estimated from the equations:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \textbf{Mean}$$

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \textbf{Standard deviation}$$

where n is the total number of samples, and X_i are the Monte Carlo sampling results.

Estimating CDF. Sometimes one wants to develop the entire CDF of a response, not just compute the probability of failure. There are two basic ways to do this. The first method is to compute the probability of failure, using the procedure defined previously, at a number of values of X. This is, in effect, a superset of estimating the probability of failure. Figure 16 demonstrates this procedure. The CDF is shown as the line graph in Fig. 15.

Data

3.88	3.41	3.42	1.84
2.79	2.97	4.2	2.61
4.02	2.86	3.59	1.95
3.12	4.05	2.27	2.46
3.4	2.08	2.19	3.1
3.79	3.52	2.82	2.07
2.37	2.48	3.63	3.71
2.9	2.1	2.57	3.32
1.74	3.9	4.14	2.81
2.6	4.32	3.7	2.53

X	[illegible]	[illegible]
1.5	0	0.0
2.0	3	0.08
2.5	11	0.28
3.0	21	0.53
3.5	27	0.68
4.0	35	0.88
4.5	40	1.00

Fig. 16 Generation of CDF from Monte Carlo results

A second approach to generating a CDF is to generate an empirical CDF. First, sort the Monte Carlo sampling results in ascending order. Then apply the formula:

$$F_i = \frac{i - 1/2}{n}$$

where i denotes the order of the sample in the sorted list, that is, first, second, and so on, n is the total number of sample points, and F_i is the probability associated with sample i. For example, the results for the first seven data points from Fig. 16 are shown in Fig. 17. The CDF plot for all the data is shown on the right-hand side of Fig. 17.

X	[illegible]
1.5	0.0
2.07	0.086
2.48	0.263
2.97	0.513
3.52	0.688
4.02	0.888
4.32	0.988

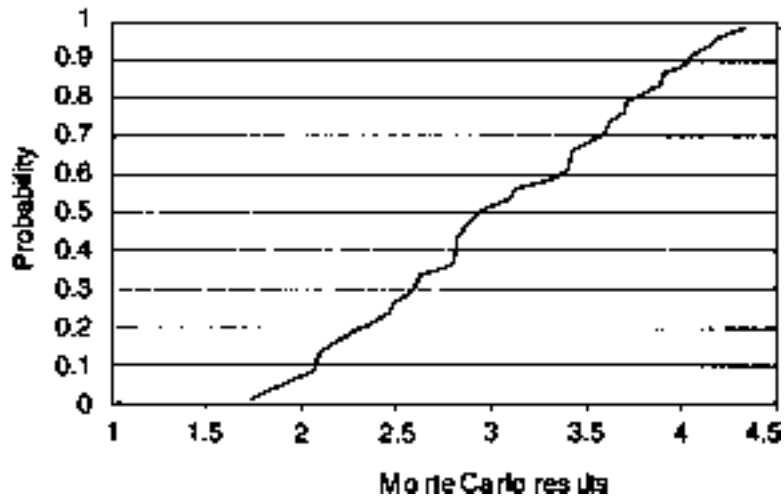


Fig. 17 Generation of CDF using empirical formula

Accuracy: Error and Confidence Estimation. A natural question arises in Monte Carlo sampling of how many samples to use. The accuracy of the probability estimate is dependent on the number of samples. It was previously shown that the probability of failure estimate from Monte Carlo sampling is itself a random variable, and that the standard deviation of the probability of failure varies as the square root of the number of samples.

The confidence bounds of the P_f can be estimated by assuming that the probability of failure is normally distributed (a good approximation). For example, Shooman (Ref 43) developed an equation that can be used to estimate the percent error in a probability estimate as:

$$\gamma = 200\sqrt{\frac{1-P_f}{nP_f}}$$

where P_f is the estimated probability of failure, n is the total number of samples, and γ is the estimated percentage error in the computed probability of failure from the true probability of failure. This equation states with 95% confidence that the percent error in the computed probability of failure, relative to the true probability of failure, is less than γ. For example, given $P_f = 0.001$ and $n = 100{,}000$. Then, the equation yields a percent error of 20%. Thus, with 95% certainty, the true probability of failure lies within 0.001 ± 20%, or P_f lies within the range (8×10^{-4}, 1.2×10^{-3}) with an expected value of 0.001.

This formula can be used to estimate the required number of samples to achieve a desired percent error as:

$$n = \left(\frac{1-P_f}{P_f}\right)\left(\frac{200}{\gamma}\right)^2$$

Note that because P_f is usually small for structural reliability analysis, the term $1 - P_f$ can usually be approximated as 1.

The required number of samples versus probability of failure for different percent error values with 95% confidence is shown in Fig. 18. The figure indicates that a very large number of samples may be required. For example, if the estimated probability of failure is 10^{-6} and a 10% error is desired, more than 100 million samples are required. This number of samples is clearly not practical if the evaluation time for the response is non-trivial.

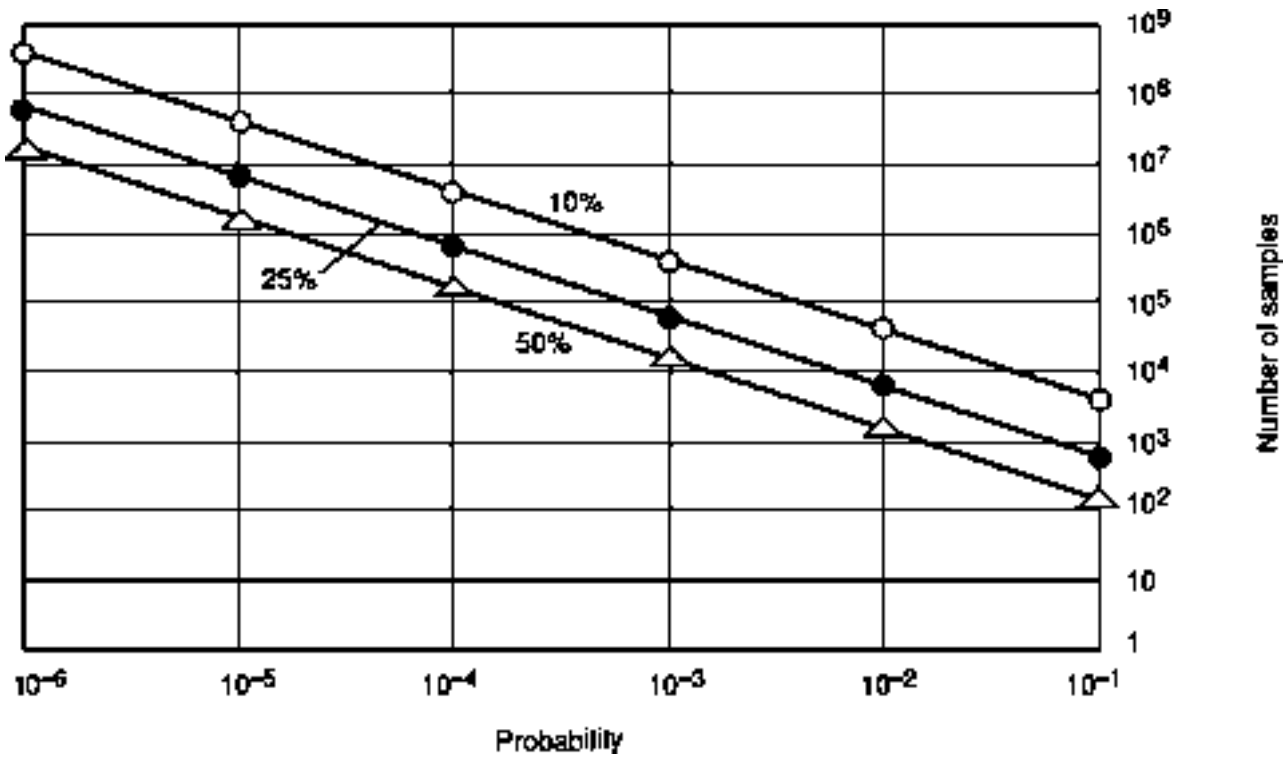


Fig. 18 Required samples versus probability for different percent error values

A simple but crude rule of thumb is to ensure that one has at least 10 failure points (providing an error of 63% with 95% confidence). The minimum number of sample points to achieve this is:

$$n = \frac{10}{P_f} \qquad (P_f < 0.5)$$

$$n = \frac{10}{1 - P_f} \qquad (P_f > 0.5)$$

References cited in this section

4. A.M. Hasofer and N.C. Lind, Exact and Invariant Second Moment Code Format, *J. Eng. Mech., (ASCE),* Vol 100 (No. EM1), 1974, p 111–121

5. R. Rackwitz and B. Fiessler, Structural Reliability under Combined Load Sequences, *J. Comput. Struct.,* Vol 9, 1978, p 489–494

6. L. Tvedt, "Two Second Order Approximations to the Failure Probability," Veritas Report, RDIV/20-004-83, Det Norske Veritas, Oslo, Norway, 1983

7. K. Breitung, Asymptotic Approximations for Multinormal Integrals, *J. Eng. Mech. (ASCE),* Vol 110, 1984, p 357–366

9. Y.-T. Wu, H.R. Millwater, and T.A. Cruse, Advanced Probabilistic Structural Analysis Method for Implicit Performance Functions, *AIAA J.,* Vol 28 (No. 9), 1990

23. S. Rao, *Reliability-Based Design,* McGraw-Hill, 1992

30. A.H.-S. Ang and W. Tang, *Probability Concepts in Engineering Planning and Design,* Vol I, John Wiley & Sons, New York, 1975

33. R.E. Melchers, *Structural Reliability Analysis and Prediction,* Ellis Horwood Limited, U.K., 1987

34. L. Tvedt, Distribution of Quadratic Forms in Normal Space with Application to Structural Reliability, *J. Eng. Mech.* (*ASCE*), Vol 116, 1990, p 1183–1997

35. G. Box and N. Draper, *Empirical Model-Building and Response Surfaces,* John Wiley & Sons, Inc., 1987

36. R. Gunst and R. Mason, "How to Construct Fractional Factorial Experiments," American Society for Quality Control, Statistics Division, ASQC Quality Press, 1991

37. R. Mason, R. Gunst, and J. Hess, *Statistical Design and Analysis of Experiments,* John Wiley & Sons, Inc., 1989

38. D. Montgomery, *Design and Analysis of Experiments,* 2nd ed., John Wiley & Sons, Inc., 1984

39. G. Box and D. Cox, An Analysis of Transformations, *J. R. Stat. Soc.,* 1964, p 211–243, discussion. p 244–252

40. D. Knuth, *The Art of Computer Programming,* Vol II, Addison Wesley Longman, Inc., 1997

42. B. Shipley, *Stochastic Simulation,* John Wiley & Sons, 1987

43. M. Shooman, *Probabilistic Reliability: An Engineering Approach,* McGraw-Hill Companies, 1968

Advanced Concepts

The probabilistic analysis of structures is a broad subject. Several advanced concepts not discussed previously are presented as follows. These include importance sampling, an efficient, but specialized sampling method; time-variant reliability, reliability as a function of time; system reliability, consideration of multiple limit states; risk analysis, integrating consequences of failure; and target reliabilities, typical reliability values.

Importance Sampling

Importance sampling involves repeated sampling of the limit state, such as Monte Carlo sampling; however, the sampling is biased toward the "important" region of the design space, that is, the region where the chances of seeing a failure are greatest. This can lead to a much more efficient solution than Monte Carlo sampling. Efficiency of two orders or more is often obtained.

The term "importance sampling" is usually used generically, in that there are many variations of importance sampling (Ref 10, 11, 12, 13). Sometimes, an importance sampling method is specifically developed for a specific problem of interest. For example, in Ref 44, a search is performed to determine the range of initial crack sizes that cause failure within a specific number of cycles. Sampling of initial crack sizes is then restricted to this range. This results in a much more efficient analysis than Monte Carlo sampling.

The "price to pay" in an importance sampling solution method is that some knowledge of the limit state is required in order to focus the samples in the important region. Typically, this involves a search procedure that itself requires some computational time. More important, for ill-behaved problems, it may be difficult, if not impossible, to locate the failure region. Because of this, importance sampling methods are generally not as robust as Monte Carlo sampling.

Time-Variant Reliability

Over a period of time, structural integrity may degrade, typically due to fatigue and/or corrosion. Figure 19 illustrates a fatigue process in which there is a growing crack, resulting in a lowering of the ultimate strength with increasing time. Not only does ultimate strength degrade, but also there is uncertainty regarding its value.

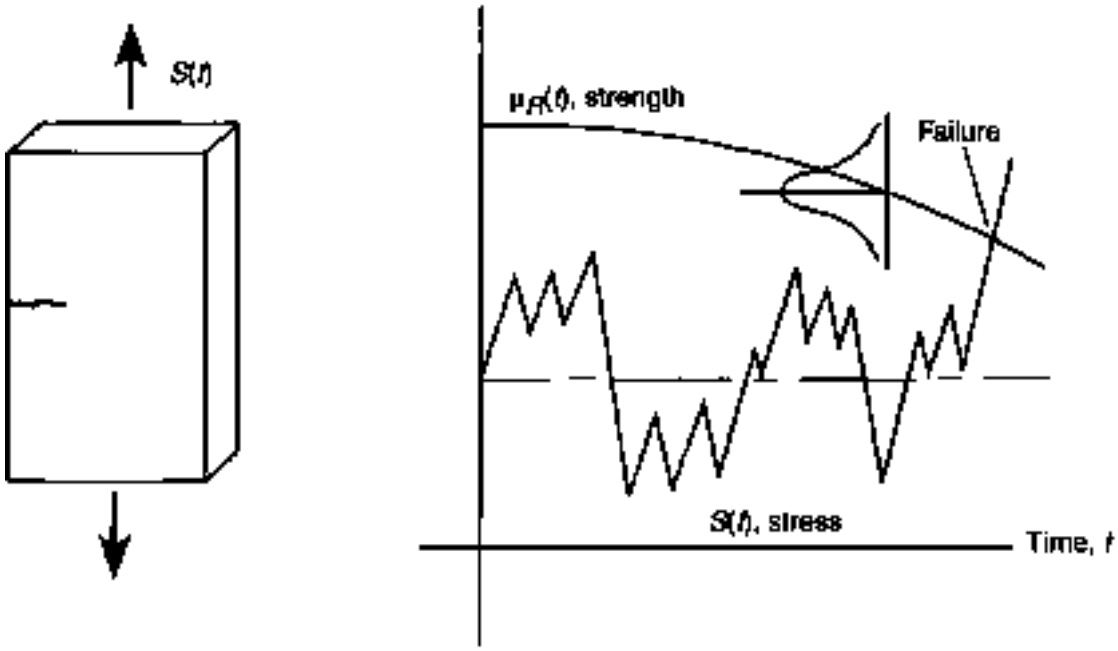


Fig. 19 Illustration of the first-passage problem

The long-term loading is, in general, a random process. Failure occurs the first time that stress exceeds strength, as shown. In random process theory, this is known as the *first-passage problem.*

The time to failure is T, which is a random variable because of the randomness in strength and the random stress process. Let T_s denote the intended service life of the component. The probability of failure is $P_f = P(T \leq T_s)$. That is, the component breaks before the end of its intended life.

In general, this is a very difficult probability problem with limited solutions. In practice, some form of simulation (Monte Carlo sampling) is typically necessary to solve this problem.

System Reliability

System reliability is more complex than component reliability for two reasons. First, the definition of failure for the system must be defined, and second, the evaluation of the probability of failure of the system must be computed. It is perhaps obvious but should be stated that multiple component probabilistic analyses must be performed during a system probabilistic analysis.

The definition of failure for the system may be a complex, time-consuming process. A complicated structural system, for example, a gas turbine engine, may have many different events and combinations of events that can contribute to failure of the system.

There are several formal methods for defining the system reliability; reliability block diagrams, fault trees, and event trees are most common (Ref 45). For example, a fault tree is a top-down method represented graphically (see Fig. 20) using a logical structure of events connected with AND and OR gates. AND gates connect events that must *both* occur to trigger a higher event. OR gates connect events for which failure in *any* event triggers a higher event.

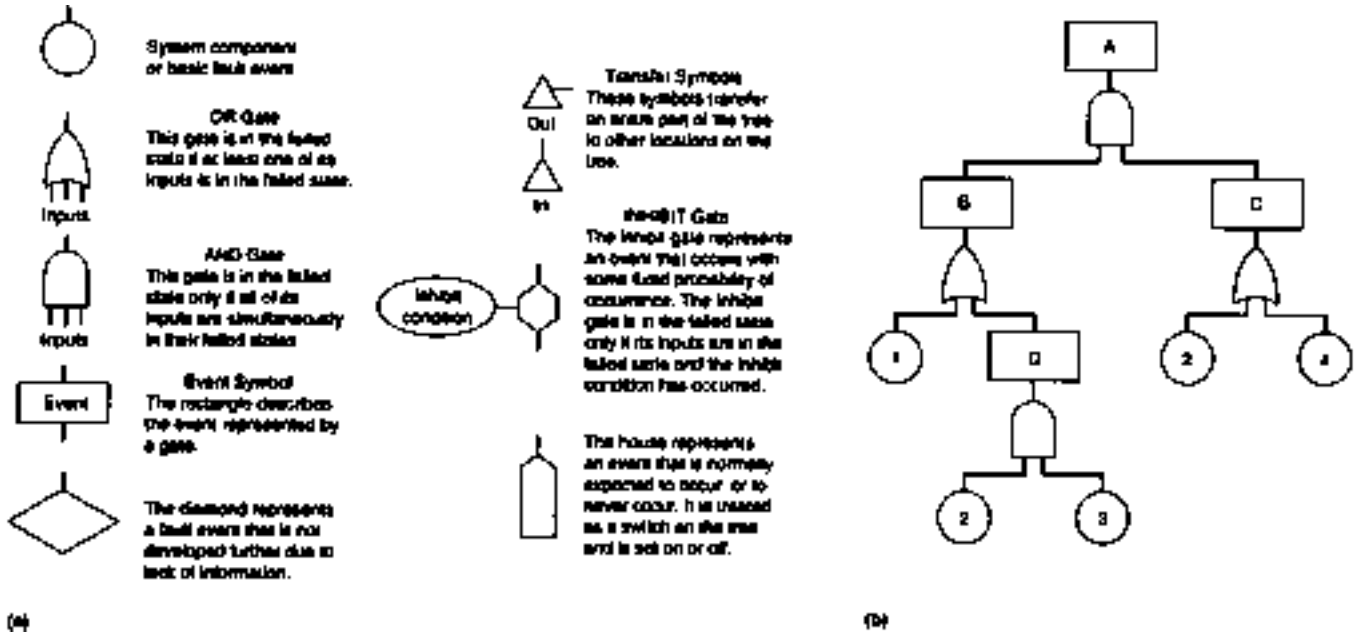


Fig. 20 Fault tree analysis. (a) Symbols. (b) Typical fault tree construction. Source: Ref 46

Structural systems are often composed of or contain “series” or “parallel” systems. A series system fails if any one of its components fails, for example, a chain. A series system arises when considering multiple failure modes of a structure; for example, the structure may fail due to yielding, fatigue, or corrosion. The structure is considered failed if *any* failure mode occurs. Series systems are modeled with an OR gate, as shown in Fig. 21.

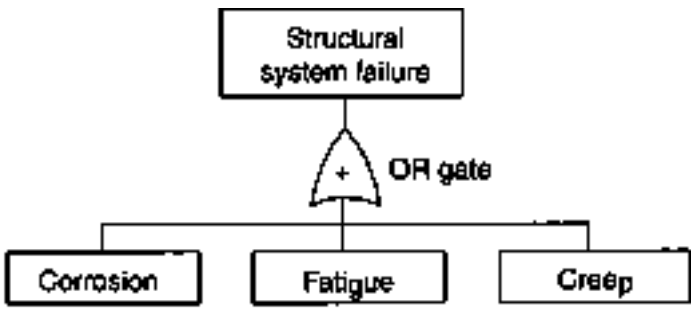


Fig. 21 Example of an OR gate used to model a series system

Parallel systems occur in structures with redundancy, and all components must fail, as shown in Fig. 22—failure of the structure occurs when *all* members have failed. Parallel systems are modeled with an AND gate, (Fig. 23). Note that failure of a structural member in a parallel system is often accompanied by load redistribution. In this case, the individual event probabilities should be updated, which may significantly complicate the analysis.

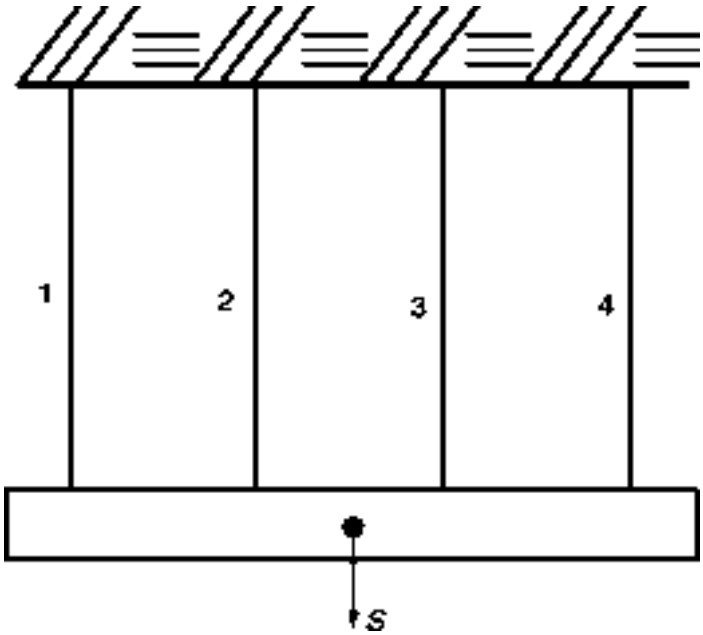


Fig. 22 Example of a parallel system

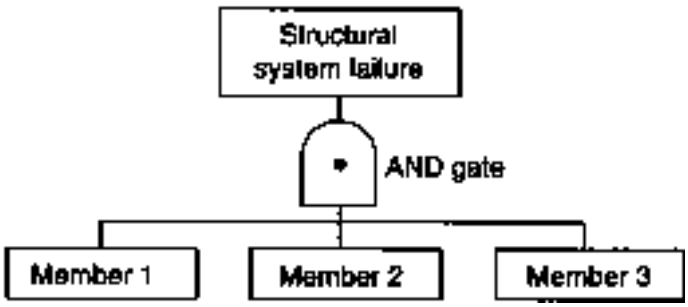


Fig. 23 Example of an AND gate used to model a parallel system

The evaluation of the system reliability given a fault tree, for example, typically uses Boolean algebra to combine probabilities of lower events to determine the probability of failure of the immediate higher event based on the AND and OR gates. However, structural system reliability often involves correlated component failure probabilities; for example, random variables common to multiple failure modes, such as loading, and standard Boolean logic-based methods may be significantly in error. As a result, bounding methods, which give bound on the probability of failure, are often used, as described subsequently.

Bounding Methods. The evaluation of the probability of failure for structural systems is significantly more complicated than for single components. As a result, bounding methods are often used to compute the upper and lower bounds of the probability of failure. Bounding methods are usually either first- or second-order based. First-order bounds are given subsequently for series and parallel systems and are fairly straightforward to apply (Ref 47). However, these bounds may be too wide for practical use. Second-order bounds are more accurate but more complicated and require additional calculations (Ref 48, 49).

Series Systems. If the failure modes are independent, then an upper bound can be derived as follows:

$$P_s = \prod_{i=1}^{k} P_{s_i}$$

where P_s is the probability of system survival, P_{s_i} is the probability of survival for each failure mode, and k is the number of failure modes.

Substituting $P_{f_i} = 1 - P_{s_i}$ provides the upper bound to the probability of failure of the system:

$$P_s = \prod_{i=1}^{k} \left(1 - P_{f_i}\right)$$

and

$$P_f = 1 - \prod_{i=1}^{k}\left(1 - P_{f_i}\right)$$

If the failure modes are fully correlated, then a lower bound is $P_f = \max[P_{f_i}]$.
Thus, the bounds on the probability of failure of the system are:

$$\max\left[P_{f_i}\right] \le P_f \le 1 - \prod_{i=1}^{k}\left(1 - P_{f_i}\right)$$

Parallel Systems. Bounds for parallel systems are:

$$\prod_{i=1}^{N} P_{f_i} \le P_f \le \min\left[P_{f_i}\right]$$

where the lower bound corresponds to independent failure modes and the upper bound corresponds to fully correlated failure modes.

Risk Analysis

The Concept of Risk. While there have been different uses for the term “risk,” a common application now is to define risk as the expected cost of failure:

$$\text{Risk} = \text{expected failure cost} = P_f C_f$$

where P_f is the probability of failure and C_f is the cost of failure.
This popular concept is often used in the decision-making process. The dangers are that one has to believe in one's P_f estimate in an actuarial sense, and all costs must be considered.
Costs that would be included in C_f are loss of the component or system, loss of revenue, possible injuries or fatalities, liability for damages, and losses associated with bad publicity resulting from the failure. To perform a risk analysis, it is necessary to put monetary value on all of these items.
Economic Value Analysis. In the development of a component, as an alternative to defining a target reliability, the design can be based on economic value analysis (EVA), that is, choose that design for which the total life-cycle cost is minimized. The simplest form of EVA is illustrated in Fig. 24.

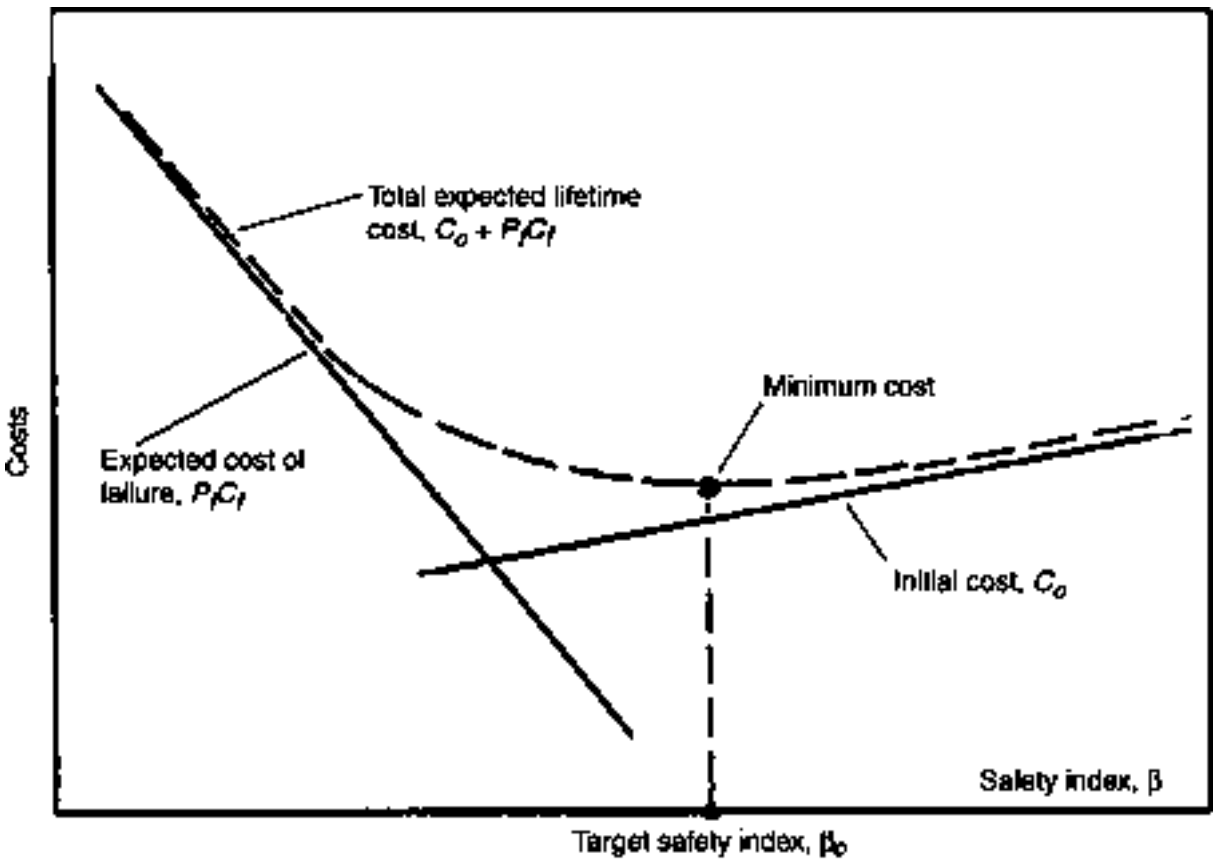


Fig. 24 Total expected life-cycle cost as a function of safety index

Target Reliabilities

Upon computing the safety index or probability of failure, one needs to ask what is acceptable, or safe. Specifically, one needs to define a target safety index, β_o, or maximum allowable probability of failure, p_o. Thus, an acceptable or safe design would satisfy $\beta > \beta_o$ and $P_f < p_o$. Note that β_o and p_o are related through the standard normal distribution function $P_f = \Phi(-\beta)$.

The value chosen depends on many factors but principally the consequences of failure and the presence and effectiveness of a maintenance program.

Table 6 presents some general guidelines for target reliabilities. Choices in specific cases are often decided by committees upon reviewing details of their physical situation.

Table 6 General guidelines for target reliabilities

Consequences of failure	Safety index, β_o	Probability of failure, P_f
Low (easy to fix)	2.0	2.3×10^{-2}
Moderate (general structural and mechanical components)	3.0	1.3×10^{-3}
Serious (loss of life and/or expensive repairs)	4.75	10^{-6}
Catastrophic loss of commercial aircraft	6.0	10^{-9}

References cited in this section

10. A. Harbitz, An Efficient Sampling Method for Probability of Failure Calculations, *Struct. Saf.*, Vol 3, 1986, p 109–115

11. C.G. Bucher, Adaptive Sampling—An Iterative Fast Monte Carlo Procedure, *Struct. Saf.*, 1988, p 119–126

12. P. Bjerager, On Computational Methods for Structural Reliability Analysis, *Struct. Saf.*, Vol 9, 1990, p 79–96

13. Y.-T. Wu. Computational Methods for Efficient Structural Reliability and Reliability Sensitivity Analysis, *AIAA J.*, Vol 32, 1994, p 1717–1723

44. Y.-T. Wu, M.P. Enright, H.R. Millwater, G.G. Chell, C.J. Kuhlman, and G.R. Leverant, "Probabilistic Methods for Design Assessment of Reliability with Inspection (DARWIN)," Proc. 41st Structures, Structural Dynamics, and Materials Conf., 3–6 April 2000 (Atlanta, GA), AIAA/ASME/ASCE/ASC

45. A.H.-S. Ang and W. Tang, *Probability Concepts in Engineering Planning and Design*, Vol II, John Wiley & Sons, New York, 1984

46. J.B. Fussell and D.P. Wagner, Fault Tree Analysis as a Part of Mechanical Systems Design, *Engineering Design*, Proceedings of the Mechanical Failures Prevention Group, 25th Meeting, NBS 487, National Bureau of Standards, 1976, p 289–308

47. C.A. Cornell, Bounds on the Reliability of Structural Systems, vol. 93, *J. Struct. Div. (ASCE)*, 1967

48. E.G. Kounias, Bounds for the Probability of a Union with Applications, *Ann. Math. Stat.*, Vol 39 (No. 6), 1968, p 2154–2158

49. O. Ditlevsen, Narrow Reliability Bounds for Structural Systems, *J. Struct. Mech.*, Vol 7 (No. 4), 1979, p 453–472

Harry R. Millwater, Jr., The University of Texas at San Antonio; Paul H. Wirsching, The University of Arizona

Case Histories: Examples of the Use of Probabilistic Analysis

Example 1: Probabilistic Damage Tolerance Analysis of Gas Turbine Rotors. In 1989, a United Airlines DC-10 in transit from Denver to Chicago experienced failure of the center engine. The titanium compressor disk burst and severed the hydraulics of the plane, making the plane difficult to control. Heroically, the pilot was able to crash land the plane at the Sioux City, IA airport, saving over 180 passengers, but the accident still resulted in 112 fatalities. A subsequent investigation revealed that the cause of the disk rupture (Fig. 25) was the presence of a large fatigue crack near the bore emanating from a hard alpha (HA) defect. An HA defect can result from occasional upsets during the vacuum melting process of the titanium. These nitrogen-rich alpha titanium anomalies are brittle and often have associated microcracks and microvoids (Fig. 26).

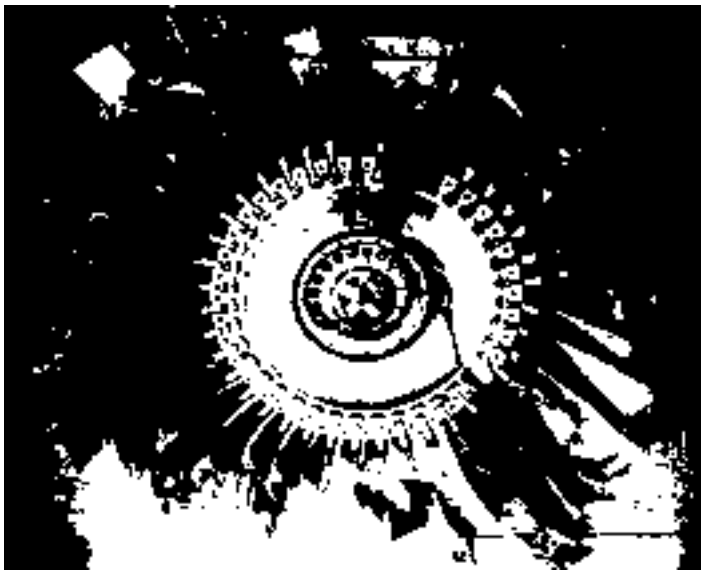


Fig. 25 Ruptured disk

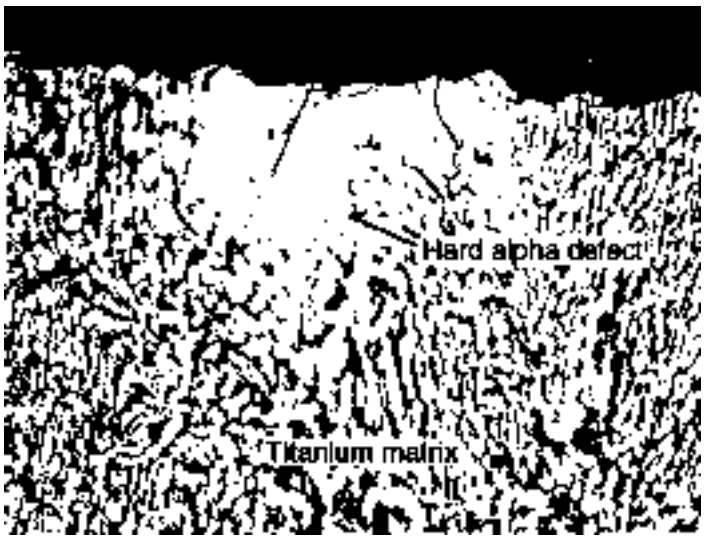


Fig. 26 Example of a hard alpha particle

In a report issued by the Federal Aviation Administration (FAA) after the accident, it was recommended that a probabilistic damage tolerance approach be implemented to explicitly address HA anomalies, with the objective of enhancing, not replacing, conventional rotor life management methodology, which is founded on the assumption of nominal material and manufacturing conditions. The approach adopted and summarized here is based on probabilistic fracture mechanics. Probabilistic methods are essential for proper treatment of this problem, due to the extremely rare occurrence of HA particles in rotor-grade titanium.

The Aerospace Industries Association Rotor Integrity Subcommittee, an industry group composed of commercial gas turbine original equipment manufacturers, has been working on this issue in consort with the FAA by developing methodology and recommendations for a probabilistic treatment of HA in titanium rotors (Ref 50, 51).

The necessary ingredients to quantify the probability of fracture of a titanium disk are briefly summarized here. More detail may be found in Ref 25, 26 and 52. A computer code, DARWIN (Design Assessment of Reliability With INspection), has been developed under FAA funding that is specifically tailored to perform a probabilistic analysis of titanium rotors subject to HA (Ref 53). DARWIN integrates finite-element stress analysis, fracture mechanics analysis, nondestructive inspection simulation, and probabilistic analysis to assess the probability of rotor fracture. It computes the probability of fracture as a function of flight cycles, considering random defect occurrence and location, random inspection schedules, and several other random variables using either Monte Carlo sampling or importance sampling methods. A summary of the elements of DARWIN is shown in Fig. 27.

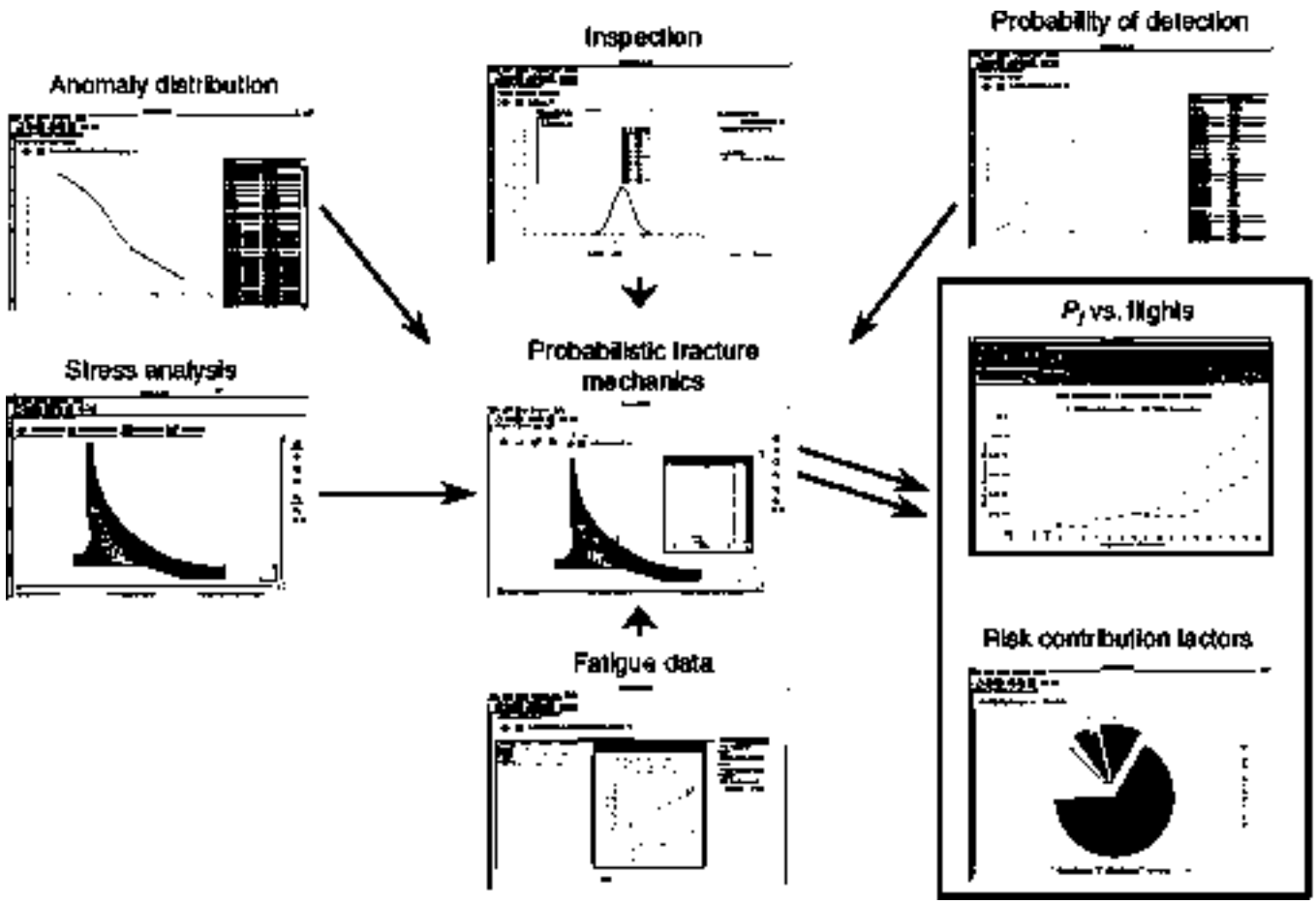


Fig. 27 Overview of DARWIN

A critical ingredient for performing a realistic probability of fracture calculation is the anomaly distribution. The anomaly distribution defines the likelihood of having a HA defect in a disk and its size. An example anomaly distribution is shown in Fig. 28. The figure provides the number of HA defects of a given size expected in one million pounds of titanium. For example, from the figure, the total number of defects greater than 100 square mils per one million pounds of titanium is approximately three.

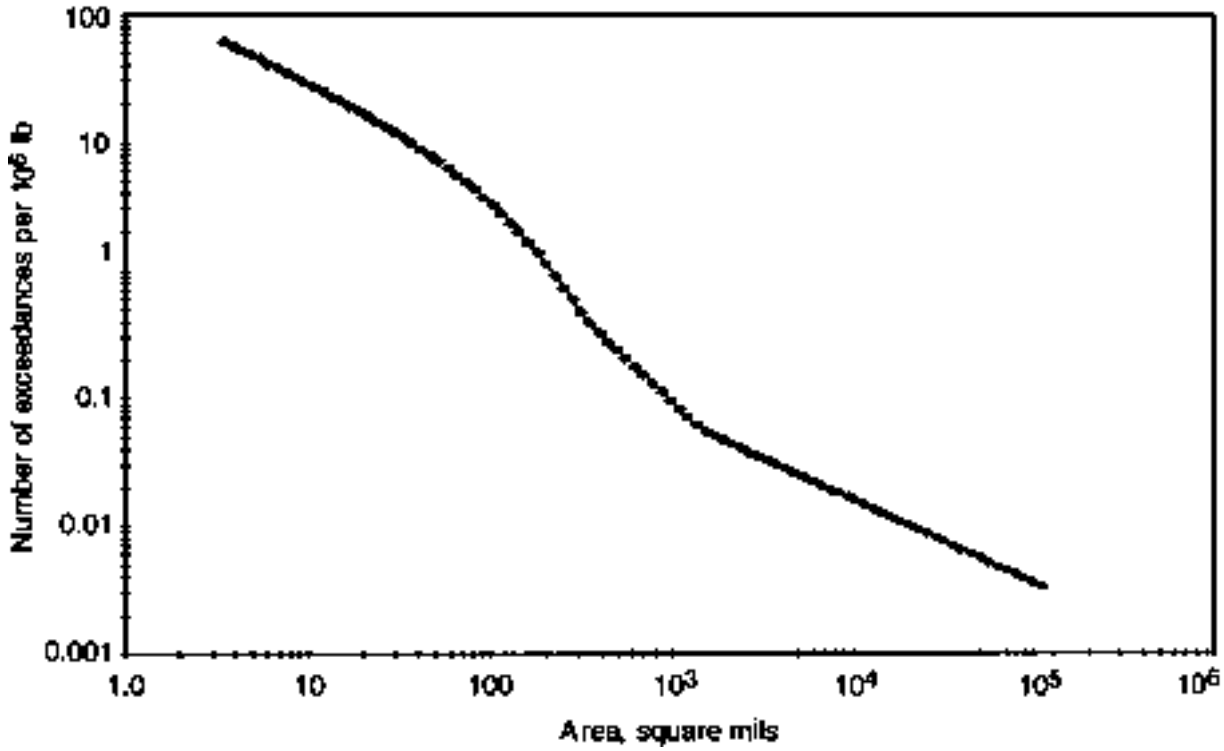


Fig. 28 Example of an exceedance distribution of hard alpha

Defect distributions for different melt and inspection conditions were developed by the industry through a combination of data analysis of titanium finds, modeling, and calibration with field experience (Ref 50).

In 2001, the FAA issued an advisory circular, AC 33.14-1, stating that all new titanium rotors should be designed with a probability of fracture per flight of 10^{-9} or less per disk (design target risk) and the cumulative probability of fracture of all titanium disks on an engine be 5×10^{-9} or less (Ref 51). These probability levels are such that the incidence rate due to HA is expected to be reduced by a factor of 3 or more compared to present levels.

The probability of fracture per flight as a function of flights is shown in Fig. 29 both with and without inspection (an inspection and removal of defective disks is simulated at 10,000 cycles). This example indicates that, for 20,000 flight cycles, this design would not pass the design target risk limit of 10^{-9} without inspection, but that the design does pass with inspection.

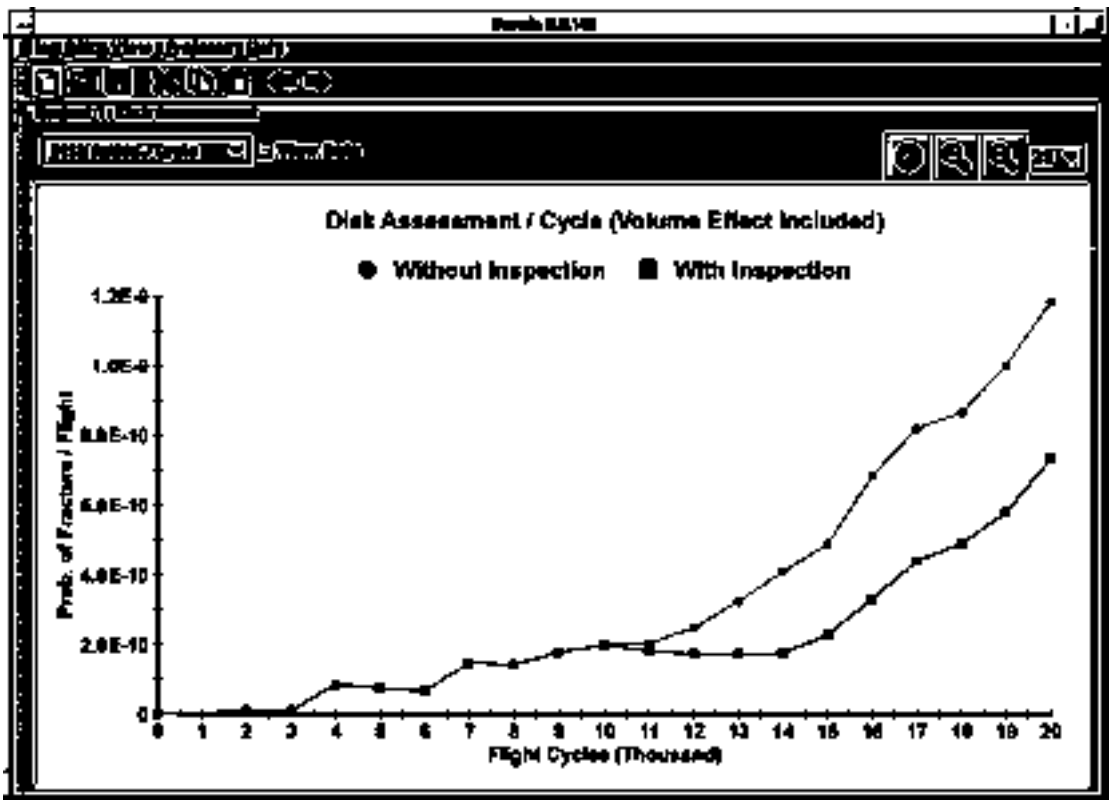


Fig. 29 Probability of fracture per flight

The relative contributions of the probability of fracture of different regions of the disk are defined in the risk contribution factors shown in Fig. 30. These factors indicate the regions of the disk that contribute the most to the total probability of failure.

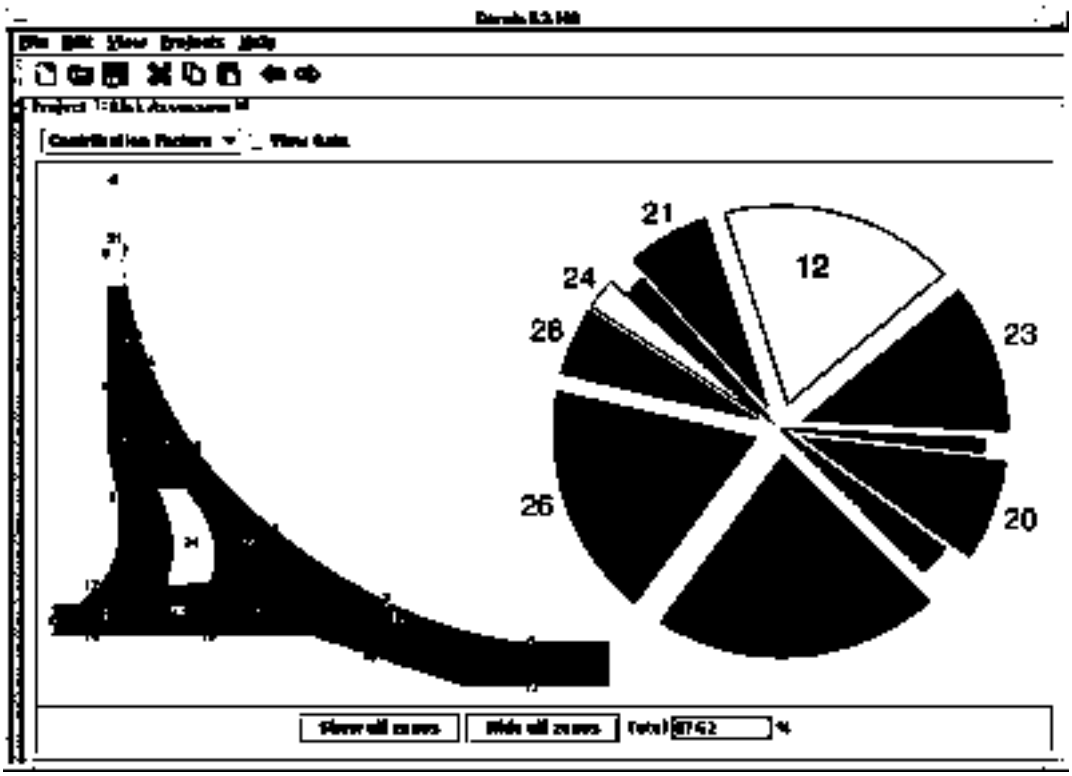


Fig. 30 Risk contribution factors

Example 2: Development of Recommended Practice API 579. Historically, the American Society of Mechanical Engineers (ASME) and American Petroleum Institute (API) design codes for new pressure vessels do not address the fact that equipment degrades while in service and that structural deficiencies due to degradation from the original fabrication may be found during subsequent investigations. Fitness-For-Service (FFS) API recommended practice 579 (Ref 54) has been recently developed to perform FFS testing to demonstrate the structural integrity of an in-service component containing a crack or flaw. Results of a FFS assessment can be used to make run-repair-replace decisions to help ensure that pressurized equipment containing detected flaws can continue to operate safely. The FFS algorithm is deterministic, yet the major variables—stress, fracture toughness, and crack size—possess significant uncertainty. Conservative estimates of these variables (characteristic values in the safe side of the distribution) can be made to ensure an acceptable safety margin. The problem is that such an approach is known to produce overly conservative results.

The alternative is a reliability approach in which design factors subject to significant uncertainty are treated as random variables. A maximum allowable probability of failure is specified, and the partial safety factors (PSFs) are derived using FORM to ensure this level of reliability. Partial safety factors are applied to nominal values of the key variables in the FFS algorithm, thus providing the appearance of a conventional safety check expression.

In this section, the process is summarized, and a simple example is presented. For simplicity, the geometry factor is assumed to be constant, and secondary stresses are ignored. More detailed descriptions are provided in Ref 55 and 56.

Critical Variables. The process starts with a definition of the random variables. Development of the PSFs requires only that the coefficients of variation of each design factor are known. Following are the assumptions that were made in the first effort.

For stress (S), three levels of uncertainty are defined (C_X = COV of the variable X):

- C_s = 0.30 approximate loads and stress analysis using simplified formulas
- C_s = 0.20 more refined loads and stress analysis
- C_s = 0.10 refined loads and stress analysis; probably using FEA
- Distribution family: EVD (extreme value distribution)

To determine crack depth (a), an inspector finds a crack that is then measured. Given a measurement, there is uncertainty in the actual depth of the crack. Studies have been performed on inspector performance; the distribution has been constructed from the results. Characteristically small and large crack depths were defined. They are:

- Small: $\mu_a = 0.10$ $C_a = 0.30$
- Large: $\mu_a = 0.30$ $C_a = 0.20$
- Distribution family: lognormal

For normalized fracture toughness $\hat{K} = K/\mu_K$ a three-parameter Weibull was fit to characteristic data. The CDF of this three-parameter model is:

$$F_{\hat{k}}(\hat{k}) = 1 - \exp\left[\left(\frac{\hat{k}-\delta}{\beta}\right)^{\alpha}\right] \quad \hat{k} > \delta$$

where $\hat{K}$ is normalized fracture toughness; the mean is 1.0, and the standard deviation is the coefficient of variation. The shape parameter $\alpha = 4.0$; the scale parameter $\beta = 0.8826$; and the location parameter $\delta = 0.20$.
The next step is to determine target reliability. Three cases are considered, depending on the consequences of failure:

$p_o = 10^{-6}$	$\beta_o = 4.75$	Serious consequences
$p_o = 10^{-3}$	$\beta_o = 3.09$	Moderate consequences
$p_o = 0.0228$	$\beta_o = 2.0$	Low consequences

The geometry factor is $Y(a) = 1.0$.
Limit State. First define the relevant variables. The load ratio is:

$$L_r = \frac{S}{R_y}$$

where S = primary stress, and R_y = yield strength (considered here to be constant).
Define the toughness ratio as:

$$K_r = \frac{K_p}{K}$$

where K_p = stress-intensity factor of primary stress, and K = fracture toughness:

$$K_p = Y(a)\sqrt{\pi a}S$$

$Y(a)$ is the geometry factor.
The limit state is shown in the failure assessment diagram (FAD) shown in Fig. 31.

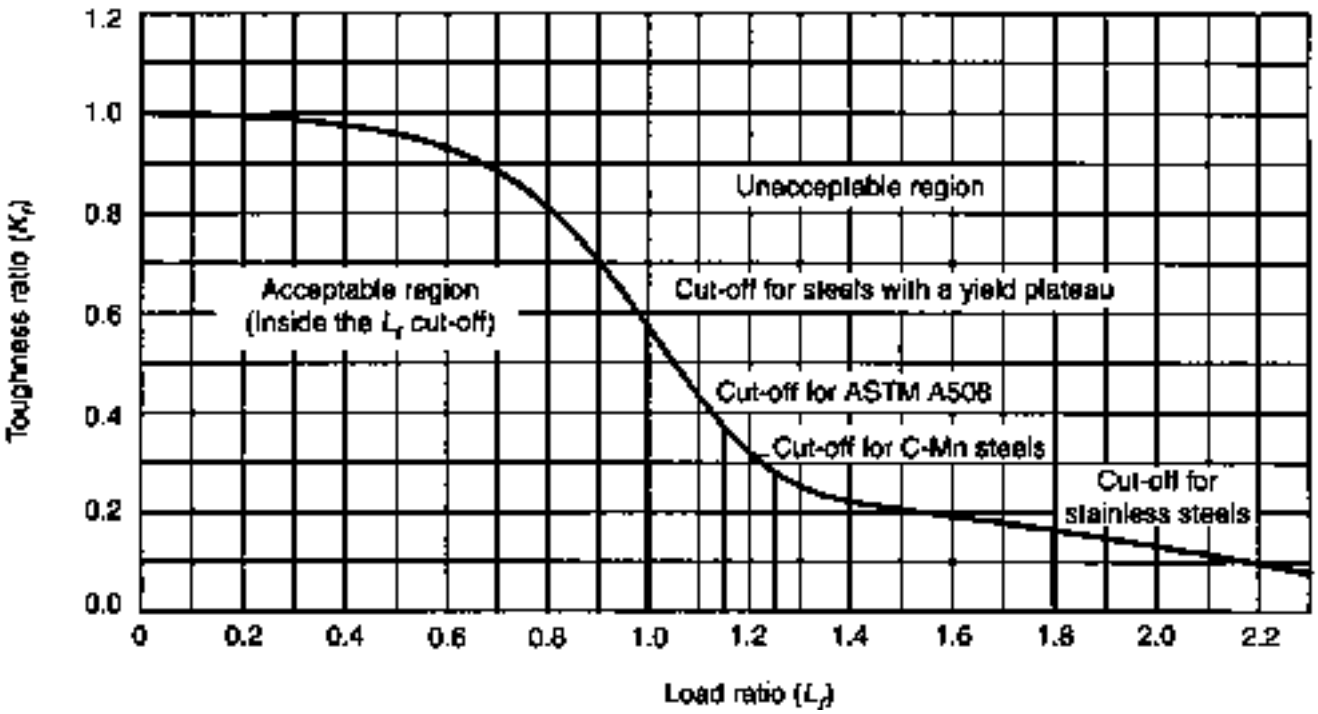


Fig. 31 Failure assessment diagram. Source: Ref 54

For purposes of the PSF analysis, the limit state can be approximated as:

$$K_r' = \left(1 - S_r^{2.5}\right)^{0.20}$$

K_r' is strength relative to K_r. The limit state function is then:

$$g = K_r' - K_r$$

Derivation of Partial Safety Factors. There is a PSF associated with each of the three random variables (K, a, and S). A FORM process (see the section "First/Second-Order Methods" is this article) is run to compute the MPP that corresponds to the safety index, β_o, specified. This is an iterative process, because FORM algorithms compute β_o directly. Program PSF has been developed by the Materials Properties Council to automate the process. The PSFs are defined as:

$$\gamma_S = \frac{S^*}{S_n}, \quad \gamma_a = \frac{a^*}{a_n}, \quad \text{and} \quad \gamma_K = \frac{K^*}{K_n}$$

where * indicates the MPP value in the basic coordinates, and the subscript n indicates nominal value. The nominal value is often a point in the safe side of the distribution of the variable. In the subsequent example, it is chosen to be the mean value for simplicity.

Safety Check Expression. To perform a graphical safety check, compute:

$$K_r = \frac{Y(a)(\gamma_S S_n)\sqrt{\pi(\gamma_a a_n)}}{\gamma_K K_n} \qquad L_r = \frac{\gamma_S S_n}{R_y}$$

where a_n is the observed crack depth. Then refer to the FAD shown previously in Fig. 31. If the point lies inside, then the component is safe.

For example, during a plant shutdown for maintenance, the FWM Power Company discovers a crack of measured depth of a = 5 mm (0.20 in.) in a high-pressure pipe. A FFS safety check is required. They assume that their estimate of the peak operating stress of 205 MPa (30 ksi) is reasonably accurate (C_s = 0.10). They also estimate that the mean fracture toughness is $165\text{MPa}\sqrt{m}$ ($150\text{ksi}\sqrt{\text{in.}}$) and that the yield strength is R_y = 415 MPa (60 ksi). The consequences of failure of this pipe are considered to be moderate ($p_o = 10^{-3}$).

Tables are being constructed for API 579 using a computer program (Program PSF) developed by the Materials Properties Council. Using the data of this case study plus the assumed value of C_a = 0.20 and the computed value of $B = \mu_K/R_y$ = 2.5, Program PSF produces the following PSFs (letting nominal values = means): γ_s = 1.46, γ_K = 0.78, and γ_a = 1.06.

Compute K_r and L_r:

$$K_r = \frac{(1)(1.46)(30)\sqrt{\pi(1.06)(0.2)}}{(0.78)(150)} = 0.31$$

$$L_r = \frac{(1.46)(30)}{60} = 0.73$$

Referring to the FAD of Fig. 31, the point lies comfortably in the safe region.

A complete analysis for API 579 should include considerations of secondary stresses, the appropriate geometry factor function, the plasticity interaction factor, and a function $L(a)$ to account for the reduced section over which the tensile stress acts. In addition, the PSFs need to be established as a function of $B = \mu_K/R_y$.

References cited in this section

25. G.R. Leverant, D.L. Littlefield, R.C. McClung, H.R. Millwater, and Y.-T. Wu, "A Probabilistic Approach to Aircraft Turbine Rotor Material Design," Paper 97-GT-22, IGTI Congress, 1997

26. G.R. Leverant, R.C. McClung, H.R. Millwater, and M.P. Enright, "A New Tool for Design and Certification of Aircraft Turbine Rotors," Paper GT-2002-30303, ASME Turbo Expo 2002, 3–6 June 2002 (Amsterdam, The Netherlands), American Society of Mechanical Engineers

50. Aerospace Industries Association (AIA) Rotor Integrity Subcommittee Subteam, The Development of Anomaly Distributions for Aircraft Engine Titanium Disk Alloys, 38th American Institute of Aeronautics and Astronantics/American Society of Mechanical Engineers/American Society of Civil Engineers/AHS/ASC SDM Conf. 1997, p 2543–2553

51. "Damage Tolerance for High Energy Turbine Engine Rotors," Advisory Circular 33.14-1, Federal Administration Association, Jan 2001

52. Southwest Research Institute, AlliedSignal, General Electric, Pratt & Whitney, Rolls-Royce Allison, Scientific Forming Technologies, Federal Aviation Administration Report DOT/FAA/AR-00/64, Grant 95-G-041, May 2000

53. H.R. Millwater, S.H. Fitch, Y.-T. Wu, D.S. Riha, M.P. Enright, G.R. Leverant, R.C. McClung, C.J. Kuhlman, G.G. Chell, and Y. Lee, "A Probabilistically-Based Damage Tolerance Analysis Computer Program for Hard Alpha Anomalies in Titanium Rotors," Paper 2000-GT-0421, ASME Turbo Expo, IGTI, 2000

54. "API 579: Fitness-for-Service," API recommended practice 579, American Petroleum Institute, Washington, D.C., 2000

55. D.A. Osage, K.S. Shipley, P.H. Wirsching, and A.E. Mansour, "Application of Partial Safety Factors for Pressure Containing Equipment," Proc. ASME Pressure Vessel and Piping Conf., 23–27 July 2000 (Seattle, WA), American Society of Mechanical Engineers

56. P.H. Wirsching, H.P. Nguyen, D.A. Osage, and A.E. Mansour, "Probability-Based Fitness for Service for Pressure Vessels and Piping," Proc. Eighth ASCE Specialty Conf. Probabilistic Mechanics and Structural Reliability, 24–26 July 2000 (Notre Dame University), American Society of Civil Engineers

Probabilistic Analysis Software

Several software programs are specifically designed to perform probabilistic analysis. Some programs are general purpose, in that they contain a variety of probabilistic algorithms; for example, CALREL, NESSUS, PROBAN, PROFES, STRUREL, and UNIPASS include links to external analysis programs such as finite-element programs and allow the user to solve an arbitrary limit state. These programs may also have system reliability capabilities. Other programs are focused on a particular application, such as probabilistic fatigue, for example, DARWIN and PROF. Other programs allow the user to solve an arbitrary limit state but are focused on Monte Carlo sampling, for example, CRYSTAL BALL and @RISK.

There are specific programs for analyzing system reliability problems. These programs typically involve a fault tree, event tree method, or reliability block diagram for organizing the definition of system failure. Programs such as QRAS, SAPPHIRE, RAPTOR, MEADEP and BLOCKSIM fall in this category.

Many mathematics/engineering programmable programs contain necessary routines that can be used to perform probabilistic analysis, for example, IMSL, MATHCAD, MATHEMATICA, MATLAB, IDL, PROFIT, and others. Thus, one can program a probabilistic analysis algorithm within one of the environments. Also, many statistical packages have useful statistical functions, such as response surface evaluation. Finally, a spreadsheet program can often be used for Monte Carlo sampling.

Source code that can perform many of the necessary functions for computing the probability of failure is also available. For example, the GNU Scientific Library contains C-source code for generating random variables from a large list of PDFs, statistical routines, and several forms of Monte Carlo sampling, including importance sampling. Several Web sites contain many FORTRAN routines for random variable generation and statistical functions. The textbook *Numerical Recipes in FORTRAN* (Ref 41) has several chapters related to random number generation and statistics.

Reference cited in this section

41. W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in FORTRAN: The Art of Scientific Computing,* 2nd ed., Vol 1, University of Cambridge, 1992

References

1. A.M. Freundenthal, Safety of Structures, *Trans. ASCE*, Vol 112, 1947, p 125–180

2. A.M. Freudenthal, J.M. Garrelts, and M. Shinozuka, The Analysis of Structural Safety, *J. Struct. Div. (ASCE)*, Vol 92 (No. ST2), Feb 1966

3. C.A. Cornell, A Probability-Based Structural Code, *J. Am. Concr. Inst.*, Vol 66 (No. 12), 1969

4. A.M. Hasofer and N.C. Lind, Exact and Invariant Second Moment Code Format, *J. Eng. Mech., (ASCE)*, Vol 100 (No. EM1), 1974, p 111–121

5. R. Rackwitz and B. Fiessler, Structural Reliability under Combined Load Sequences, *J. Comput. Struct.*, Vol 9, 1978, p 489–494

6. L. Tvedt, "Two Second Order Approximations to the Failure Probability," Veritas Report, RDIV/20-004-83, Det Norske Veritas, Oslo, Norway, 1983

7. K. Breitung, Asymptotic Approximations for Multinormal Integrals, *J. Eng. Mech. (ASCE)*, Vol 110, 1984, p 357–366

8. Y.-T. Wu and P.H. Wirsching, New Algorithm for Structural Reliability Estimation, *J. Eng. Mech.*, Vol 113 (No. 9), 1987, p 1319–1336

9. Y.-T. Wu, H.R. Millwater, and T.A. Cruse, Advanced Probabilistic Structural Analysis Method for Implicit Performance Functions, *AIAA J.*, Vol 28 (No. 9), 1990

10. A. Harbitz, An Efficient Sampling Method for Probability of Failure Calculations, *Struct. Saf.*, Vol 3, 1986, p 109–115

11. C.G. Bucher, Adaptive Sampling—An Iterative Fast Monte Carlo Procedure, *Struct. Saf.*, 1988, p 119–126

12. P. Bjerager, On Computational Methods for Structural Reliability Analysis, *Struct. Saf.*, Vol 9, 1990, p 79–96

13. Y.-T. Wu. Computational Methods for Efficient Structural Reliability and Reliability Sensitivity Analysis, *AIAA J.*, Vol 32, 1994, p 1717–1723

14. T.V. Galambos and M.K. Ravindra, Properties of Steel for Use in LRFD, *J. Struct. Eng. (ASCE)*, Vol 104 (No. 9), 1978, p 1459–1468

15. F. Moses, "Implementation of a Reliability-Based API RP2A Format," final report, API PRAC 83-22, American Petroleum Institute, 1985

16. F. Moses, "Development of Preliminary Load and Resistance Design Document for Fixed Offshore Platforms," final report, API-PRAC 95-22, American Petroleum Institute, 1986

17. C. Sundararajan, Ed., *Probabilistic Structural Mechanics Handbook, Theory and Industrial Applications*, Chapman & Hall, 1995

18. L. Geschwindner, R. Disque, and R. Bjorhovde, *Load and Resistance Factor Design of Steel Structures*, Prentice Hall, 1994

19. "Probabilistic Structural Analysis Methods for Select Space Propulsion System Components (PSAM)," 1st–5th annual reports, NASA Contract NAS3-24389, Southwest Research Institute, 1985–1990

20. "Probabilistic Structural Analysis Methods for Select Space Propulsion System Components (PSAM), final report, NASA Contract NAS3-24389, Southwest Research Institute, 1995

21. K.R. Rajagopal, J. Keremes, H. Ho, and G. Orient, "A Probabilistic Approach to the Evaluation of Fatigue Damage in a Space Propulsion System Injector Element," 33rd Structures, Structural Dynamics, and Materials Conf., April 1992 (Dallas, TX), AIAA/ASME/ASCE/AHS/ASC

22. J.E. Shigley and C.R. Mischke, *Mechanical Engineering Design,* 6th ed., McGraw-Hill, 2000

23. S. Rao, *Reliability-Based Design,* McGraw-Hill, 1992

24. A.P. Berens, P.W. Hovey, and D.A. Skinn, "Risk Analysis for Aging Aircraft Fleets, Vol 1—Analysis," WL-TR-91-3066, Flight Dynamics Directorate, Wright-Patterson Air Force Base, OH 1991

25. G.R. Leverant, D.L. Littlefield, R.C. McClung, H.R. Millwater, and Y.-T. Wu, "A Probabilistic Approach to Aircraft Turbine Rotor Material Design," Paper 97-GT-22, IGTI Congress, 1997

26. G.R. Leverant, R.C. McClung, H.R. Millwater, and M.P. Enright, "A New Tool for Design and Certification of Aircraft Turbine Rotors," Paper GT-2002-30303, ASME Turbo Expo 2002, 3–6 June 2002 (Amsterdam, The Netherlands), American Society of Mechanical Engineers

27. M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions,* 2nd ed., John Wiley & Sons, 1993

28. S.M. Shelby, Ed., *Standard Mathematical Tables,* 23rd ed., CRC Press, 1973

29. J. Benjamin and C. Cornell, *Probability, Statistics, and Decisions for Civil Engineers,* McGraw-Hill, 1970

30. A.H.-S. Ang and W. Tang, *Probability Concepts in Engineering Planning and Design,* Vol I, John Wiley & Sons, New York, 1975

31. M. Stephens, "Goodness of Fit with Special Reference to Tests for Exponentiality," Technical Report 262, Department of Statistics, Stanford University, Stanford, CA, 1977

32. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* Applied Mathematics Series, Vol 55, National Bureau of Standards, 1964

33. R.E. Melchers, *Structural Reliability Analysis and Prediction,* Ellis Horwood Limited, U.K., 1987

34. L. Tvedt, Distribution of Quadratic Forms in Normal Space with Application to Structural Reliability, *J. Eng. Mech.* (*ASCE*), Vol 116, 1990, p 1183–1997

35. G. Box and N. Draper, *Empirical Model-Building and Response Surfaces,* John Wiley & Sons, Inc., 1987

36. R. Gunst and R. Mason, "How to Construct Fractional Factorial Experiments," American Society for Quality Control, Statistics Division, ASQC Quality Press, 1991

37. R. Mason, R. Gunst, and J. Hess, *Statistical Design and Analysis of Experiments,* John Wiley & Sons, Inc., 1989

38. D. Montgomery, *Design and Analysis of Experiments,* 2nd ed., John Wiley & Sons, Inc., 1984

39. G. Box and D. Cox, An Analysis of Transformations, *J. R. Stat. Soc.*, 1964, p 211–243, discussion. p 244–252

40. D. Knuth, *The Art of Computer Programming,* Vol II, Addison Wesley Longman, Inc., 1997

41. W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in FORTRAN: The Art of Scientific Computing,* 2nd ed., Vol 1, University of Cambridge, 1992

42. B. Shipley, *Stochastic Simulation*, John Wiley & Sons, 1987

43. M. Shooman, *Probabilistic Reliability: An Engineering Approach*, McGraw-Hill Companies, 1968

44. Y.-T. Wu, M.P. Enright, H.R. Millwater, G.G. Chell, C.J. Kuhlman, and G.R. Leverant, "Probabilistic Methods for Design Assessment of Reliability with Inspection (DARWIN)," Proc. 41st Structures, Structural Dynamics, and Materials Conf., 3–6 April 2000 (Atlanta, GA), AIAA/ASME/ASCE/ASC

45. A.H.-S. Ang and W. Tang, *Probability Concepts in Engineering Planning and Design*, Vol II, John Wiley & Sons, New York, 1984

46. J.B. Fussell and D.P. Wagner, Fault Tree Analysis as a Part of Mechanical Systems Design, *Engineering Design*, Proceedings of the Mechanical Failures Prevention Group, 25th Meeting, NBS 487, National Bureau of Standards, 1976, p 289–308

47. C.A. Cornell, Bounds on the Reliability of Structural Systems, vol. 93, *J. Struct. Div. (ASCE)*, 1967

48. E.G. Kounias, Bounds for the Probability of a Union with Applications, *Ann. Math. Stat.*, Vol 39 (No. 6), 1968, p 2154–2158

49. O. Ditlevsen, Narrow Reliability Bounds for Structural Systems, *J. Struct. Mech.*, Vol 7 (No. 4), 1979, p 453–472

50. Aerospace Industries Association (AIA) Rotor Integrity Subcommittee Subteam, The Development of Anomaly Distributions for Aircraft Engine Titanium Disk Alloys, 38th American Institute of Aeronautics and Astronantics/American Society of Mechanical Engineers/American Society of Civil Engineers/AHS/ASC SDM Conf. 1997, p 2543–2553

51. "Damage Tolerance for High Energy Turbine Engine Rotors," Advisory Circular 33.14-1, Federal Administration Association, Jan 2001

52. Southwest Research Institute, AlliedSignal, General Electric, Pratt & Whitney, Rolls-Royce Allison, Scientific Forming Technologies, Federal Aviation Administration Report DOT/FAA/AR-00/64, Grant 95-G-041, May 2000

53. H.R. Millwater, S.H. Fitch, Y.-T. Wu, D.S. Riha, M.P. Enright, G.R. Leverant, R.C. McClung, C.J. Kuhlman, G.G. Chell, and Y. Lee, "A Probabilistically-Based Damage Tolerance Analysis Computer Program for Hard Alpha Anomalies in Titanium Rotors," Paper 2000-GT-0421, ASME Turbo Expo, IGTI, 2000

54. "API 579: Fitness-for-Service," API recommended practice 579, American Petroleum Institute, Washington, D.C., 2000

55. D.A. Osage, K.S. Shipley, P.H. Wirsching, and A.E. Mansour, "Application of Partial Safety Factors for Pressure Containing Equipment," Proc. ASME Pressure Vessel and Piping Conf., 23–27 July 2000 (Seattle, WA), American Society of Mechanical Engineers

56. P.H. Wirsching, H.P. Nguyen, D.A. Osage, and A.E. Mansour, "Probability-Based Fitness for Service for Pressure Vessels and Piping," Proc. Eighth ASCE Specialty Conf. Probabilistic Mechanics and Structural Reliability, 24–26 July 2000 (Notre Dame University), American Society of Civil Engineers

Nondestructive Evaluation and Life Assessment

R.B. Thompson, Center for Nondestructive Evaluation, Iowa State University

Introduction

NONDESTRUCTIVE TESTING (NDT) and nondestructive evaluation (NDE) involve the use of noninvasive measurement techniques to gain information about defects and various properties of materials, components, and structures—information that is needed to determine their ability to perform their intended function and prevent failure. As defined by ASTM (Ref 1), "NDT is the development and application of technical methods to examine materials or components in ways that do not impair future usefulness and serviceability in order to detect, locate, measure and evaluate discontinuities, defects and other imperfections; assess integrity, properties and composition; and measure geometrical characteristics." Although the words "nondestructive inspection," "testing," and "evaluation" are occasionally used somewhat synonymously, the word "evaluation" has become increasingly associated with those methods for which quantitative interpretation of the test results in terms of characteristics of the material or defects is possible. Nondestructive evaluation is used in that context in this section. Attention is restricted to structural components, but applications exist in a wide range of other industries such as electronic, agricultural, food processing, biomedical, and so forth.

Quantitative interpretation is an important issue, because at the completion of every test, an accept/reject decision must be made regarding the suitability of the component for future service. This has led to some important semantic issues, and the words *defect, flaw, discontinuity,* and *imperfection* have distinct meanings in the NDE literature. Again following ASTM (Ref 1), definitions used are:

discontinuity.

A lack of continuity or cohesion; an intentional or unintentional interruption in the physical structure of configuration of a material or component

imperfection.

A departure of a quality characteristic from its intended condition

flaw.

An imperfection or discontinuity that may be detectable by nondestructive testing and is not necessarily rejectable

defect.

One or more flaws whose aggregate size, shape, orientation, location, or properties do not meet specified acceptance criteria and are rejectable

The development of nondestructive testing and evaluation techniques has occurred primarily in the last 100 years (Ref 2, 3), with progress often being driven by technological developments in other fields. For example, the development of pulsed electronic circuitry in the 1940s created the foundation for many of the pulse-echo ultrasonic instruments in use today. Advances in instrumentation have provided the opportunity to gain more information about flaws and material properties, for example, by the formation of higher resolution images. These advances in measurement capability have been paralleled by changes in practices for the design of structures. Thinking has evolved from the notion that no flaws can be tolerated in structures and equipment to the recognition that some level of material imperfection will generally be present, but that this can be tolerated if it can be established that the severity of the condition will not lead to failure over some specified period of service. The job of testing thus becomes more quantitative; it is not expected that all imperfections in a material will be found. Instead, it is desired that all imperfections above a certain size (selected in the context of the material), the loads that it is experiencing, and the period of time before the next opportunity for inspection will be found. In the terminology of ASTM, it is only those "flaws" that are considered to be "defects." Nondestructive evaluation is used throughout the lifecycle, and includes the inspection of new, aging, and in-service structures and equipment. With the increasing performance demands imposed by the highly competitive economic climate of today, it is often necessary to extend the service life or expectancy. Thus, it is critical to understand how safe and reliable that equipment or structure might be over this additional time of service.

References cited in this section

1. "Standard Terminology for Nondestructive Examinations, Metals Test Methods and Analytical Procedures," E 1316–00a, *Annual Book of ASTM Standards* Vol 3.03, ASTM, 2001, p 645–683

2. *Nondestructive Testing Handbook,* 2nd ed., P. McIntire, Ed., American Society for Nondestructive Testing (ASNT), Columbus, OH, 10 volumes published in the period 1982–1996

3. L. Cartz, *Nondestructive Testing,* ASM International, 1995

Common Measurement Techniques

Many measurement techniques are employed in NDE. Those most widely used are visual, liquid penetrant, magnetic particle, eddy current, ultrasonic, and radiographic testing. Table 1 (Ref 4) provides a brief definition of these methods (other than visual) and their uses and limitations. Figure 1 shows photographs of some actual inspection situations.

Table 1 Common nondestructive evaluation methods

Method	Description	Uses	Limitations
Liquid penetrant	Penetrant solution containing visible dye or fluorescent particles is drawn into surface-breaking cracks and voids by capillary action. A developer is applied to draw the penetrant out of the flaw, giving a visible indication.	Used on metal, glass, and ceramics to locate *surface-connected* discontinuities. Relatively easy to use and does not require elaborate equipment	Cannot detect discontinuities that do not break the surface of a test part. Requires properly trained personnel with considerable application experience
Magnetic particle	A magnet or electrical current creates magnetic fields in a specimen. Magnetic particles applied to the surface indicate the location of a field leakage, signaling a discontinuity.	Used on metal that can be magnetized (ferromagnetic) to detect surface or near-surface discontinuities. Relatively easy to use. Portable equipment is available for field testing.	Cannot be used on metals that cannot be magnetized. Requires properly trained personnel with considerable application experience
Ultrasonic	Ultrasonic waves are injected into material. Measurements of the time or frequency response allow detection of defects and changes in material properties.	Used on almost any material to detect surface and subsurface discontinuities. Also used for measuring material thickness and material property changes.	Cost varies from inexpensive portable equipment to high precision, automated, stationary systems. Requires properly trained personnel with considerable application experience
Eddy current	Time-varying electrical signals induce eddy currents in a specimen. Discontinuities that alter the path of the induced currents are indicated by change in coil impedance.	Used on conductive materials to detect surface and near-surface discontinuities and heat treatment condition. Plating and nonmetallic coating thickness measurements are also possible.	Inspection depth limited to less than one inch. Requires properly trained personnel with considerable application experience
Radiographic	Electromagnetic rays (x-rays and gamma rays) penetrate the test object, and variations in x-ray absorption are recorded using either film or real time detectors.	Used on nearly all material to detect surface or subsurface discontinuities. Also useful for thickness or material property change measurements. A permanent record of inspection is possible.	Higher initial cost. Potential safety hazard to personnel. Requires properly trained personnel with considerable application experience

Fig. 1 Typical inspections. (a) Automated inspection installation for the fluorescent penetrant inspection of large workpieces, such as castings. The installation incorporates a complex roller conveyor system. (b) Retirement-for-Cause (RFC) inspection facility

More details on these and other techniques as currently practiced can be found in extensive handbooks that have been prepared by the American Society for Nondestructive Testing and ASM International (Ref 2, 5), as well as a number of textbooks (Ref 3, 6, 7). In addition, there is much ongoing research to advance these techniques, as continuously documented in a number of technical journals and conference proceedings.

References cited in this section

2. *Nondestructive Testing Handbook,* 2nd ed., P. McIntire, Ed., American Society for Nondestructive Testing (ASNT), Columbus, OH, 10 volumes published in the period 1982–1996

3. L. Cartz, *Nondestructive Testing,* ASM International, 1995

4. *Introduction to Nondestructive Evaluation,* Iowa Demonstration Laboratory, Center for Nondestructive Evaluation, Iowa State University

5. *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989

6. R. Halmshaw, *Nondestructive Testing,* Edward Arnold, London, 1987

7. D.E. Bray and R.K. Stanley, Nondestructive Evaluation, A Tool, *Design, Manufacturing and Service, Revised Edition,* CRC Press, Boca Raton, FL, 1997

Nondestructive Evaluation and Life Assessment

R.B. Thompson, Center for Nondestructive Evaluation, Iowa State University

Life Assessment Strategies

As has been noted in previous Sections of this Volume, there is a wide range of material/component life assessment strategies appropriate to different industries and damage mechanisms, and these can be implemented within either a probabilistic or deterministic framework. Fatigue is perhaps the most widely encountered damage mechanism (Ref 8), and over time, two major design philosophies have been employed: safe-life (based on either stress versus life-cycle curves or strain versus life-cycle curves) (Ref 9) and damage tolerant (based on fracture mechanics) (Ref 9, 10, 11). Situations in which inspection is “not a regularly employed practice, impractical, unfeasible, or occasionally physically impossible … are prime candidates for the applications of safe-life techniques when coupled with the appropriate technologies to demonstrate the likelihood of failure to be sufficiently remote” (Ref 9). In damage tolerant approaches, inspection

generally plays an integral role. These two philosophies are perhaps best illustrated by examples of their application to a specific problem, the approach employed by the U.S. Air Force (USAF) in managing fatigue in metal aircraft structures (Ref 12).

Engineers use two different approaches for the design of metal aircraft structures to resist fatigue and thus ensure flight safety. The so-called safe-life approach is a probabilistic based method. The safe-life of a structure is that usage period in flight hours when there is a low probability that the strength will degrade below its design ultimate value due to fatigue cracking. The determination of the safe-life of an aircraft depends primarily on the results of a full-scale fatigue test of the structure. The number of simulated flight hours of operational service successfully completed in the laboratory is the "test life" of the structure. The safe-life also depends on the expected distribution of failures. The distribution of failures provides the basis for factoring the test life. The factor is called the *scatter factor*. The distribution of failures may be derived from past experience from similar aircraft or from the results of design development testing preceding the full-scale fatigue test. The test life is divided by the scatter factor to determine the safe life. The scatter factor (usually in the interval from two to four) is supposed to account for material property and fabrication variations in the population of aircraft.

The other method used for the design of metal aircraft structures to resist fatigue and thus ensure flight safety is the damage tolerant approach. Damage tolerance is the attribute of a structure that permits it to retain its required residual strength for a period of usage without repair after the structure has sustained specified levels of fatigue, corrosion, and/or accidental or discrete source damage. The damage tolerance approach as currently applied is primarily a deterministic method. However, many aspects of it are based on probabilistic methods, the most notable of which, perhaps, is the "inspectable" flaw size. The inspectable flaw size is the size that will be detected with a specific probability and confidence. The USAF uses 90% probability and 95% confidence for its inspectable-flaw criterion. Inspections are performed using a fracture mechanics analysis to determine the extent of growth of the inspectable crack to critical length. The crack propagates rapidly at critical length and may result in catastrophic loss of the aircraft. The inspection is typically performed at one-half the time for the inspectable crack to grow to critical length. When a crack is found by NDE, the structure is repaired or modified to eliminate the defect.

For the case of fatigue, the damage tolerant approach, which had found a wide range of applications (Ref 13, 14, 15, 16, 17), is based on the physical understanding of damage evolution summarized in Fig. 2. The growth and/or fracture of a crack is assumed to be controlled by the stress intensity factor, K_{Ic} that defines the singularity in stress at the tip of the crack and is a function of the applied loads and the crack size. If the stress intensity factor exceeds a material property known as the fracture toughness, failure is imminent. If the load is cyclic and the maximum stress intensity factor is less than the fracture toughness, the crack will grow at the rate that is controlled by the excursions in the cyclically varying stress intensity factor until the fracture criteria is reached.

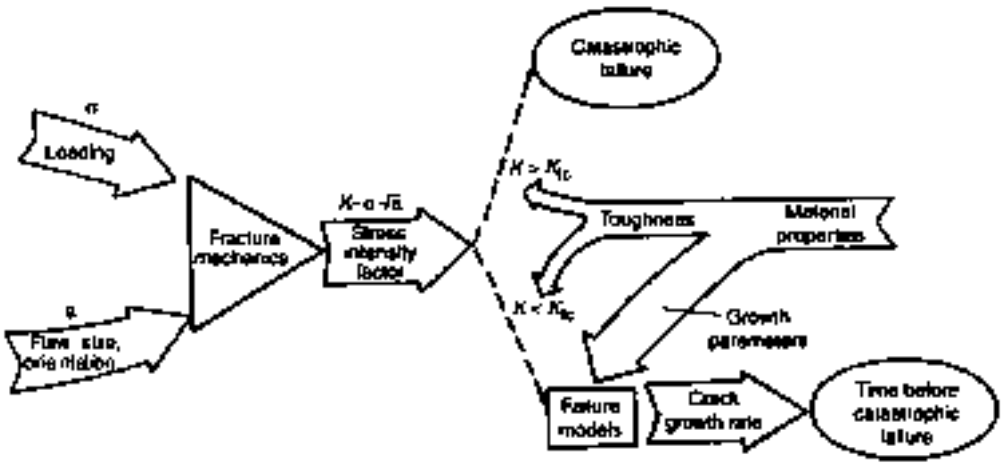


Fig. 2 Fracture mechanics concepts governing the prediction of failure under conditions of cyclic fatigue

References cited in this section

8. *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996

9. D.W. Cameron and D.W. Hoeppner, Fatigue Properties in Engineering, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 15–26

10. D. Broeck, Concepts of Fracture Control and Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 410–419

11. D. Broeck, The Practice of Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 420–426

12. J.W. Lincoln, private communication

13. M.P. Kaplan and T.A. Wolff, Life Extension and Damage Tolerance of Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 557–565

14. T. Swift, Damage Tolerance Certification of Commercial Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 566–576

15. M.P. Kaplan and J.W. Lincoln, The U.S. Air Force Approach to Aircraft Damage Tolerant Design, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 577–588

16. W.D. Cowie, Fracture Control Philosophy, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989 p 666–673

17. W.D. Rummel, G. L. Hardy, and T. D. Cooper, Applications of NDE Reliability to Systems, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 674–688

Nondestructive Evaluation and Life Assessment

R.B. Thompson, Center for Nondestructive Evaluation, Iowa State University

The Role of NDE

Figure 3 conceptually illustrates the interplay of NDE and fracture mechanics in the damage tolerant approach. It is assumed that the service-induced damage in a part will increase in time. For the case of fatigue, the measure of damage is crack size. However, the same concepts would apply to high-temperature creep, corrosion, wear, or some other degradation mechanism. The symbol *D* will be used in this and some subsequent figures to denote the generality of the concepts. In the safe-life approach, one simply defines the service life to be sufficiently short that the probability of failure due to this natural evolution of damage to a critical value, D_F, is at an acceptably low level. However, in the damage tolerant approach, the use of inspection allows a reduction of this conservatism of the safe-life approach. It is recognized that flaws will be present, and NDE techniques are expected to remove those components from service that contain a flaw whose size is greater than the inspectable flaw (i.e., a *defect* in the previously cited terminology of ASTM).

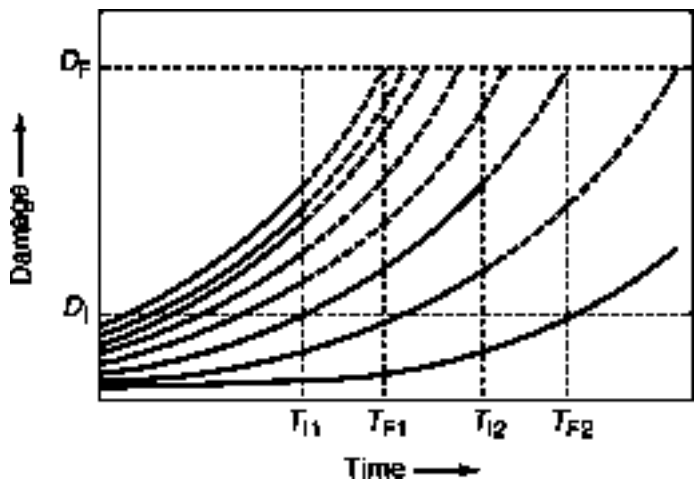


Fig. 3 Illustration of the damage tolerance approach to life assessment. Curves indicate progression of damage from different nucleating conditions, with broken segments representing regimes in which a perfect NDE technique would remove the component from service

In the preparation of Fig. 3, it has been assumed that there is a distribution of damage levels (or damage nucleating conditions) within the material and that the inspection has perfect differentiation, that is, that it can perfectly identify all damage levels greater than some threshold (D_I) for removal from service. The curves in Fig. 3 (neglecting for the moment the difference between the solid and dashed segments) show the evolution of damage with time. The time, T_{F1}, is taken to be the time at which the first failure (or an acceptably low probability of failure) would occur. Inspection would then be conducted at an earlier time, T_{I1} (dictated by the desired safety margin—a factor of 2 in the previously described metal aircraft example). The hypothesized perfect inspection would then allow all components containing damage greater than D_I to be removed from service. The damage evolution avoided by the inspection is denoted by the dashed segments of the curves. The expected life of the remaining components would then be greater. As indicated, the process could be repeated additional times, extending the life of remaining components.

Probability of Detection. More realistic approaches recognize that no NDE technique is perfect. Given a family of nominally identical flaws (i.e., identical from the perspective of failure analysis), there will be different NDE responses due to such effects as morphology of the flaw, influences of the nearby microstructure, and variations in measurement systems/human operators. A metric of an inspection is the degree to which it can correctly identify for removal from service those components whose damage exceeds the intended threshold level, D_I. Formally, this metric is known as the probability of detection (POD). As sketched in Fig. 4, the POD is based on the concept that, for a given flaw size, there will be a distribution of flaw signals. However, it must be recognized that, in the absence of any flaw, there will also be a distribution of noise signals from benign discontinuities (part geometry, grain boundaries, etc.), and these signals could lead to false alarms if they are sufficiently large. The POD, or any other measure of the effectiveness of a technique, will obviously depend on the criteria for acceptance or rejection of a part. For example, in the case of a pulse-echo ultrasonic inspection, a received electrical signal at a time, T, after the wave was launched indicates that there is a discontinuity at a distance $vT/2$ from the transducer, where v is the velocity of sound in the material under inspection. Such a signal is generally interpreted as being due to a flaw if its amplitude exceeds a predetermined threshold, often deduced from the strength of signals reflected from a known reference reflector such as a flat-bottomed or side-drilled hole. Lowering the threshold will increase the POD for a given flaw size. However, it will also increase the probability of false alarms (PFA), because it will be more likely that noise signals from geometrical or microstructural features produce signals exceeding the threshold. Thus, the POD and PFA are defined, respectively, by the distributions of signals that would be produced by flaws of the same nominal size and the distributions of noise in the absence of a flaw, each as compared to the threshold, or accept/reject criteria (Fig. 4, Ref 17). The threshold is selected to achieve sufficiently high POD (thereby avoiding failed parts and their consequences) while keeping the PFA sufficiently low (avoiding an unacceptably high number of good parts being removed from service). This decision must be made within the context of the overall life-management program for a particular component, with key ingredients being the relative consequences of component failure and rejection of serviceable components.

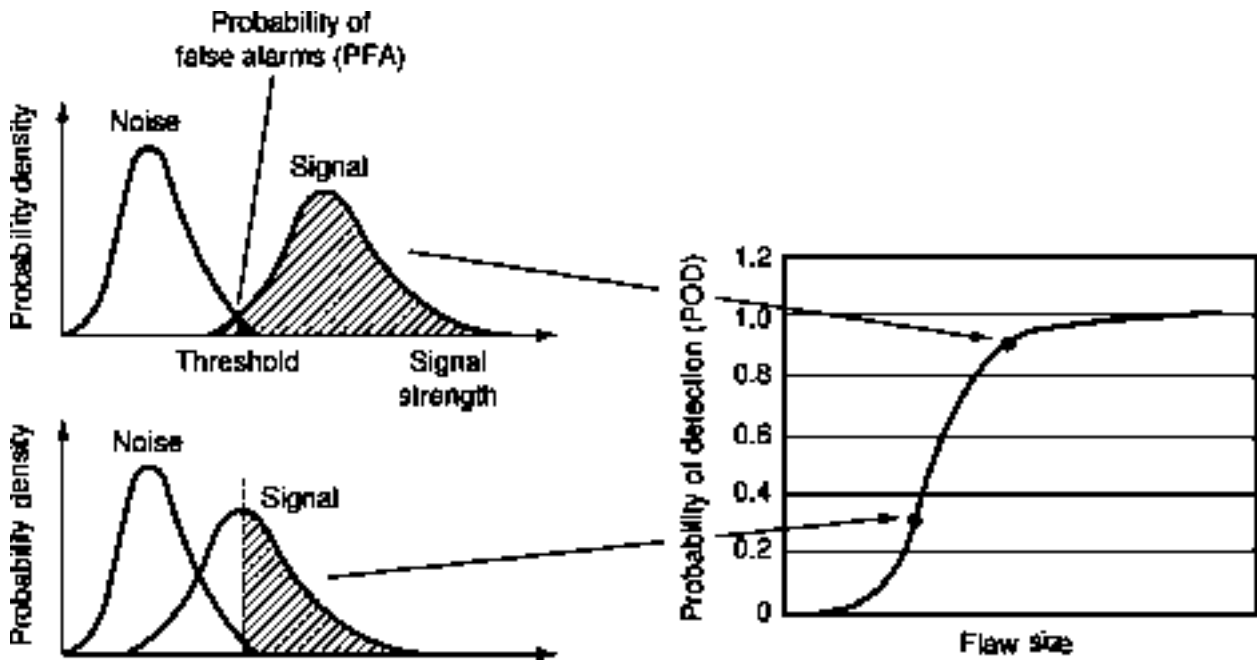


Fig. 4 Probability of detection concepts (POD) and probability of false alarm (PFA) are determined by fractions of signal and noise distributions above a threshold. Signal distribution generally shifts to higher levels as flaw size increases, leading to the sigmoidal POD curve. Source: Ref 22

In recent years, an increasing level of attention has been given to the determination of POD for various techniques (Ref 18). Several empirical procedures are available (Ref 17, 18, 19, 20). In the hit/miss approach, inspectors are presented a set of samples containing flaws of various sizes, and the fraction of the flaws that are detected is recorded. In the flaw response approach, the actual amplitude of the observed signal is recorded. In each case, a maximum likelihood analysis is performed to extract estimates of the POD and confidence levels from the data (Ref 19, 20). The difficulty in either of these approaches is the need to construct a set of realistic test samples each time that a POD study is conducted. This difficulty has motivated an increase in research directed toward using insights gained from physics-based models of the inspection process to allow the maximum amount of information to be extracted from the minimum number of samples (Ref 21, 22). As an example of empirically determined POD curves, Fig. 5 presents results for various geometrical features in aircraft engines, as inspected by the retirement-for-cause (RFC) inspection systems (developed in the 1980s and used to inspect parts for the F100 engine at the San Antonio Air Logistics Center at Kelley Air Force Base, San Antonio, TX) (Ref 17).

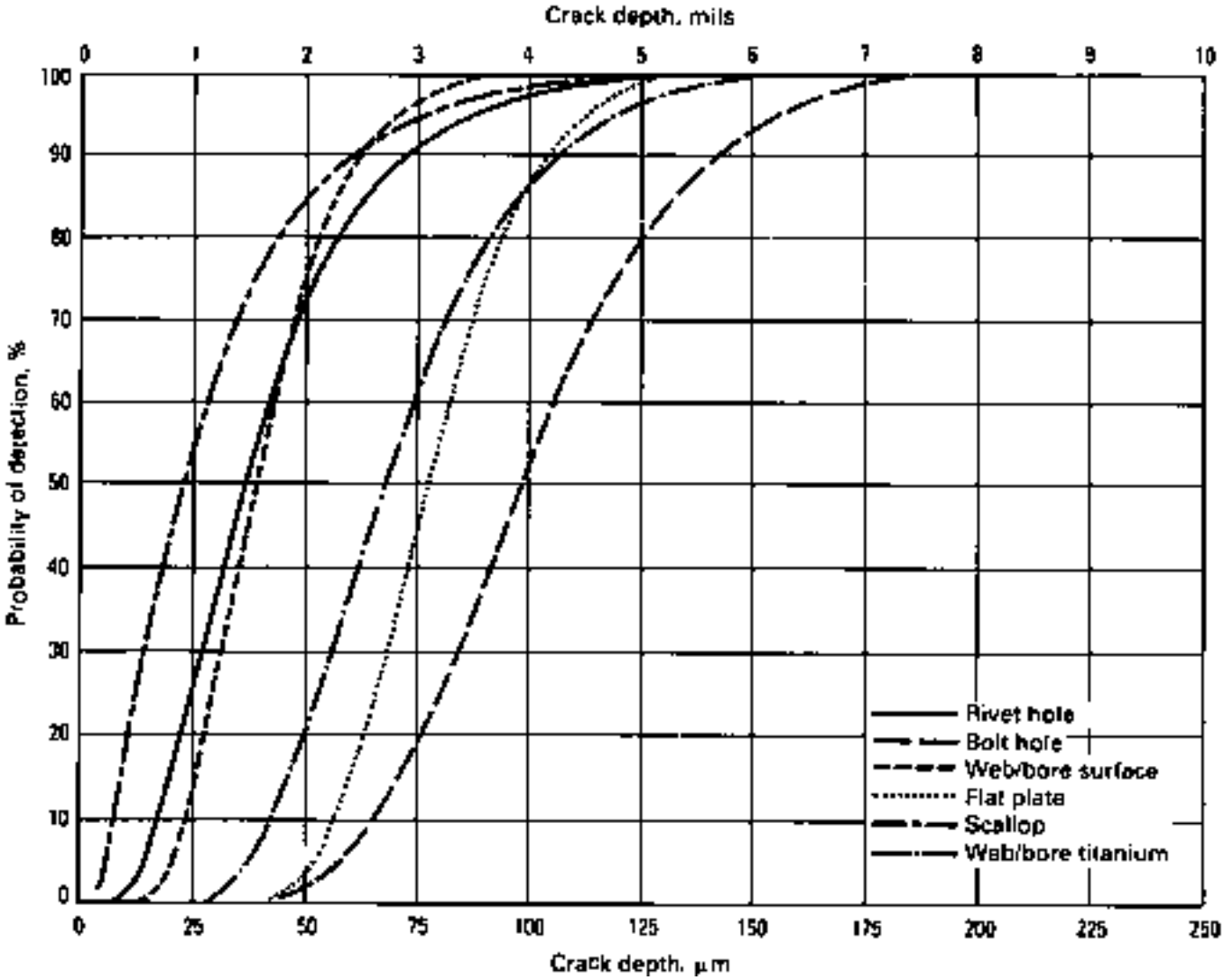


Fig. 5 Retirement-for-Cause (RFC) inspection system probability of detection (POD) curves for various geometrical features. Source: Ref 17

A useful compilation of the results of many POD studies conducted in the aerospace industry over a period of many years can be found in Ref 23. These studies show a wide range of results, which suggests that the way in which the inspection is conducted plays a crucial role in determining the POD. For example, when the inspection is conducted manually, there is the opportunity for a variety of human factors to come to bear. These depend not only on the skill of the individual, but

the conditions under which the inspection is performed, non-work-related factors that may cause performance to vary from day to day, and other issues. In contrast, automatic inspections are considered to yield much more reproducible results. A commonly held opinion is that an automated inspection does no better than can be done under good conditions by a skillful inspector. However, it introduces a much higher level of uniformity and, frequently, more speed into the inspection process. The key question in reliability studies is not how small of a flaw can be detected, but rather how large of a flaw might be missed. Automation directly addresses this issue.

It would be desirable to identify a particular crack size that could be detected by each method. Unfortunately, such specification is not easy, because the POD can be strongly influenced by geometrical features, material, and conditions under which the inspection is performed (as can be seen by comparing the curves in Fig. 5). A review of data that was discussed earlier (Ref 23) shows that there is a wide range of flaw sizes for which the 90% POD value is reached for any given technique. Table 2 presents the maximum and minimum flaw sizes reported for each of the five major techniques, based on a wide range of studies under different conditions (Ref 23). The fact that more than an order of magnitude of difference exists between the best and the worst amply illustrates the cautions that must be exercised in making general statements about sensitivity. It should be emphasized, however, that many of the poorer results were obtained in early studies and should not be taken as representative of the best practice of today. Other important factors are flaw morphology and operator performance.

Table 2 Range of flaw sizes for which 90% POD has been reported in various studies, conducted over the past 25 years

Nondestructive evaluation method	Flaw sizes, mm (in.)	
	min	max
Liquid penetrant	0.43 (0.017)	17.78 (0.700)
Magnetic particle	0.99 (0.039)	6.02 (0.237)
Ultrasonic	0.36 (0.014)	6.73 (0.265)
Eddy current	0.56 (0.022)	19.05 (0.750)
Radiographic	0.61 (0.024)	18.52 (0.729)

POD, probability of detection. Source: Ref 23

The influence of POD on damage distributions is shown in Fig. 6. A damage distribution is shown by a solid line as it might appear after some preinspection period of service. The broken lines indicate how this would be modified by (a) a perfect NDE technique that rejects all and only those damage states exceeding the specified threshold, and (b) a more realistic, “typical” NDE technique. Fracture-mechanics based life assessments, based on these two post-inspection distributions, would obviously yield different results.

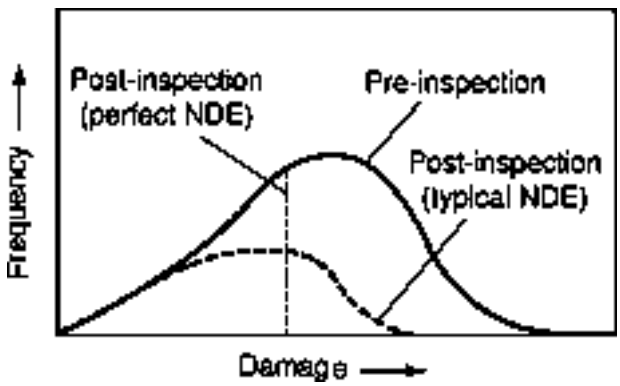


Fig. 6 Pre-inspection and post-inspection damage distributions for two NDE techniques

Inspection Intervals. A key aspect of a damage tolerant fracture control program is the interval between inspections. This depends on the specific values of the level of damage that will cause fracture, the rate of evolution of damage, and the probability of detection of damage. Inspection intervals should be selected such that the probability of failure before the next inspection is below an acceptable level. Improving the POD of the inspection will allow the interval between inspections to be increased, as is suggested by the differences in the two post-inspection defect distributions shown in Fig. 6.

As an example, consider the application of damage tolerance to military aircraft engines (Ref 16), a part of the USAF Engine Structural Integrity Program (ENSIP). In this example, components are designed for crack growth so that the safety limit (the time for assumed initial flaws to grow and cause failure) exceeds two times the required inspection interval (e.g., the planned depot visit interval). Figure 7 shows the size of a worst case flaw, as a function of time, for this

deterministic approach. After each inspection interval, the flaw size is reset to an initial size determined by the inspectable flaw size.

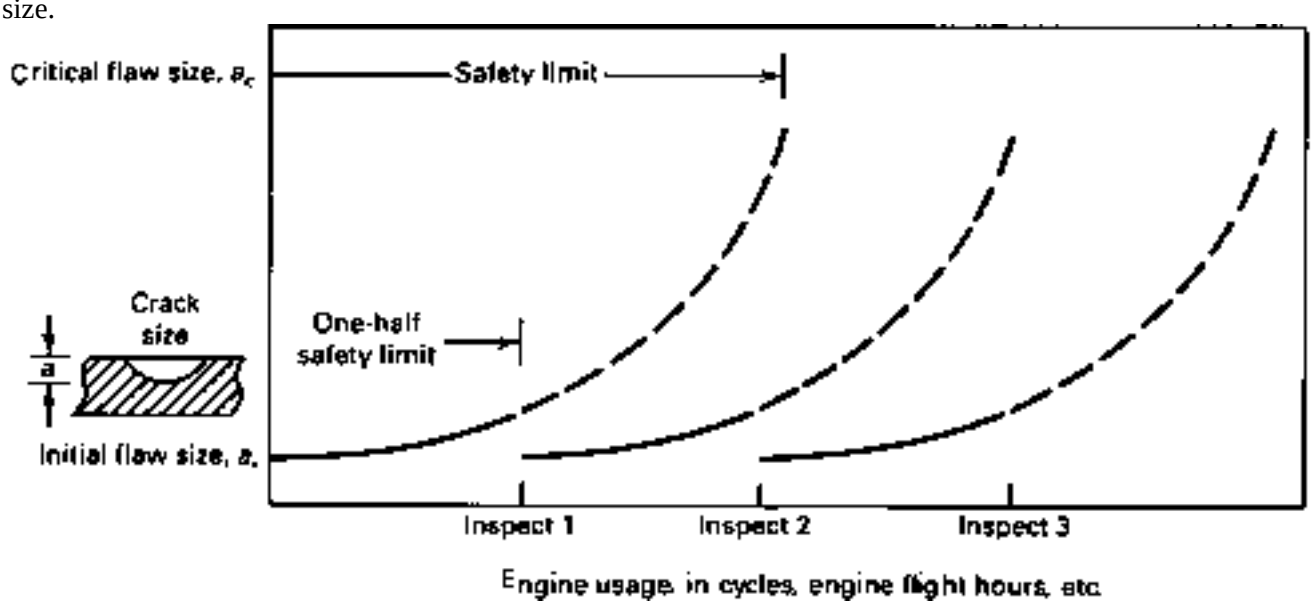


Fig. 7 Damage tolerance approach to life management of cyclic-limited engine components. The safety limit or residual life is the time for the initial flaw to grow and cause failure. The size of the initial flaw, a_i, is based on the inspection method or material defect distribution (for embedded defects). Source: Ref 16

In this example, it is essential that the inspection interval be compatible with the overall engine maintenance plan. This is driven by a number of practical issues. Obvious is the fact that the operation of the fleet is less disrupted if inspection can be conducted while the engine is being serviced in other ways. Less obvious is the fact that, if the engine were disassembled for the purpose of inspection alone, one would have to take into account the possibility of accidentally inducing damage in the tear down and assembly operations. Similar situations exist in the nuclear power industry, when the inspection interval is linked to other operations such as refueling.

There is an obvious motivation to extend the inspection interval. An important objective in the design and operation of structures and equipment is to reduce costs by increasing the maintenance intervals, and it is necessary to be able to adjust the inspection intervals to match. Even if inspection is not linked to maintenance, there is an obvious practical advantage to increasing these intervals. Figure 8 conceptually summarizes several ways in which this can be realized. For a more detailed discussion, the reader is encouraged to consult the original reference (Ref 10).

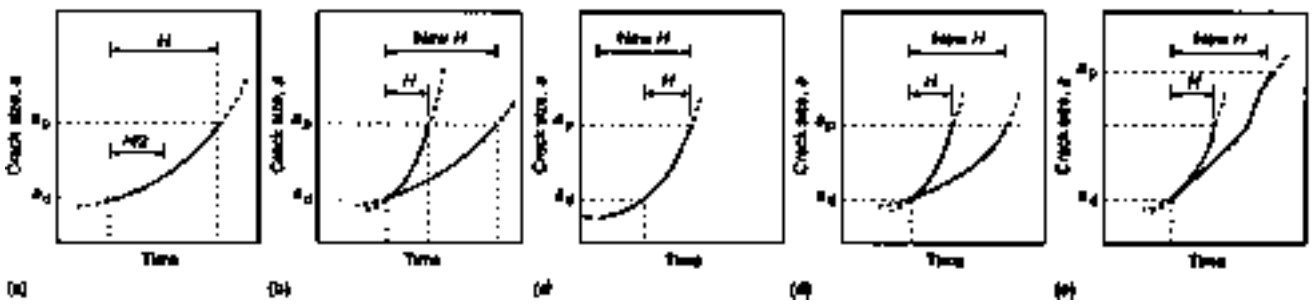


Fig. 8 Ways to increase inspection interval. (a) Base case. (b) Use of better material. (c) Use of more sensitive inspection method. (d) (Detail) design with lower stress. (e) Redundancy (fail-safe) or crack arresters. Here, *H* is the safety limit that is taken to be twice the inspection interval. Source: Ref 10

Fully Probabilistic Approaches. There is an increasing interest in moving towards fully probabilistic approaches to design and life management, and Fig. 9 summarizes the role of NDE. The life of a component depends on the damage distribution that existed in the component at the time when it entered service and the rate at which that damage evolved under the conditions of the service, modified by the extent to which components with excess damage were removed from

service as a result of exhibiting signals greater than the accept/reject criteria during various inspections. The POD is the metric whereby the efficacy of improved NDE is taken into account. For the case of aircraft engines, a series of interagency workshops describes the progress in developing these fully probabilistic approaches (Ref 24).

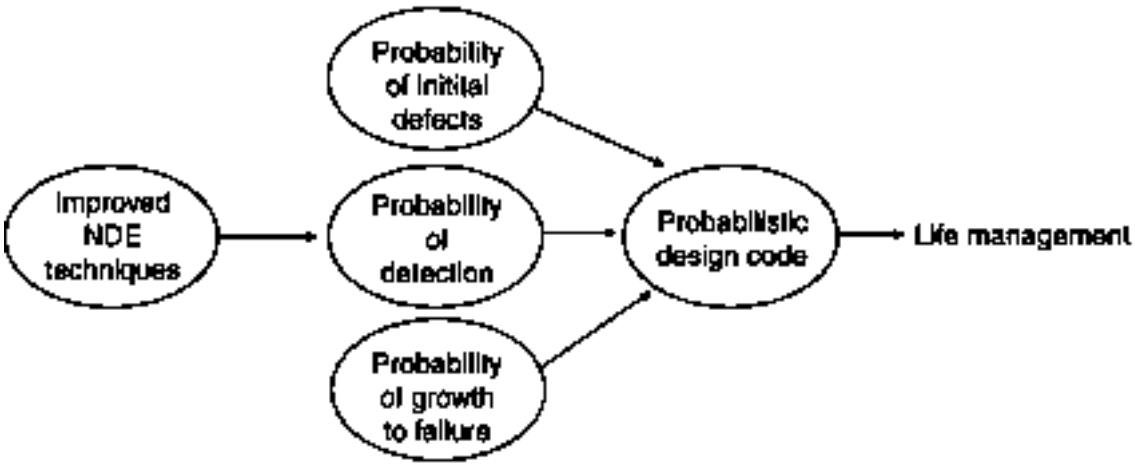


Fig. 9 Role of NDE, as quantified by probability of detection (POD), in fully probabilistic life management

References cited in this section

10. D. Broeck, Concepts of Fracture Control and Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 410–419

16. W.D. Cowie, Fracture Control Philosophy, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989 p 666–673

17. W.D. Rummel, G. L. Hardy, and T. D. Cooper, Applications of NDE Reliability to Systems, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 674–688

18. R. Singh, Three Decades of NDT Reliability Assessment and Demonstration, *Materials Evaluation,* Vol 59, 2001, p 856–860

19. A.P. Berens, NDE Reliability Data Analysis, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 689–701

20. F.W. Spencer, Estimating Probability of Detection Curves from Regression Data, *Materials Evaluation,* Vol 59, 2001, p 866–870

21. J.N. Gray, T.A. Gray, N. Nakagawa, and R.B. Thompson, Models for Predicting NDE Reliability, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 702–715

22. R.B. Thompson, Using Physical Models of the Testing Process in the Determination of Probability of Detection, *Materials Evaluation,* Vol 59, 2001, p 861–865

23. G.A. Matzkanin and W.D. Rummel, *NDE Capabilities Data Book,* NTIAC-DB-97-02, 1997

24. *Proc. 4th Annual FAA/Air Force Workshop on the Application of Probabilistic Methods to Gas Turbine Engines,* (Dayton, OH), Universal Technology Corporation, 2000

Non-Aerospace Applications

Most of the examples given in the previous section were taken from the aerospace industry. However, these same general concepts are applied in a variety of industries, depending on the details of the failure mechanisms operative in a particular component or structure and the context of its application. For example, the decision of whether to use probabilistic or deterministic approaches, whether to try to tightly link a careful quantification of POD to life management models or set accept/reject thresholds based on conservative safety margins, and so forth are very industry specific and change with time. Although the approaches may initially seem quite different, the differences are very logical consequences of the intended service. For example, the way that reliability would be handled in a space vehicle designed for one flight with catastrophic consequences of failure would be quite different from a pressure vessel that would leak before it breaks. The different ways in which these decisions have been made in the off-shore pipeline, nuclear power plant, and aerospace industries were discussed at a late-1990s European-United States workshop on reliability (Ref 25). In addition to industry specific approaches, differences in the philosophies in the two geographical regions were also evident.

As an example, consider the differences in the approaches adopted in aircraft and nuclear power applications. The military aviation industry has a well-developed damage tolerant philosophy using the so-called slow crack growth procedure for monolithic structures, as discussed. However, the details of the philosophy in the nuclear industry have been somewhat different. As noted by Doctor (Ref 26), "In the mid-60s, the designers of nuclear reactors in the USA concluded that the environment was so potentially harmful to humans that the best strategy was to design the reactors so that they would never fail and thus, would never require in-service inspection (ISI). As a consequence, there were no considerations given for ISI regarding design, material selection, access, etc. The philosophy was to perform NDE at the fabrication stage to insure a high quality component. The designers focused on the main area of concern at the time, namely fatigue, and this was the main design basis—insure failure did not occur by this mechanism. Since that time other degradations of failure have occurred. …"

Dealing with these other degradations drives much of the ISI work in the nuclear industry today, which is a layer in the "defense in depth" philosophy. Again following Doctor (Ref 26), "During the design stage, postulated accidents and both anticipated and unanticipated operating conditions are considered. The plant is then designed, built, and operated so as to maintain structural and leak-tight integrity of components important to safety. No one action, system, or component is depended upon to maintain safety; but a number of actions, systems, and components with multiple backups taken together—defense in depth—are responsible for maintaining high reliability and safety of operating nuclear power reactors. In most cases, a single component can fail without endangering public health and safety."

Preservice and in-service inspections are conducted within the context of this philosophy, and requirements are given by Title 10 of the Code of Federal Regulations Part 50, as endorsed by the ASME Boiler and Pressure Vessel Code. Initially, these codes specified that inspections be conducted on an arbitrarily specified fraction of the components of a reactor, motivated by the goal of identifying generic "failures," including those that might be associated with unexpected damage mechanisms. However, as experience was gained, this led to the identification of particular components in which "failure" was most likely to occur. For example, once a new mechanism of damage such as intergranular stress-corrosion cracking was identified, much more specific and localized tests were adopted. Currently, there is a transition underway from deterministic to a risk-informed (or risk-based) approach based on a probabilistic methodology (Ref 27). In this approach, the selection of ISI is driven by two factors: the risk posed by the failure of a particular component and the capability of the technique in question to detect the damage of interest. It is the latter that drives the quantification of the NDE. In the early days of the nuclear industry, the inspection procedures followed from the ASME code requirement. However, a number of international round-robin tests demonstrated reliability problems with this approach, and in 1989 the code changed from a prescriptive to a performance-demonstration-based approach (Ref 28). Performance demonstrations are not intended to measure the reliability of a particular operator. Rather, they provide a screening test to describe the capability of an inspector population based on success in detection and flaw depth sizing tests. This approach essentially shifted the responsibility of reliable procedure development from the ASME code to the NDE engineers. Within the context of risk-informed, in-service inspection, it can be argued that the level of performance that must be demonstrated varies significantly from component to component (Ref 29).

References cited in this section

25. *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997

26. S.R. Doctor, The Role of NDE in Maintaining and Extending the Life Cycle of Engineering Components, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 33–40

27. K.R. Balkey and O.J.V. Chapman, The ASME Research Task Force on Risk-Based In-service Inspection, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 169–184

28. T. Taylor, Impact of Performance Demonstration on U.S. Codes and Standards, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 59–66

29. F. Ammirato, NDE Performance Demonstration in the US Nuclear Power Industry-Applications, Costs, Lessons Learned, and Connections to NDE Reliability, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 115–122

Implementation Guidance

Implementation of NDE, damage tolerance, and related life-assessment strategies is governed by a combination of corporate practice and regulatory guidance that is industry specific. Corporations in many industries utilize NDE and damage tolerance concepts to guide the design and maintenance programs for a variety of structural components, with the full details often being proprietary and influenced by internal issues of risk tolerance, economic competition, and so on. However, in a number of industries in which the consequences of accidents are large, implementation must take cognizance of a variety of external documents.

Within the USAF, significant changes occurred in the mid-1990s (Ref 30). Prior to Acquisition Reform, the Aircraft Structural Integrity Program (ASIP) and how it was applied to the design of aircraft was governed by MIL-STD-1530, while the Engine Structural Integrity Program (ENSIP) was defined in MIL-STD-1783. As a result of Acquisition Reform, these respectively became MIL-HDBX-1530 and MIL-HDBK-1783, issued for guidance only. Documents pertaining to DoD specification and standards (Ref 31) and policies and procedures (Ref 32) are available on the Internet.

In the commercial aviation industry, the governing documents are FAA Advisory Circulars. For example, A.C.33.14-1 describes damage tolerance for high energy turbine engine rotors. Therein, it is noted that this material is advisory, not mandatory. It describes an acceptable means, but not the only means, for demonstrating compliance. FAA Advisory Circulars may be found by consulting the FAA homepage, http://www.faa.gov/.

As noted previously, the ASME Boiler and Pressure Vessel Code plays a central role in the nuclear industry, and numerous additional documents exist in other industries.

References cited in this section

30. T.D. Cooper, private communication

31. ASSIST Online, in Acquisition Streamlining and Standardization Information System (ASSIST) [database online] (Philadelphia, Department of Defense Single Stock Point for Military Specifications and Standards, DODSSP, 2002, [cited April 2002]), available at http://dodssp.daps.mil/assist.htm

32. Defense Standardization Program, in Department of Defense Resource Locator System, 2002 [cited April 2002], available at http:www.dsp.dla.mil/

Conclusions

The role of NDE in life assessment is continuously evolving and will likely see modification before the next revision of this Volume is published. Currently, it can be speculated that there will be a continued increase in the use of probabilistic approaches, which allow some of the conservatism associated with fixed safety factors in deterministic approaches to be circumvented through a more accurate assessment of the condition of individual components and of the efficacy of the inspection, and that improved techniques to gather the desired input information, such as POD and flaw distributions, will continue to be developed.

Nondestructive Evaluation and Life Assessment
R.B. Thompson, Center for Nondestructive Evaluation, Iowa State University

References

1. "Standard Terminology for Nondestructive Examinations, Metals Test Methods and Analytical Procedures," E 1316–00a, *Annual Book of ASTM Standards* Vol 3.03, ASTM, 2001, p 645–683

2. *Nondestructive Testing Handbook,* 2nd ed., P. McIntire, Ed., American Society for Nondestructive Testing (ASNT), Columbus, OH, 10 volumes published in the period 1982–1996

3. L. Cartz, *Nondestructive Testing,* ASM International, 1995

4. *Introduction to Nondestructive Evaluation,* Iowa Demonstration Laboratory, Center for Nondestructive Evaluation, Iowa State University

5. *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989

6. R. Halmshaw, *Nondestructive Testing,* Edward Arnold, London, 1987

7. D.E. Bray and R.K. Stanley, Nondestructive Evaluation, A Tool, *Design, Manufacturing and Service, Revised Edition,* CRC Press, Boca Raton, FL, 1997

8. *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996

9. D.W. Cameron and D.W. Hoeppner, Fatigue Properties in Engineering, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 15–26

10. D. Broeck, Concepts of Fracture Control and Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 410–419

11. D. Broeck, The Practice of Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 420–426

12. J.W. Lincoln, private communication

13. M.P. Kaplan and T.A. Wolff, Life Extension and Damage Tolerance of Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 557–565

14. T. Swift, Damage Tolerance Certification of Commercial Aircraft, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 566–576

15. M.P. Kaplan and J.W. Lincoln, The U.S. Air Force Approach to Aircraft Damage Tolerant Design, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 577–588

16. W.D. Cowie, Fracture Control Philosophy, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989 p 666–673

17. W.D. Rummel, G. L. Hardy, and T. D. Cooper, Applications of NDE Reliability to Systems, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 674–688

18. R. Singh, Three Decades of NDT Reliability Assessment and Demonstration, *Materials Evaluation,* Vol 59, 2001, p 856–860

19. A.P. Berens, NDE Reliability Data Analysis, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 689–701

20. F.W. Spencer, Estimating Probability of Detection Curves from Regression Data, *Materials Evaluation,* Vol 59, 2001, p 866–870

21. J.N. Gray, T.A. Gray, N. Nakagawa, and R.B. Thompson, Models for Predicting NDE Reliability, *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989, p 702–715

22. R.B. Thompson, Using Physical Models of the Testing Process in the Determination of Probability of Detection, *Materials Evaluation,* Vol 59, 2001, p 861–865

23. G.A. Matzkanin and W.D. Rummel, *NDE Capabilities Data Book,* NTIAC-DB-97-02, 1997

24. *Proc. 4th Annual FAA/Air Force Workshop on the Application of Probabilistic Methods to Gas Turbine Engines,* (Dayton, OH), Universal Technology Corporation, 2000

25. *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997

26. S.R. Doctor, The Role of NDE in Maintaining and Extending the Life Cycle of Engineering Components, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 33–40

27. K.R. Balkey and O.J.V. Chapman, The ASME Research Task Force on Risk-Based In-service Inspection, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 169–184

28. T. Taylor, Impact of Performance Demonstration on U.S. Codes and Standards, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 59–66

29. F. Ammirato, NDE Performance Demonstration in the US Nuclear Power Industry-Applications, Costs, Lessons Learned, and Connections to NDE Reliability, *Proc. European-American Workshop on Determination of Reliability and Validation Methods on NDE,* Deutsche Gesellschaft fur Zerstorungsfreie Prufung, Berlin, 1997, p 115–122

30. T.D. Cooper, private communication

31. ASSIST Online, in Acquisition Streamlining and Standardization Information System (ASSIST) [database online] (Philadelphia, Department of Defense Single Stock Point for Military Specifications and Standards, DODSSP, 2002, [cited April 2002]), available at http://dodssp.daps.mil/assist.htm

32. Defense Standardization Program, in Department of Defense Resource Locator System, 2002 [cited April 2002], available at http:www.dsp.dla.mil/

Fatigue-Life Assessment

Mitchell P. Kaplan and Timothy A. Wolff, EAD Aerospace, Inc.

Introduction

THE DETERMINATION OF THE LIFE of an engineering structure is based on two precepts. These are knowledge of the structure itself and knowledge of how that structure is loaded. The underlying information required for fully understanding these two singularly complicated items is significant. Additionally, the knowledge that is required spans many different disciplines. These disciplines, in the case of aircraft, can include aerodynamics, fluid flow, mechanics, mechanisms, structures, metallurgy, materials science, corrosion, inspection methods, statistics, testing methods, and design. In order to predict the life of the component, therefore, the determination of rational inspection intervals requires the knowledge of several engineering disciplines. One needs to know the future loadings of the structure so one may anticipate (calculate) its life and then determine the appropriate inspection intervals to preclude catastrophic failure.

The individuals tasked with the determination of the component life depend on the knowledge and results of those who can supply the aforementioned information. These analysts work with knowledgeable people in each of these disciplines to ensure the necessary information is calculated and forwarded to them in a timely manner.

Why are these many disciplines required? Why is the analyst so dependent on them? This article examines them in the case of aircraft. However, automobiles and ships require almost the same disciplines. The small differences are indicated.

A flying aircraft experiences both lift and drag. These loads on the aircraft are dependent on the design of the external parameters of the aircraft. These parameters include geometry, weight, shape of the wing, geometry of the fuselage, whether it is pressurized, location and geometry of the empennage, the power plant of the aircraft, and anticipated usage. The same may be said for the other two vehicles, except for the particular geometry associated with the aircraft and the vertical dimension of flight. It should be noted that these other vehicles also experience lift and drag. In aircraft, air is also considered a fluid, keeping the aircraft aloft, impinging on the structure. The motion of the vehicle through the fluid and the effect of the fluid on the vehicle also induce loads.

The particularly interesting aspect of this is that these loads manifest themselves onto the structure. In any of these structures, these may be thought of as external loads. These loads are placed onto the structure by the operating environment. For an aircraft, it would be velocity, altitude, roll, yaw and pitch angles, velocities and acceleration, maneuver, gust, and so forth. For an automobile it would be velocity, road roughness, yaw angle, velocity and acceleration, and so forth. For bridges, it would be type, span, loading, wind velocity, and so forth.

The next area is the determination of these external loads into internal loads. The internal loads are loads that manifest themselves throughout the structure. One assumes a particular point in time and has a load condition (all the external variables are fixed), and from that one can calculate the loads going throughout the structure. These internal loads are then transformed to stresses. This may be accomplished using equations or computational mechanics (finite-element analysis). One needs to know the geometries and the materials as well as the processing history and mechanical properties of the materials.

Thus far, this article has briefly discussed the mechanics, the structures, and, to a lesser extent, the metallurgy/materials science portion of the investigation. The mechanisms also need to be discussed. The same word can apply to the operating mechanisms or to the variability of the loading. Is the component statically loaded, or is it dynamically loaded? The loading mechanism determines the design parameters one must consider and the material itself. Many structures are exposed to dynamic loading that can cause fatigue failure to exist on the structure. If this is the case, the selection of the material becomes extremely important.

Finally, the operating environment is extremely important. The temperature and the chemical atmosphere determine the operational life of the structure. Again use the aircraft example and consider environmental degradation. Aircraft are either in flight or on the ground. Generally in flight, the aircraft is subjected to cold temperatures and corrosion is insignificant. However, when the aircraft is parked, the temperature and location of the airfield are significant. If the airfield is located in a dry environment, little corrosion will occur. However, if the airfield is located in a humid area by the sea, then much corrosion can occur. Another variable is the cargo. If the aircraft is carrying either cargo that has corrosive characteristics (cartage of live animals) or locations in the aircraft (heads, galleys, etc.), then the location of the airfield is much less important.

The final portion of the study that is critical in determining inspection interval is the methodology used for inspecting the structure, called nondestructive inspection (NDI). Significant strides have been made in this area. *Nondestructive Evaluation and Quality Control,* Volume 17 of *ASM Handbook* (Ref 1) discusses these advances in detail. It should be noted that on aircraft metallic components eddy current, ultrasonic, x-ray, dye-penetrant, and visual inspections are regularly accomplished.

Reference cited in this section

1. *Nondestructive Evaluation and Quality Control*, Vol 17, *ASM Handbook*, ASM International, 1987

Fatigue-Life Assessment

Mitchell P. Kaplan and Timothy A. Wolff, EAD Aerospace, Inc.

Fatigue Crack Growth Variables

The fundamental variables involved in any life assessment are those that describe the effects and interaction of material behavior, geometry, and stress history on the life of a component. For fracture mechanics methods, material behavior is described by fracture toughness and crack growth rate data, and geometry is defined through a dimensionless β factor. For traditional fatigue methods such as the stress-life method, an *S-N* curve, which plots stress versus number of cycles to failure, is obtained through fatigue testing and is used to define a material-dependent endurance limit. The effects of geometry are established through a fatigue notch factor. Both methods define the stress history within a spectrum that describes the frequency and order of the stresses expected in service. It is important to remember that the stresses expected in service are limit stresses and not ultimate stresses, which have been multiplied by a safety factor to meet static strength requirements.

Stress-Life Variables

Constant-Amplitude Loading and Endurance Limit. A constant-amplitude loading cycle is described in Fig. 1 by several parameters that are defined in terms of the minimum and maximum stresses applied during the loading cycle. The mean stress, σ_m, and the alternating stress, σ_a, are defined as:

$$s_m = \frac{s_{max} + s_{min}}{2}$$

$$s_a = \frac{s_{max} - s_{min}}{2}$$

An additional parameter that is useful for characterizing the type of loading is the *R*-ratio:

$$R = \frac{s_{min}}{s_{max}}$$

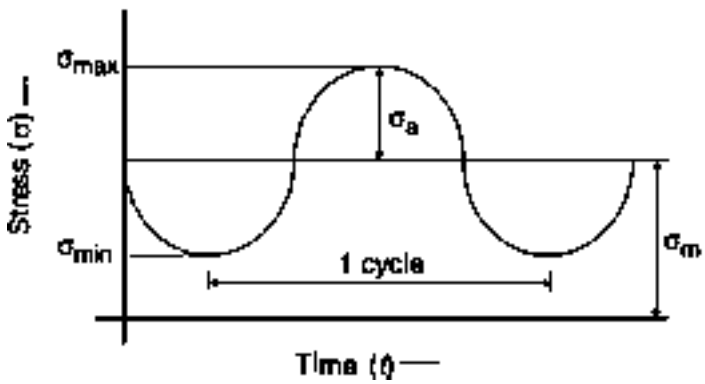


Fig. 1 Constant-amplitude loading cycle

A negative *R*-ratio indicates a tension-compression loading cycle. For an *R*-ratio of -1, the loading is fully reversed and the mean stress is consequently 0. The results of fatigue tests run at *R* = -1 are plotted on an *S-N* curve, as shown in Fig. 2, to establish an endurance limit for the material. For low-strength steels, the endurance limit is established by the portion of the *S-N* curve that becomes asymptotic as the number of cycles becomes large. Stress levels below the endurance limit are considered nondamaging. The endurance limit for low-strength steels is roughly 50% of ultimate tensile strength. For high-strength steels and aluminum, which do not exhibit a clearly defined asymptote, the endurance limit is defined as the alternating stress level occurring at 10^7 cycles. The concept of an endurance limit is useful in fatigue design, and values for a variety of materials can be found in reference books (Ref 3, 4, 5). By keeping the expected stress levels below the

endurance limit, a structural component can theoretically be designed for “infinite” life. In practice, however, the mean stress is not always 0. In such cases, the endurance limit can be inferred from an *S-N* curve that presents data for different levels of mean stress.

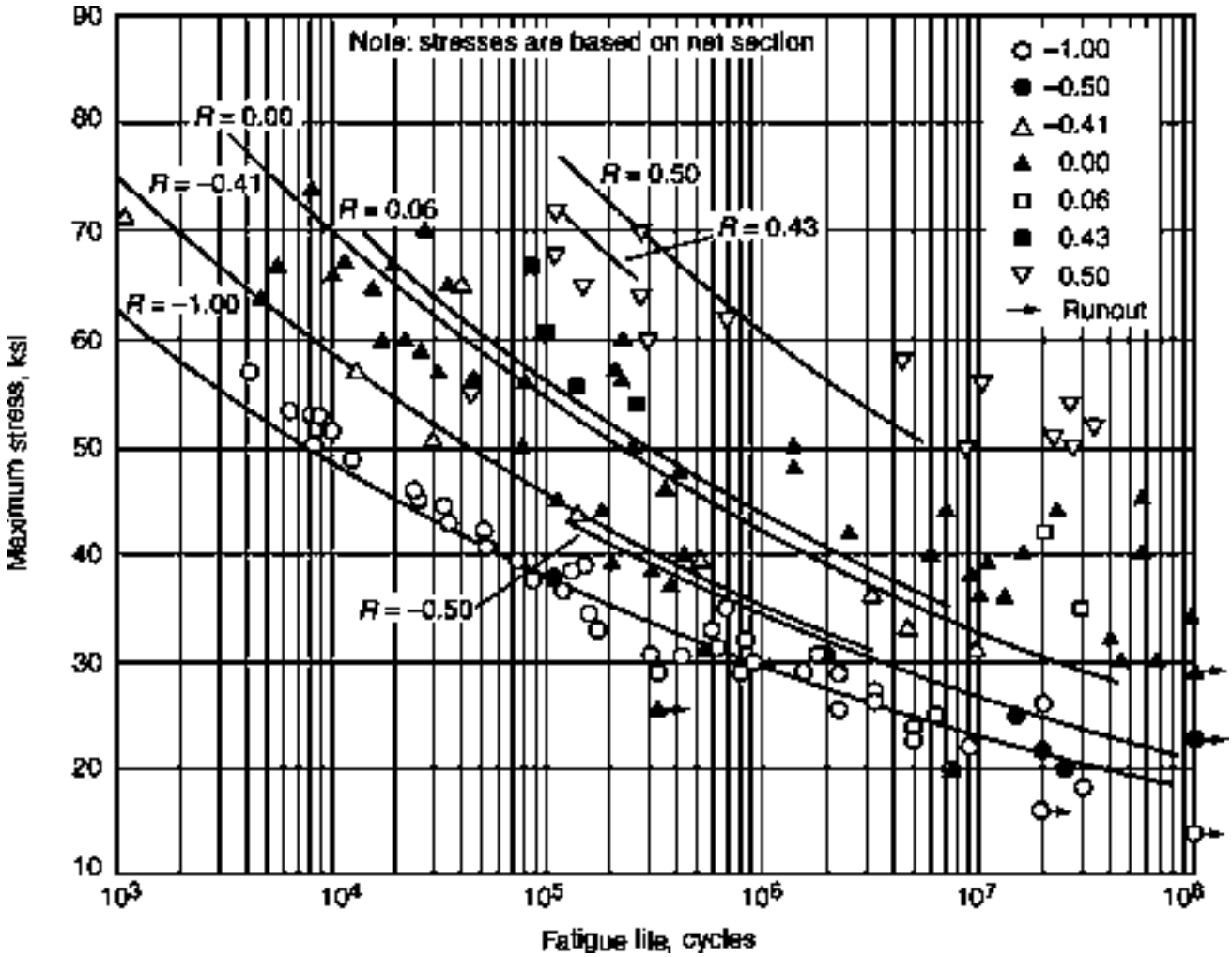


Fig. 2 *S-N* curve for unnotched 2024-T4 aluminum alloy bar. Source: Ref 2

Variable-Amplitude Loading and Cumulative Damage. While constant-amplitude loading is useful for defining fatigue properties and comparing test data, variable-amplitude loading, such as that shown in Fig. 3, is more likely to be encountered by structural components in actual service. In order to calculate the life of a component subject to variable loading, a method is needed that relates constant-amplitude fatigue test data to a random stress history. The Palmgren-Miner cumulative damage rule (Ref 7, 8), which has the form:

$$\sum_{i=1}^{m} \frac{n_i}{N_i} = 1$$

provides a simplified approach to this problem. N_i represents the total life of a component for a given stress cycle, n_i represents the number of stress cycles actually sustained by the component for the given stress cycle, and the upper limit, m, is equal to the number of unique stress cycles in the stress spectrum. The fraction of damage done by each unique stress cycle is added together to obtain a cumulative damage fraction. The reciprocal of this fraction is equal to the number of stress spectra that the component can withstand before failure. The Palmgren-Miner rule essentially reduces a variable stress spectrum into “blocks” of constant-amplitude stresses to obtain a fatigue life estimate. To accomplish this, two primary assumptions are required. First, the relationship between n_i and N_i is assumed to be linear. Second, the fatigue life is assumed to be independent of the order in which the blocks of stresses are applied. These assumptions are valid only under a very limited set of circumstances. For this reason, life estimates obtained from the Palmgren-Miner rule are divided by a large safety factor.

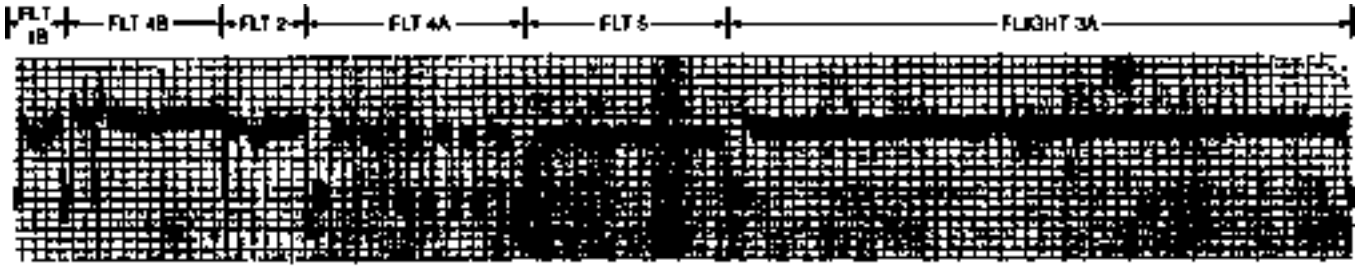


Fig. 3 Variable-amplitude loading data. Source: Ref 6

Stress Concentration and Fatigue Notch Factor. High-stress regions in a structural component are typically located around geometric details, such as a hole, fillet radius, or notch. These local high-stress regions are described in terms of the remote or far-field stress of the component by multiplying the remote stress by a stress concentration factor. Stress concentration factors for various geometric details are derived from the theory of elasticity and/or obtained experimentally. Fatigue tests have shown that the local notch stresses predicted by the stress concentration factor are usually higher than those measured from testing. Therefore, in order to differentiate the theoretical stress concentration factor from the actual measured stress concentration, the term fatigue notch factor is used to describe the local notch stresses that occur during fatigue loading. For axial or bending fatigue tests, the fatigue notch factor is obtained by dividing the endurance limit of an unnotched test specimen by the endurance limit of a notched test specimen.
The fatigue notch factor can also be defined in terms of notch sensitivity, q:

$$K_f = q(K_t - 1) + 1$$

where K_f is the fatigue notch factor and K_t is the stress concentration factor for the given notch geometry. Notch sensitivity is a measure of how close the experimental results for a particular geometry approach the theoretical results. Figure 4 displays notch sensitivity versus notch radius for various metals, based on (Ref 9):

$$q = \frac{1}{1 + (A_k / r)}$$

where A_k is a material constant and r is the notch radius. As the notch radius becomes large, the notch sensitivity factor approaches one, resulting in a fatigue notch factor that is equal to the stress concentration factor. For ferrous alloys, A_k can be approximated by the empirical relationship (Ref 10):

$$A_k = \left(\frac{300}{S_a}\right)^{1.8} x10^{-3} in$$

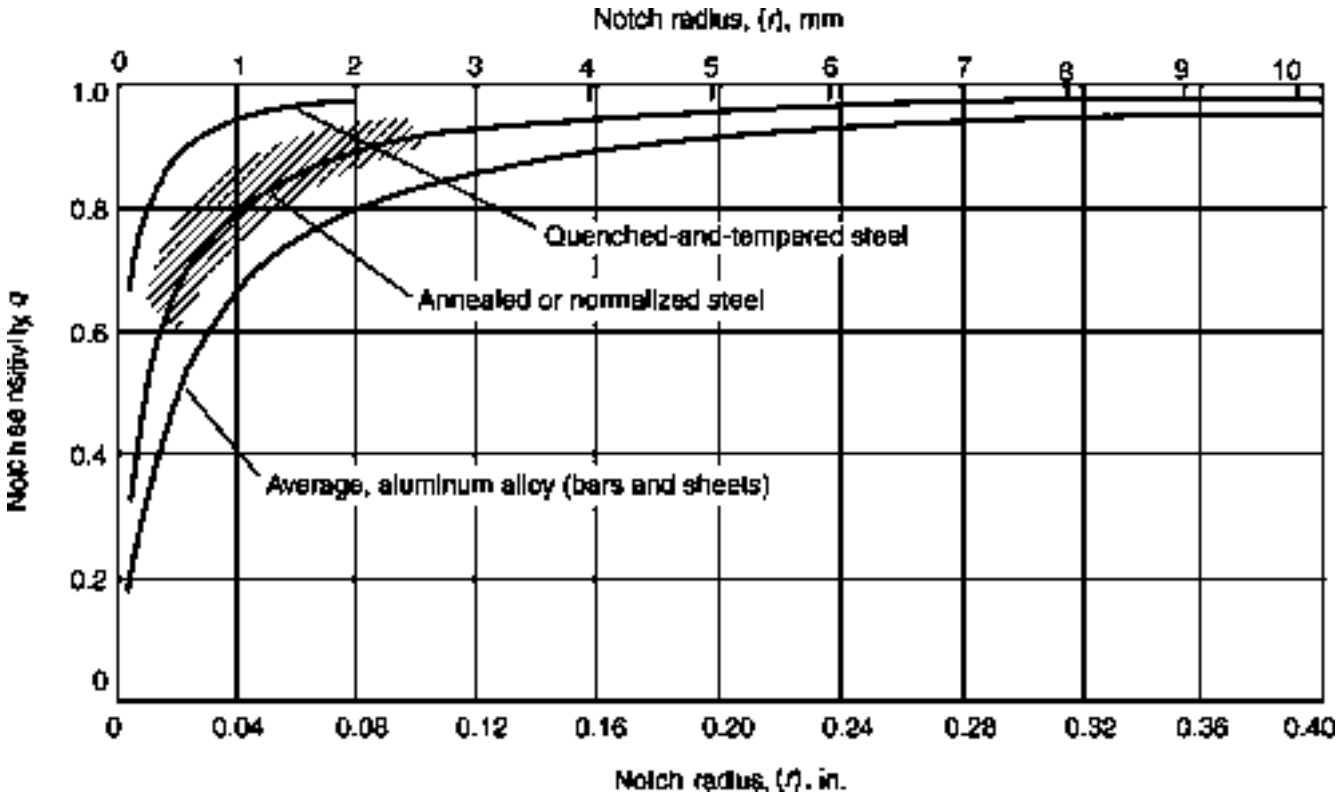


Fig. 4 Notch sensitivity versus notch radius for various metals. Approximate values (note shaded band). Not verified for deep notches thickness/radius. Source: Ref 9

Peterson (Ref 9) provides A_k values of 0.0025 for quenched-and-tempered steel, 0.01 for annealed or normalized steel, and 0.02 for aluminum alloy sheets and bars. It should be noted that the relationship for notch sensitivity is material dependent and should not be used if there is any doubt as to its applicability. Under such circumstances, it is conservative to assume that the fatigue notch factor is equal to the full value of the stress concentration factor.

Fracture Mechanics Variables

Stress-Intensity Factor and Fracture Toughness. The stress-intensity factor describes the stress distribution at a crack tip. For a small edge crack in a thin flat plate subject to a uniform tensile stress, the stress-intensity factor can be determined from (Ref 11):

$$K_1 = s\sqrt{p\,a}$$

where K_I is the stress-intensity factor, σ is the applied stress, and a is the crack length. The subscript on the stress-intensity factor refers to the mode of crack loading. There are three modes of loading (Ref 12). For a crack in a flat plate, mode I loading corresponds to a tensile stress on the edges of the plate that tends to open the gap between the opposing crack surfaces. Mode II loading corresponds to shear stresses that tend to make the opposing crack surfaces slide relative to each other in the plane of the plate. Mode III loading corresponds to shear stresses that make the opposing crack surfaces slide relative to each other perpendicular to the plane of the plate. Stress-intensity factors can be defined for all three modes of crack loading, but mode I loading is typically more critical for crack growth and life estimation.

For the loading conditions described previously, the crack in the flat plate has a critical length (a_c) that will result in fracture of the plate. This length is dependent on the magnitude of the applied stress and the fracture toughness, K_c, of the plate material:

$$K_c = s\sqrt{p\,a_c}$$

Fracture toughness is, in effect, a measure of the critical stress-intensity factor a material can withstand before immediately fracturing. Under static loading conditions, a flat plate with a crack will not fail if the combination of applied stress and crack length produces a stress intensity that is less than the fracture toughness of the material. However, loading conditions are often dynamic, resulting in cyclic stresses that can cause the crack to grow. If the crack grows slowly and in small increments, the crack growth is considered stable. Stable crack growth is only possible if the stress-intensity factor is less than the fracture toughness of the plate. Over time, as the crack continues to grow longer, it will reach a length that, when combined with the applied stress, will result in a stress intensity that is greater than the fracture toughness of the plate. When this occurs, the crack will grow in a very rapid and unstable manner, resulting in fracture of the plate.

As with most material properties, fracture toughness values are established through material testing and depend on the requisite test parameters, such as type of loading, temperature, specimen geometry, and so forth. Figure 5 demonstrates the effect of plate thickness, which is governed by the stress state in the plate. For thin plates, the prevailing stress state is plane stress and fracture toughness is represented by K_c. As the plate thickness increases, the stress state transitions to plane strain conditions, where fracture toughness is represented by K_{Ic} to differentiate it from the plane stress value, which is typically higher. Because measured fracture toughness values are very sensitive to the test conditions, it can be difficult to obtain the true value. For this reason, measured fracture toughness values that are known to be lower than the true value are reported as the apparent fracture toughness, K_{app}. Table 1 presents fracture toughness results for 1.3 mm (0.050 in.) thick 2024-T3 aluminum.

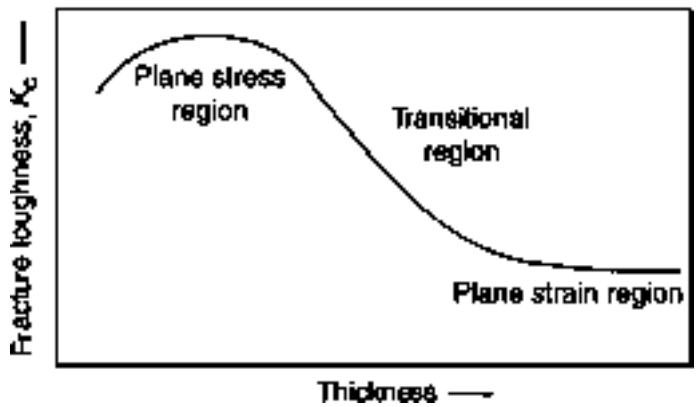


Fig. 5 Effect of thickness on fracture toughness

Table 1 Plane stress and transitional fracture toughness data for 0.050 in. thick 2024-T3 aluminum alloy sheet

Specimen No.	Crack length, in.		Stress, ksi		Toughness, $Ksi\sqrt{in}$			Net stress ratio		
	Initial	Final	Maximum (S_c)	Onset (S_o)	Apparent (K_{app})	Critical (K_c)	Onset (K_o)	Apparent	Critical	Onset
Specimen width, 8.0 in.										
1	2.660	3.600	31.30	19.10	68.73[(a)]	85.36[(a)]	41.94	0.890	1.080	0.543
2	4.770	5.750	19.20	17.10	68.28[(a)]	88.25[(a)]	60.81[(a)]	0.902	1.295	0.804
Specimen width, 24.0 in.										
3	7.000	13.500	27.60	20.00	96.64	159.57[(a)]	70.03	0.739	1.197	0.536
4	7.160	11.000	28.30	16.20	100.48	135.67[(a)]	57.52	0.765	0.991	0.438
5	14.270	18.000	16.50	12.30	101.30	141.83[(a)]	75.52	0.772	1.252	0.576
6	2.320	7.400	42.20	22.10	81.03[(a)]	152.94[(a)]	42.43	0.886	1.158	0.464
Average value					**99.47**	**…**	**61.37**			
Standard deviation					**2.49**	**…**	**14.70**			

Crack orientation is long transverse; buckling of crack edges restrained. Yield strength, 52.7 ksi; ultimate strength, 71.5 ksi. Tested at 70 °F.

(a) Net section stress exceeds 80% of yield strength. Value not included in average or standard deviation.

Source: Ref 13

Based on the preceding discussion, it should be apparent that the stress-intensity factor has a profound effect on the growth of cracks. In fact, the stress-intensity factor is directly related to the elastic energy release rate, *G*, which is a measure of the amount of energy available for crack growth:

$$G_I = \frac{K_I^2}{E} \quad \text{for plane stress conditions } (\sigma_{zz} = 0)$$

$$G_I = \frac{(1-\nu^2)K_I^2}{E} \quad \text{for plane strain conditions } (\epsilon_{zz} = 0)$$

where *E* is the modulus of elasticity and ν is Poisson's ratio. For cracks in ductile metals undergoing stable crack growth, the energy released during crack extension is consumed in the creation of a plastic zone around the new crack tip. If there is not enough energy to form a new plastic zone, the crack will not grow. Many relationships have been proposed in an attempt to describe this stable crack growth quantitatively in terms of the stress-intensity factor. However, no single relationship has universal applicability. The various equations are limited to describing the crack growth behavior of specific materials or for specific ranges of stress-intensity factor values. The Forman equation (Ref 14) correlates well with metals, especially aluminum:

$$\frac{da}{dN} = \frac{C\Delta K^n}{(1-R)K_c - \Delta K}$$

da/dN represents the change in crack length per stress cycle, otherwise known as the crack growth rate, where each stress cycle consists of a minimum and maximum stress. The ratio of the minimum to maximum stress is represented by *R*. *C* and *n* are material-dependent coefficients. ΔK represents the stress-intensity range, which is equal to the difference between the maximum stress-intensity factor and the minimum stress-intensity factor (K_{max} - K_{min}). The maximum and minimum stress-intensity factors are obtained from the corresponding maximum and minimum stresses for each stress cycle.

Fatigue crack growth relationships, such as the Forman equation, are solved iteratively to obtain an estimate of the number of stress cycles to failure for a given spectrum. The material coefficients, *C* and *n*, are calculated from the slope of crack growth rate data curves, which plot *da/dN* versus ΔK. The data for these curves are obtained from tests that are run for a particular *R*-ratio and may exhibit several regions with different slopes. For this reason, it is important to determine the material coefficients from that portion of the curve corresponding to the expected stress-intensity range. If the expected stress-intensity range falls below the stress-intensity threshold, ΔK_{th}, very little crack growth will occur. In

practice, crack growth is assumed to occur only if ΔK is greater than ΔK_{th}. An approximate value for ΔK_{th} can be identified from the indicated stress-intensity range for the minimum data point on the crack growth rate data curve.

Beta Factor and Crack Geometry. The dimensionless factor, β, allows the effects of geometry to be included in the equation for stress-intensity factor:

$$K_I = \sigma\sqrt{\pi a}\beta$$

Beta solutions are available for many different combinations of geometry and applied loading (Ref 15). For a flat plate with a small edge crack and infinite width, β = 1. For a flat plate with a finite but large width relative to the size of the edge crack, β = 1.12. Not all β solutions are exact, however. For a flat plate with a circular hole and a diametrical crack at the edges of the hole subject to a remote tensile stress, as shown in Fig. 6, the β factor is a function of crack length and hole radius, $\beta = f(a/r)$. Several solutions for this function exist, with the following equation proposed by Newman (Ref 16):

$$\beta(a/r) = 1 - 0.15\lambda + 3.46\lambda^2 - 4.47\lambda^3 + 3.52\lambda^4$$

where

$$\lambda = \frac{1}{1+(a/r)}$$

It should be noted that since the β factor is a function of crack length for this geometry, solving for the critical crack length becomes an iterative process. An initial estimate of the critical crack length is necessary so the β factor can be evaluated. The calculated β factor is then used to determine the resulting critical crack length, which is then compared to the initial estimate. This process is repeated until the two values converge. As long as the underlying assumption of linear elastic conditions holds true, β factor solutions for more complex geometries can be obtained through the method of superposition.

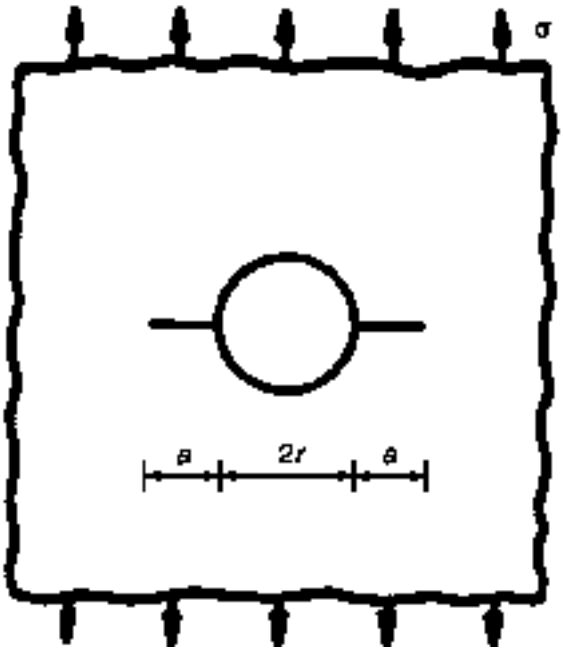


Fig. 6 Diametrical crack at a hole subject to remote tensile loading. Source: Ref 15

When choosing a β factor solution, it is necessary to establish the expected crack geometry. In the preceding example for a circular hole with a diametrical crack, it has been implicitly assumed that the crack is through the thickness of the plate, which limits the resulting crack growth to one dimension. Beta factor solutions also exist for elliptical surface and internal flaws, although the solutions are more complex because the crack is now three-dimensional. Figure 7 shows the geometry for a semielliptical crack under tension loading. The stress-intensity factor solution for this case was developed by Irwin (Ref 17):

$$K = 1.1\sigma\sqrt{\frac{\pi a}{Q}}\left(\left(\frac{a}{c}\right)^2\cos^2\theta + \sin^2\theta\right)^{1/4}$$

where

$$Q = \Phi^2 - 0.212\left(\frac{\sigma}{\sigma_{ys}}\right)^2$$

and an empirical relationship for Φ^2 is:

$$\Phi^2 = 1 + 1.464\left(\frac{a}{c}\right)^{1.65} \quad \text{for } \frac{a}{c} \le 1$$

$$\Phi^2 = 1 + 1.464\left(\frac{c}{a}\right)^{1.65} \quad \text{for } \frac{a}{c} > 1$$

The remaining variables are also defined in Fig. 7. The choice of crack geometry can have a major impact on the calculated life. In a thin plate, a surface flaw will likely become a through crack before failure occurs, making the through crack a conservative choice when obtaining a life estimate. For a thick plate, a through crack may not be a conservative choice if the critical crack length for a surface flaw in the plate is less than the thickness of the plate. Failure of the plate would occur before the surface flaw becomes a through crack.

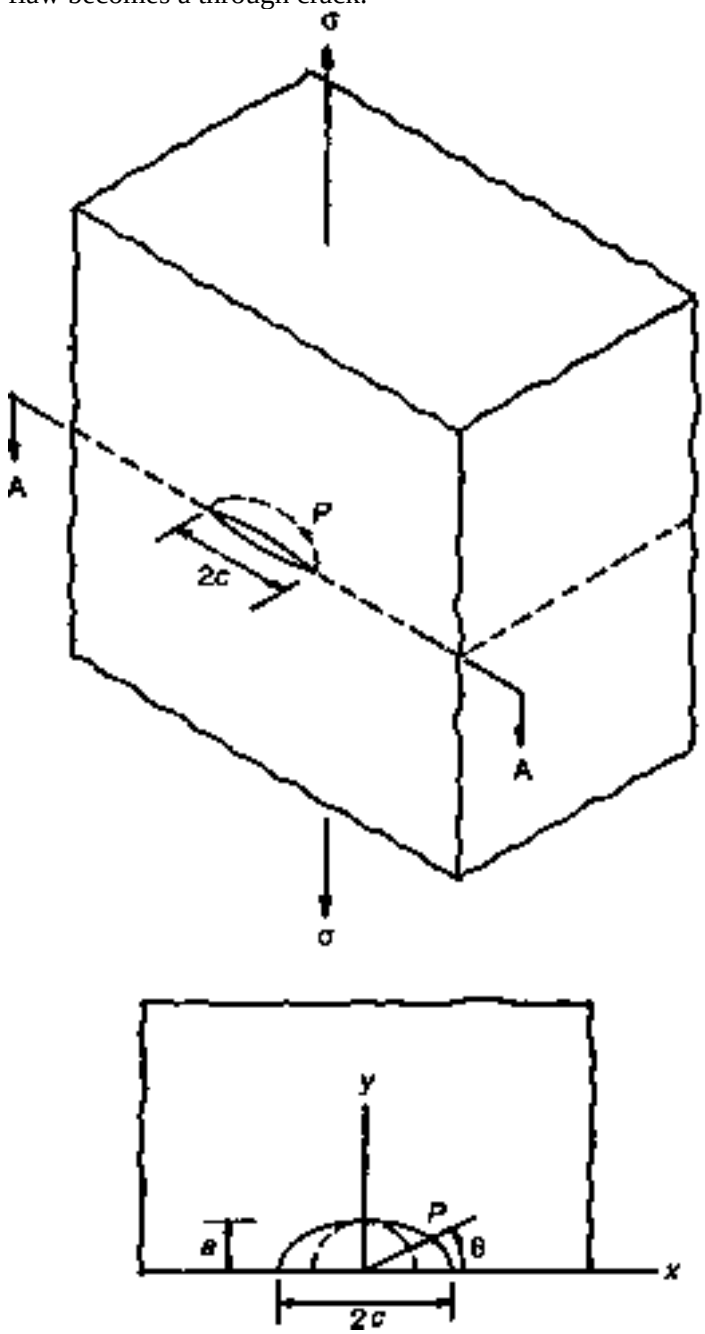


Fig. 7 Semielliptical crack in a thick solid subject to remote tensile loading. Coordinates of point *P*: x = *c* cos θ; Y = *a* sin θ. Source: Ref 15

Stress Spectrum

Increased awareness of the effects of load interaction has demonstrated the importance of defining a realistic loading spectrum for fatigue and damage-tolerance analysis and test. This has led to much effort being expended to quantitatively define how the variation of key spectrum parameters affect crack growth life (Ref 18). The key question remains how closely a design spectrum must simulate real life usage to accurately predict component life.

The first spectra were all block type. It should be noted that the most simplistic block spectrum would have all cycles at a single stress. More complex block spectra would have several different stresses with many cycles at each of the stresses. These block spectra may be stress increasing, stress decreasing, or have the stresses first increase and then decrease. Figure 8 shows these different types of spectra. There were several reasons for using the block spectra, including: it is easier to program the test equipment, and analytical fatigue techniques such as the Palmgren-Miner Rule did not recognize complex stress sequencing.

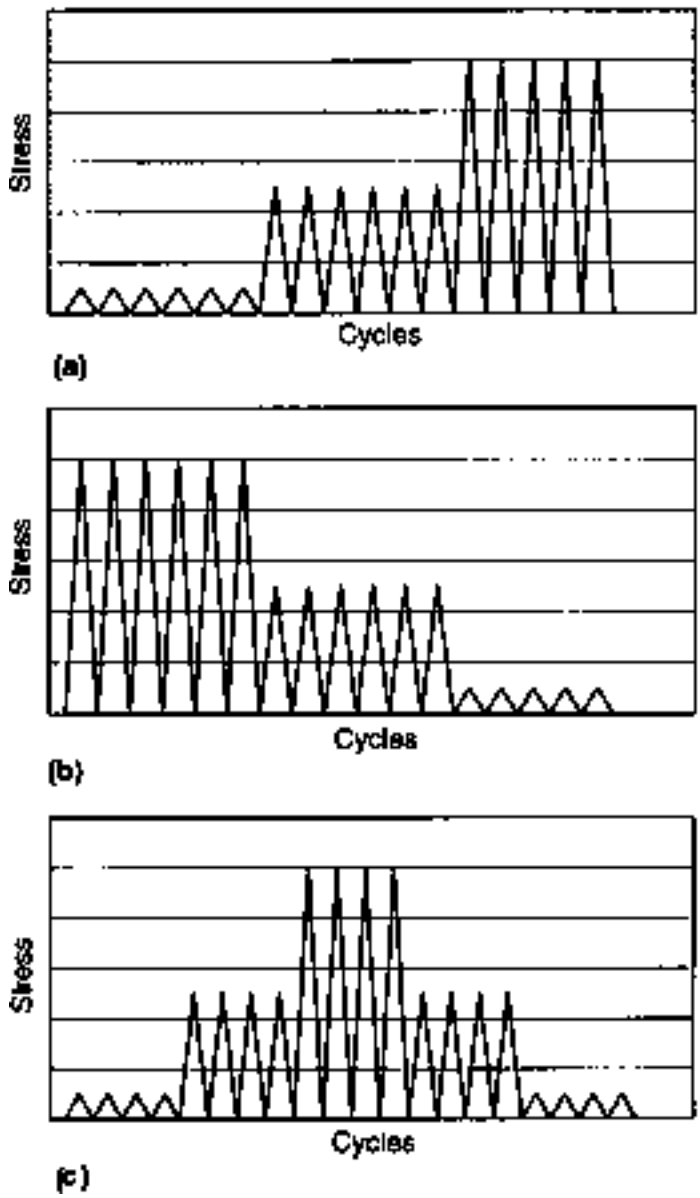


Fig. 8 Different types of block spectra. (a) Low-high. (b) High-low. (c) Low-high-low

Unfortunately, the use of the block spectrum did not accurately reflect the fleet experience, and it should be noted that the fleet experience indicated a shorter life. Therefore, it was necessary to go to a more complex test spectrum to improve the accuracy of the test results when compared with the lives experienced by the fleet. The more complex stress spectrum was the flight-by-flight spectrum. There are no equipment complexities: personal computers and the computers attached to the laboratory test equipment allow the investigator to use any type of spectra desired.

Problems then occur with the derivation of the stress spectra that best reflects actual use. To define a spectrum that is used to determine component life, care must be taken to ensure ease of analysis and test coupled with realism. For any component (aircraft, automobile, ship, or bridge), the type of loading(s) must be defined and quantified. In all these instances, there are two major types of loads—operator-induced loads and environmentally induced loads. Operator-induced loads may be maneuver loads for aircraft, automobiles, and ships, and transport loads (mass and velocity) for bridges. The environmental loads may be gust loads, wind loads, and wave motion for aircraft, automobiles, bridges, and ships, respectively. The operator-induced loads are determinate, while the environmental loads are probabilistic in nature. Since in many cases these loads occur simultaneously, combining them in a meaningful and consistent manner remains a challenge. It is generally left to the investigator to determine the means of accomplishing this combination.

To determine the fatigue-loading spectrum, it is imperative to know the anticipated use of the product. There may be regulations or historical precedents that determine the method of defining design life. In the case of aircraft, one can get the maneuver loads for the aircraft based on a load factor exceedance diagram. This diagram is shown in Fig. 9. From this diagram, one can determine the load factor of the aircraft and the number of times that aircraft will experience that load factor per 1000 h of flight. Statistical data also exist for gust loading. Data have been collected defining gust velocity as a function of altitude. Thus, if one knows the altitude of the aircraft and the time at that altitude, a joint probability analysis may be accomplished to determine the load factor per unit time. With that information, a method of combining the two individual input data may be accomplished.

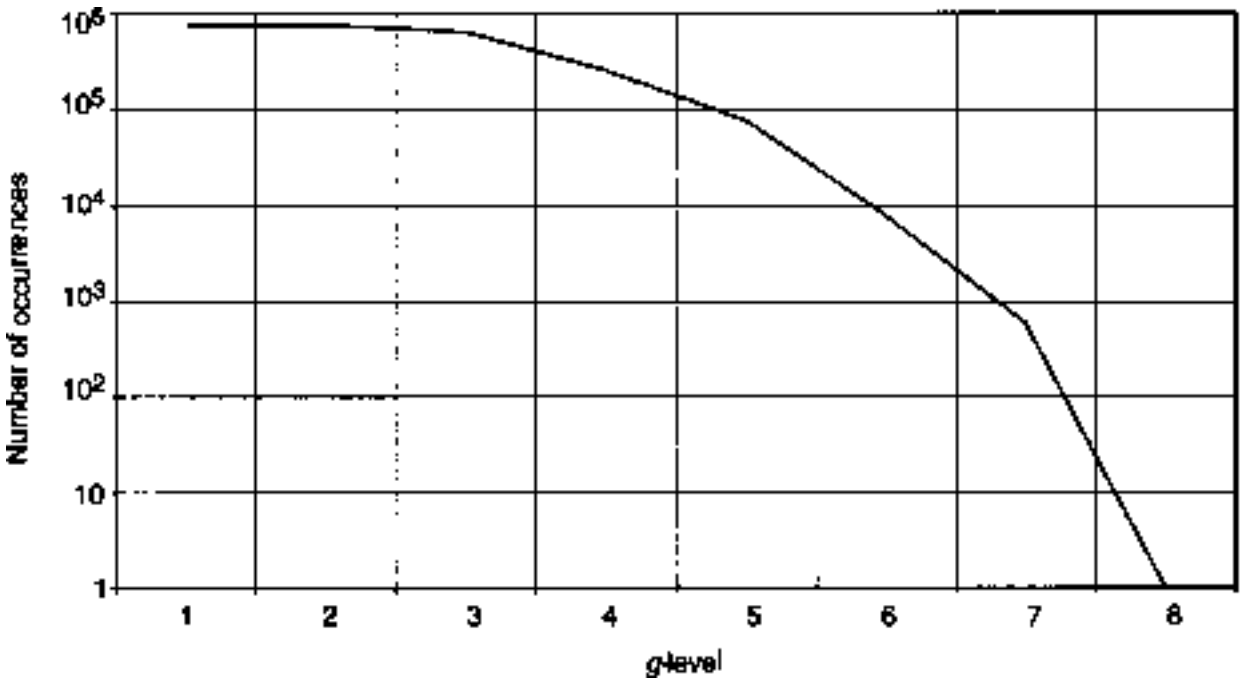


Fig. 9 Load factor exceedance diagram

Once one has obtained the load factor exceedance curves, one needs to change the load factor to a stress. In the case of an aircraft, the stress depends significantly on the weight of the aircraft—and that weight may be fuel, cargo, passenger, stores, and so forth. From the mission statement, one can determine the flight profile (velocity, altitude, weight, mission assignment) and the loads/stresses associated with them. Loads, hence stresses, for the 1 g and negative load factors are determined in a similar manner.

Then one must couple the stresses. The higher (or lower) load occurs for a finite time and then the load returns to its rest or 1 *g* condition. One can then get minimum and maximum points for each stress couplet. If these points are combined in a continuous manner, then a stress-time history can be obtained. After coupling the stresses, a cycle-counting procedure must be performed.

The techniques of crack growth or fatigue analysis demand that a spectrum be input as a series of whole cycles where a cycle consists of a maximum stress, minimum stress, and maximum stress. Thus, a whole cycle is defined as a load excursion with constant mean stress across it. The block-type spectrum consists of, by definition, constant mean stress cycles and thus can be used in the crack growth analysis directly. The random spectra, either flight by flight or mission segment by mission segment, do not consist of whole cycles, but of half cycles (maximum to minimum or minimum to maximum). Thus, it is necessary to use a cycle-counting technique that will redefine the random spectrum in terms of whole cycles with constant mean stress. The range-pair-range or rain-flow method couples the maximum positive stress with the maximum negative stress, the next highest maximum stress with the next lowest minimum stress, and so forth, continuing in this manner until all the stresses have been accommodated. Analysis indicates that there is a significant difference in life between using no counting method and either the rain-flow or range-pair-range method. The counted

spectrum gives a more conservative result (Ref 18). Furthermore, the results of these analyses demonstrate that when the spectrum is either rain-flow or range-pair-range counted, the calculated life agrees with the experimental result.

References cited in this section

2. *Metallic Materials and Elements for Aerospace Vehicle Structures,* MIL-HDBK-5F, Department of Defense, Nov 1990, p 3–104

3. N.H. Polakowski and E.J. Ripling, *Strength and Structure for Engineering Materials,* Prentice-Hall, 1966, p 490

4. G.E. Dieter, *Mechanical Metallurgy,* 3rd ed., McGraw-Hill, 1986, p 379

5. S. Suresh, *Fatigue of Materials,* Cambridge University Press, 1991, p 126–130

6. P.R. Abelkis, "Effect of Transport/Bomber Loads Spectrum on Crack Growth," AFFDL-TR-78-134, Air Force Flight Dynamics Laboratory and Air Force Wright Aeronautical Laboratories, 1978, p 20

7. A. Palmgren, Die Lebensdauer von Kugellagern (The Life Span of Ball Bearings), *Z. Ver. Deutsch. Ing.,* Vol 68, p 339–341

8. M.A. Miner, Cumulative Damage in Fatigue, *J. Appl. Mech.,* Vol 12, p 159–164

9. R.E. Peterson, *Stress Concentration Factors,* John Wiley & Sons, 1974, p 10, 11

10. M.R. Mitchell, Fundamentals of Modern Fatigue Analysis for Design, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 242

11. G.R. Irwin, *Handbuch der Physik,* Vol 6, p 551–590

12. D. Broek, *Elementary Engineering Fracture Mechanics,* Kluwer Academic Publishers, 4th rev. ed., 1991, p 77–81

13. *Damage Tolerant Design Handbook,* MCIC-HB-01, Part 1, Battelle Columbus Laboratories, 1975, p 5.1–21

14. R.G. Forman, V.E. Kearney, and R.M. Engle, Numerical Analysis of Crack Propagation in a Cyclic-Loaded Structure, *J. Basic Eng. (ASME Trans.),* Vol 89D, 1967, p 459

15. J.P. Gallagher et al., *USAF Damage Tolerant Design Handbook: Guidelines for the Analysis and Design of Damage Tolerant Aircraft Structures,* AFWAL-TR-82-3073, U.S. Air Force, May 1984

16. J.C. Newman, "Predicting Failures of Specimens with Either Surface Cracks or Corner Cracks at Holes," TN D-8244, NASA Langley Research Center, 1976

17. G.R. Irwin, Crack Extension for a Part Through Crack in a Plate, *J. Appl. Mech.,* Dec 1962, p 651–654

18. J.A. Reiman, M.A. Landy, and M.P. Kaplan, Effect of Spectrum Type on Fatigue Crack Growth Life, *Fatigue Crack Growth under Spectrum Loads,* STP 595, American Society for Testing and Materials, 1976, p 187–202

Fatigue-Life Assessment

Mitchell P. Kaplan and Timothy A. Wolff, EAD Aerospace, Inc.

Safe-Life versus Damage-Tolerance Approach

Safe-Life Approach. Life-assessment procedures have evolved over time, typically as the result of lessons learned from catastrophic structural failures. The safe-life approach to structural life assessment was developed in the 1950s in response to the loss of several commercial and military aircraft, even though they had been designed to meet static strength requirements. It soon became evident that the repeated loading of highly stressed structural details could, over time, lead to localized cracking and eventual fatigue failure. The safe-life approach attempts to maintain safety by designing structural components to have a fatigue life longer than their expected use or by replacing them long before they suffer failure. Component fatigue tests, which simulate the stress history expected in actual service, provide a representative sample of the number of stress cycles required for failure. This number is then divided by a suitable safety factor to obtain a replacement time for the component. For aircraft components, such as landing gear, control surfaces, or engine rotor blades, the replacement time is usually defined in terms of allowable flight hours or landings, after which the component must be discarded and replaced with a new one. Airframes that reach their "safe-life" are simply retired from service.

The safe-life approach successfully addresses many fatigue problems, but has several shortcomings. Fatigue testing is typically conducted on components that are in pristine condition. As a result, the component replacement time is based on the assumption that the component contains no flaws. If a component receives a defect due to manufacturing or maintenance, such as a material void or surface scratch, the life of the component could be much shorter than the replacement time indicated by fatigue testing. Under the safe-life approach, catastrophic failure is still an inherent possibility. From an economic standpoint, the safe-life approach is costly as well. Just as a component could have a fatigue life much shorter than the indicated replacement time, a component could also have a fatigue life that is much longer. Because component fatigue lives are divided by large safety factors to obtain their replacement time, many components are replaced even though they show no evidence of fatigue cracking.

Damage-Tolerance Approach. As the shortcomings of the safe-life approach became known in the early 1970s (Ref 19), the U.S. Air Force adopted a new method of defining component failure called damage tolerance, which redefined the basis for protecting aircraft from catastrophic structural failure. Rather than assess the life of a component based on empirical methods as in the safe-life approach, the damage-tolerance approach employs the rational methods of fracture mechanics. The fatigue life of a component is based on the calculated number of stress cycles required for an assumed initial crack to grow to its critical size. The damage-tolerance approach embraces the premise that every component is potentially flawed. Typically, these flaws are assumed to exist at important structural details, such as holes or sharp corners, where the accompanying high stresses accelerate crack growth. Using an appropriate crack growth rate equation, the initial flaw is grown mathematically through iterative calculations that take into account the expected stress history and the material crack growth properties and geometry of the component. The resulting life estimate is then divided by a safety factor (usually two or three) to obtain the inspection interval for the component. Dividing by a safety factor ensures that the component will be inspected at least once before a flaw could reach its critical size.

Inspections and reliable crack detection are a mandatory part of the damage-tolerance approach. A component cannot be damage tolerant without them. It is important that the selected inspection method be compatible with the chosen initial flaw size and the geometric details of the component. For instance, the smallest crack length that eddy-current inspection techniques are able to detect is approximately 0.1 mm (0.004 in.) (Ref 20). If the calculated crack size at the first inspection is less than this value, it is unlikely that a crack, if present, will be found. Therefore, the initial flaw size must be large enough to ensure the first inspection crack size will be detectable. It is just as important that a crack not be missed through inspector error. For this reason, first inspection crack sizes are also adjusted for the probability of detection. For eddy-current inspection, the smallest crack length that can be found reliably is approximately 1 mm (0.04 in.).

The damage-tolerance approach has several advantages over the safe-life approach. While safe-life components are replaced whether or not they contain cracks, components designed for damage tolerance are repaired or replaced only if a crack is found during an inspection. Also, because inspections are based on fracture mechanics models that describe the physical behavior of cracks, the hefty scatter factors characteristic of the safe-life approach are not needed. Combined, these advantages provide a strong economic incentive in favor of the damage-tolerance approach. From a safety standpoint, the damage-tolerance approach makes sense as well by offering protection against random or environmental damage sustained by a component over the course of its lifetime. In contrast, the safe-life approach assumes a component is undamaged from manufacture to replacement. However, the damage-tolerance approach is not appropriate for all circumstances. Highly stressed components manufactured from materials with low fracture toughness, such as those found in aircraft landing gear, typically have very short critical crack lengths. The resulting inspection intervals are also very short, making frequent inspection expensive and impractical. For these situations, the safe-life approach is used.

References cited in this section

19. *Aircraft Structural Integrity Requirements,* MIL-STD 1530, U.S. Air Force

20. S. Shanmugham and P.K. Liaw, Detection and Monitoring of Fatigue Cracks, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 210

Material Behavior

There are primarily three different metallic materials used in the design and manufacture of structural components: steel, aluminum, and titanium. Steel, the most commonly used structural alloy, is used in ships, buildings, bridges, automobiles, pressure vessels, and aircraft to name but a few. Aluminum is increasingly being used in automobiles and is the primary material used in aircraft. Titanium is used most sparingly, but has been used in several aircraft.

Why does a designer choose a particular material? There are many reasons, including material suitability, strength, corrosion resistance, stiffness, machinability, cost, toughness, fatigue resistance, weldability, weight, experience, and so forth. Usually the choice of the particular alloy (steel, aluminum, or titanium) can be made very rapidly. One would build neither a tank of aluminum nor an aircraft of steel. The usage generally defines the material. However, within each material group, the selection process is similar.

If one is to build a pressure vessel, one is reasonably sure it will be constructed from steel. Because of the stress state that occurs in a pressure vessel, the material must be welded. Additionally one would like a material that has high strength, yet high toughness. Additionally, the material must be fatigue resistant because it will be exposed to cyclic loading. Depending on what is contained within the pressure vessel, the sensitivity to a corrosive environment may play a large role. If the pressure vessel were to be used in a nuclear environment, the use of stainless steel becomes increasingly important. One can see how the design process and the usage environment narrow the material choices.

An axle would also be designed from steel. In this instance machinability, strength, toughness, fatigue resistance, notch insensitivity, and corrosion resistance would be important. Hardenability is also important. Axles are generally surface hardened. Because there are millions of axles made every year, cost cannot be ignored. Again, the material choice is narrowed due to the anticipated use of the material.

One could take examples from bridges, aircraft, nuclear reactors, offshore drilling platforms, and the like. The more knowledge and information one has at the beginning of the design process, the more quickly one can define the type of material to be used. This cannot be overstated. The designer must work with the engineer to define the anticipated use of the particular component. This knowledge allows for the selection of the appropriate material that can fulfill the function the component must serve. However, because one can define many aspects of the material, the selection of the particular material may become more complex. In addition, the chemistry and processing history of the material becomes extremely important.

Chemistry and processing history can include melting practices, remelting practices, additives to the melt for particular desirable properties, cleanliness of the material, interstitial content, rolling practices, heat treatment methods, forming methods, grain flow, surface hardening methods, plating practices, plating materials, surface finish, bend radii, welding, drilling, reaming, and so forth.

Materials engineers are asked to determine why materials fail. To answer that question, they must first determine how materials fail. They must continue backward on that path to understand how and why the particular alloy was selected and then how it was used. The design for that component assumed a particular usage. The materials engineers must determine if that assumption was accurate. They need to examine the assumptions that went into the anticipated use and then determine how close that assumption was to reality. They must examine the material as well. Was there some environmental effect that was not correctly anticipated, that is, filiform corrosion, galvanic corrosion, septic tank usage? Was the corrosion coating adequate? Did the usage include some environmental contamination that was not anticipated?

Other articles in this Volume address these questions. The point of this discussion is to demonstrate the need for these questions. The materials engineer must work as a member of a team with designers, stress analysts, aerodynamicists, chemists, and others who can illuminate some aspect of the problem. Other volumes in the *ASM Handbook* series give much information concerning these various concerns. However, there are other sources of information as well outside the disciplines covered in this series. The engineer must expose himself or herself to them also.

Retardation and Spectrum Load Effects

Retardation and spectrum load effects are based on the same phenomenon, the plasticity inherent in the specimen because of fatigue loading. If one assumes a crack, one has, in principle, an infinite stress riser. If one loads the component, then the stress in the area ahead of the crack increases dramatically. In a simplistic manner (plane stress), the stress cannot increase beyond the elastic limit (yield strength) until the entire area affected by the crack tip is plastically deformed. In a more complex case (plain strain), the stress can increase beyond the ultimate strength of the material.

Look first at the fatigue case. This is easier to understand and is directly applicable to that of fracture mechanics. The method is attributable to Neuber (Ref 21). If one assumes a typical fatigue specimen with a $K_t = 3$ and loads the specimen in tension, then the area adjacent to the stress concentration actually increases in stress three times as fast as the nominal stress. If the stress is increased beyond the yield strength of the material, plasticity occurs. If the load is now decreased, the stress will decease proportionally to the product of the load and the stress-concentration factor. Since plasticity has occurred, it is probable that the minimum load (minimum stress) is in the compressive zone. If cycling were to take place, then the maximum load would be low and the minimum load in compression. This would extend the life of the component. Figure 10 demonstrates this quite well. It should be noted that Fig. 10(a) shows the effect from a load perspective, while Fig. 10(b) and (c) show the same from a stress perspective. Thus, by having a high load introduced into the load regime, the life of the component may be extended.

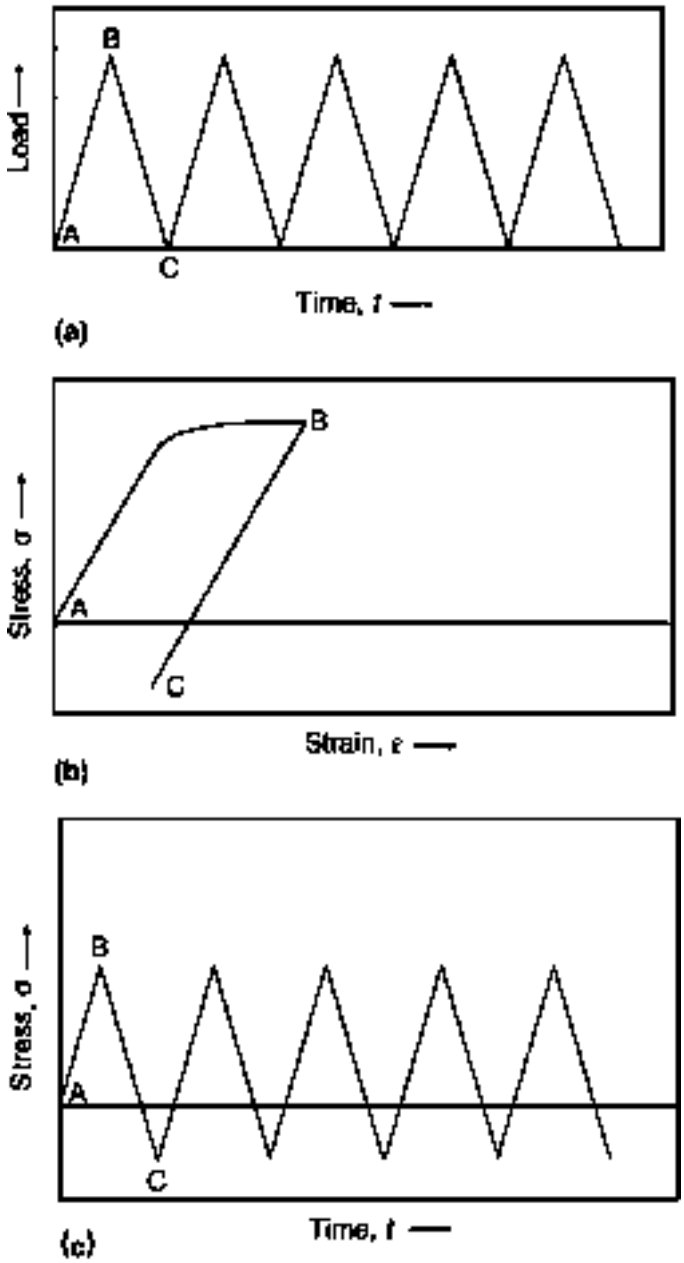


Fig. 10 Spectrum load effects. (a) Load versus time. (b) Stress versus strain. (c) Stress versus time

Retardation is similar to that discussed previously. In the case of fracture mechanics, there is an existing crack. Because the crack has an extremely small radius, the stress riser caused by the crack is significantly high. Taking this to its logical limit, if the radius at the crack tip were 0, that crack would generate an infinite stress riser. The stress just ahead of the crack can only increase at this high rate until the material reaches its yield strength. Then the plastic zone ahead of the crack would grow. The size of the plastic zone depends on the strength and the geometry of the material. If the material were soft (high ductility, large reduction of area), then the plastic zone would be larger. Additionally, if the geometry allowed for a plane stress condition, a large plastic zone would occur. Conversely, a plane strain condition would result in a smaller plastic zone.

As the component is cycled, the crack grows. There will always be a plastic zone ahead of the crack. However, for constant loading the crack can easily grow through the existing plastic zone. The size and the geometry of the plastic zone are determinate. Please refer to Dugdale (Ref 22) for one method of calculating the plastic zone size. When a high load is applied to the structure, the plastic zone is large. In all cases, the crack grows through the plastic zone. Because the material has been plastically deformed, there is a compressive residual stress in the zone when the crack is closed. This compressive residual stress must be accommodated when calculating the crack growth rate because the growth rate is dependent on the stress-intensity factor, the stress-intensity factor is dependent on the stress, and the stress is dependent on the compressive residual stress. As in the case of fatigue, the compressive residual stress may be calculated. The compressive residual stress is subtracted from both the maximum stress and the minimum stress. A new *R*-ratio is determined, and a new ΔK stress-intensity factor is defined. The crack growth rate is then determined from this ΔK stress-intensity factor and the crack growth determined. After the crack grows through the plastic zone, the effect of retardation on the component life is insignificant.

There are several different methods, that is, equations, to account for crack retardation. The most popular fall into two camps. One is the Generalized Willenborg Model (Ref 23) and the other uses the concept of a strip yield model and is based on crack closure (Ref 24). The underlying assumption is that cracks grow only when the crack is open; a determination must be made when the crack closes. This model accommodates this thesis. The Generalized Willenborg Model is an analytically defined method with one experimentally defined parameter. The equation is:

$$K_i^{eff} = K_i - K_R$$

where

$$K_R = \phi K_R^W = K_{max}^{oL}\sqrt{1 - \frac{a_i - a_{oL}}{r_{poL}}} - K_{max,i}$$

where a_i is the instantaneous crack size, a_{oL} is the crack size at the occurrence of the overload, r_{poL} is the yield zone that occurs as a result of the overload, K_{max}^{oL} is the maximum stress-intensity factor of the overload, $K_{max_i}^{oL}$ is the maximum stress intensity for the current cycle, and K_R^W is the difference in stress-intensity factor between that required to produce the plastic zone equal to (Z_{OL} - Δa) and the maximum applied stress-intensity factor.

As shown previously, $K_R = \phi K_R$ where K_R is the effective stress ratio and:

$$\phi = \frac{1 - (\Delta K_{th}/\Delta K)}{(R_{SO} - 1)}$$

In this equation, R_{SO} is the shut-off value of the stress ratio. This is an empirically determined material-dependent parameter and defines whether the crack will grow. If this value were 2.3 or greater (2000 series aluminum alloys) or 2.65 or greater (7000 series aluminum alloys), crack growth would cease because the crack growth program would preselect the stress-intensity factor to the threshold stress-intensity factor. It should be noted that at the threshold stress-intensity factor, crack growth ceases.

It can be seen from these equations that the stress-intensity factor is reduced as a function of the overload and the plastic zone. It should be noted that because of the overload, the maximum and minimum stresses are changed; the *R*-ratio is changed as well. The new *R*-ratio, R_{eff} is defined as:

$$R_{eff} = \frac{K_{min} - K_R}{K_{max} - K_R} = \frac{K_{min,eff}}{K_{max,eff}}$$

The plastic zone may be calculated through:

$$Z_{OL} = \frac{\pi}{8}\left(\frac{K_{max}}{\alpha_g \sigma_{ys}}\right)^2$$

where Z_{OL} is the plastic zone size, σ_{ys} is the yield stress, and α_g is a constraint factor.
The constraint factor (Ref 25) for one-dimensional crack models may be defined as:

$$\alpha_g = 1.15 + 1.4e^{-0.95(K_{max}/\sigma_{ys}\sqrt{t})}$$

where t is the thickness of the material.
For two-dimensional cases, the limit value for α_g is 1.15 (plane stress conditions) or 2.55 (plane strain conditions).
As stated earlier, another approach is that of crack closure. These methods are based on the work of Elber (Ref 26). He hypothesized that crack growth only occurred when the cracks were open; that is, a positive stress-intensity factor could be determined using crack opening displacement (COD) methods. He noted that during the crack growth cycle, the crack would close prior to attainment of the minimum load. As a result, ΔK would be lessened (becoming ΔK_{eff}), and the crack growth that occurred in the specimen would be slower. Many investigators came up with retardation schemes using this approach (Ref 27, 28).

References cited in this section

21. H. Neuber, Theory of Stress Concentrations for Shear Strained Prismatical Bodies with Arbitrary Non-Linear Stress Strain Law, *J. Appl. Mech.,* Vol 28, p 544–550

22. D.S. Dugdale, Yielding of Steel Sheets Containing Slits, *J. Mech. Phys. Sol,* Vol 8, 1960, p 100–108

23. J.P. Gallagher et al., “A Generalized Development of Yield Zone Models,” AFFDL-TM-74-28-FBR, Wright Patterson AFB Laboratory, 1974

24. H.J. Ten Hoeve and A.U. de Koning, “Implementation of the Improved Strip Yield Model into NASGRO Software—Architecture and Detailed Design Document,” NLR CR 95312L, National Aerospace Laboratory NLR, 1995

25. Fatigue Crack Growth Computer Program, NASGRO Version 3.0, JSC-22267B, National Aeronautics and Space Administration and Southwest Research Institute, 2001

26. W. Elber, *The Significance of Crack Closure,* STP 486, American Society of Testing and Materials, 1971, p 230–242

27. W. Elber, Mathematical Modeling of Crack Growth Interaction Effects, *Fatigue Crack Growth under Spectrum Loads,* STP 595, American Society for Testing and Materials, 1976, p 157–171

28. P.D. Bell and A. Wolfman, Equivalent Constant Amplitude Concept for Crack Growth under Spectrum Loading, *Fatigue Crack Growth under Spectrum Loads,* STP 595, American Society for Testing and Materials, 1976, p 236–250

Fatigue-Life Assessment

Mitchell P. Kaplan and Timothy A. Wolff, EAD Aerospace, Inc.

Crack Growth Software Packages

A large number of specialized software programs exist that provide fatigue and damage-tolerance life-assessment capabilities. Finite-element software packages also provide analysis modules with varying life-assessment capabilities. AFGROW (Ref 29) and NASGRO (Ref 30) are two full-featured fracture mechanics programs that have versions available to the public. Each program features an extensive material property database and stress-intensity solution library. Additionally, AFGROW, which is currently developed and maintained by the U.S. Air Force, has a refined user interface and very detailed on-line help topics. AFGROW may be obtained at no cost to the user, and it is regularly

updated by the developer. NASGRO, which is currently developed and maintained by NASA and Southwest Research Institute, is available for purchase.

References cited in this section

29. AFGROW software Web page, U.S. Air Force Research Laboratory, http://fibec.flight.wpafb.af.mil/fibec/afgrow.html

30. NASGRO Fracture Mechanics and Fatigue Crack Growth Analysis Software Web site, Southwest Research Institute, http://www.nasgro.swri.org/

Fatigue-Life Assessment

Mitchell P. Kaplan and Timothy A. Wolff, EAD Aerospace, Inc.

Case Studies

Although two of the following simple examples of fatigue life assessment are taken from the aerospace industry, the overall analytical process is the same in other industries. The third example is not defined as a fatigue life assessment and is taken from the marine environment. It should be noted that this design is not particular to this environment, but discusses a design approach that resulted in premature failure. The first example, for a torque link bolt on a fixed nose landing gear, demonstrates the safe-life approach, which employs the stress-life method of fatigue life assessment. The second example, for a circular penetration of a pressurized fuselage, describes the damage-tolerance approach. The third example is that of static loading and is based on an actual incident.

Example 1: Assessment of a Torque Link Bolt on a Fixed-Nose Landing Gear. This example demonstrates the safe-life approach.

Step 1: Develop the Stress Spectrum. The torque link bolt is loaded only when the aircraft nose wheel turns. Thus, the significant loading on the torque link bolt occurs while the aircraft is taxiing before takeoff and after landing. The torque link bolt can encounter eight basic loading conditions during taxi: a turning condition, a nose wheel yaw condition, and six towing conditions. By using data acquisition equipment to record the strains on the torque link bolt for a typical taxi event, the maximum and minimum strains for each loading condition can be defined. These strains are then converted to stresses using transfer functions that take into account the component geometry. The variable-amplitude stress data must then be cycle counted and ordered into stress pairs to obtain the number of occurrences per flight. A simplified listing of the maximum and minimum stresses for each loading condition on the torque link bolt is shown in Table 2 along with the number of occurrences.

Table 2 Stress levels for the different loading conditions experienced by a torque link bolt (example 1)

Loading condition	Stress levels, ksi		Mean stress, ksi	Occurrences per flight	Occurrences per 50 flights
	Maximum	Minimum			
Turning	28	0	14	6	300
Wheel yaw	51	0	25.5	0.02	1
Towing 1	46	0	23	0.10	5
Towing 2	62	0	31	0.10	5
Towing 3	36	0	18	0.10	5
Towing 4	67	0	33.5	0.10	5
Towing 5	53	0	26.5	0.10	5
Towing 6	56	0	28	0.10	5

Step 2: Determine if Local Stress Effects are Significant. The torque link bolt has radial lubrication holes drilled into the shank that have a diameter of 0.157 in. These lubrication holes result in a stress concentration factor of 2.0 (Ref 9). The torque link bolt material is 4130 steel, which has an ultimate tensile strength of 180 ksi. The notch sensitivity is:

$$q = \frac{1}{1+(A_k/r)} = \frac{1}{1+(0.0025\ \text{in.}/0.0785\ \text{in.})}$$
$$= 0.97$$

where A_k is obtained from the empirical relationship:

$$A_k = \left(\frac{300}{S_u}\right)^{1.8} \times 10^{-3}\ \text{in.} = \left(\frac{300}{180}\right)^{1.8} \times 10^{-3}\ \text{in.}$$
$$= 0.0025\ \text{in.}$$

The resulting fatigue notch factor is:

$$K_f = q(K_t - 1) + 1 = 0.97(2.0 - 1) + 1 = 1.97$$

This value is not appreciably lower than the stress-concentration factor. Therefore, the value for the fatigue notch factor will be taken as 2.0.

Step 3: Interpret the S-N Data. Figure 11 shows *S-N* data for notched (K_t = 2.0) 4130 steel under axial loading. An equivalent stress equation (Ref 31) that fits the data is:

$$\log(N) = 17.1 - 6.49 \log (S_{eq})$$

where $S_{eq} = S_{max}(1 - R)^{0.86}$. Since R = 0 for all loading equations, $S_{eq} = S_{max}$. For the turning loading condition, S_{max} = 28 ksi. The number of fatigue cycles the torque link bolt can sustain at this stress level before failing is:

$$\log(N) = 17.1 - 6.49 \log (28)$$

where $N = 51 \times 10^6$ cycles. Since runout of the fatigue data occurs at 10×10^6 cycles, N should satisfy the condition:

$$N \leq N_{runout}$$

Therefore, for the turning loading condition, N will be set equal to 10×10^6 cycles. The calculated cycles to failure for each loading condition are presented in Table 3.

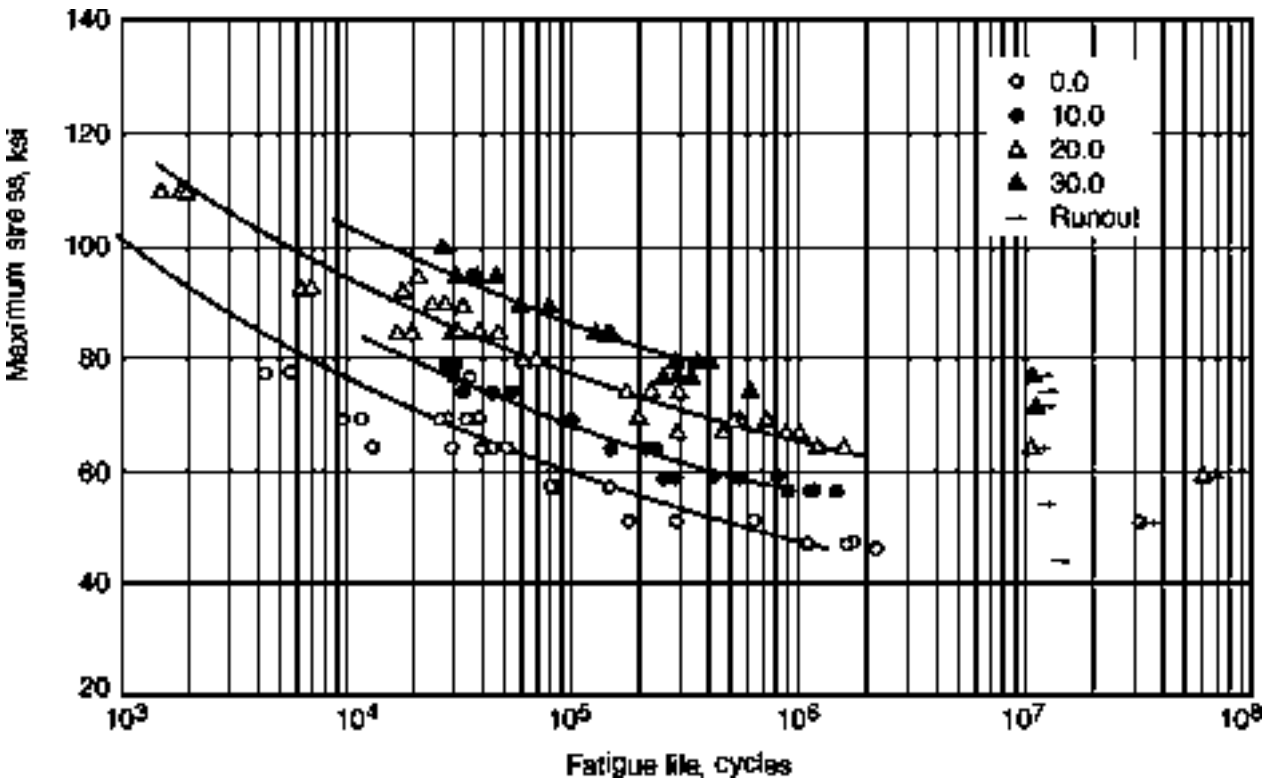


Fig. 11 *S-N* curve for notched (K_t = 2.0) 4130 alloy steel sheet. Stresses are based on net section. Source: Ref 31

Table 3 Calculated cycles to failure for the different loading conditions for a torque link bolt (example 1)

Loading condition	No. of cycles to failure
Turning	10,000,000
Wheel yaw	1,000,000
Towing 1	2,000,000
Towing 2	293,000
Towing 3	10,000,000
Towing 4	177,000
Towing 5	812,000
Towing 6	568,000

Step 4: Calculate the Cumulative Damage and Apply a Scatter Factor. The fatigue life of the torque link bolt is estimated using the Palmgren-Miner Rule:

$$\sum_{i=1}^{8} \frac{n_i}{N_i} = 1$$

The cumulative damage sustained by the torque link bolt per 50 flights is calculated, using the loading spectrum defined in Table 2:

$$\frac{300}{10{,}000{,}000} + \frac{1}{1{,}000{,}000} + \frac{5}{2{,}000{,}000} + \frac{5}{293{,}000} + \frac{5}{10{,}000{,}000} + \frac{5}{177{,}000} + \frac{5}{812{,}000} + \frac{5}{568{,}000} = 9.427 \times 10^{-5}$$

The inverse value of the cumulative damage provides an estimate of the number of loading spectra that can be sustained before failure:

$$\frac{1}{9.427 \times 10^{-5}} = 10{,}607$$

10,607 × (50 flights/spectrum) = 530,350 flights before failure

Applying a typical safe-life scatter factor of four, the torque link bolt would have a fatigue life of 132,587 flights. After this number of flights, the torque link bolt would have to be replaced. In practice, the nose landing gear components would probably all be replaced at the same time, determined by the component with the shortest estimated life.

Example 2: Assessment of Circular Penetration in a Pressurized Fuselage. This example demonstrates the damage-tolerance approach.

Step 1: Develop the Stress Spectrum. The pressure differential acting on the fuselage of an aircraft varies with altitude. At sea level, the pressure differential is 0. At higher altitudes, the pressure differential increases in order to maintain a comfortable atmosphere for the passengers and crew. An aircraft is generally pressurized to an altitude of 2400 m (8000 ft). Of course, every pressurized fuselage is designed for a limit pressure differential that is maintained by an overflow valve. While the altitude of an aircraft varies during a typical flight, resulting in a corresponding variance in the pressure differential, it is conservative to assume that a typical flight consists of a single pressurization cycle. Figure 12 displays a series of flights that each consist of a constant-amplitude loading cycle where the pressure differential is assumed to be 0 at takeoff and to reach its maximum value at maximum altitude. For a remote tensile hoop stress of 105 MPa (15 ksi) at maximum altitude and a minimum stress of 0 when the aircraft is on the ground, the stress ratio, *R*, will be 0.

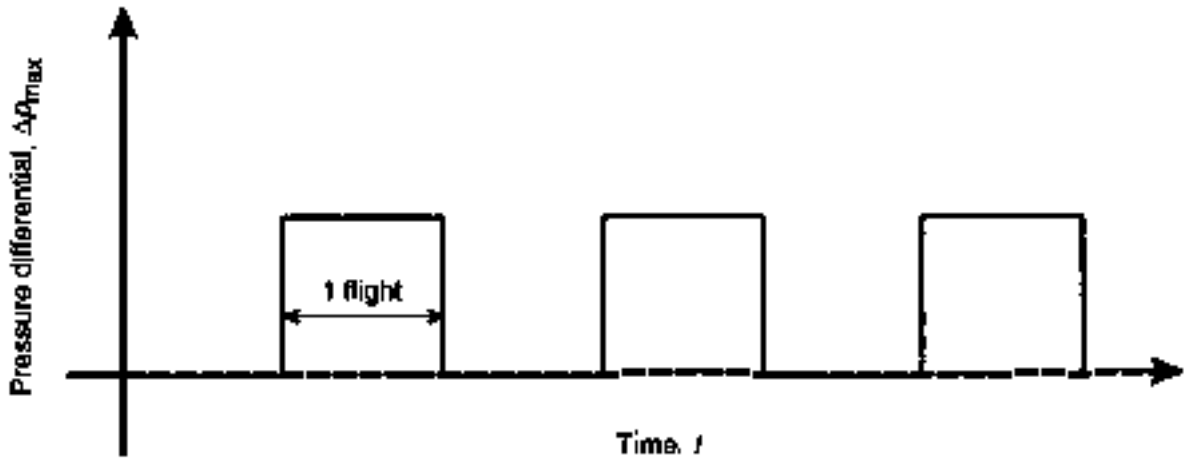


Fig. 12 Simplified fuselage pressurization spectrum (example 2)

Step 2: Choose an Appropriate Crack Geometry and Obtain the Stress-Intensity Factor Solution. The diameter of the penetration is 3.00 in. The chosen crack geometry is a radial through crack at the edge of the hole subject to an applied tensile stress. Since the fuselage is relatively large compared to the hole diameter, the following solution developed by Newman (Ref 16) is used:

$$\beta(a/r) = 0.707 - 0.18\lambda + 6.55\lambda^2 - 10.54\lambda^3 + 6.85\lambda^4$$

where

$$\lambda = \frac{1}{1+(a/r)}$$

The stress-intensity factor solution is:

$$K_1 = \beta(a/r)\sigma\sqrt{\pi a}$$

Step 3: Obtain da/dN Data and Choose an Appropriate Fatigue Crack Propagation Model. The Paris equation (Ref 32) is chosen to describe the crack growth since the *R*-ratio is a constant (i.e., constant-amplitude loading):

$$\frac{da}{dN} = C\Delta K^n$$

where *C* and *n* are material coefficients that describe a straight line on a logarithmic *da/dN* versus ΔK curve. Provided the crack growth rate data follow a relatively straight line over the expected stress-intensity range, the Paris equation will be an acceptable model. For an initial crack length of 0.05 in. and a remote stress of 15 ksi, the expected stress-intensity range is:

$$\lambda = \frac{1}{1+(a/r)} = \frac{1}{1+(0.05\text{ in.}/1.50\text{ in.})} = 0.9677$$

$$\beta(a/r) = 0.707 - 0.18(0.9677) + 6.55(0.9677)^2 - 10.54(0.9677)^3 + 6.85(0.9677)^4 = 3.122$$

$$\begin{aligned}\Delta K = \beta\sigma\sqrt{\pi a} &= (3.122) \\ &\quad \times (15{,}000\text{ psi})\sqrt{\pi(0.05\text{ in.})} \\ &= 18{,}560\text{ psi}\sqrt{\text{in.}}\end{aligned}$$

The fuselage skin is manufactured from 0.050 in. thick 2024-T3 aluminum. Figure 13 shows crack growth rate data for 2024-T3 aluminum at a stress ratio of $R = 0$. The material coefficients, *C* and *n*, for the Paris equation will be obtained from the equation of a line for the portion of the *da*/*dN* curve that corresponds to a stress-intensity range of roughly 20 $\text{ksi}\sqrt{\text{in.}}$:

$$\log\left(\frac{da}{dN}\right) = \log C + n\log\Delta K$$

Also note that the expected stress-intensity range is well above the stress-intensity threshold, ΔK_{th}, which is approximately 5 $\text{ksi}\sqrt{\text{in.}}$ as indicated by the data on Fig. 13. Therefore, crack growth will occur. If the stress-intensity

range were less than the threshold, no crack growth would occur under the assumed conditions, making further analysis unnecessary.

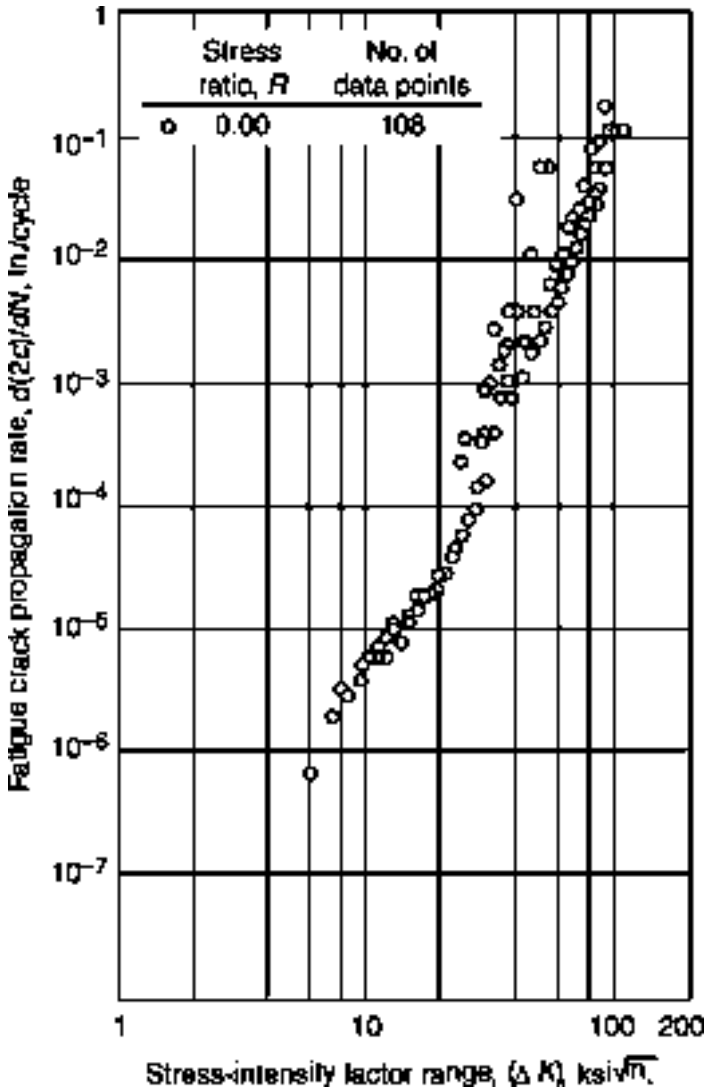


Fig. 13 Crack growth rate data for 2024-T3 aluminum at *R* = 0 (example 2). Source: Ref 33

Step 4: Obtain Data for Fracture Toughness. Fracture toughness data for 0.050 in. 2024-T3 aluminum sheet is shown in Table 1, where the average apparent fracture toughness, K_{app}, is shown as 99.47 $\text{ksi}\sqrt{\text{in.}}$ Since none of the values obtained for the critical fracture toughness, K_c, meet the validity requirements for the test, the value for K_{app} is conservatively used instead.

Step 5: Calculate the Critical Crack Length. The critical crack length is given by:

$$a_c = \frac{1}{\pi}\left(\frac{K_{app}}{\beta\sigma}\right)^2$$

Since β is a function of crack length, an iterative procedure must be used to obtain the critical crack length. The equation for the β factor converges to 0.707 as the crack length becomes large. Therefore, an initial β factor value of 0.707 is assumed. The resulting critical crack length is:

$$a_c = \frac{1}{\pi}\left(\frac{99.47\ \text{ksi}\sqrt{\text{in.}}}{(0.707)(15\ \text{ksi})}\right)^2 = 28\ \text{in.}$$

The β factor for this crack length is:

$$\lambda = \frac{1}{1+(a/r)} = \frac{1}{1+(28\ \text{in.}/1.50\ \text{in.})} = 0.05$$

$$\beta(a/r) = 0.707 - 0.18(0.05) + 6.55(0.05)^2 - 10.54(0.05)^3 + 6.85(0.05)^4 = 0.713$$

The critical crack length for a β factor of 0.713 is:

$$a_c = \frac{1}{\pi}\left(\frac{99.47\ \text{ksi}\sqrt{\text{in.}}}{(0.713)(15\ \text{ksi})}\right)^2 = 27.5\ \text{in.}$$

which is in good agreement with the initial guess. The critical crack length is taken as 27.5 in.

Step 6: Calculate the Number of Cycles to Failure and Determine Inspection Intervals. Table 4 displays the first few iterations of the Paris equation using values of $C = 4.328 \times 10^{-17}$ and $n = 2.873$. The calculated crack growth rate is added to the crack length after each cycle, which continues until either the critical crack length is reached or the loading spectrum has been repeated a predetermined number of times. The crack growth curve is shown in Fig. 14. The final result for the life estimate of the 3.00 in. fuselage penetration is 9250 pressurization cycles until failure. This value is divided by a factor of 2 to obtain the inspection interval of 4625 pressurization cycles. Dividing by a factor of 2 ensures that the penetration will be inspected at least once before failure could occur. The crack length at an inspection interval of 4625 cycles is 3.81 in., which is certainly detectable using an eddy-current or even visual inspection method.

Table 4 Initial iterations of crack growth calculations for a circular penetration in a pressurized fuselage (example 2)

Cycle	Crack size, in.	β	σ_{max}, ksi	σ_{min}, ksi	K_{max}, $\text{ksi}\sqrt{\text{in.}}$	K_{min}, $\text{ksi}\sqrt{\text{in.}}$	ΔK, $\text{ksi}\sqrt{\text{in.}}$	*da/dN*, in./cycle × 10^{-5}
0	0.050000	3.12247	15	0	18.563	0	18.563	7.945
1	0.050079	3.12209	15	0	18.576	0	18.576	7.961
2	0.050159	3.12171	15	0	18.588	0	18.588	7.976
3	0.050239	3.12133	15	0	18.600	0	18.600	7.992
4	0.050319	3.12094	15	0	18.613	0	18.613	8.007
5	0.050399	3.12056	15	0	18.626	0	18.626	8.023
6	0.050479	3.12017	15	0	18.638	0	18.638	8.038
7	0.50559	3.11978	15	0	18.651	0	18.651	8.054

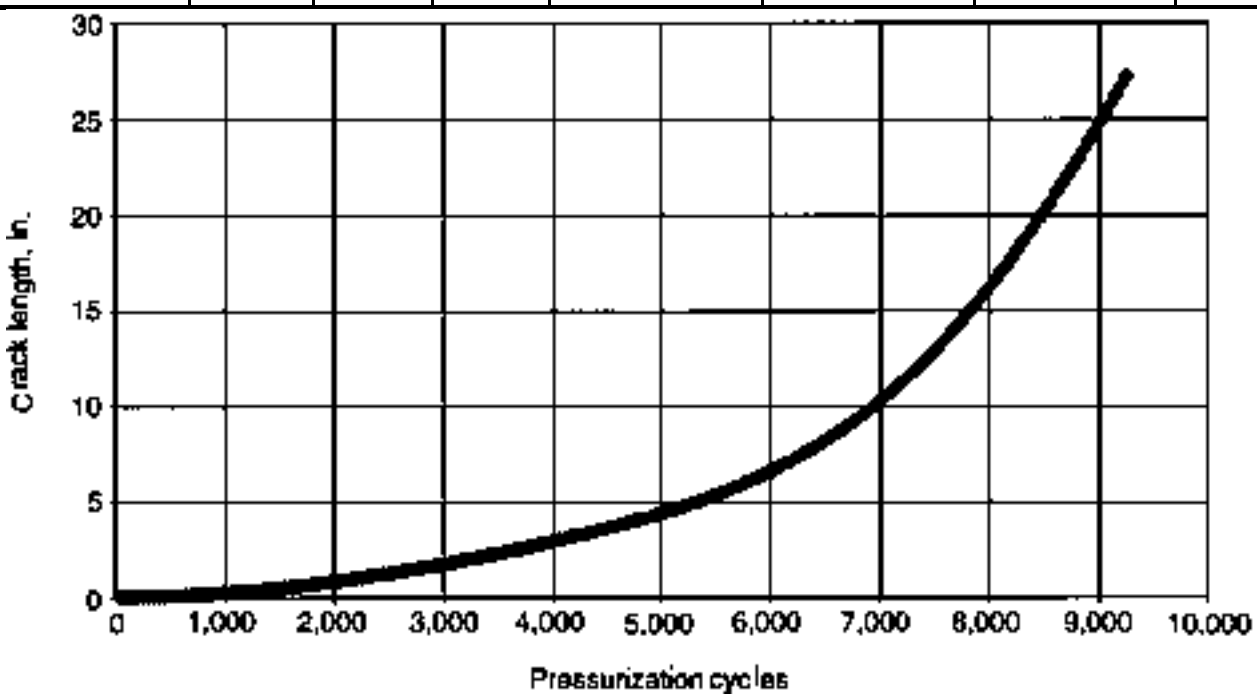


Fig. 14 Crack growth curve for fuselage penetration (example 2)

Conclusions from Examples 1 and 2. Examples 1 and 2 show how the effects of material behavior, component geometry, and stress history are systematically combined to arrive at an estimate of component life. Although the approach may differ, the combination of these effects is a recurring theme in fatigue life assessments. Both examples also demonstrate the similarities and differences between the safe-life and damage-tolerance methodologies. The results of the safe-life approach are very dependent on the availability and applicability of *S-N* data. If, for the torque link bolt, the notch factor and *R*-ratio had been 2.3 and 0.2, respectively, uncertainty in the final result would increase since the available *S-N* data

do not explicitly cover these values. To address this uncertainty, tests could be run to gather *S-N* data for the exact notch factor and *R*-ratio conditions or a larger scatter factor could be used. In contrast, the damage-tolerance approach, although dependent on crack growth rate data, is much more flexible in handling variability in stress ratio and geometry. The β factor solution for the fuselage penetration can easily be adapted to any size hole, provided the underlying assumptions of infinite width (i.e., no edge effects) and linear elastic behavior remain valid. Crack growth equations, like the Forman equation, offer similar flexibility with respect to the *R*-ratio.

Example 3: Design Analysis of a Pressure Vessel Hatch Cover. This particular example was adapted from Ref 34. A cement-carrying barge was unloading its cargo on the Mississippi River when a hole in the main discharge line developed allowing the cement to leak. The compressors were switched off, and gate valves were opened to reduce pressure. To increase the depressurization rate, several hatch covers were opened. These were opened by the workers kicking the locking cams or "dogs" with their feet. As a particular worker kicked the second locking cam open, the hatch cover failed with sufficient explosive force to throw both it and the worker into the water. The hatch cover was never recovered; however, the other hatch covers on the barge were examined. Two of the hatch covers that were examined had failed hinges while two others had their hinges intact. A sketch of the hatch cover may be seen in Fig. 15.

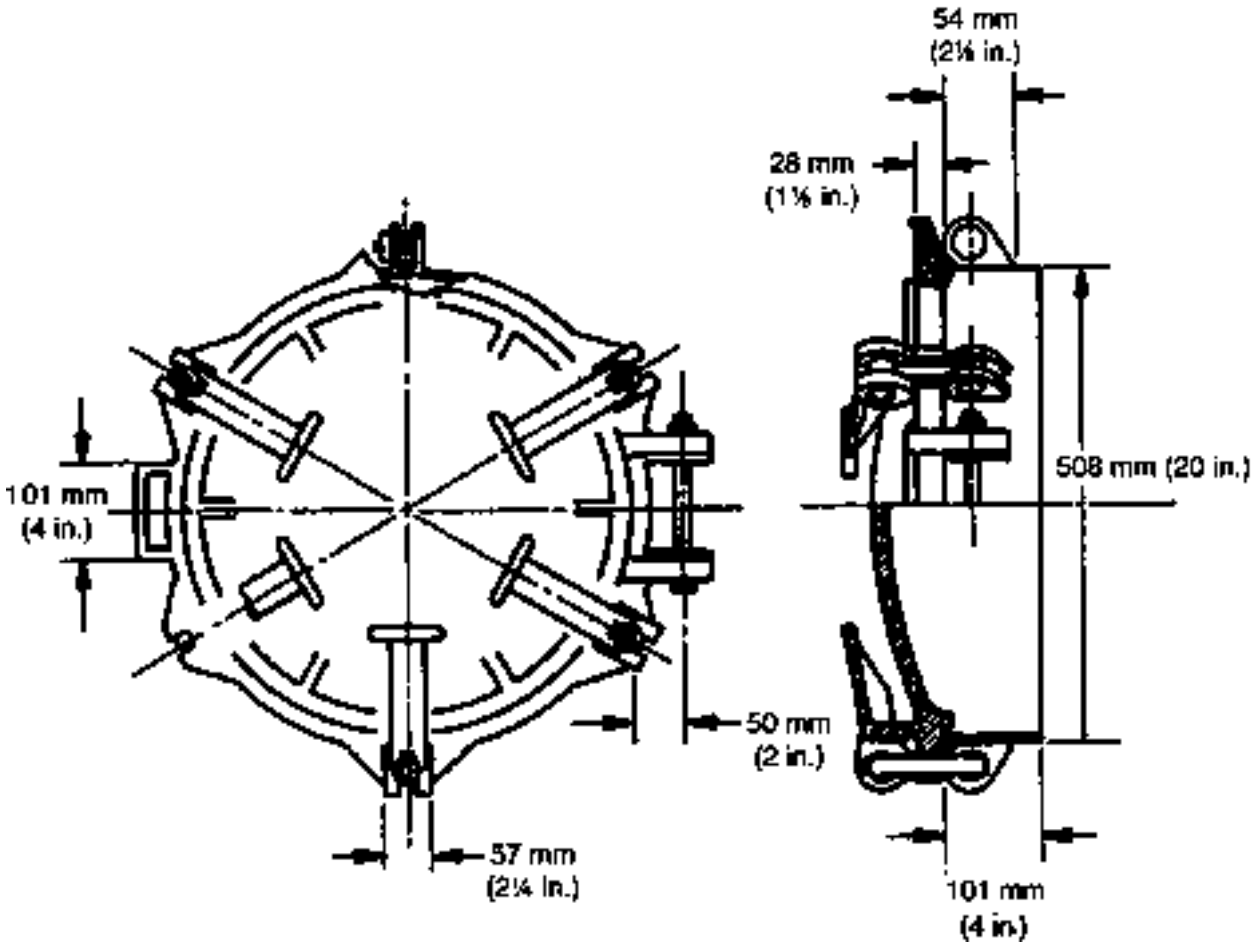


Fig. 15 Barge hatch cover (example 3)

The examination consisted of visual inspection (prior to and after cleaning), chemical analysis, manufacturing of tensile and fracture toughness specimens, testing of these specimens, a microstructural analysis, and an examination using a scanning electron microscope (SEM). In addition, a finite-element analysis of the hatch cover was accomplished. Finally, strain gages were placed on the hatch covers and the hatch covers loaded (hydrostatically pressurized) to correlate the analytical study.

The hatch covers were manufactured from A356-T6 aluminum and were 508 mm (20 in.) in diameter and 6.4 mm (0.25 in.) thick. There were six latches on each hatch cover and a neoprene seal (O-ring) to maintain pressure. Five tests were conducted (Table 5), four with four dogs attached and one with all dogs attached. The pressure was the independent variable.

Table 5 Test results for a pressure vessel hatch cover (example 3)

Run No.	Pressure		Locking cams	Hinge condition
	kPa	psi		
1	275.6	40	All attached	Not intact
2	275.6	40	Four attached	Intact
3	103.4	15	Four attached	Intact
4	275.6	40	Four attached	Not intact
5	103.4	15	Four attached	Not intact

The results of the analysis and tests agreed reasonably well. They showed that the stresses increased dramatically when two of the locking cams were opened. The stresses adjacent to the zones adjacent to the hinge were significant. Because this material is 356-T6 Al, it is relatively brittle. The fracture toughness was approximately $13.8\,\mathrm{MPa}\sqrt{m}$ $(12.5\,\mathrm{ksi}\sqrt{\mathrm{in.}})$. Microstructural studies showed a typical cast microstructure with shrinkage porosity. Examination with an SEM showed evidence of overload and porosity.

Regarding the design of the barge, it is known that cement is particularly abrasive and during aerated flow, it can pack itself tightly into crevices that hamper valve movements. A lack of pressure valves on the barge causes the employees to remain ignorant of the depressurization rate. This can lead to the workers trying to increase the speed at which depressurization occurs by kicking the dogs open.

The hatch covers themselves had some design concerns. They were mounted on the hatch in a fashion that they were never resting on the deck surface or in a protective cradle. When they were open, a load was imposed upon the hinge and the supporting structure. In this position, they were susceptible to damage from any traffic in the area. Of the 18 hatch covers on the barge, 11 had broken hinges. The design mechanism allowed the latch to be opened regardless of the pressure contained within. As stated earlier, if one of the latches were opened, the stress on the adjacent latch increased substantially. Figure 16 shows the stress intensity as a function of loading condition. It can be seen that if the hinge had failed and two dogs were kicked open, probability of failure was significant.

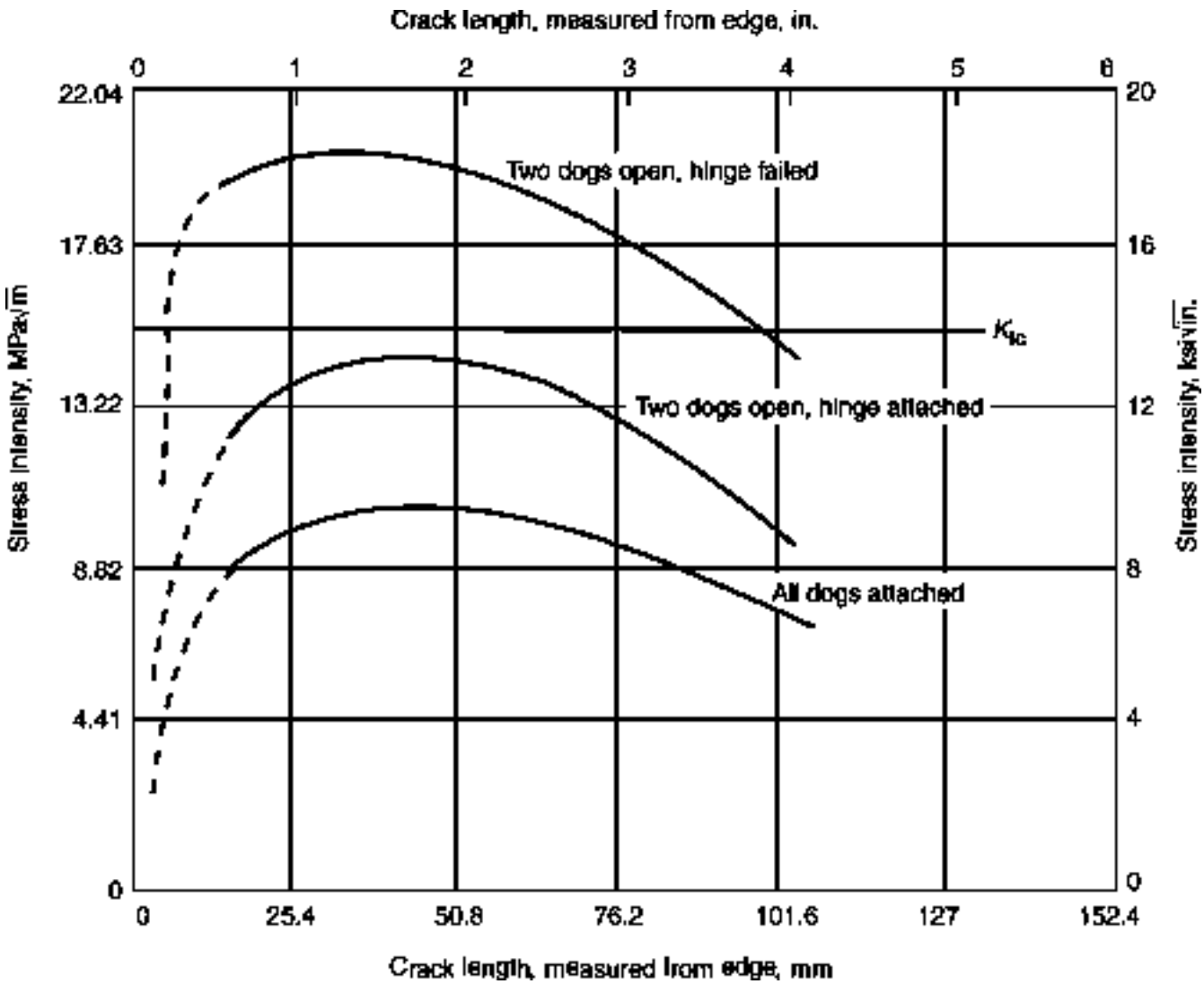


Fig. 16 Stress intensity versus loading condition for hatch cover failure (example 3)

The purpose of this example is to demonstrate how usage, material selection, system design, and component design all play a role in having a product that is useful and safe. Furthermore, if fracture mechanics methodology had been used, the choices may have been different. One is preprogrammed to believe that fracture mechanics methodology is used only for sophisticated design. This example shows the folly of that thought.

References cited in this section

9. R.E. Peterson, *Stress Concentration Factors,* John Wiley & Sons, 1974, p 10, 11

16. J.C. Newman, "Predicting Failures of Specimens with Either Surface Cracks or Corner Cracks at Holes," TN D-8244, NASA Langley Research Center, 1976

31. *Metallic Handbook and Elements for Aerospace Vehicle Structures,* MIL-HDBK-5F, Nov 1990, p 2–25, 2–29

32. P.C. Paris, The Fracture Mechanics Approach to Fatigue, *Fatigue—An Interdisciplinary Approach,* Syracuse University Press, 1964, p 459–464

33. *Damage Tolerant Design Handbook,* MCIC-HB-01, Part 2, Battelle Columbus Laboratories, 1975, p 8.1–14

34. M.P. Kaplan, T. Willis, and R.L. Barnett, A Pressure Vessel Hatch Cover Failure—A Design Analysis, *Case Histories Involving Fatigue and Fracture Analysis,* STP 918, American Society for Testing and Materials, 1986, p 46–64

References

1. *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1987

2. *Metallic Materials and Elements for Aerospace Vehicle Structures,* MIL-HDBK-5F, Department of Defense, Nov 1990, p 3–104

3. N.H. Polakowski and E.J. Ripling, *Strength and Structure for Engineering Materials,* Prentice-Hall, 1966, p 490

4. G.E. Dieter, *Mechanical Metallurgy,* 3rd ed., McGraw-Hill, 1986, p 379

5. S. Suresh, *Fatigue of Materials,* Cambridge University Press, 1991, p 126–130

6. P.R. Abelkis, "Effect of Transport/Bomber Loads Spectrum on Crack Growth," AFFDL-TR-78-134, Air Force Flight Dynamics Laboratory and Air Force Wright Aeronautical Laboratories, 1978, p 20

7. A. Palmgren, Die Lebensdauer von Kugellagern (The Life Span of Ball Bearings), *Z. Ver. Deutsch. Ing.,* Vol 68, p 339–341

8. M.A. Miner, Cumulative Damage in Fatigue, *J. Appl. Mech.,* Vol 12, p 159–164

9. R.E. Peterson, *Stress Concentration Factors,* John Wiley & Sons, 1974, p 10, 11

10. M.R. Mitchell, Fundamentals of Modern Fatigue Analysis for Design, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 242

11. G.R. Irwin, *Handbuch der Physik,* Vol 6, p 551–590

12. D. Broek, *Elementary Engineering Fracture Mechanics,* Kluwer Academic Publishers, 4th rev. ed., 1991, p 77–81

13. *Damage Tolerant Design Handbook,* MCIC-HB-01, Part 1, Battelle Columbus Laboratories, 1975, p 5.1–21

14. R.G. Forman, V.E. Kearney, and R.M. Engle, Numerical Analysis of Crack Propagation in a Cyclic-Loaded Structure, *J. Basic Eng. (ASME Trans.),* Vol 89D, 1967, p 459

15. J.P. Gallagher et al., *USAF Damage Tolerant Design Handbook: Guidelines for the Analysis and Design of Damage Tolerant Aircraft Structures,* AFWAL-TR-82-3073, U.S. Air Force, May 1984

16. J.C. Newman, "Predicting Failures of Specimens with Either Surface Cracks or Corner Cracks at Holes," TN D-8244, NASA Langley Research Center, 1976

17. G.R. Irwin, Crack Extension for a Part Through Crack in a Plate, *J. Appl. Mech.,* Dec 1962, p 651–654

18. J.A. Reiman, M.A. Landy, and M.P. Kaplan, Effect of Spectrum Type on Fatigue Crack Growth Life, *Fatigue Crack Growth under Spectrum Loads,* STP 595, American Society for Testing and Materials, 1976, p 187–202

19. *Aircraft Structural Integrity Requirements,* MIL-STD 1530, U.S. Air Force

20. S. Shanmugham and P.K. Liaw, Detection and Monitoring of Fatigue Cracks, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 210

21. H. Neuber, Theory of Stress Concentrations for Shear Strained Prismatical Bodies with Arbitrary Non-Linear Stress Strain Law, *J. Appl. Mech.,* Vol 28, p 544–550

22. D.S. Dugdale, Yielding of Steel Sheets Containing Slits, *J. Mech. Phys. Sol,* Vol 8, 1960, p 100–108

23. J.P. Gallagher et al., "A Generalized Development of Yield Zone Models," AFFDL-TM-74-28-FBR, Wright Patterson AFB Laboratory, 1974

24. H.J. Ten Hoeve and A.U. de Koning, "Implementation of the Improved Strip Yield Model into NASGRO Software—Architecture and Detailed Design Document," NLR CR 95312L, National Aerospace Laboratory NLR, 1995

25. Fatigue Crack Growth Computer Program, NASGRO Version 3.0, JSC-22267B, National Aeronautics and Space Administration and Southwest Research Institute, 2001

26. W. Elber, *The Significance of Crack Closure,* STP 486, American Society of Testing and Materials, 1971, p 230–242

27. W. Elber, Mathematical Modeling of Crack Growth Interaction Effects, *Fatigue Crack Growth under Spectrum Loads,* STP 595, American Society for Testing and Materials, 1976, p 157–171

28. P.D. Bell and A. Wolfman, Equivalent Constant Amplitude Concept for Crack Growth under Spectrum Loading, *Fatigue Crack Growth under Spectrum Loads,* STP 595, American Society for Testing and Materials, 1976, p 236–250

29. AFGROW software Web page, U.S. Air Force Research Laboratory, http://fibec.flight.wpafb.af.mil/fibec/afgrow.html

30. NASGRO Fracture Mechanics and Fatigue Crack Growth Analysis Software Web site, Southwest Research Institute, http://www.nasgro.swri.org/

31. *Metallic Handbook and Elements for Aerospace Vehicle Structures,* MIL-HDBK-5F, Nov 1990, p 2–25, 2–29

32. P.C. Paris, The Fracture Mechanics Approach to Fatigue, *Fatigue—An Interdisciplinary Approach,* Syracuse University Press, 1964, p 459–464

33. *Damage Tolerant Design Handbook,* MCIC-HB-01, Part 2, Battelle Columbus Laboratories, 1975, p 8.1–14

34. M.P. Kaplan, T. Willis, and R.L. Barnett, A Pressure Vessel Hatch Cover Failure—A Design Analysis, *Case Histories Involving Fatigue and Fracture Analysis,* STP 918, American Society for Testing and Materials, 1986, p 46–64

Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

Introduction

ELEVATED-TEMPERATURE FAILURE MECHANISMS and metallurgical instabilities reduce life or cause loss of function or operating time of high-temperature components. In addition, once a failure occurs from creep, fatigue, or an embrittlement degradation phenomenon at high temperature, the analysis team is often confronted with the question: How long will similar components last? Or, when should the next inspection be performed? To address these questions, life

assessment methodologies are often used. The failure analyst should be aware of some of the life assessment methodologies used to address these important operational and reliability questions.

Life assessment methods for elevated-temperature failure mechanisms and metallurgical instabilities that reduce life or cause loss of function or operating time of high-temperature components are described in this article. The review is intended for the purpose of informing failure analysts of methodologies and tools for the more common life-limiting, elevated-temperature failure mechanisms. Gas turbine blades and piping and tubing components were selected to demonstrate typical life assessment approaches for common elevated-temperature failure mechanisms, although the failure mechanisms and life assessment methods discussed in this article can be applied to other elevated-temperature components. The information in this article is taken from firsthand experiences and the available and referenced technical literature. The life assessment methods discussed include:

- Life fraction rule
- Parameter-based assessments
- Thermal-mechanical fatigue (TMF)
- Coating evaluations
- Hardness testing
- Microstructural evaluations
- Creep cavitation damage assessment
- Oxide-scale-based life prediction
- High-temperature crack growth methods

Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

Definition of Damage, Life, and Failure Criteria

In any component, the failure criteria need to be defined and established. Failure does not always involve fracture or rupture. Progressive damage of rotating blades and high-temperature components under operating conditions leads to exhaustion of life, thus leading to failure. Damage may be defined as a “progressive and cumulative change acting to degrade the structural performance of the load-bearing component or components that make up the plant” (Ref 1). Life may be defined as the “period during which a component can perform its intended function safely, reliably, and economically” (Ref 2). With some modifications, the definitions used in Ref 2 for fossil-power steam plant components can be used to define failure and life of elevated-temperature components; that is, component life is expended when:

- Design life has elapsed.
- Calculations predict life exhaustion.
- Service time has reached some arbitrarily chosen fraction of calculated or experimental failure life.
- Previous failure statistics indicate high probability of failure.
- Frequency of repair renders continued operation uneconomical.
- Nondestructive inspection reveals cracking.
- Surface degradation from corrosion, including coating degradation, is excessive.
- Grain-boundary attack and/or pitting by oxidation/hot corrosion is excessive.
- Foreign object damage is severe.
- Destructive sampling and testing indicate life exhaustion.
- Excessive deformation has occurred due to creep, causing distortion and unfavorable changes in clearances.
- Sudden and complete fracture occurs.

References cited in this section

1. L.F. Coffin, Damage Evaluation and Life Prediction for High-Temperature Gas Turbine Materials, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials*, EPRI AP-4477, Project 2382-3, Electric Power Research Institute, April 1986, p 1.1–1.17

2. R. Viswanathan and R.B. Dooley, Creep Life Assessment Techniques for Fossil Power Plant Boiler Pressure Parts, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials*, Electric Power Research Institute, April 1986, p 2.1–2.28

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

Life-Limiting Elevated-Temperature Failure Mechanisms

When conducting a failure investigation, it is important to identify the failure mechanism. Likewise, to perform a meaningful life assessment, it is important to know the failure mechanism. If the failure mechanism for crack initiation and crack propagation is not properly identified, then the life assessment is meaningless. Once the failure mechanism is identified, then it can be determined whether a life assessment can be performed. Life assessments are often conducted on time-dependent failure mechanisms. The principal types of elevated-temperature failures are stress rupture, creep, low- or high-cycle fatigue, thermal fatigue, and coating degradation in gas turbines. For high-temperature tubing and piping components, embrittlement phenomena can occur at high temperatures due to metallurgical instabilities; for example, carbide coarsening, sigma-phase formation, temper embrittlement, and so on can facilitate rapid brittle fracture at low temperatures during transient conditions. As a rule, the type of failure is established by examination of the fracture surfaces and comparison of operating conditions with available data on elevated-temperature fractures. In addition to fractographic evaluation, optical microscopy analysis may be required when conditions have been such that a change in metallurgical structure of the failed component has occurred. A brief description follows for some elevated-temperature failure mechanisms. A more complete definition can be found in this Volume under the appropriate failure mechanism.

Stress Rupture. One of the common mechanisms of failure for tubes and piping is stress rupture (Ref 3). Because an increase in either stress or temperature can reduce the time to rupture, attention must be given to both factors during investigation of a failure by a stress-rupture mechanism. Stress-rupture testing, which determines the time of failure at a specific stress and temperature, is often used to understand stress-rupture failures. The significant subcategories of stress-rupture failure mechanisms are short-term overheating and high-temperature creep.

The term *overheating failure* is often misused but generally means a failure resulting from operation of a component at a temperature higher than expected in design. Time at temperature is an important factor, and these types of failures are often called short-term and long-term overheating failures.

A short-term overheating failure is one in which a single incident or a small number of incidents exposes the component to an excessively high temperature (hundreds of degrees above normal), to the point where deformation or yielding occurs. Overheating results from abnormal conditions such as loss of coolant flow and excessive temperature excursion.

In general, short-term overheating failures involve considerable deformation in the form of metal elongation and reduction in wall area or cross section. Such failures often are characterized as having knife-edge fracture surfaces, in the case of tubes. Figure 1(a) (Ref 4) illustrates the elongation and deformation normally encountered with short-term overheating. For example, in tubes, wall thinning and local bulging precede the actual fracture, because the strength of the material is reduced at the higher temperature. A fishmouth appearance, with thin-edge fracture surfaces at considerable swelling, is typical for a ferrite steel tube that has failed before its temperature has exceeded the upper critical temperature (Ac_3). If, however, the tube temperature was high enough to transform the iron in the steel from ferrite to austenite, there is no noticeable “necking down,” or reduction, in wall thickness of the fracture edges. There is still metal elongation and tube swelling, so that measurement of the tube diameter shows an increase. A metallurgical analysis of the microstructure of the steel should be performed to confirm that the tube temperature prior to failure was high enough to transform the ferrite to austenite.

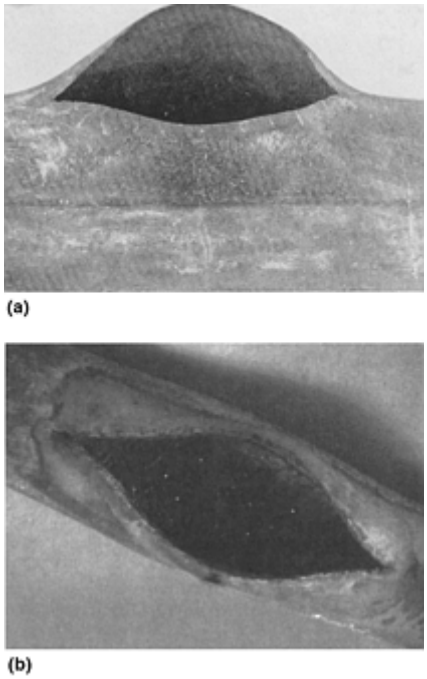


Fig. 1 Typical short-term overheating and long-term creep failures. (a) Typical thin-lip, short-term overheating failure of a 9.5 cm (3.75 in.) outside diam by 8.7 mm ($\frac{11}{32}$ in.) wall tube. Scaling caused the 13 cm (5 in.) knife-edge rupture. (b) Typical thick-lip, long-term creep failure

High-Temperature Creep. Failures can result from high-temperature creep. Metal degradation and permanent deformation occur with time, depending on the actual temperature and stress levels. If temperatures and stresses exceed design-selection values, the component exhibits a higher creep rate and fails earlier than expected. High-temperature-creep failures sometimes are called “long-term” or “extended” overheating failures. Such a failure results from a relatively continuous extended period of slight overheating (differential between design and actual operating temperatures), a slowly increasing level of temperature or stress, or accumulation from several periods of excessive overheating. The creep damage occurs along the grain boundaries of the material and is aligned 90° from the direction of applied tensile stress. Creep deformation results in little or no reduction in wall thickness but produces measurable creep elongation or increases in diameter in ferritic steel tubes.

For a tube, high-temperature creep usually results in a longitudinal fracture on the heated side of the tube. The extent of the fracture may vary and have different physical appearances. A small fracture forms a blister-type opening, whereas a large fracture exhibits a wide, gaping, fishmouth-type appearance. The fracture surface has thick edges or thick lips, because the creep damage creates linkups of individual voids and black oxide-filled cracks. Secondary cracking adjacent to the main fracture is extensive and is a positive indication of creep, although the absence of longitudinal cracks in the brittle iron oxide scale does not mean that creep swelling has not occurred. Failures of intermediate duration at moderately high temperatures exhibit some deformation and wall-thickness reduction. Figure 1(b) shows a high-temperature-creep failure due to overheating over a substantial length of the tube. The fishmouth opening exposes the thick-edge fracture surfaces.

Creep is time-dependent strain, which can cause failures when a component operating at high temperatures under steady load is so deformed or distorted that it can no longer perform its intended function. Creep deformation is typically characterized by time-dependent strain curves, as discussed in more detail in the article “Creep and Stress Rupture Failures” in this Volume. Most (but not all) creep curves show three distinct stages, as briefly described in the following. Creep strength (i.e., the resistance to creep deformation) of a metal is determined by a variety of factors, including composition, melting and deoxidation practice, grain size (inherent or actual), heat treatment, and loading magnitude.

Creep curves are used for life assessment evaluations, and the three stages are shown schematically in Fig. 2. Following initial elastic strain resulting from immediate effects of the load, there is a region of increasing plastic flow at decreasing strain rate (first-stage, or primary, creep). This first-stage creep represents an adjustment within the metal, which occurs in the first few moments after initial loading. The creep rate then decreases as crystallographic imperfections undergo realignment, which then leads to secondary creep.

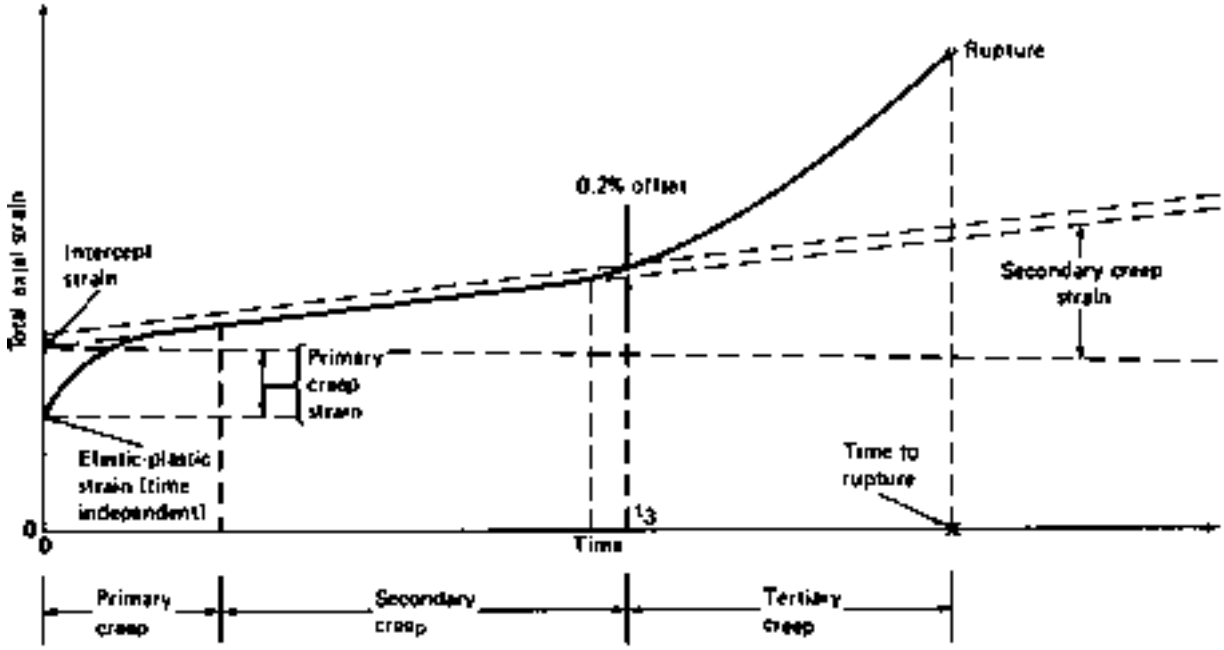


Fig. 2 Schematic tension-creep curve showing the three stages of creep

Next is a region of fairly constant strain rate (second-stage, secondary, or steady-state creep). Secondary, or second-stage, creep is an equilibrium condition between the mechanisms of work hardening and recovery. A constant but relatively low creep rate characterizes this stage. This constant or steady-state creep rate generally is the minimum creep rate and is widely used in engineering studies. Secondary creep is essentially a transition between primary and tertiary creep, even though it usually occupies the major portion of the time duration of a creep test. On a strain-time curve, this transition zone in creep-resistant materials is sufficiently flat, so that the minimum creep rate is applicable to virtually all of the secondary creep.

Finally there is a region of drastically increased strain rate, which rapidly extends to fracture (third-stage, or tertiary, creep). This third and final stage of creep (Fig. 2) refers to the region of increasing rate of extension, which is followed by fracture. Tertiary creep occurs when reduction in cross-sectional area resulting from cracking, necking, or both accelerates the creep rate. Environmental conditions that decrease the cross-sectional area may also increase the rate. Tertiary creep may also result from metallurgical changes, such as recrystallization under load, which promote rapid deformation, accompanied by work hardening that is insufficient to retard the increased flow of metal. Microscopic discontinuities, such as grain-boundary voids and cracks, develop and grow during tertiary creep. As these voids and cracks interact and join, stress rupture may occur (Ref 3), as described in more detail in the article “Creep and Stress Rupture Failures” in this Volume.

Elevated-Temperature Fatigue. In service, the steady loads or stresses to which components are subjected often are accompanied by mechanically induced cyclic loads responsible for failure by fatigue. The effect of temperature on fatigue strength is important, because fatigue strength and tensile strength decrease with increasing temperature. However, the precise relationship between temperature and fatigue strength varies widely. It is often dependent on the alloy, the temperature to which it is subject, and the duration of peak load. The variation of the fatigue crack growth as a function of temperature is shown in Fig. 3. For a more detailed description of this effect, see Ref 5.

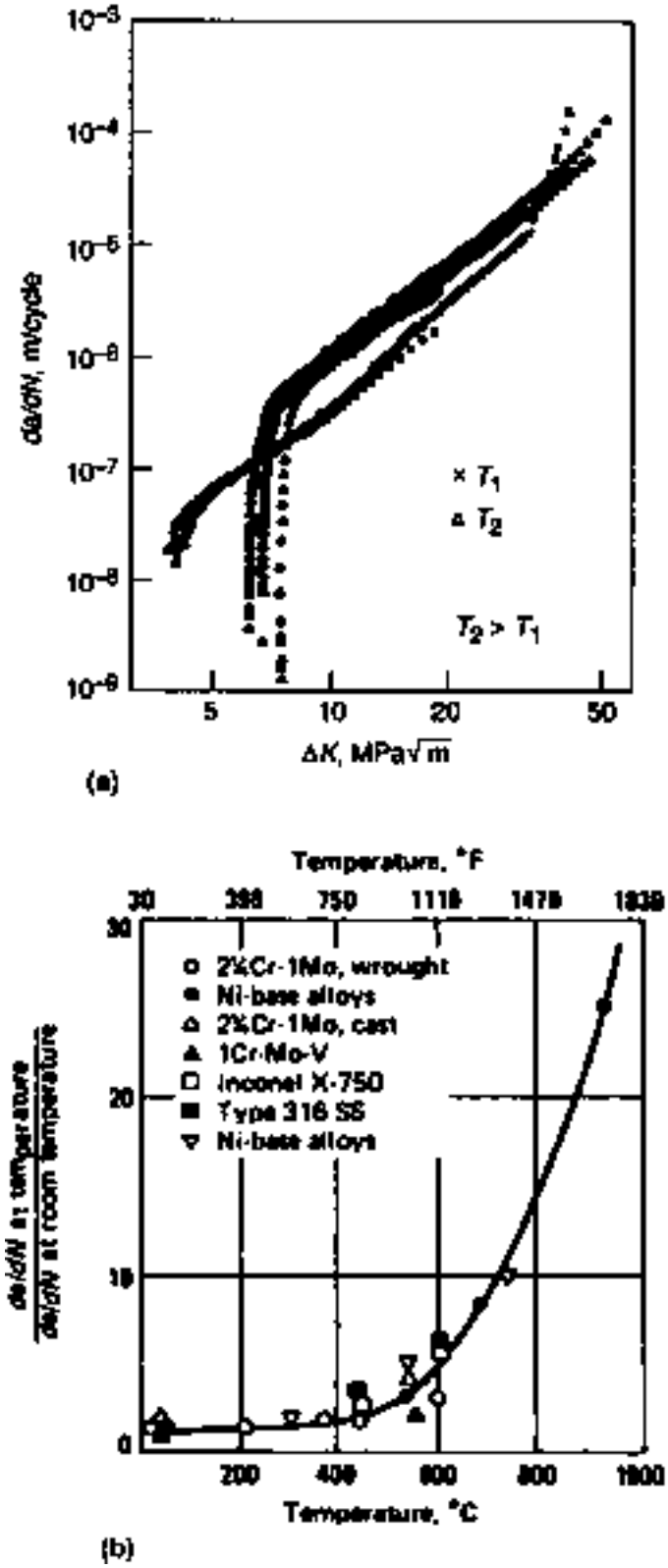


Fig. 3 Variation of the fatigue crack growth rate as a function of temperature. (a) Temperature effect on fatigue crack threshold and growth rates. (b) Variation of fatigue crack growth rates at ΔK of 30 MPa$\sqrt{m}$ (27 ksi$\sqrt{in}$)

Fatigue life assessment can be estimated using three different methods. The method selected depends on whether one is evaluating the fatigue crack initiation or crack propagation or even total fatigue life. For fatigue crack initiation, the alternating stress (*S*) versus the number of cycles (*N*) to fail is typically represented by an *S-N* curve. For fatigue life, accounting for the elastic and plastic behavior, strain-life approach may be considered. For fatigue crack propagation, the crack growth rate (*da/dN*) versus the stress-intensity value (ΔK), typically represented by a *da/dN* versus ΔK plot, is used. Later in this article is a discussion on consideration of elevated-temperature crack growth. Greater details on the ways to determine fatigue behavior are discussed in the article "Fatigue-Life Assessment" in this Volume and in Ref 6.

Thermal Fatigue. Mechanical conditions are not the only sources of cyclic loading, which can contribute to fatigue failure. Transient thermal gradients within a component can induce plastic strains; if these thermal gradients are applied repeatedly, the resulting cyclic strain can induce failure. This process is known as thermal fatigue. Thermal fatigue is often considered a low-cycle failure mechanism that occurs due to operating conditions. Thermal fatigue can be defined as the gradual deterioration and eventual cracking of a material by alternating heating and cooling, during which free thermal expansion is partially or fully constrained. The constraint of a part or material does not allow free expansion to occur.

For example, thermal strains are generated in the airfoils, because the outer surfaces change temperature more rapidly than the metal within. The effects of thermal fatigue may develop in airfoils after a relatively short service life (10,000 to 100,000 cycles). In turbine airfoils, creep strains are superimposed on cyclic thermal strains and thus account for a further reduction in life expectancy.

Thermal-fatigue cracks usually initiate at the surface, then propagate oriented perpendicular to the surface and the plane of maximum stress. They may occur singly or in multiples but often occur in multiples. Because the crack initiates externally, the amount of corrosion or oxidation along the surface of a thermal-fatigue crack is inversely proportional to the depth of the crack. Thermal-fatigue cracks usually progress transgranularly. However, in the presence of intergranular oxidation, which may help to initiate the cracks, TMF cracks tend to propagate along the grain boundaries. Examples of such cracks in land-based gas turbine blades are shown in Fig. 4. In TMF life prediction methods, the wave form, particularly the phasing between strain and temperature, is extremely important in determining life. The wave form is established by the temperature or strain versus the time. More information is presented later in this article.

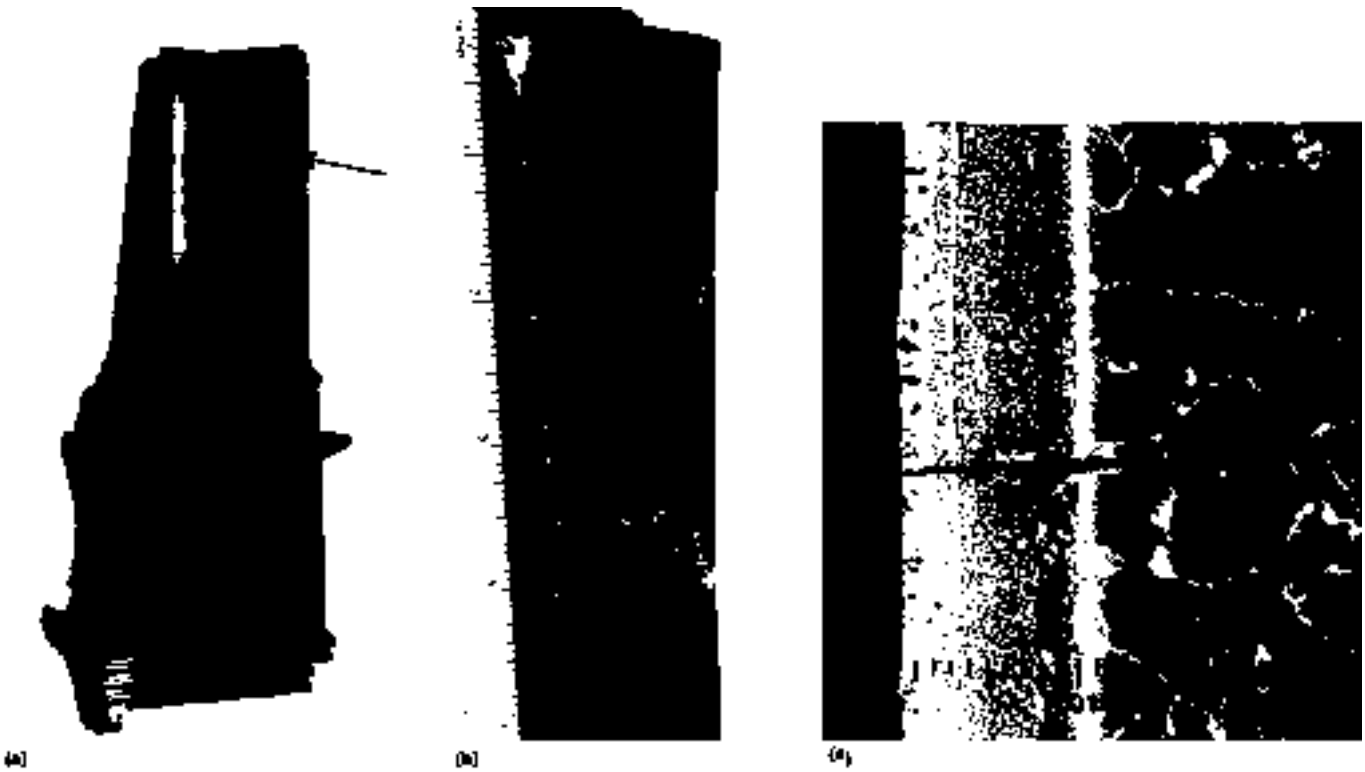


Fig. 4 Examples of thermal-mechanical fatigue cracking and oxidation in a first-stage turbine blade

References cited in this section

3. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components,* ASM International, 1989, p 188–190, 214, 228, and 369

4. D.J. Benac, *Elevated-Temperature Failures,* Lesson 7, ASM Practical Failure Analysis Course, 2002

5. A.F. Liu, High Temperature Life Assessment, *Fatigue and Fracture* Vol 19, *ASM Handbook,* ASM International, p 520–526

6. N.E. Dowling, Estimating Fatigue Life, *Fatigue and Fracture.* Vol 19, *ASM Handbook,* ASM International, 1996, p 250–262

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

Metallurgical Instabilities

Metallurgical instabilities form when superalloy turbine blades and steel tubes or piping are exposed to elevated temperatures for long times. Typical metallurgical instabilities for turbine blades include carbide coarsening, gamma-prime formation, and hot corrosion. For steel alloys used for tubes and piping, carbide spheroidization and coalescence, sigma-phase formation, sensitization, and temper embrittlement reduce the life expectancy of a component.

Steel-Based Alloys

Steel-based alloys, such as carbon steels, low-alloy steels, and stainless steels, are used at elevated-temperature applications for tubing and piping. The steel-based alloys perform well if operated within the limitations of the material. If steel is exposed at temperatures and stresses that exceed the normal recommended operating conditions, then metallurgical instabilities can occur that reduce the life of a component. A few of the more common instabilities are discussed briefly.

Carbide Changes. The gradual change of the carbide microstructure under prolonged heating may lead to a reduction of creep-strength properties of as much as 25% or more of carbon steel, carbon-molybdenum steel, and certain of the low-alloy chromium-molybdenum steels (Ref 7). The effect of temperature in changing the form of carbide in plain low-carbon steel is as follows. In low-carbon steel before service, the carbide is present as pearlite, which consists of alternate lamellae of iron carbide and ferrite. Following heating, the pearlite starts to spheroidize, sometimes known as divorce of pearlite, and continued heating causes complete spheroidization and finally coalescence into large particles of iron carbide. Still further heating may cause final reversion of the iron carbide to graphite, which may result in loss of impact and ductility and reduction of tensile strength properties. Figure 5 shows spheroidization and graphitization that can occur in carbon steel when exposed to heating at 540 °C (1000 °F). The addition of chromium together with other strong carbide formers, such as tungsten, vanadium, molybdenum, niobium, and titanium, tends to stabilize the carbide constituent so that it does not ultimately revert to graphite, but such carbides undergo progressive spheroidization and coalescence under suitable temperature environments.

Fig. 5 Representative microstructures of carbon steel tubes. (a) Lamellar pearlite of a tube before service. (b) Spheroidization of iron carbide (Fe_3C) in steel tube after exposure to long heating at 540 °C (1000 °F). (c) Graphitization that occurred in a carbon steel component

The change of the lamellae of iron carbide to altered carbide microstructures goes on quite slowly at temperatures below 540 °C (1000 °F) but occurs with increasing rapidity as temperatures increase above 540 °C (1000 °F), arriving at a

maximum just below the Ac_1 transformation temperature. At the Ac_1 temperature (which is approximately 705°C, or 1300 °F), carbon or carbide begins to go into solution in the ferrite.

Examination of the carbide microstructure, the correlation to stress-rupture test results, and hardness data can be useful in establishing the reduction of life in properties at the corresponding operating temperatures and stresses. For more details, see the section “Life Assessment Methods for Power Plant Piping and Tubing” in this article.

Sigma-Phase Formation. Sigma phase sometimes is found in austenitic stainless steel weld metal as well as in gas turbine nickel-base superalloys. Although sigma phase can occur in the as-welded condition in stainless steels, it is more commonly encountered after long exposures at temperatures of 595 to 870 °C (1100 to 1600 °F). Sigma is a brittle, intermetallic, chromium-rich phase that forms in austenitic stainless steels after prolonged exposure to temperatures in the 595 to 930 °C (1100 to 1700 °F) range. Only alloys containing more than approximately 16.5% Cr are affected. The effects of sigma are variable, depending in large part on the amount, size, and distribution of the phase present. Materials with appreciable sigma exhibit a loss in notch tensile strength, ductility, and impact toughness at room temperature. Operation of equipment with sigma at service conditions may not be a problem; however, to avoid failures during shutdown, it is imperative that procedures be in place to minimize thermal shock and differential thermal contraction stresses. Sigma phase may embrittle austenitic stainless steel at room temperature; however, it ordinarily is not a problem at operating temperatures above 650 °C (1200 °F). Figure 6 shows a crack between delta ferrite and sigma phase that occurred in type 347 stainless steel component (Ref 3). Gas turbine superalloys exhibit lower impact strength even at high temperatures, due to sigma-phase formation, and become highly susceptible to foreign object damage (FOD) and domestic object damage (DOD).

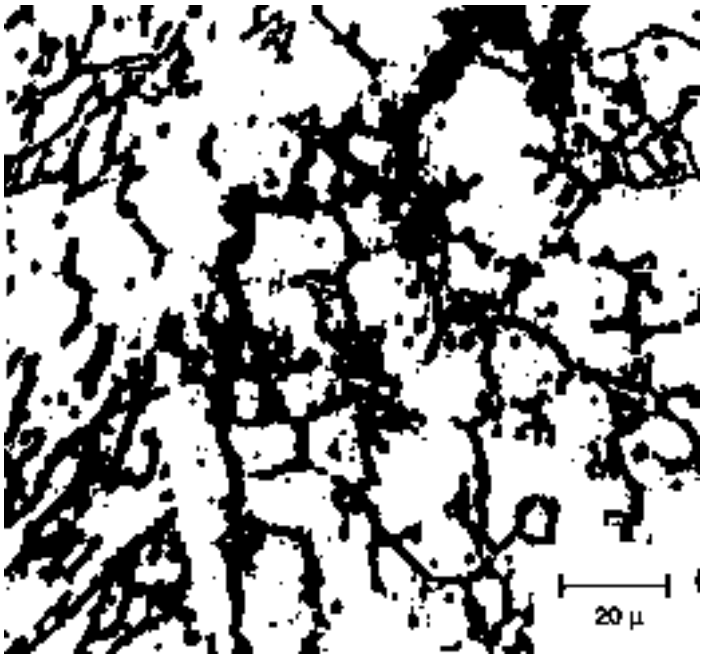


Fig. 6 Crack propagation through delta ferrite and sigma phases in type 347 stainless steel. Source: Ref 3

Sensitization. Certain austenitic stainless steel and higher-nickel alloys are susceptible to sensitization when held for long periods in the 480 to 870 °C (900 to 1600 °F) temperature range. Sensitization refers to the precipitation of a continuous grain-boundary network of chromium-rich carbides, rendering adjacent areas depleted of the necessary chromium to resist aqueous corrosion and high-temperature oxidation phenomenon. When elevated-temperature service in the sensitizing range is anticipated, stabilized or low-carbon grades of stainless steel are specified. The stabilized grades (type 321 or 347) work by adding strong carbide-forming elements, such as niobium or titanium, to the stainless steel chemistry. These stabilizers preferentially form carbides in place of chromium, preventing chromium depletion in service. The other approach is to reduce the carbon level of the stainless steel to minimize the degree of sensitization in service. Susceptible stainless steel, such as type 304 and type 316 austenitic grades, are exposed to sensitizing conditions in refineries. Problems often arise during down periods, when this equipment, containing sulfide scales, is exposed to steam (moisture) and air.

Temper embrittlement occurs predominantly in alloy steels subjected to a temperature range that results in embrittling from elements of phosphorus, tin, antimony, and arsenic in the steel. Consequently, the material becomes less ductile, and the ductile-to-brittle transition temperature shifts toward the highest temperatures. For example, studies conducted on 1CrMoV rotors that were exposed in the range of 370 to 425 °C (700 to 800 °F) showed the greatest loss of ductility and shift in the fracture transition temperature (Fig. 7).

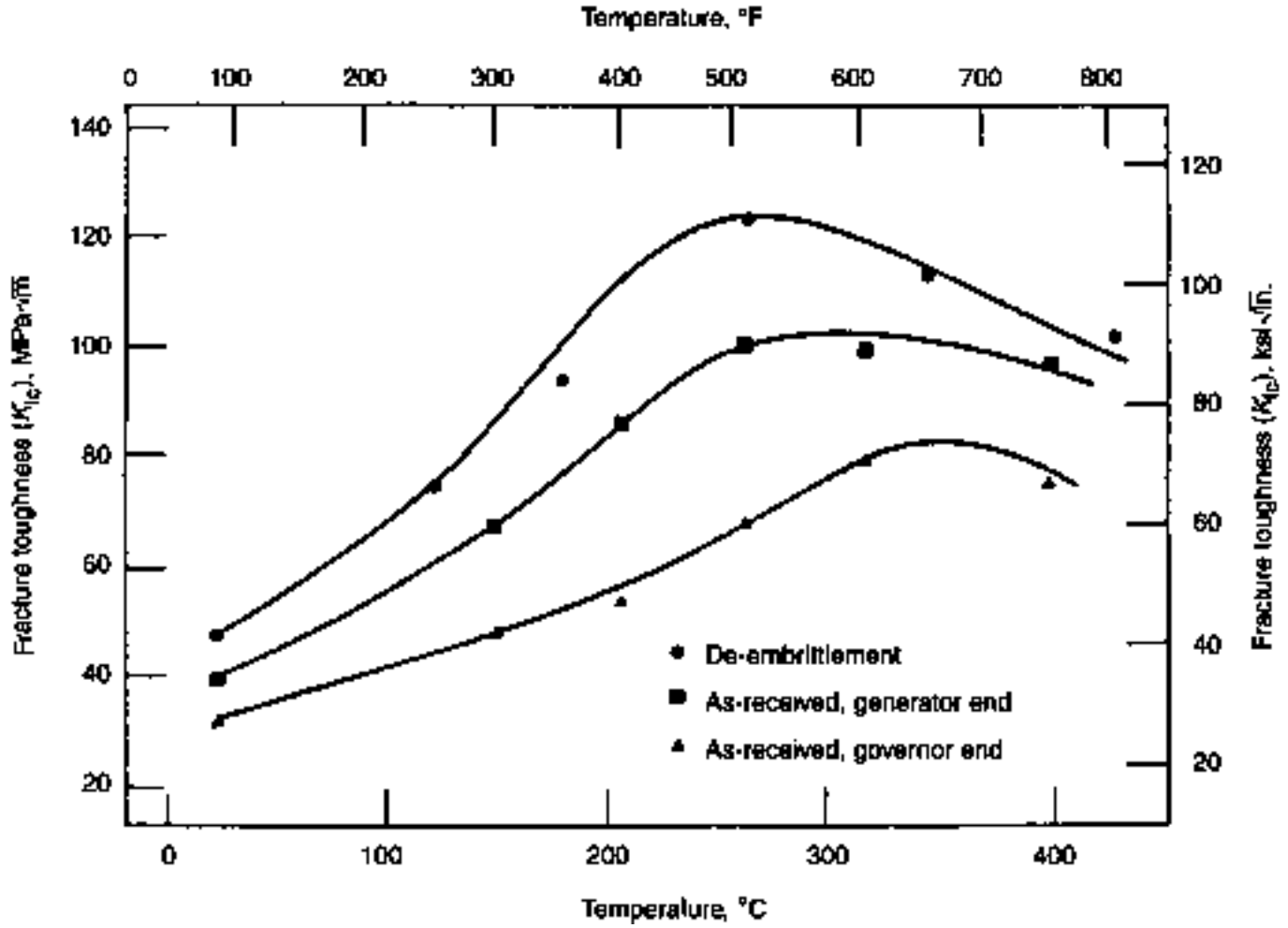


Fig. 7 Plot showing the effect of temper embrittlement on the fracture toughness of a 1CrMoV steel. Source: Ref 8

Hydrogen Attack. Although not a common occurrence in gas turbines, high-temperature exposure of carbon and low-alloy steels used for piping and pressure vessels to high-pressure hydrogen leads to a special form of degradation known as hydrogen attack. In contrast to hydrogen embrittlement, which degrades toughness at low temperatures and imposes restrictions on start-up and shutdown procedures, hydrogen attack leads to a degradation of material properties at the operating temperature.

Some equipment involves the use of, or production of, hydrogen at high pressures and at temperatures of 230 °C (450 °F) or above. These service conditions lead to the deterioration of metal components and result in failure, notably of pressure vessels and pipes. Under certain conditions of temperature and hydrogen partial pressure, atomic hydrogen permeates the steel and reduces iron carbide (Fe_3C) to form methane (CH_4). The methane does not diffuse from the metal, and its pressure may exceed the cohesive strength of the metal and cause fissuring between grains. When fissuring occurs, the ductility of the metal is significantly and permanently lowered. The severity of hydrogen attack increases with increasing temperature and hydrogen partial pressure.

Usually, hydrogen attack occurs in three stages: (1) atomic hydrogen diffuses into the metal, (2) decarburization occurs (in steel), and (3) intergranular fissuring occurs. A metal that has undergone only the first stage of hydrogen attack suffers a loss in ductility that is considered temporary, because ductility can be restored by heating. However, when the internal pressure generated by the entrapped methane exceeds the strength of the metal and results in formation of the fissure, permanent, irreversible embrittlement results. Consequently, the term *hydrogen attack* denotes the permanent embrittlement that occurs in the second and third stages.

Hydrogen attack is a decarburization reaction. If the attack is confined to the surface, it is known as surface attack. If the attack occurs internally, the resultant product—that is, methane—is unable to escape, forms bubbles, and leads to permanent internal damage. The methane bubbles nucleate as carbides grow under methane pressure and link up to form fissures and cracks, as shown in Fig. 8 (Ref 4). Such internal hydrogen cracks occurring in large steam turbine forgings are known as “hydrogen flakes” due to their cornflakelike appearance.

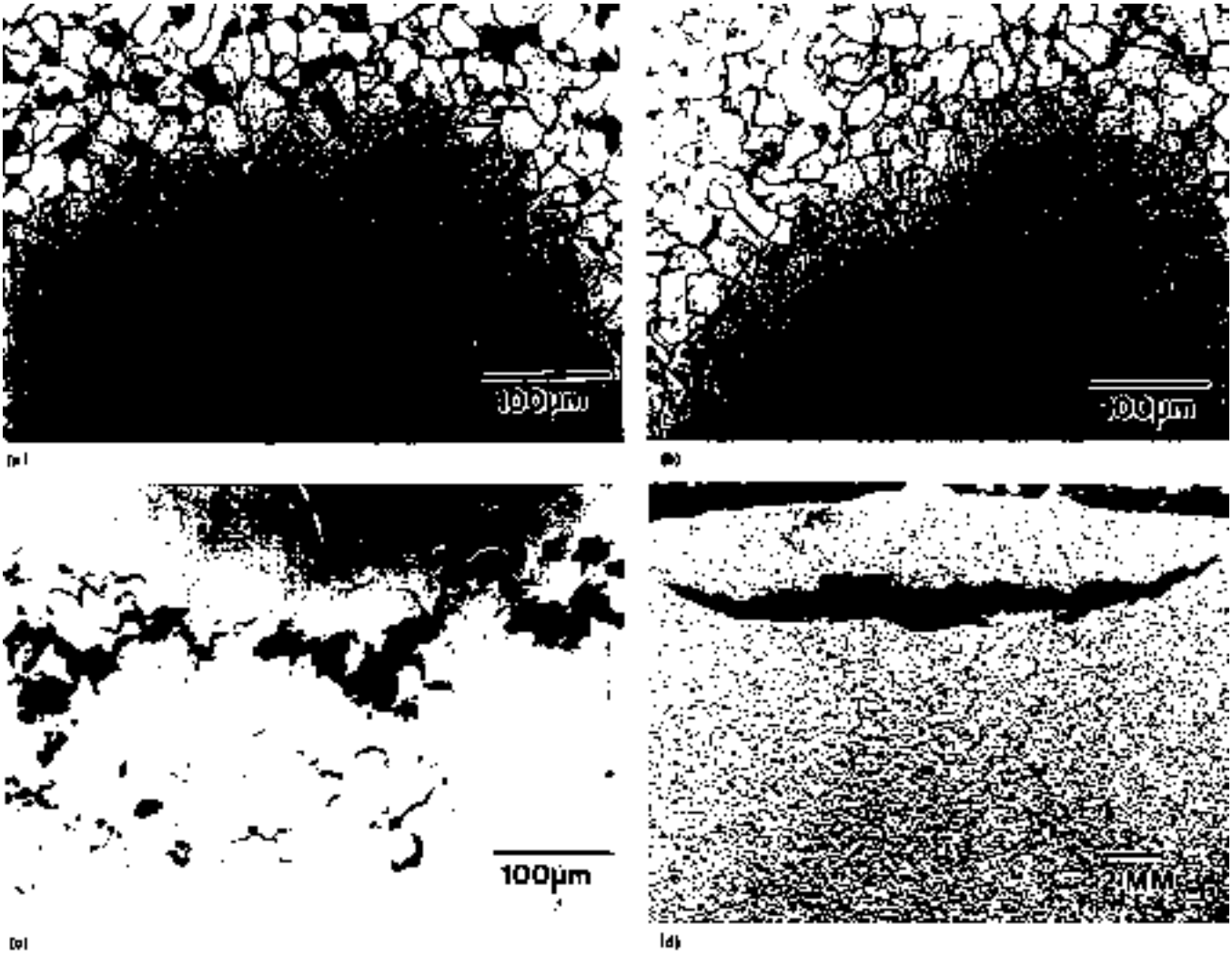


Fig. 8 Hydrogen-damaged refinery platformer line (carbon steel, 0.5% Mo). (a) Undamaged microstructure. (b) Decarburization region caused by hydrogen depleting the iron carbides. (c) Microfissuring at inclusions. (d) Hydrogen blister caused by methane gas formation. (a) and (b), nital etch. (c) and (d), unetched

Nickel-Base Superalloys

To conduct life assessment of turbine blades, it is important to understand gas turbine blade metallurgy (Ref 9). This section provides a general overview of nickel-base superalloys used for rotating turbine blades in industrial gas turbines. Several detailed reviews of the subject are available (Ref 10, 11, 12) and so the topic is discussed only briefly here, with specific reference to those features, that are changed by service exposure.

During extended service, the high temperatures and stresses and exposure to operating environments result in changes to the microstructure and surface composition of the alloy. Because these features influence the material properties, corresponding changes in strength and oxidation/corrosion resistance occur. Typical nickel alloy metallurgical instabilities include gamma-prime overaging, carbide overaging, creep damage, oxidation/hot corrosion damage, and coating degradation.

Strengthening Mechanisms. Almost all of the alloys used in industrial gas turbine blades belong to the family of precipitation-hardened alloys. High creep, tensile, and fatigue strengths are achieved in superalloys by a variety of strengthening mechanisms. The most important mechanisms, both in terms of their contribution to alloy strength and in terms of changes occurring during service, are the effects of gamma prime and carbide precipitation. Other mechanisms include solid-solution strengthening effects in the matrix and grain-boundary strengthening effects.

Gamma-Prime Precipitation. The principal strengthening mechanism in nickel-base blade alloys is the precipitation of gamma prime. The common microstructural feature in these alloys is the presence of coherent precipitates of gamma-prime phase in austenitic nickel-chromium matrix. The gamma-prime phase is a face-centered cubic compound with a composition A_3B, where typically A is nickel and, to a lesser extent, cobalt; and B is aluminum and, to a lesser extent, titanium, niobium, and tantalum. The precipitated gamma-prime particles act to obstruct dislocation motion and therefore

provide resistance to deformation. The strengthening gamma-prime particle size and morphology in nickel-base superalloys is not stable over long periods at engine operating temperatures (Ref 13). After heat treatment, the gamma-prime phase is very nearly at equilibrium with the matrix, and therefore, little additional precipitation or growth of gamma prime occurs from supersaturation of the matrix. Gamma-prime overaging, shown in Fig. 9, can occur, which compromises the properties and reduces the life of turbine blades.

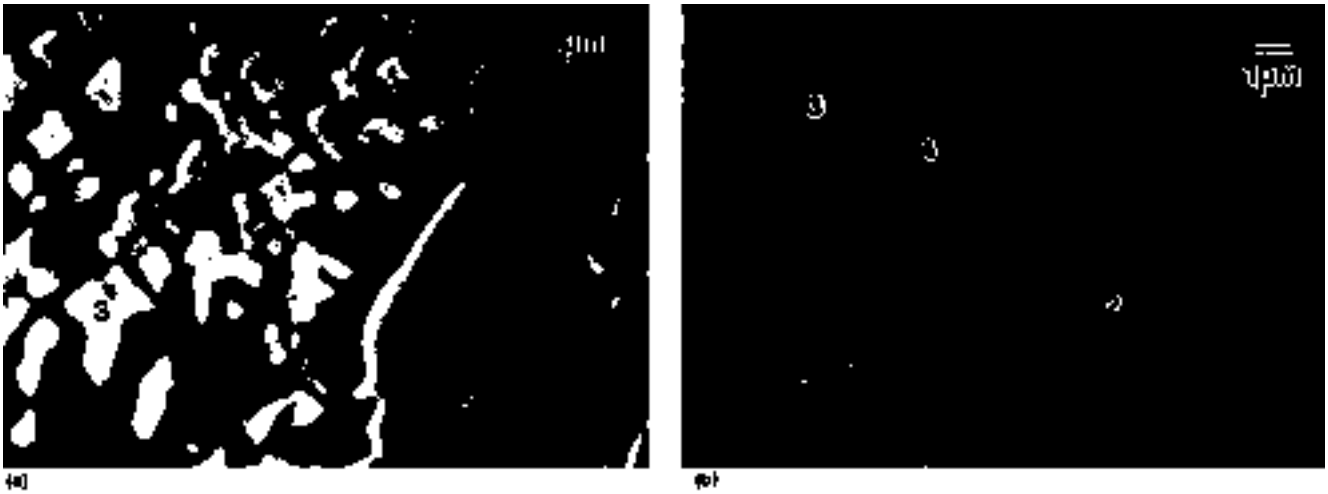


Fig. 9 Gamma-prime overaging in a nickel-base alloy turbine blade material. (a) SEM micrograph of the blade material, showing the breakdown of the eutectic gamma prime (5) and the spreading of the coarse gamma prime. Smaller particles of fine aging gamma prime (4), which would appear between the coarse gamma prime, are completely dissolved (indicated by the arrow). (b) SEM micrograph of a known, good blade material standard showing the (1) eutectic gamma prime, (2) metal carbide particle, (3) coarse gamma prime, and (4) fine aging gamma prime

Carbide Precipitation and Overaging. In addition to the gamma-prime phase, all turbine blade alloys form carbide precipitates of two principal types. Primary carbides, which form when carbon reacts with strong carbide-forming elements such as tantalum and titanium during the solidification of the molten alloy, have the formula MC, where M is the metallic element. Typically, such carbides have a blocky or scriptlike morphology, as shown in Fig. 10. Secondary carbides are formed at lower temperatures from the remaining carbon dissolved in the matrix and from the breakdown of MC carbides.

Fig. 10 Metal carbide (MC) and grain-boundary film in a Waspaloy forging. The grain-boundary carbide films substantially reduce stress-rupture life. Transmission electron micrograph, 3400×

The cause of degradation of impact properties of superalloys is microstructural instability at prolonged high-temperature exposure. The principal microstructural changes associated with embrittlement are transformation of the metastable MC carbides to the stable $M_{23}C_6$ grain-boundary carbides, and their growth leads to a continuous carbide film that embrittles the grain boundaries. Taking replicas (Ref 14, 15) from the surface of blades and comparing the aged microstructure with microstructures of known impact toughness facilitates assessment of the degree of embrittlement.

Oxidation and Hot Corrosion Resistance. The operating environment of a gas turbine blade contains significant levels of oxygen and smaller amounts of other elements, such as sulfur, that react readily with metals at operating temperatures. Hot corrosion is a severely accelerated form of environmental attack, which occurs by a combination of oxidation and sulfidation in contaminated operating environments. The most common contaminants, which lead to hot corrosion, are sodium and sulfur. However, other metallic impurities, such as potassium, vanadium, lead, and molybdenum, can lead to accelerated attack. Two distinct hot corrosion mechanisms have been recognized: Type 1 hot corrosion occurs between 800 and 925 °C (1470 and 1700 °F), while Type 2 hot corrosion occurs between 600 and 750 °C (1110 and 1380 °F). In both cases, the protective oxide layer is melted by a fluxing reaction, allowing greatly accelerated oxidation and sulfidation to occur. The mechanisms by which the fluxing occurs vary, depending on the contaminant type and concentration. Figure 11 shows hot corrosion attack of nickel-base alloy René 77 (Ref 4).

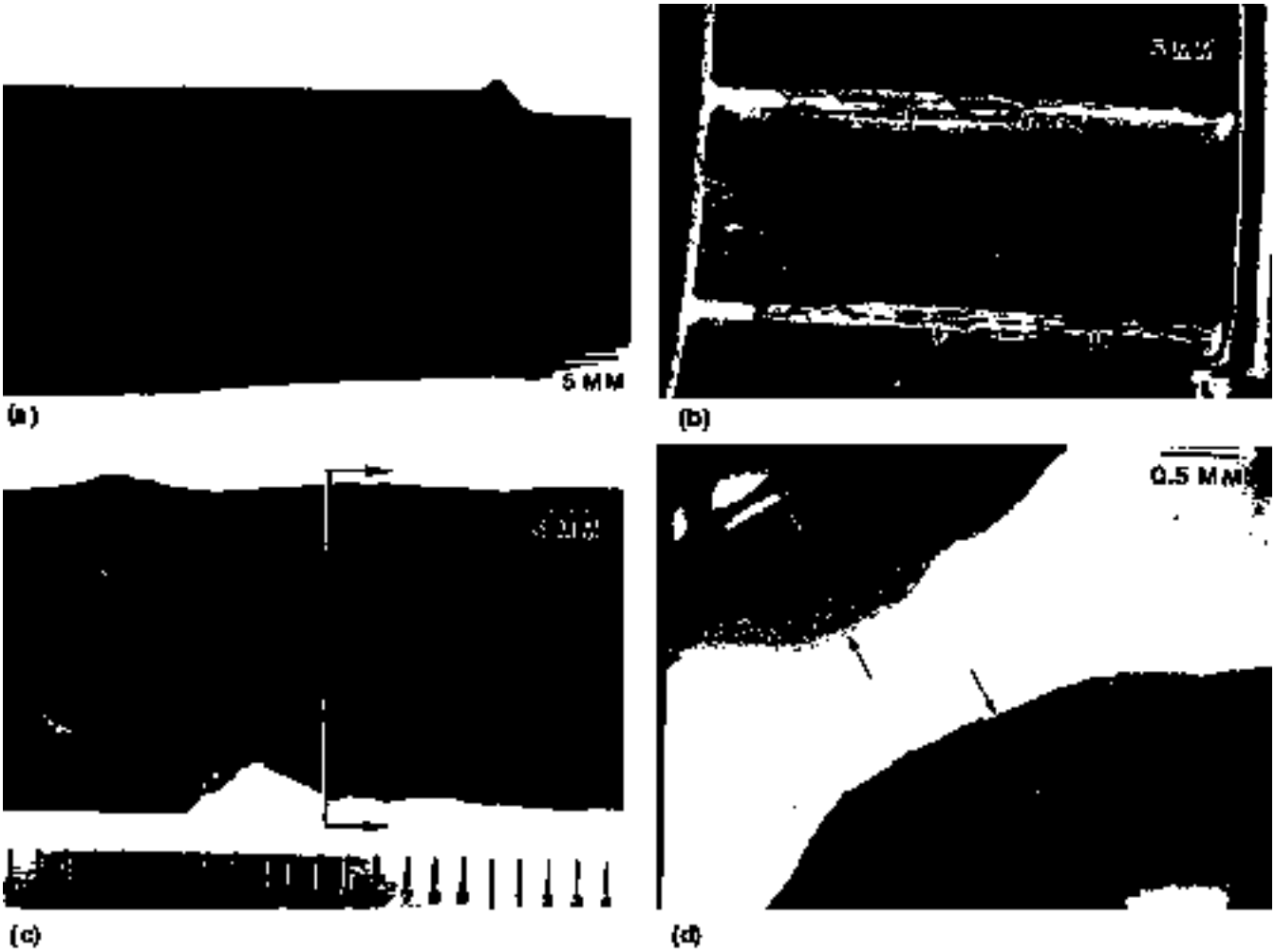


Fig. 11 Hot corrosion attack of René 77 nickel-base alloy turbine blades. (a) Land-based, first-stage turbine blade. Notice deposit buildup, flaking, and splitting of leading edge. (b) Stationary vanes. (c) A land-based, first-stage gas turbine blade that had type 2 hot corrosion attack. (d) A metallurgical cross section showing the layered corrosion attack of the blade shown in Type 1 and (c)

The use of protective coatings can greatly increase the lives of turbine blades for operating conditions in which the oxidation or hot corrosion resistance of the base material is unacceptable (Ref 16, 17). Two generic coating types are used for blade applications: diffusion coatings and overlay coatings. Both types of coatings result in a surface layer enriched in oxide-forming elements to promote formation of a protective oxide layer. However, the protective coatings themselves are subject to degradation under engine operating conditions and thus have limited lives. Life assessment of coatings is discussed later in this article.

References cited in this section

3. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components,* ASM International, 1989, p 188–190, 214, 228, and 369

4. D.J. Benac, *Elevated-Temperature Failures,* Lesson 7, ASM Practical Failure Analysis Course, 2002

7. "Properties of Carbon and Alloy Seamless Steel Tubing for High-Temperature and High Pressure Service," Technical Bulletin 6G, The Babcock and Wilcox Company, 1955

8. R. Viswanathan and R.I. Jafee, Toughness of CrMoV Steels for Steam Turbine Rotors, *J. Eng. Mater. Techol. (Trans., ASME),* Vol 105, Oct 1983, p 286

9. V.P. Swaminathan and P. Lowden, “Gas Turbine Blade Life Assessment and Repair Guide,” EPRI Report GS-6544, 1989

10. C.T. Sims, N.S. Stoloff, and W.C. Hagel, Ed., *Superalloys II,* John Wiley & Sons, 1987

11. M.J. Donachie, Jr., Ed., *Superalloys Source Book,* American Society for Metals, 1984

12. R.F. Decker, “Strengthening Mechanisms in Nickel Base Superalloys,” Steel Strengthening Mechanisms Symposium, May 1969 (Zurich)

13. R.A. Stevens and P.E.J. Flewitt, The Effects of Gamma Prime Precipitate Coarsening during Isothermal Aging and Creep of the Nickel-Base Superalloy IN 738, *Mater. Sci. Eng.,* 1979

14. A.G. Pard and D. Kramer, Replication Microscopy for Estimating Mechanical Degradation of Gas Turbine Blades, *Proc. Replication Technology,* (San Antonio), 1984, p 99–101

15. A.G. Pard, “Service Embrittlement of Gas Turbine Blade Alloys Udimet 710 and IN-738,” EPRI GS-6441, EPRI Project RP2775-1, Electric Power Research Institute, 1989

16. A. McMinn, “Coatings Technology for Hot Components of Industrial Combustion Turbines: A Review of the State of the Art,” EPRI AP-5078, Electric Power Research Institute, Feb 1987

17. “Evaluation of Protective Coatings for Hot Components of Combustion Turbines,” EPRI AP-4194, Electric Power Research Institute, Aug 1985

Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

Gas Turbine Blade Life Assessment

Gas turbine blades often require a life assessment after the turbine has been in operation because of an overtemperature situation or possible end of blade usefulness. Combustion turbine hot section rotating blades are fabricated from nickel-base superalloys and operate at high temperatures and aggressive environments. Both metallurgical and mechanical property degradation of the blading material occurs during service, which can limit the useful service life of blades. It is important to assess the condition of blades periodically and to estimate the remaining life to provide guidelines for replacement or reconditioning of the blades. A properly managed life assessment and refurbishment program can extend the life of present and future turbines by many years, thus avoiding major capital expenditure for the utilities (Ref 9).

A number of land-based turbines currently in operation have been in service in excess of 50,000 h. Thus, there is a growing need in the industry to assess the remaining life of aging components. Operating practices vary widely among utilities and can be different from the manufacturer's recommended practice. Both above-design (e.g., rapid cycling) and below-design (e.g., degrading) operation are common.

To accomplish reliability, proper techniques and tools for assessing the expended as well as remaining life of blades are required. Another important consideration is the safety of plant and personnel. Failure of blades in gas turbines can lead to consequential damage of other components downstream and to catastrophic failure of the turbine. Accurate life assessment methods and techniques can help avoid such occurrences.

Trends in turbine use increase the need for improved life assessment techniques. There is an increasing trend in the use of gas turbines in cogeneration and in combined cycle plants, operating in base load or intermediate load configurations. These types of duty cycles require high-reliability components with greater longevity than expected of machines operating under peak duty. There is also a trend toward increasing turbine inlet temperatures to increase turbine efficiency. Time-dependent damage mechanisms such as creep, corrosion, as well as TMF can become more important in this context.

Useful service life of gas turbine blades depends on the time it requires to initiate engineering-size cracks, or irreparable loss of metal occurs by FOD or corrosion. Cracks may initiate due to the exhaustion of creep life (which may manifest in the form of numerous creep cavities at critical locations of a blade), due to the accumulation of cyclic damage under TMF

conditions, or due to the deterioration of the protective coating under hot corrosion or TMF. Once a crack initiates, it may propagate rapidly, causing failure of the blade under high-cycle vibratory loading conditions. Thus, for the purpose of estimating expended and remaining life of blades, it is assumed that terminal condition is reached when it is determined that a crack is initiated. Various methods are available to determine blade life under a specified failure mode, that is, creep, high-cycle fatigue, low-cycle fatigue, TMF, and so on. Methods that combine synergistic effects of the various damage mechanisms are scarce.

Gas Turbine Engine Component Operating Conditions

Because operating temperatures range from ambient to above 1205 °C (2200 °F), gas turbine engine components are made of a variety of metals. Steels and titanium alloys are used for the relatively cool components, such as those in the fan and low-pressure compressor sections. Nickel-base, iron-nickel, and iron-base heat-resisting alloys are used for "warm" parts, such as shafts, turbine disks, high-pressure compressor disks, and cases. Nickel-base and cobalt-base heat-resisting alloys are used for hot parts, such as burners, turbine blades, and vanes.

Turbine blades and vanes are designed for high strength at elevated temperatures. First-stage turbine vanes are exposed to hot gases from the burner, with or at temperatures in excess of 1205 °C (2200 °F); consequently, airfoil metal temperatures may exceed 980 °C (1800 °F). Figure 12 shows some heat-damaged turbine blades. Creep is a problem, in that blades and vanes will stretch. Stretch is acceptable, provided that it remains within the limits established by engine design. Occasionally, however, components are exposed to excessive temperatures and stretch beyond limits, which may lead to premature component failures.

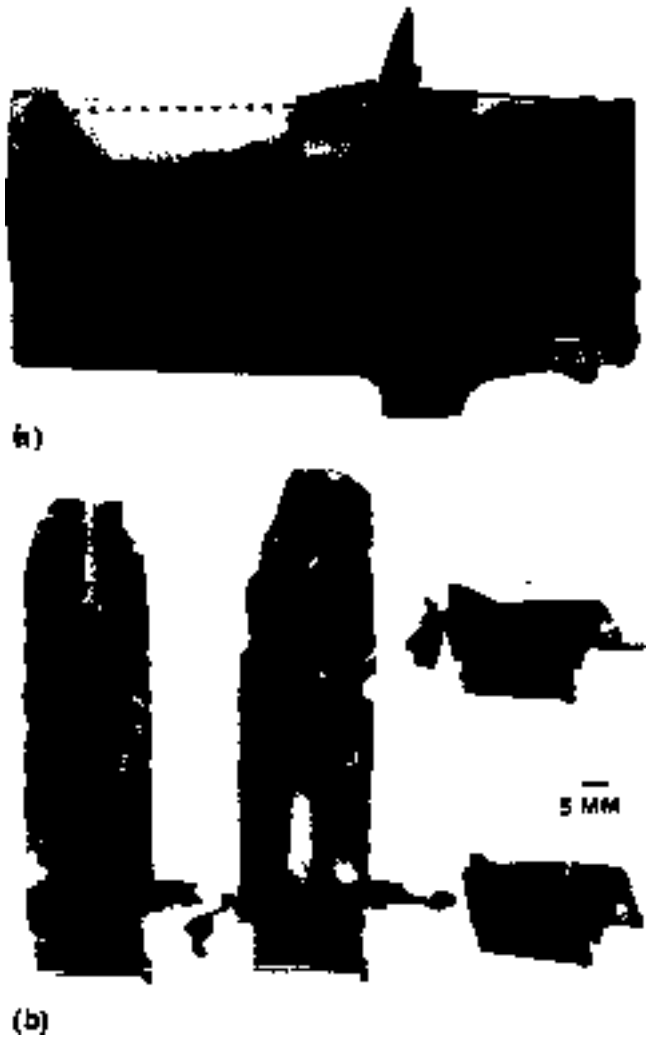


Fig. 12 Heat-damaged turbine blades. (a) Heat-damaged first- or second-stage turbine blade (A), which remained intact but with a darkened appearance. It is common to have blades that appear to be in relatively good condition but with an underlying overtemperature condition. (b) Two third-stage blades (B) with significant airfoil damage from the overtemperature condition, which had a detrimental effect on their

microstructures. The fourth-stage blades (C) ruptured as a result of lower fracture toughness and creep resistance, which is caused by elevated-temperature operation in gas turbine engines.

Because stress-rupture failures normally initiate at the grain-boundary interfaces, a logical approach is to reduce the grain-boundary interfaces or completely eliminate grain boundaries. Directional-solidification (DS) and single-crystal (SC) alloys were developed to increase high-temperature strength and minimize failures. Directional-solidification alloys for gas turbine blades have columnar crystals grown with their long axes parallel to the length of the airfoil, which is the direction of principal stress. Directional-solidification alloys were introduced into military and commercial aircraft gas turbines in the 1960s and 1970s, respectively. In the 1980s, DS alloys were used in land-based turbines. In the early 1980s, SC alloys were used for commercial test engines. In the 1990s, SC alloys were used in land-based turbine applications. Table 1 shows some of the land-based turbine alloys used for turbine blades (buckets), vanes, and nozzles (Ref 3, 4).

Table 1 Land-based gas turbine alloys used for blades, vanes, and nozzles

Alloy designation	Description
Alloys for blades and buckets	
Udimet 520	Wrought nickel-base alloy
Udimet 500	Cast nickel-base alloy
IN 738	Conventionally cast nickel-base alloy, high carbon
IN 738LC	Conventionally cast nickel-base alloy, low carbon
DS CM 247 LC	Directionally cast nickel-base alloy, low carbon
GTD 111	Conventionally and directionally cast nickel-base alloy
SC René N5	Single-crystal nickel-base alloy
PWA 1483	Single-crystal nickel-base alloy
Alloys for vanes and nozzles	
IN 939	Conventionally cast nickel-base alloy
IN 738LC	Conventionally cast nickel-base alloy, low carbon
FSX-414	Conventionally cast cobalt-base alloy
ECY-768/X-45	Conventionally cast cobalt-base alloy
DS CM 247 LC	Directionally cast nickel-base alloy, low carbon
PWA 1483	Single-crystal nickel-base alloy

Gas Turbine Blade Life Assessment Methods

Properties that control the integrity and useful life of blades deteriorate with time in service, and it is imperative that the condition of blades is assessed periodically. The “remaining life” is also estimated to determine what actions, if any, need to be taken to ensure continued safe and reliable operation of the turbine.

Remaining life assessment involves providing guidelines for when to replace or recondition the blades and for selecting inspection intervals for reevaluation. Thus, condition/remaining life assessment is a continuous process as the turbine ages. Because several blade failure mechanisms may be operative depending on operating conditions, methods are needed to predict remaining blade life for each mechanism and combinations thereof.

Various life prediction methodologies have been developed, and many are still under active development for application in jet engines as well as land-based gas turbines. Figure 13 illustrates the general requirements for the in-service assessment of land-based gas turbine blades for remaining service life. The steps in the flow diagram of Fig. 13 are described in following article sections, “Inspection and Evaluation” and “Remaining Life Assessment.” This guideline is intended to provide guidance and steps to assess blade life. This guideline does not replace what the turbine manufacturer or turbine operator recommends (Ref 9).

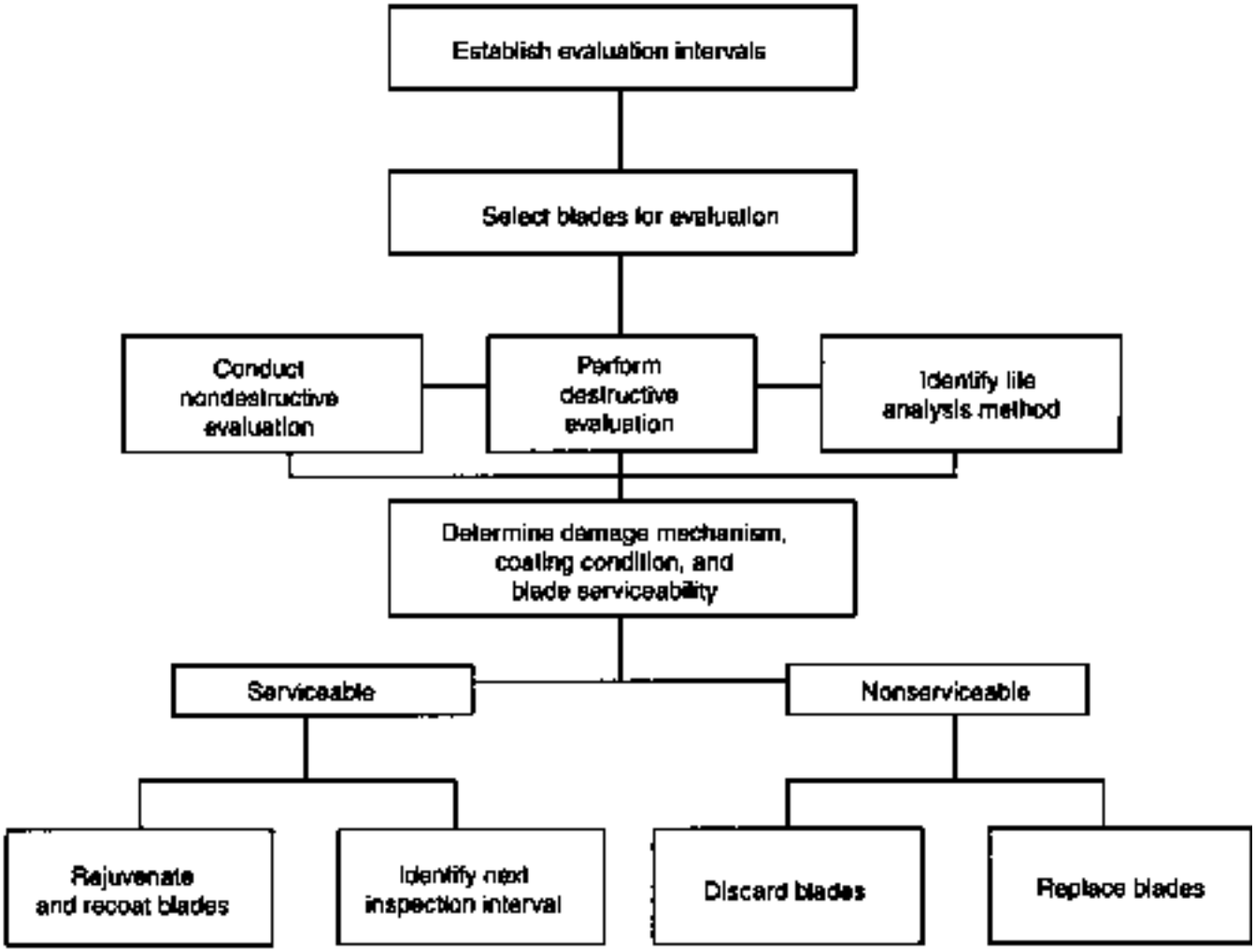


Fig. 13 Flow diagram for remaining life assessment of gas turbine blades

Inspection and Evaluation

Inspection and evaluation includes the assessment of mechanical damage, surface degradation, microstructural deterioration, and creep damage. Mechanical damage is generally detected by nondestructive testing techniques, while the balance of the damage is best detected by destructive examination.

Step 1: Establish Evaluation Intervals. Step 1 involves establishing the evaluation intervals based on the design life and service history of the blade. For example, nondestructive evaluation (NDE) for mechanical damage (dents, nicks, rubbing, erosion, and cracking) should be performed at the interval recommended by the engine manufacturer, or, in the absence of such a guideline, a determined factor of the design life of the blade, such as 20% of the design life.

Step 2: Conduct NDE. The NDE approach continuously monitors turbine operation, and a rigorous inspection program is imposed to periodically check for cracks by NDE and for microstructural damage by microscopic examination of the surface material. Various NDE methods for inspection (Ref 18) are discussed in the article “Nondestructive Evaluation and Life Assessment” in this Volume. It is paramount that the NDE method identifies the known damage or defect. It is likewise necessary that any suspect damage or defect be confirmed as present either by microstructural evaluation or destructive testing. The following guidelines and steps for NDE are suggested:

- Follow manufacturer's inspection procedures, if available.
- Clean the blade surfaces by an appropriate mechanical or chemical method, ensuring that cooling holes are not plugged.
- Visually and dimensionally inspect the blades for seal wear, tip rub, rubs, FOD or DOD, cracking, surface erosion or coating loss, hot corrosion, or oxidation.
- Inspect the entire surface of the blades using an approved method and type of fluorescent liquid penetrant.
- Blades with cracks, nicks, or dents are considered not serviceable. They can be returned to a serviceable condition by dressing within the engine manufacturer's allowances. If such allowances are not available, limits may be calculated based on stress and temperature distributions in the blade.

- Tip rubs and seal rubs are serviceable, provided they do not affect cooling air flows. However, they can adversely affect engine performance. Estimates of the performance penalty associated with worn blades can be made based on wear measurements.
- Blades with damage extending beyond the dressing limits or welding can repair excessive wear, provided the extent of welding does not exceed the recommended limits.
- Blades that are suspect or designated as inspection blades should have the microstructure evaluated by use of replication.

Step 3: Conduct Destructive Evaluation. Destructive evaluation involves mechanical testing, microstructural characterization, and coating evaluations of blades in service. Destructive evaluation results of blades in service are compared to base-line blades or material data. Destructive evaluation should be performed at the interval recommended by the engine manufacturer, or, in the absence of such a guideline, a determined factor of the design life of the blade. For example, destructive evaluation may be performed at 50 to 80% of the design life initially. Subsequent evaluations may be performed at 20 to 50% of the design life, depending on the results of the previous evaluation.

Selection of Sample Blade. The blade removed from the set for evaluation should be chosen to be as representative as possible of the balance of the set (that is, it should have the same service and manufacturing history). Records should be made of replacement blades, so they are not used in subsequent life assessments.

Surface Evaluation. Following are guidelines for evaluating the surface of the sample blade (Ref 9):

- Metallographic sections through the blade surface should be examined from the upper, middle, and lower airfoil at the leading and trailing edges and any other areas that appear visually to have surface damage (Fig. 14). Internal surfaces of cooled parts should be examined in addition to the external surfaces.
- In uncoated blades, the total depth of environmental attack should be measured, including the depth of any oxide layer, the depth of internal oxidation, and the depth of the alloy-depleted zone.
- The user should establish a criterion of acceptance. For example, uncoated blades may be considered reusable if the total depth of environmental attack is less than 0.05 mm (0.002 in.). Blades with attack depths greater than 0.05 mm (0.002 in.) may also be used, if the damage can be removed by dressing within the manufacturer's allowable limits. For improved life, the use of a protective coating should be considered in such instances.
- On coated blades, the remaining coating thickness, excluding any diffusion layer, should be compared in the best and worst sections of the blade. The original thickness of coating should be estimated based on the remaining thickness in the best area, and the depth of attack should be judged by the remaining thickness in the worst areas.
- Coating life should be considered exhausted if less than $\frac{1}{3}$ of the original coating thickness remains unattacked. Such blades may be reused by reapplication of the protective coating. If base metal attack has occurred, the damaged areas must be removed by dressing within the manufacturer's allowable limits prior to recoating. Blades with attack deeper than the manufacturer's allowable dressing limits are unrepairable.
- Remaining coating life on coatings with more than $\frac{1}{3}$ of the original coating thickness remaining may be estimated by assuming a linear rate of attack; for example, a coating with $\frac{1}{3}$ of its thickness attacked has had $\frac{1}{2}$ of its life exhausted.

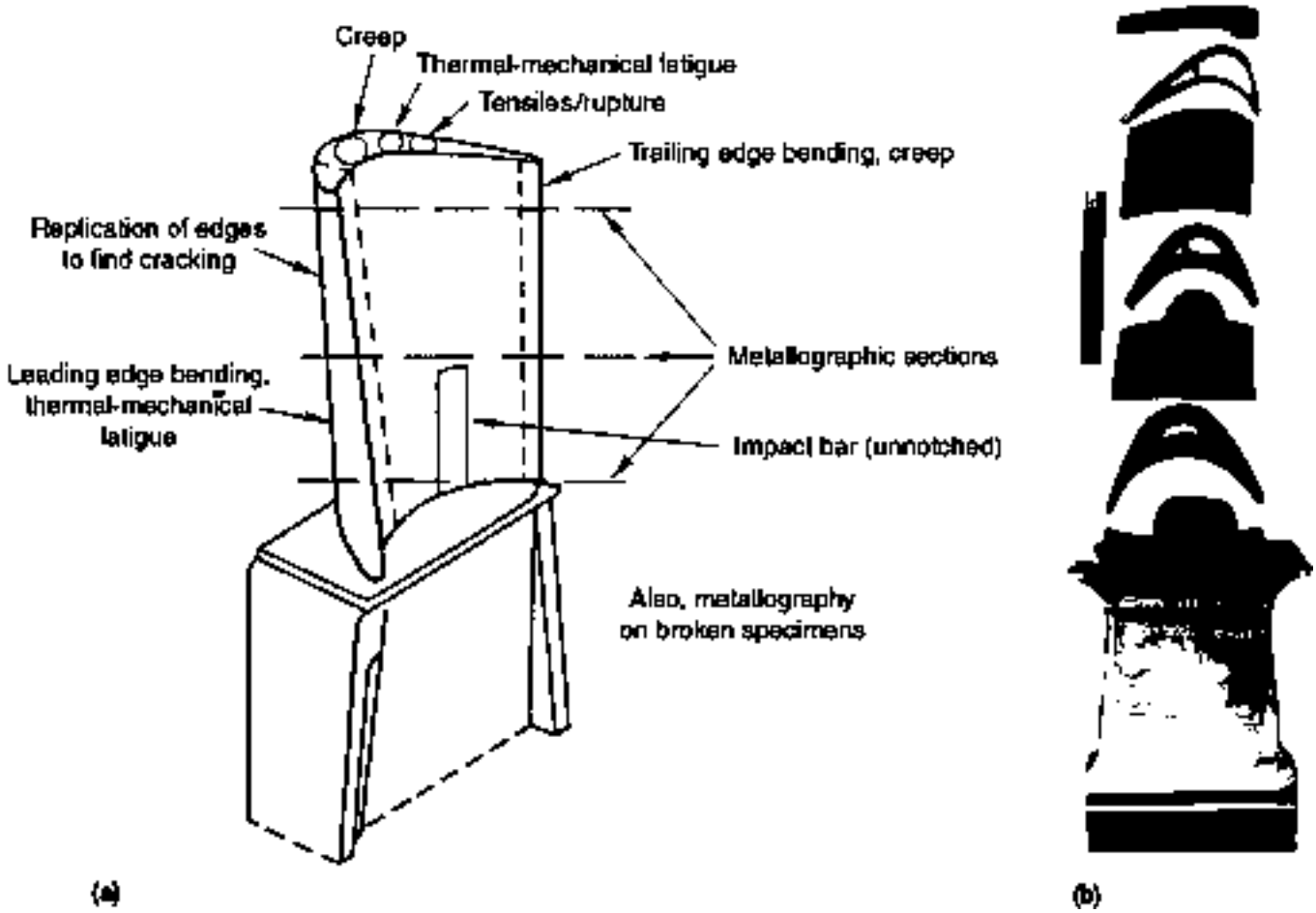


Fig. 14 Sectioning of turbine blades for metallographic examination. (a) Typical locations for cross sectioning of turbine blades. (b) View of Sectioned blade

Microstructural Evaluation. Following are guidelines for analyzing the microstructure of the sample blade:

- Metallurgical samples from the upper, middle, and lower airfoil and root section of the blade and from any other location that appears visually to have been at higher temperature should be prepared for metallographic examination. Sections should be oriented in a direction parallel to the stress axis.
- All samples (especially the longitudinal samples) should be examined by optical microscopy for the presence of intergranular creep voids. Any blades in which creep voids are identified are unserviceable, unless repaired or rejuvenated.
- All samples should be examined by optical microscopy for microstructural instabilities. Changes from the new material microstructure can be identified by comparing airfoil samples with those from the root. In particular, the presence of brittle phases, grain-boundary carbide aging, and gamma-prime aging, if visible, should be noted.
- The microstructure of the gamma-prime phase should be examined by a scanning electron microscope technique. Samples should be prepared from the areas showing the most severe instability (usually the mid or upper airfoil) and from the root. Comparing the structures in these two areas, the degree of service aging can be identified. The diameter of the particles should be measured to determine the amount of particle growth that has occurred. In alloys with duplex gamma-prime microstructures, the absence of fine, secondary gamma-prime particles and the agglomeration of the primary particles should be noted, if present.
- Microstructural analysis is generally not suitable as a sole method for identifying serviceability. Normally, the results are used to clarify the results of mechanical testing. For example, significant microstructural degradation combined with low creep properties indicates temperature-driven damage accumulation, while low creep properties in the absence of degradation indicate high-stress, lower-temperature damage.

Remaining Life Assessment

To predict the remaining life of a blade, one needs to know the operating environment, temperature distribution, and loads on the blade; macroscopic response of the component to the applied load and environment (i.e., stress, strain, temperature, corrosion, erosion, etc.); and microscopic response of the material to the applied load and environment (i.e., material

behavior and applicable life-prediction techniques). Thus, any failure prediction model should take into account the physical damage mechanism(s) that lead to the final failure of the blade. Continuous damage models that lend themselves to analytical procedures do not always describe failure in a physically realistic manner. Furthermore, the material performance information is usually derived from simple tests and standard procedures under controlled conditions.

Determining Remaining Creep Life. Several assessment techniques can be used to evaluate remaining creep life. The most suitable technique depends on the availability of appropriate data. Some of the techniques widely used for gas turbine blades are:

- *Qualification testing:* The blades are tested using the same tests used to qualify lots of new material. This technique is widely applicable, because very little background data is necessary; however, it is limited to providing a pass/fail indication of serviceability rather than a remaining life estimate. The same parameters used to qualify new material should be used. Samples that fail to meet the minimum new material stress-rupture life are considered unserviceable and should be retired or refurbished by hot isostatic pressing and/or reheat treatment. Blades that are returned to service should be reexamined at a determined percentage of the design life of the blade.
- *Life fraction rule:* Remaining life is determined by comparing stress-rupture properties of new material with those obtained after service exposure. A conservative estimate of service stress and higher-than-service temperature should be used.
- *Parameter-based approaches:* These methods are based on a creep-property database developed by testing several specimens to obtain data from both long- and short-term tests. Two such methods that are widely used are the Larson-Miller parameter and the Monkmon-Grant parameter.

Serviceability and remaining life are assessed differently, depending on the methodology used.

Life Fraction Rule. The remaining life is calculated according to the equation:

$$\frac{t_s}{T_s} + \frac{t_t}{T_t} = 1 \qquad \text{(Eq 1)}$$

where t_s is time in service, t_t is time to rupture in accelerated testing, T_s is time to rupture in service conditions, and T_t is time to rupture of virgin material in accelerated test conditions.

Samples should be removed from the blade airfoil in the creep life-limiting section, as determined from the stress and temperature profiles in the blade. The orientation of samples should be parallel to the stress axis. Samples should also be removed from the much cooler root shank, so that unaged (original state) material can be tested for comparison purposes.

Samples should be creep or stress-rupture tested as appropriate for the creep life assessment technique to be used. Testing should be performed in accordance with ASTM E 139. A minimum of two samples should be tested from the airfoil and at least one from the shank.

It was proposed that for any sequence of stress and temperature combinations, the life fractions could be linearly summed, with failure occurring when the sum of the fractions is unity. A similar rule proposed by Lieberman (Ref 19) uses strain fraction instead of time. This rule is expressed by the following equation:

$$\sum_i \varepsilon_i/\varepsilon_{if} = 1 \qquad \text{(Eq 2)}$$

where ε_i is strain accumulated at stress σ_i and temperature T_i, and ε_{if} is time to failure at σ_i and T_i.

An important assumption in applying either of the previous two summation rules is that the material behavior is not history-dependent. For many years the life fraction rule was assumed to be true, and procedures were developed to continuously monitor remaining lives of components based on it. However it has been found that the damage does not generally sum to unity as stated by the equation. Thus in reality the equation should be expressed in the following form for general application:

$$\sum_i t_i/t_{if} = k \qquad \text{(Eq 3)}$$

where k is a variable. Generally, k is a function of the sequence of loading at constant temperature for an increasing stress sequence, $k \ll 1$, while for a decreasing stress sequence, $k \rightarrow 1$ (Ref 20, 21). Comparison of actual residual life versus predicted lives by the accelerated stress tests versus accelerated temperature tests for ferritic 1Cr-0.5Mo steel is shown in Fig. 15.

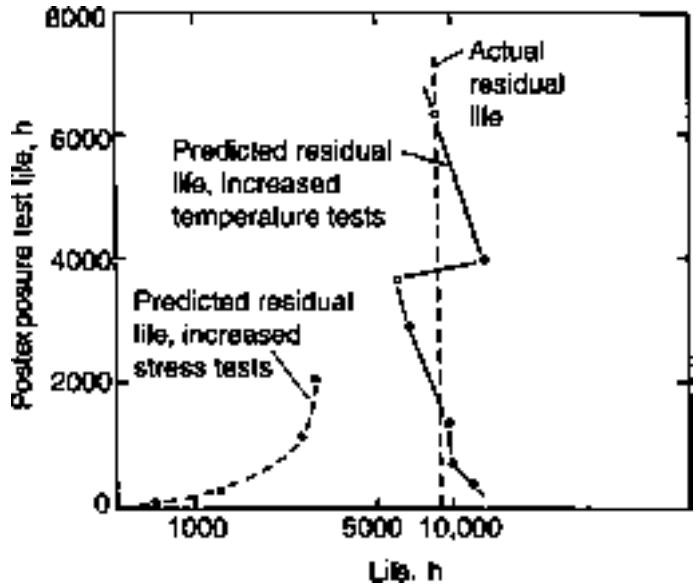


Fig. 15 Residual life predictions using the life fraction rule from increased stress and temperature tests for 1Cr-0.5Mo steel

Blades should be retired or refurbished if the t_s/T_s ratio exceeds an established value, for example, 0.6. Blades returned to service should be reexamined after half of the remaining life or half the design life, whichever interval is shorter. For blades that are considered unserviceable due to creep void formation, blades may be rejuvenated by an approved hot isostatic pressing method. For blades that are considered unserviceable due to unacceptable size of gamma-prime phase, blades may be rejuvenated by an approved reheat treatment.

The life fraction rule is discussed further in the section "Life Assessment Methods for Power Plant Piping and Tubing" in this article.

Time-Temperature Parameter Approaches. Design of gas turbine blades is based on extrapolation of relatively short-time test data to the planned lifetime of the component. The shortcomings and uncertainties associated with the prediction of long-time creep behavior using such data is discussed in this section. Numerous time-temperature parametric methods to correlate and extrapolate creep data are cited in the literature. The most popular time-temperature parametric methods are the Larson-Miller parameter (LMP) (Ref 22), the Orr, Sherby, and Dorn (OSD) parameter (Ref 23), the Manson-Haferd parameter (Ref 24), and the minimum commitment parameter (Ref 25).

The time-temperature parameters are used to correlate the creep-rupture test variables of temperature and time with stress. They are also used to correlate stress with temperature and minimum strain rate. Properly developed correlations can be used for the following purposes in the gas turbine hot section blade evaluation:

- To represent creep data obtained from testing in a compact format, which allows analytical treatment and analysis and to extrapolate (or interpolate) the data when experimental data are not available
- To provide a simple way to compare the creep behavior of various superalloys and rate them in a relative manner
- To extrapolate to time ranges on the order of expected blade lives, because it is impractical to perform such long-term tests due to time and cost limitations

A time-temperature parameter, $P(t, T)$, is an analytical expression containing both time (t) and temperature (T) in such a way that any combination of these two variables that produces the same numerical value of the parameter causes failure at the same stress level, that is:

$$P(t, T) = G(\sigma) \qquad \text{(Eq 4)}$$

where G is a function of the stress level, σ.

The temperature dependency is expressed by an Arrhenius factor ($e^{-Q/RT}$) where Q is activation energy, R is gas constant, and T is temperature. Applying this factor to analyses of creep-rupture properties of low-alloy and stainless steels, Larson and Miller (Ref 3) obtained the following parameter:

$$\text{LMP} = T(C + \log t) \qquad \text{(Eq 5)}$$

where C is a material constant (usually assumed to be equal to 20), and T is the absolute temperature (273 + °C, or 460 + °F), and t is time in hours. However, the constant C in the LMP changes as a function of time. Figure 16 shows a typical Larson-Miller rupture curve.

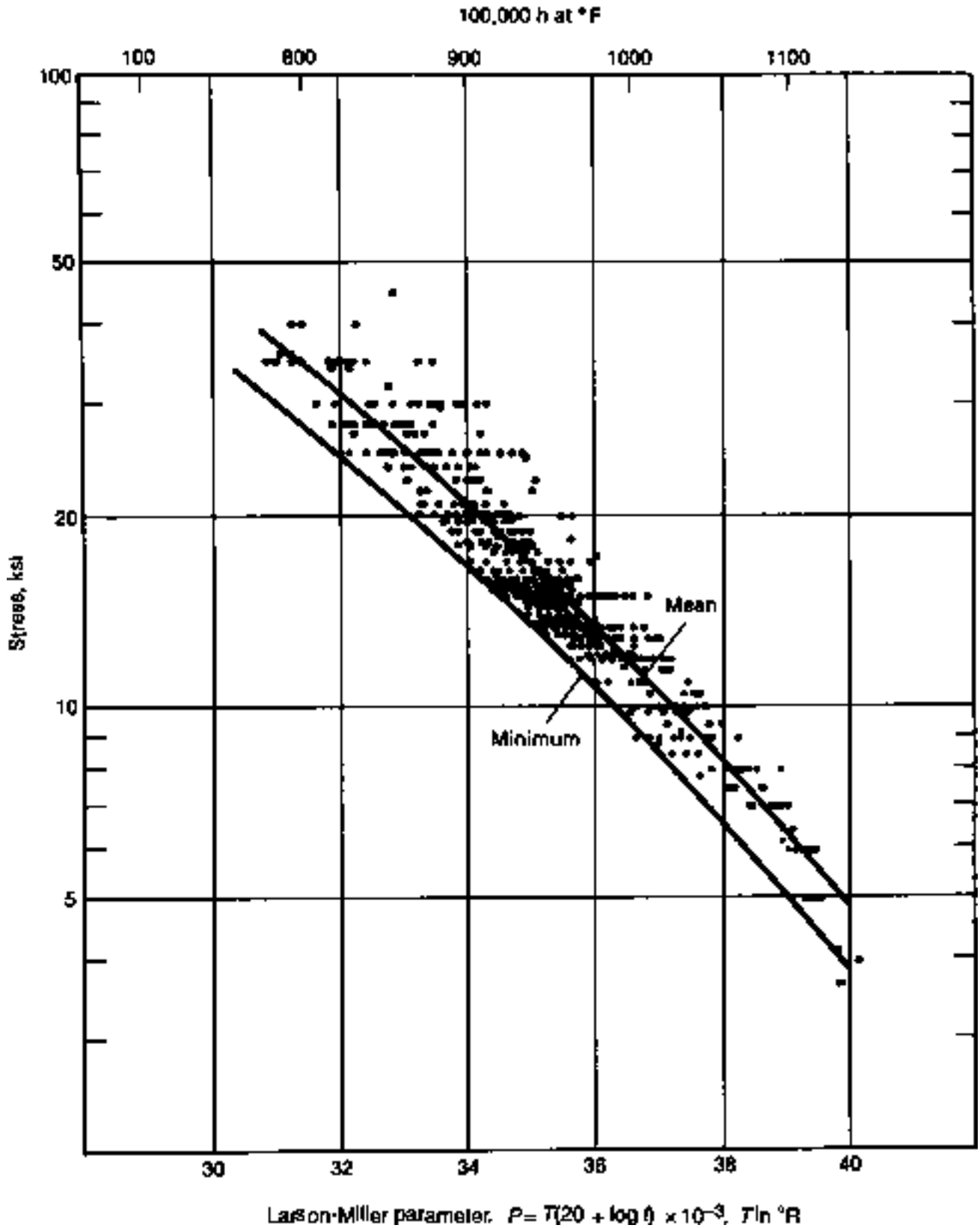


Fig. 16 Plot of Larson-Miller rupture parameter and stress for annealed $2\frac{1}{4}$Cr-1Mo steel

In addition to the LMP, other parameters may be considered. Predictions using the OSD and LMP methods are illustrated in Fig. 17 for IN 738. The actual test data is up to 20,000 h. Both parameters are excellent in predictions up to 100,000 h at the two lower temperatures. However, at higher temperature, the deviation is large, the LMP being on the nonconservative side. A sensitivity analysis using slightly different *Q*/*R* values to calculate the OSD parameter was also performed and the results plotted, as in Fig. 17(b). Approximately a 10% decrease in the *Q*/*R* value resulted in 40% less life, and a similar increase in the value of *Q*/*R* produced a proportional increase in life. The Larson-Miller master curve and isothermal curves are shown in Fig. 17(c) and (d), respectively. These authors concluded from their statistical regression analysis that the scatter in the creep-rupture life distribution is inherent to the material properties and/or to the

testing procedure. Effective fit of the data was achieved by all three parameters. However, the authors concluded that the minimum commitment method is the most reliable extrapolation method (Ref 26).

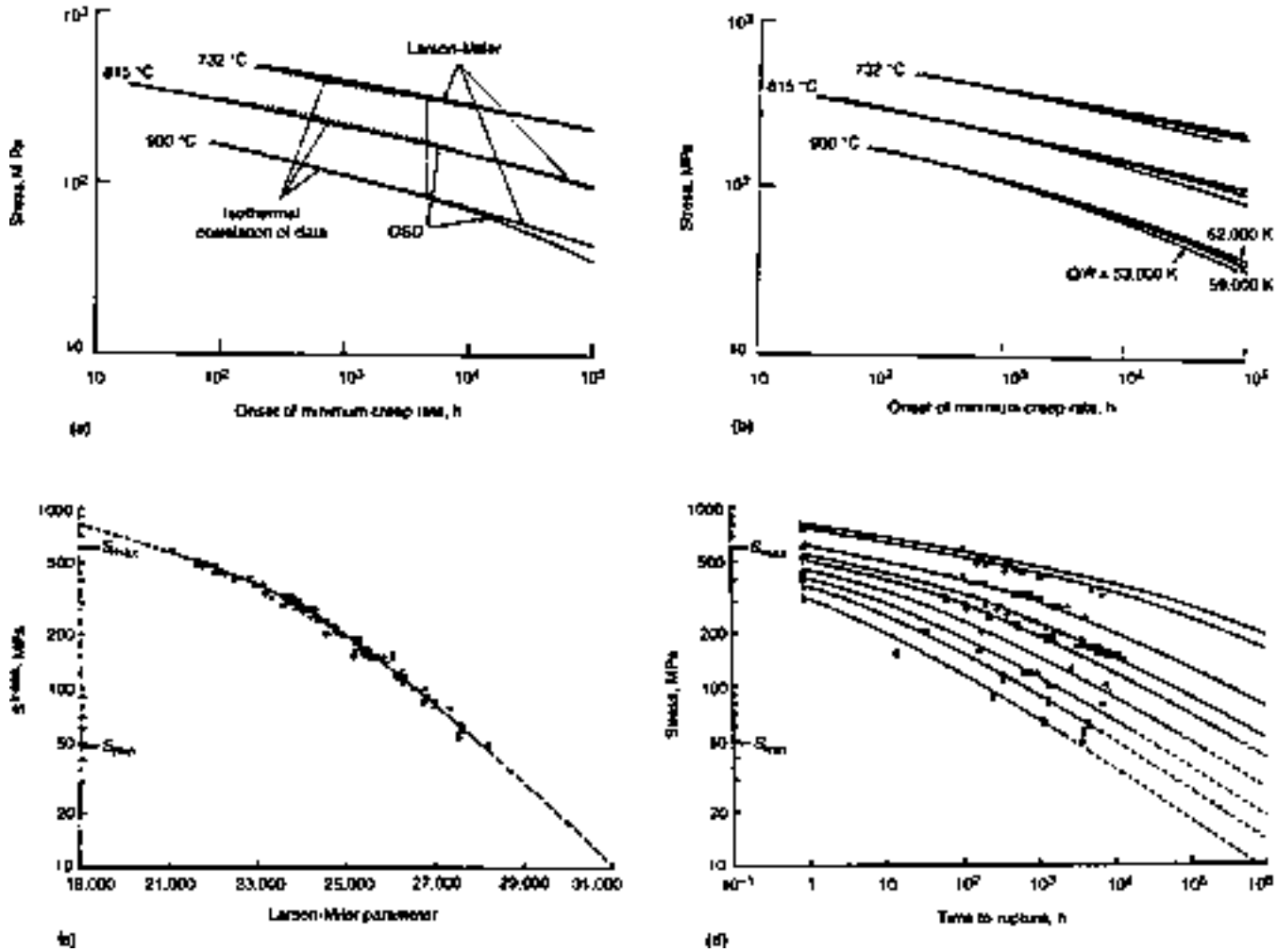


Fig. 17 Long-term creep predictions using the Larson-Miller and Orr-Sherby-Dorn (OSD) time-temperature parameters for IN 738 alloy. (a) Correlation of Larson-Miller and OSD data. (b) Effect of changing the activation energy on creep predictions. (c) Larson-Miller master curve. (d) Larson-Miller isothermal rupture curve

Impact Testing. In some cases, other mechanical tests, such as stress-rupture or impact tests, are run. For example, blades that are prone to serious embrittlement or FOD should be impact and tensile tested. Foreign object damage occurs when solid objects (debris and loose hardware) in the flow impact the rotor blades. The cost of replacing a badly impact-damaged stage of blades and vanes of a large frame engine can exceed $1,000,000. It is therefore important to monitor blades to determine when such blades become inordinately vulnerable to impact by even small foreign objects.

The most accurate way of establishing the condition and remaining life of a service-run blade is by sacrificing one or more representative blades from the row and by machining test samples from the airfoil at strategic locations. Because aging effects on impact toughness are greatest at high temperatures, the samples should be removed from the hottest regions of the blade subject to impact. Impact usually occurs in the upper half of the airfoil on the suction surface of the leading-edge side of the blade. Maximum-temperature regions can be identified by examination of microstructural test samples.

Monitoring can be accomplished by tracking impact toughness using a plot of aging parameter (Larson-Miller time-temperature parameter) versus impact toughness, as shown in Fig. 18, and scheduling a hot parts inspection when it is predicted that toughness has decreased to a minimally acceptable value (Ref 27).

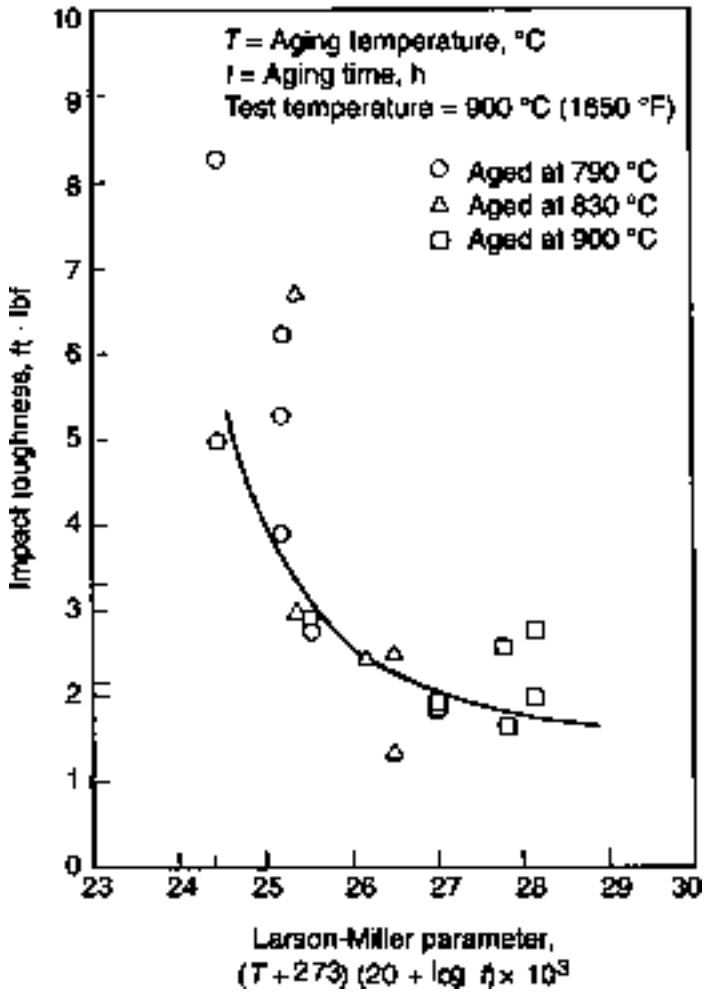


Fig. 18 Effect of aging in terms of the Larson-Miller parameter on toughness of U-710 tested at 900 °C (1650 °F)

For uncooled airfoils, full-sized Charpy V-notch (CVN) specimens can usually be machined from the maximum-temperature region of the airfoil. For cooled blades, it is important to remove samples as close to the surface as possible, because the surface runs substantially hotter than the interior. This necessitates testing subsize CVN samples. In order to calibrate subsize results, control tests must be run on identical samples removed from the root shank, where the temperatures are low enough that embrittlement does not occur. This should be done in any case because it establishes the initial versus aged condition of that particular lot of blade material.

Values from high-temperature (830 °C, or 1526 °F) impact toughness measurements using half-sized CVN specimens from cooled service-run U-710 and IN 738 blades are given in Table 2 (Ref 10). Obviously, these one-half-sized U-710 samples were much less embrittled than the full-sized samples from the aforementioned uncooled U-710 blade (Table 2), inasmuch as the half-sized specimens absorbed as much energy as the full-sized specimens. This is because they were removed from the cool interior of the blade—between cooling holes—where the temperature was probably 85 °C (150 °F) cooler than the surface of the cooled blade and the bulk of the uncooled blade.

Table 2 Impact test results at 830 °C (1525 °F) for samples taken from retired blades

Blade history	Alloy	Specimen	Absorbed energy, J (ft · lbf)	
			Individual	Average
Full-sized specimens				
9333 h, uncooled; from row 1 blade	U-710	1	2.4 (1.8)	2.8 (2.1)
		2	2.8 (2.1)	
		3	3.1 (2.3)	
		4	3.0 (2.2)	
Half-sized specimens[(a)]				
10,000 h, cooled; from row 1 blade	U-710	1	2.6 (1.9)	2.7 (2.0)
		2	2.8 (2.1)	
		3	2.8 (2.1)	

8242 h, cooled; from row 1 blade	IN 738	1	2.0 (1.5)	2.3 (1.7)
		2	2.3 (1.7)	
		3	2.7 (2.0)	

(a) The impact strength of a standard full-sized Charpy specimen is estimated to be roughly four times that of a half-sized specimen. These specimens were taken from the airfoil between cooling channels located along the chamber line of the U-710 blade and between the cooling holes and airfoil surface of the IN 738 blade.

Thermal-Mechanical Fatigue Life Assessment. The hot gas path components of a gas turbine experience a complex thermal and mechanical history during a typical cycle of operation, consisting of start-up, steady-state operation, and shutdown. Temperature gradients and mechanical constraints during this cycling give rise to cyclic thermal stresses, which can lead to fatigue damage and eventual failure. Among the rotating components, the first-stage blades are particularly prone to TMF, because they experience the most severe environment. As far as remaining TMF life assessment is concerned, postservice accelerated testing procedures are not as well established as they are for creep. Thermal-mechanical fatigue tests are highly accelerated to begin with, and hence, the effects of thermal aging are not simulated.

In TMF life prediction methods, the wave form, particularly the phasing between strain and temperature, is extremely important in determining life. The wave form is established by the temperature or strain versus the time. Based on the results of an extensive study of many materials, Kuwabara (Ref 28) found that the wave form or shape can affect the fatigue life. For example, two phases that affect life are out-of-phase (OP), minimum strain at maximum temperature, and in-phase (IP), maximum strain at maximum temperature, cycles. Based on a plot of total or inelastic strain range versus cycles to failure, the study showed that in ductile, low-strength materials, IP cycling generally results in longer lives, whereas in high-strength, low-ductility materials, OP cycling is generally better. A significant difference in lives was found between IP and OP cycling for IN-738-based alloy on a total strain range correlation (Ref 29, 30). For TMF cycles applied to test specimens of IN-738 alloy, the total strain range versus cycles to crack initiation is shown in Fig. 19(a). The data shows that at lower strain ranges, the life can differ almost by a factor of 100 when compared to isothermal cycle results.

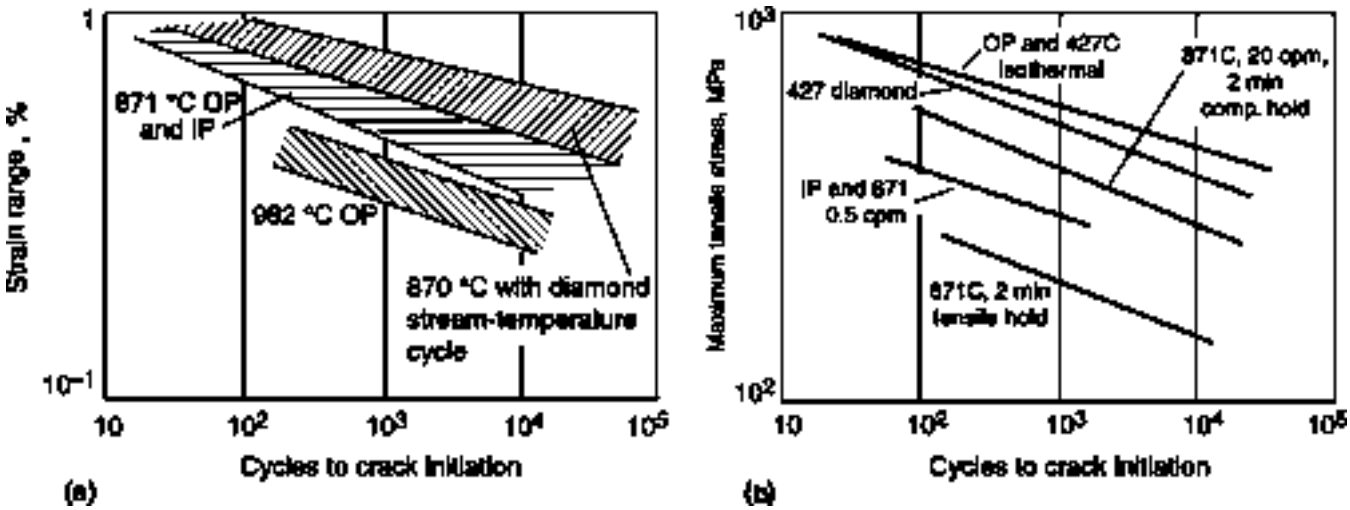


Fig. 19 Effect of in-phase (IP) and out-of-phase (OP) cycling on the fatigue life of IN-738. (a) Strain life data for the various wave forms. (b) Cycles to crack initiation based on the maximum tensile stresses

When the TMF results are plotted in terms of maximum tensile stress, instead of total strain range, a better correlation with cycles to crack initiation is obtained, as shown in Fig. 19(b). However, it is more difficult to accurately calculate maximum tensile stress in a bucket TMF cycle (because of the time-and-temperature-dependent inelastic behavior of the material) than it is to calculate strain range, which can be fairly accurately done even if fully elastic behavior is assumed.

In addition, TMF cracks initiate on the surface, where TMF properties can be seriously degraded by the environment. When diffusion or overlay coatings are used to prevent oxidation and/or hot corrosion, the TMF properties of the coating in the service environment govern TMF life.

Case History

Example 1: Evaluation of Gas Turbine Hot Section Blade Cracking under Oxidation, TMF, and Creep Conditions. This case history illustrates how to evaluate creep, oxidation, and TMF in a first-stage gas turbine blade. Often, it is necessary

to estimate the temperature and the applied stress to perform a life assessment. This case history demonstrates how that was done on a failed blade. Once the operating temperature and stresses are determined, then the life of blades with similar operating conditions can be determined. The first-stage blades in a model 501D5 gas turbine had 16 cooling holes. After 32,000 h of service, the blades exhibited cracking at the cooling holes. The blade material was wrought Udimet 520 alloy (Special Metals Corp.), with nominal composition of 57Ni-19Cr-12Co-6Mo-1W-2Al-3Ti-0.05C-0.005B. Figure 20 shows the blade and the airfoil cross section with the cooling holes. Intergranular cracks ranging from 0.13 to 0.32 mm (5 to 12.6 mils) were found at the surface of cooling holes 2 through 6. The diameter of these holes was approximately 2.5 mm (0.1 in.). The cooling holes surface was not coated. Cracks were observed by transverse metallurgical sectioning of the airfoil at approximately 60 and 90% heights, where the metal temperature is expected to be maximum. The orientation of the cracks was parallel to the blade axis and not perpendicular to the centrifugal stress (transverse) direction. Typical oxidation and intergranular cracking are shown in optical micrographs in Fig. 21.

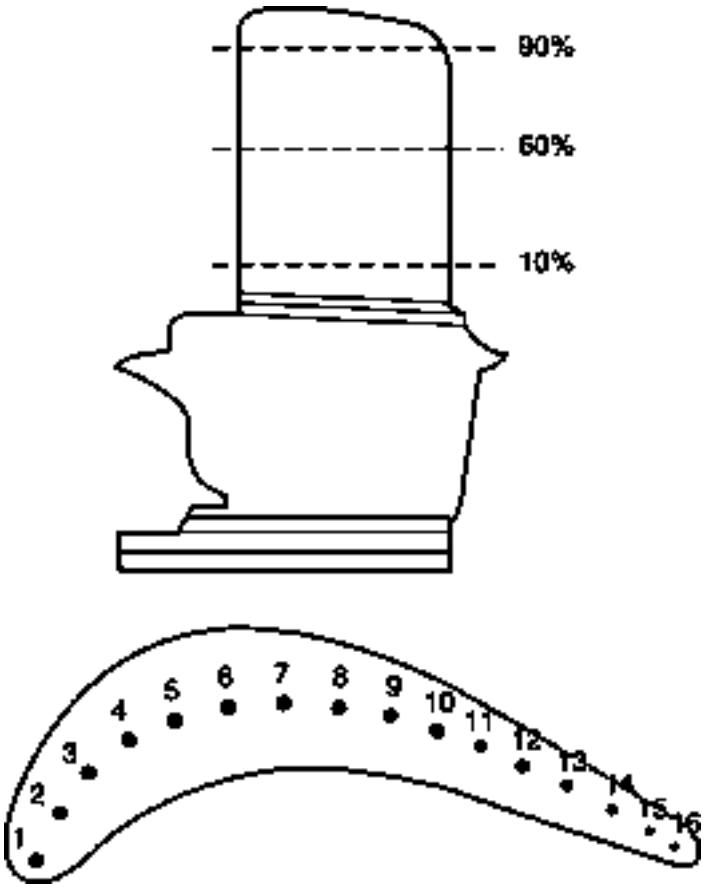


Fig. 20 Schematic of first-stage gas turbine blade that experienced cracking after 32,000 h in service. (a) Sectioning planes at three locations on the blade airfoil. (b) Cross-sectional view of the blade airfoil showing the cooling holes and numbering sequence

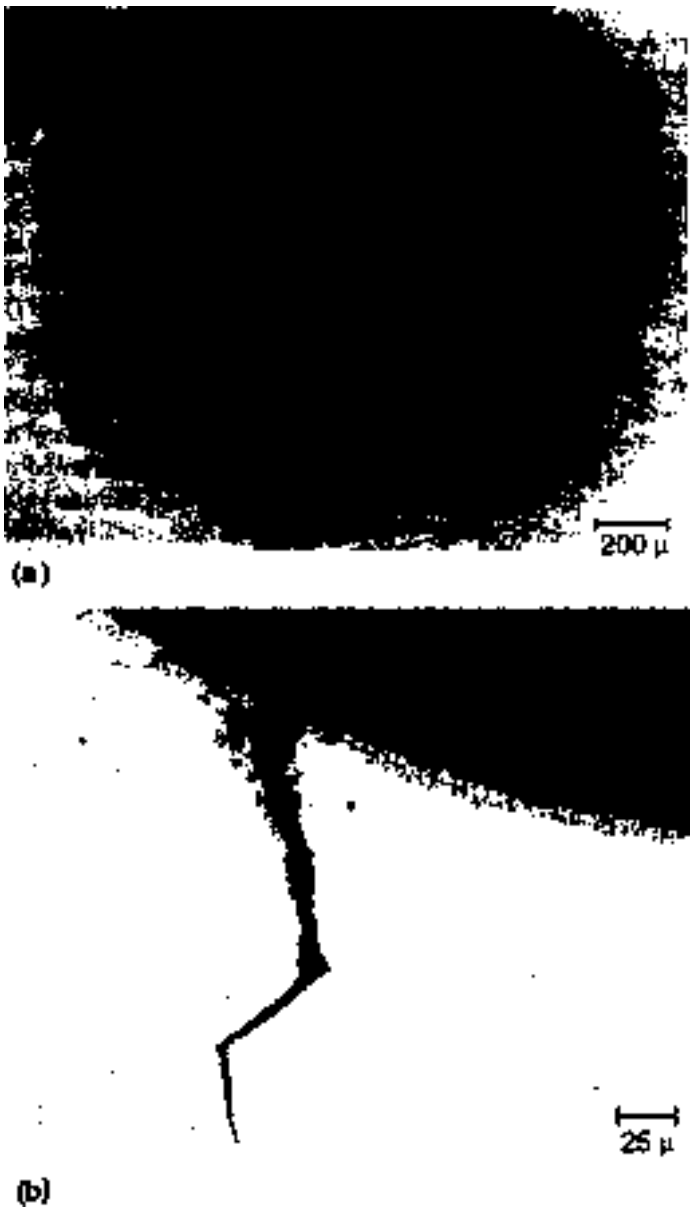


Fig. 21 Oxidation and cracking at cooling holes in a turbine blade. (a) Trailing edge cooling hole surface showing oxidation and nitridation attack on the surface after 32,000 h of operation. (b) Crack found on the surface of No. 5 cooling hole. Oxidation on the crack surface and hole surface can be noticed.

To predict the life of the blade, metal-temperature distribution and stress distribution during service are needed. Obtaining the temperature and stress distribution of the blade sometimes requires complex three-dimensional finite-element aerothermal analysis and stress analysis methods that are not practical in this situation. Thus, other, simpler methods were used to obtain this information to predict the life of these blades.

Analysis of Cooling Hole Oxidation and Temperature Prediction. Internal oxidation and nitridation of the hot section components lead to embrittlement and reduction in the life of the parts. However, an indirect method to estimate the metal temperature using the oxide measurements was developed in this study. The oxidation depth at the blade cooling hole surfaces was measured for each hole. These measurements were used to estimate the metal temperature at the various cooling hole locations. The oxidation depth is assumed to be parabolic with time, and using the Arrhenius relationship between time and temperature, that is:

$$[\log K \cdot \log (d^2/t)][T + 460] = Q/R \qquad \text{(Eq 6)}$$

where log K is constant, d is depth of the oxidation attack, t is time, T is temperature, Q is activation energy, and R is gas constant.

Using the diffusion data available from the literature, the temperature was estimated at 50 and 90% airfoil heights and plotted in Fig. 22 and 23. The details of the calculations can be found in Ref 31.

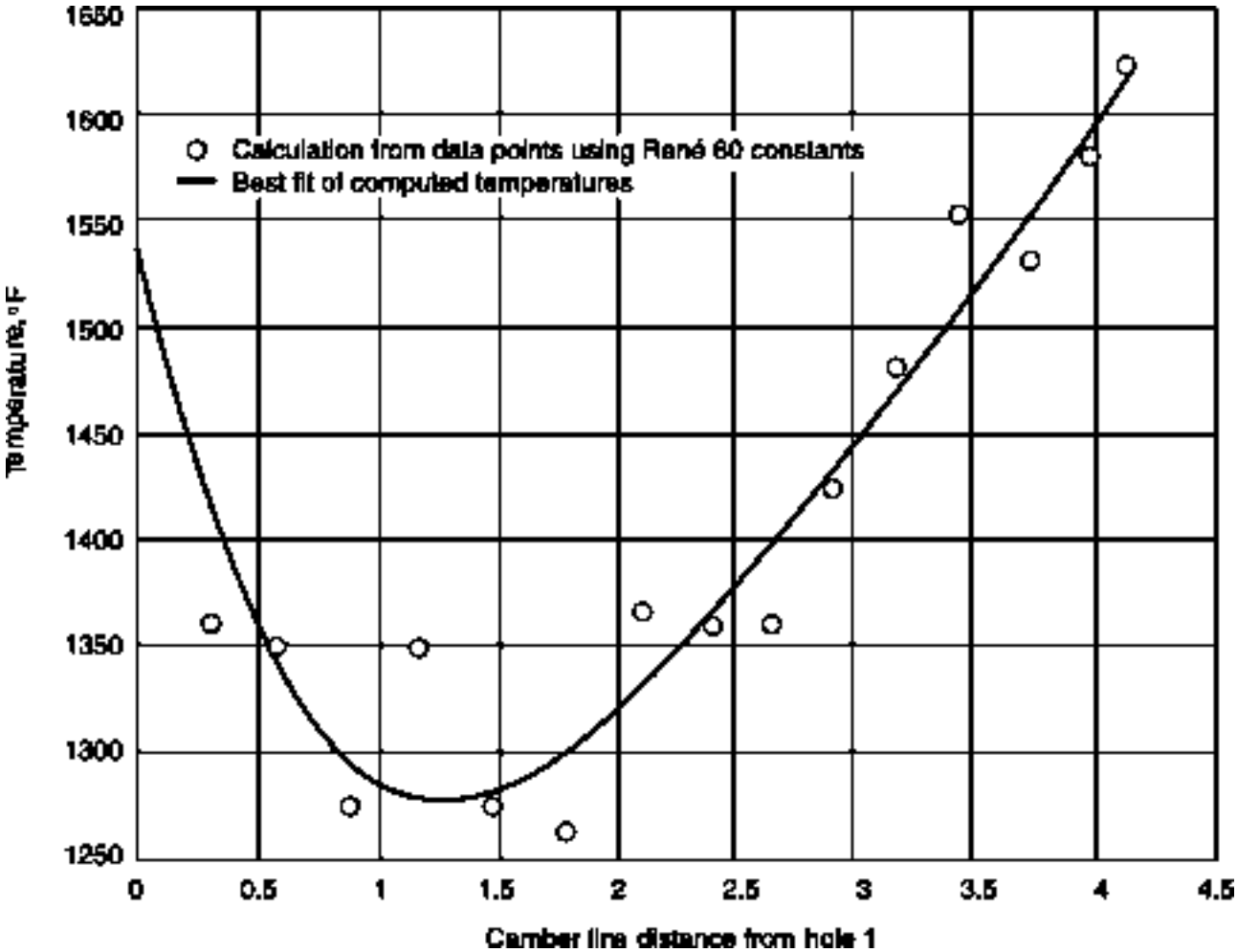


Fig. 22 Predicted temperature using oxide depth measurements at 60% airfoil height

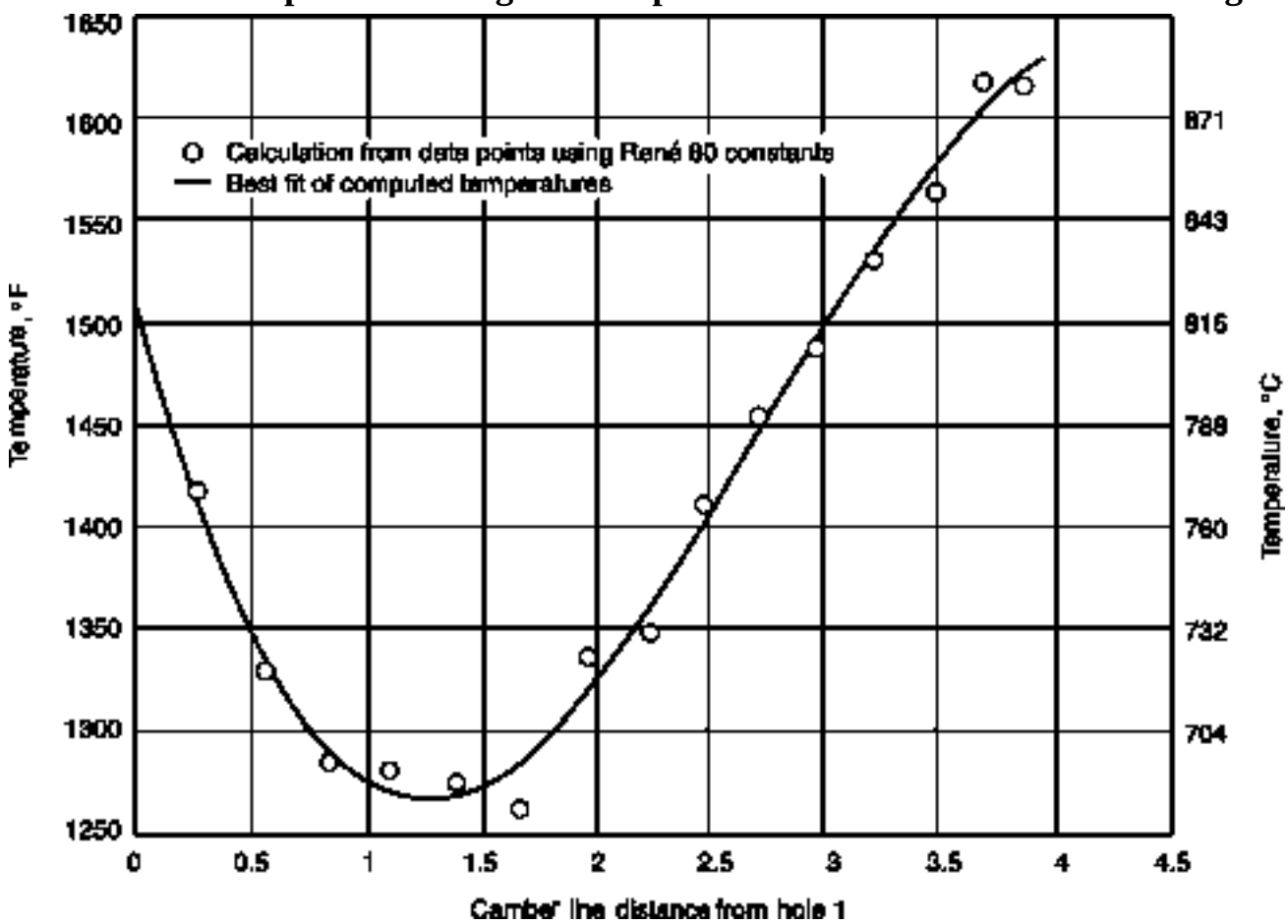


Fig. 23 Predicted temperature using oxide depth measurements at 90% airfoil height

Stress and Strain Analysis. An approximate elastic stress analysis was conducted to determine the nominal elastic stress distribution through the thickness of the airfoil. In addition, the local strain at the edge of the hole (crack initiation location) was estimated using Neuber's rule and stress-strain-creep properties of the material. The nominal elastic stress through the blade airfoil thickness was computed using a flat plat analysis model. Figure 24 shows the temperature and stress distribution at a longitudinal plane through hole 5. In addition, the temperature difference from the airfoil outer surface to the cooling hole was estimated to be approximately 189 °C (340 °F) to cause thermal stresses. At the surface of the cooling holes, high local stresses and strain can occur due to stress concentration from the holes. The stress concentration for the holes is estimated to be 2.5. It is computed that this could lead to local peak elastic stress of 1200 MPa (174 ksi).

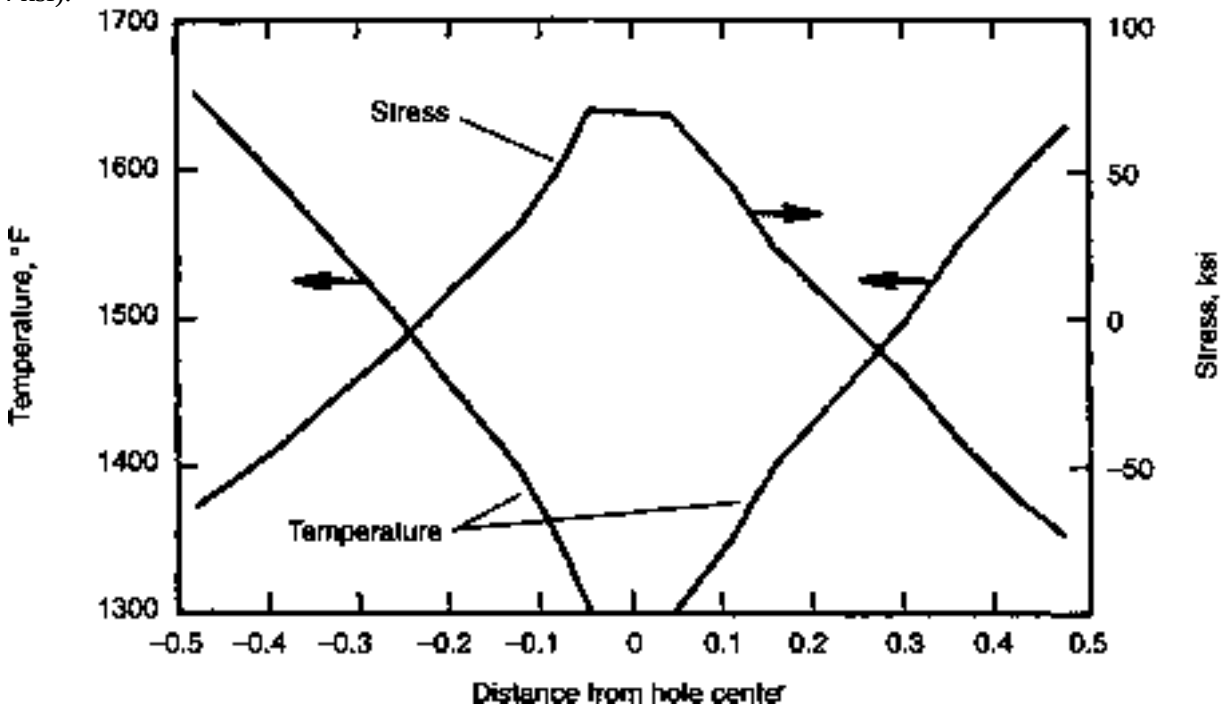


Fig. 24 Temperature and elastic stress distribution along line through hole 5

An inelastic strain analysis was estimated using Neuber's rule. The nominal and local strains are related by (Ref 31):

$$e = K_t^{[2n/(n+1)]} \qquad \text{(Eq 7)}$$

where, e is nominal strain, K_t is elastic stress-concentration factor, and n is the exponent in the power law creep.

It was calculated that the total strain at the cooling hole surface could reach 1.3%, which is large enough to initiate the cracks. Creep relaxation is predicted to occur at the estimated temperature range of 705 to 730 °C (1300 to 1350 °F). The turbine start-stop cycles contribute to TMF stress cycling. Thus, the cracking at the cooling holes was due to:

- Grain-boundary oxidation and nitridation at the cooling hole surface
- Embrittlement and loss of local ductility of the base alloy (Udimet 520)
- Temperature gradient from the airfoil surface to the cooling holes, which led to relatively high thermal stresses at the holes located at the thicker sections of the airfoil
- Stress concentration of 2.5 at the cooling hole and the presence of relatively high total strain (an inelastic strain of 1.2%) at the cooling hole surface

Thus, by the application of the methods given in this case study to estimate the metal temperature and stresses, it is possible to predict the life of turbine blades under similar operating conditions. This method is cost-effective and less time-consuming than a rigorous engineering analysis, which may require sophisticated engineering analysis models, software, and engine operating data.

Life Prediction of Coatings on Gas Turbine Blades

The principal degradation mechanisms for coatings are the same environmental mechanisms that affect uncoated superalloys. Hot corrosion and oxidation attack both occur on coated components and with the same general features as on uncoated blades (Fig. 11). Although the resistance of coatings to attack is generally higher than for uncoated

superalloys, coatings are often employed under very severe conditions of temperature or contamination. The coatings, therefore, have a limited life, after which the base metal is exposed, and accelerated attack can occur.

Land-based gas turbine blades are generally coated with overlay (MCrAlY) or diffusion coatings (plain aluminide or platinum aluminide) to improve the oxidation and hot corrosion resistance. The effectiveness of the coating deteriorates with time after exposure to high temperature and repeated heat-up during start of the engine and cool-down during the engine shutdown cycle. The active element that provides the protection on the outer surface of the coating is aluminum, which forms a protective aluminum oxide, that is, alumina, layer. When the aluminum is depleted during service exposure, the coating loses its ability to protect the substrate, because it loses its ability to form and maintain a continuous alumina layer. This leads to breaching of the coating by the environment. The loss of aluminum occurs by three mechanisms: oxidation, oxide fracture and spallation, and inward diffusion. A schematic diagram of these mechanisms is shown in Fig. 25 (Ref 32, 33).

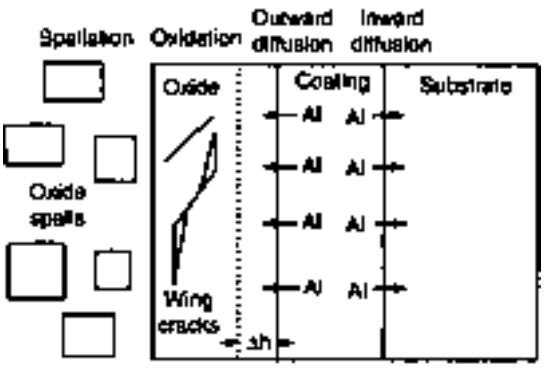


Fig. 25 Schematics of the degradation mechanisms of spalling, oxidation, and inward diffusion for coatings

A critical concentration of aluminum is needed to provide a protective oxide layer. The aluminum is stored in the coating as beta phase, which is NiAl or CoAl phase (Ref 34). This phase dissociates at the high temperatures in which the turbine blades operate—in general, 870 to 925 °C (1600 to 1700 °F)—and the aluminum diffuses to the outer surface to form the protective oxide layer. Some of this aluminum also diffuses inward to the substrate, due to the large aluminum concentration gradient across the coating/substrate interface. The volume fraction of the beta phase decreases with time and number of cycles (Ref 34). The calculated and measured values of the volume fraction of beta phase as a function of thermal cycles for platinum-aluminum-coated GTD-111 alloy substrate is shown in Fig. 26.

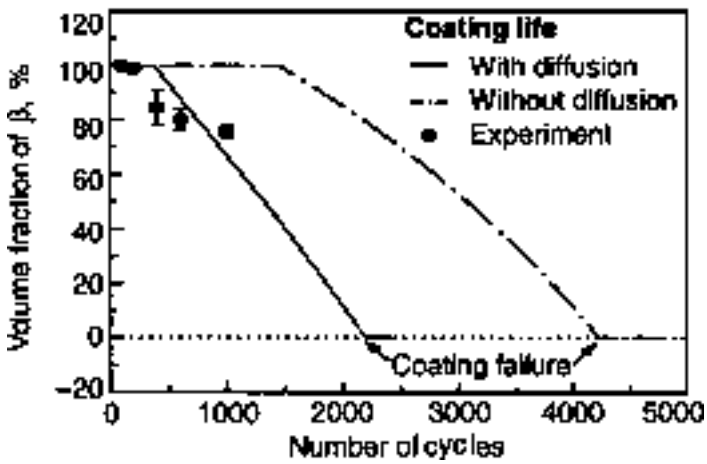


Fig. 26 Calculated and measured values of the volume fraction of beta as a function of one-hour thermal cycles from 25 to 1066 °C (75 to 1950 °F) for platinum-aluminum-coated GTD-111 blade

For the coating life prediction model development, small coupons coated with various coatings were subjected to accelerated thermal cycling in the laboratory between room temperature and a peak temperature of either 955 °C (1750 °F) or 1065 °C (1950 °F). Weight change was monitored with the number of cycles. Material constants needed for the diffusion and oxidation models were thus obtained. Coating life is predicted with and without the diffusion of aluminum and compared, as shown in Fig. 27.

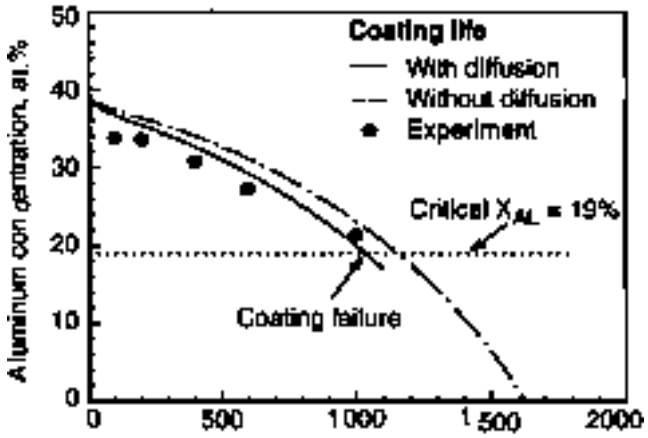


Fig. 27 Comparison of calculated and observed aluminum content as a function of one-hour thermal cycles from 25 to 1066 °C (75 to 1950 °F) for aluminide-coated GTD-111 blade

It was determined that the critical aluminum content was approximately 19% below that at which the coating becomes ineffective in providing protection, and thus, it is considered a failed coating.

Once a set of model constants has been obtained for a coating, the procedure can be used to calculate the time to failure of the coating by varying the cycle time. The calculated values of time to failure are then divided by the cycle time to obtain the number of start-up cycles to coating failure. A coating life diagram is obtained by plotting the number of start-up cycles as a function of cycle time in a log-log plot. Figure 28 illustrates this plot for a MCrAlY coating on GTD-111 alloy substrate.

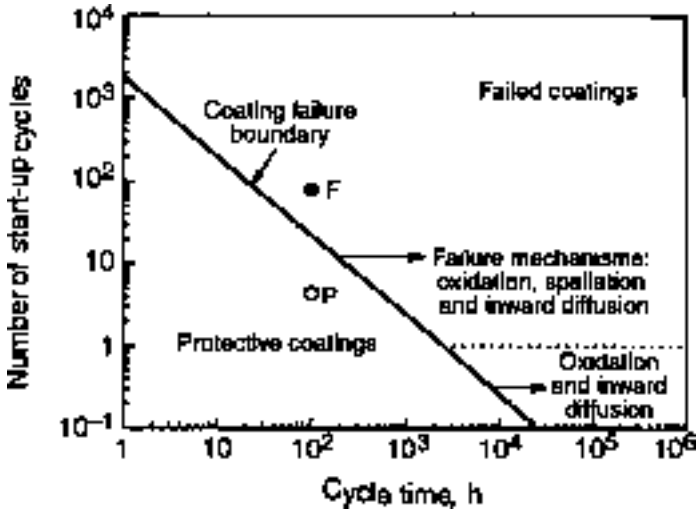


Fig. 28 Coating life diagram calculated using computer program COATLIFE. The diagram shows the protective and failed regimes of MCrAlY-coated GTD-111 blade

The coating life diagram provides a simple and rigorous means for forecasting the remaining life of a coating after it has been in service. The controlling failure mechanism along the coating failure boundary varies with cycle time and temperature. In general, oxidation and spallation dominate at short cycle times, while oxidation and inward diffusion dominate at long cycle times. At intermediate times, the coating failure is controlled by oxidation, spallation, and inward diffusion.

References cited in this section

3. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components,* ASM International, 1989, p 188–190, 214, 228, and 369

4. D.J. Benac, *Elevated-Temperature Failures,* Lesson 7, ASM Practical Failure Analysis Course, 2002

9. V.P. Swaminathan and P. Lowden, “Gas Turbine Blade Life Assessment and Repair Guide,” EPRI Report GS-6544, 1989

10. C.T. Sims, N.S. Stoloff, and W.C. Hagel, Ed., *Superalloys II,* John Wiley & Sons, 1987

18. *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989

19. Y. Lieberman, Relaxation, Tensile Strength and Failure of EI 612 and 20Khl F-L Steels, *Term. Obrabotka Met.,* Vol 4, 1962, p 6–13

20. R.V. Hart, Assessment of Remaining Creep Life Using Accelerated Stress-Rupture Tests, *Met. Technol.,* Vol 3, Jan 1976, p 1–7

21. R.V. Hart, “The Validity of Life Fraction Rule—A Review of Experimental Work,” Report SSD/NE/N. 189, CEGB, 1979

22. F.R. Larson and J. Miller, A Time-Temperature Relationship for Rupture and Creep Stresses, *Trans. ASME,* July 1952, p 765–775

23. R.L. Orr, O.D. Sherby, and J.E. Dorn, Correlation of Rupture Data for Metals at Elevated Temperatures, *Trans. ASME,* Vol 46, 1954, p 113–128

24. S.S. Manson and A.M. Haferd, NASA Report TN-2890, National Aeronautics and Space Administration, March 1953

25. S.A. Manson, “Time-Temperature Parameters—A Re-evaluation and Some New Approaches,” American Society for Metals, June 1969

26. E. Campo, D. Ferrara, and S. Quaranta, Analysis of Creep Rupture Data of Nimocast Alloy 739, *High Temperature Alloys for Gas Turbines—1982,* R. Brunetaud et al., Ed., D. Reidel Publishing Co., p 575–578

27. E.A. Crombie, W. McCall, C.G. Beck, and D.M. Moon, Degradation of High-Temperature Impact Properties of Nickel Base Gas Turbine Blade Alloy, *Proc. of the 12th CIMAC Conference,* Vol C, (Tokyo), 1977

28. K. Kuwabara, A. Nitta, and T. Kitamura, Thermal-Mechanical Fatigue Life Prediction in High Temperature Component Materials for Power Plants, *Proc. of Intl. Conf. Advances in Life Prediction,* April 1983 (Albany), American Society of Mechanical Engineers, p 131–141.

29. G.T. Embley and E.S. Russell, Thermal-Mechanical Fatigue of Gas Turbine Bucket Alloys, *First Parsons International Turbine Conference,* June 1984 (Dublin, Ireland), p 157–164

30. E.S. Russell, Practical Life Prediction Methods for Thermal-Mechanical Fatigue of Gas Turbine Buckets, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials,* EPRI AP 4477, Project 2382-3, Electric Power Research Institute, April 1986, p 3–1 to 3–17

31. V.P. Swaminathan, J.M. Allen, and G.L. Touchton, “Temperature Estimation and Life Prediction of Turbine Blades Using Post Service Oxidation Measurements,” presented at the International Gas Turbine and Aero Engine Congress & Exhibition, June 1996 (Birmingham, U.K.), American Society of Mechanical Engineers Paper 96-GT-528

32. K.S. Chan, S.N. Cheruvu, and G.R. Leverant, “Coating Life Predication under Cyclic Oxidation Conditions,” presented at the International Gas Turbine & Aero Engine Congress & Exhibition, June 1997 (Orlando, FL)

33. K.S. Chan, S.N. Cheruvu, and G.R. Leverant, Coating Life Prediction for Combustion Turbine Blades, presented at the International Gas Turbine & Aero Engine Congress & Exhibition, June 1998 (Stockholm, Sweden)

34. S.N. Cheruvu, K.S. Chan, and G.R. Leverant, "Cyclic Oxidation Behavior of Aluminide, Platinum Modified Aluminide and MCrAlY Coatings on GTD-111," presented at the International Gas Turbine & Aero Engine Congress & Exhibition, June 1998 (Stockholm, Sweden)

Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

Life Assessment Methods for Power Plant Piping and Tubing

Similar to gas turbine blades, piping and tubing in steamlines, heaters, boilers, and superheaters are subjected to elevated temperatures that can cause degradation, deformation, bulging, cracking, or bursting. Thus, it is possible for the life expectancy to be reduced. Therefore, it is often necessary that piping and tube life assessments be conducted. Some of the same methods used for gas turbine high-temperature life assessment can be applied for plant piping and tubing exposed to elevated temperatures. The following techniques for high-temperature power plant piping and components are discussed in this article (Ref 35, 36):

- Hardness testing
- Microstructural evaluations
- Creep cavitation damage assessment
- Stress-rupture tests (life fraction)
- Oxide-scale-based life prediction
- High-temperature crack growth methods

Hardness-Based Techniques. Traditionally, microstructure-based and hardness-based techniques have been used to assess the condition of piping and tubes. In the past, they have been used mainly to estimate the "equivalent temperatures" of the tubes, which are then used in conjunction with standard rupture data to calculate the fractional creep life consumed. Relationships have also been established between the degree of creep cavitation, as measured in replicas, and the creep life fraction consumed. In those instances, where oxide scale information is simply unavailable (e.g., due to spalling or chemical cleaning), the hardness- and microstructure-based technique is viable to assess the tube temperatures, provided there are comparative hardness data and microstructural comparisons. Even where oxide scales are intact, equivalent temperature estimates from hardness and microstructure can be used as additional verification of the equivalent temperature predictions based on scale thickness.

One area where hardness and microstructure have been used is in boiler tube assessments. The function of the boiler is to convert water into superheated steam, which is then delivered to a steam turbine. The types of tubes in a utility boiler include waterwall, superheater, economizer, and radiant tubes. Failure of boiler tubes is the foremost cause of forced boiler outages in the United States and other countries. It is estimated that more than $five billion per year is spent in replacement power and maintenance costs.

Boiler tube failures attributed to high-temperature environments include stress rupture, fireside corrosion, and thermal fatigue (Ref 3). Stress-rupture failures can be attributed to short-term overheating or long-term, high-temperature creep. Often, abnormal conditions cause failure such as internal blockage, loss of coolant, overfiring of boiler burners, and reduction in heat transfer due to oxide scale formation or deposit accumulation.

One way to avoid failure and increase the life expectancy of boiler tubes is to select the proper alloy for the intended high-temperature application. Boiler tube materials are selected depending on the designed maximum operating temperature and stress condition. Some of the key material properties to consider are thermal expansion coefficient, stress rupture, fatigue life, oxidation resistance, and environmental effects. A listing of some high-temperature materials and some generally accepted maximum metal-temperature limits on the outside surface is given in Table 3. These values are not design limits and may be different, depending on the boiler manufacturer, code requirements, and the anticipated oxidation resistance. As a general rule, as the chromium and molybdenum percent increases, the maximum allowable tube temperature increases.

Table 3 Maximum metal temperatures for high-temperature boiler tube materials

Tube type	Temperature, °C (°F)
Carbon steel	510 (950)
Carbon-$\frac{1}{2}$Mo (T1)	524 (975)

$1\frac{1}{4}$Cr-$\frac{1}{2}$Mo (T11)	552 (1025)
$2\frac{1}{4}$Cr-1Mo (T22)	580 (1075)
9 Cr-1Mo (T9)	650 (1200)
Stainless steel (type 304H)	704 (1300)
Stainless steel (type 347H)	704 (1300)

The strength of low-alloy steel changes with service exposure in a time-and-temperature-dependent manner. Thus, any measure of change in strength during service (e.g., change in hardness) may be used to estimate a "mean" operating temperature for the component. This information, in conjunction with stress-rupture data, can be useful to estimate the life. This hardness approach is particularly suitable when strength and hardness changes in service occur primarily as a result of carbide precipitation and growth (microstructural coarsening), and strain-induced softening can be neglected.

The effects of temperature on a material are often characterized using a LMP. For some materials, hardness changes have been described using the LMP. The tempering responses of steels at typical service temperatures, as evidenced by hardness changes influenced by time (*t*) and temperature (*T*) of exposure, often are described by the LMP, discussed in the section "Gas Turbine Blade Life Assessment" in this article. For example, typical correlations between hardness and the LMP for aging of 1Cr-$\frac{1}{2}$Mo, 2Cr-1Mo, and 9Cr-1Mo steels are given in Fig. 29 (Ref 35). If the service exposure duration and the current hardness are known, Fig. 29 can be used to estimate the equivalent temperature of the tube. A common problem in using this approach is that the initial hardness of the material is often unknown, and assumptions have to be made based on typical values obtained on similar material.

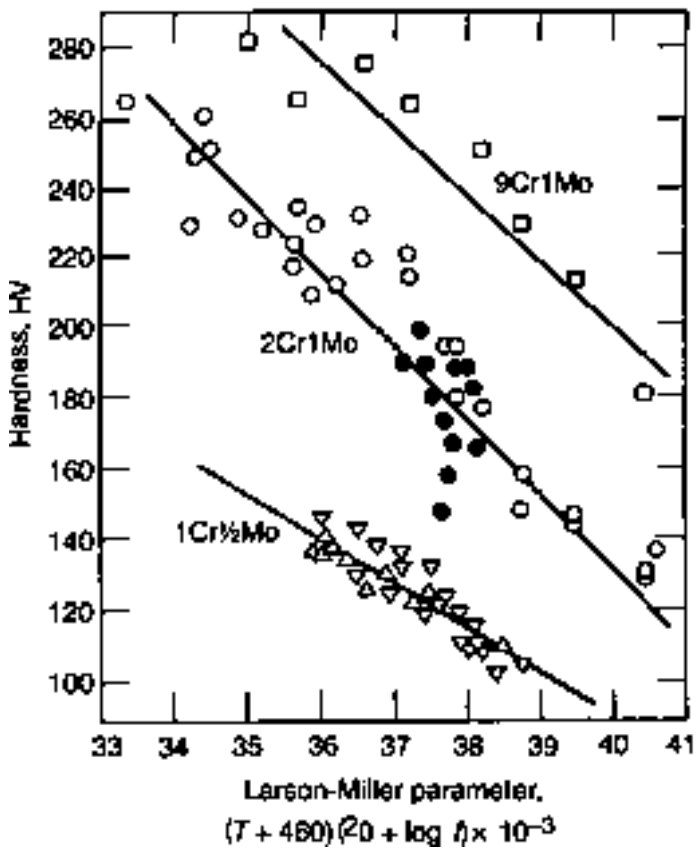


Fig. 29 Correlation between hardness and Larson-Miller parameter for alloy steels 1Cr-$\frac{1}{2}$Mo, 2Cr-1Mo, and 9Cr-1Mo

The hardness method may be used nondestructively for estimation of exterior-surface operating temperature. When performed destructively, this method can provide tube-wall operating thermal gradients. This method is relatively inexpensive, requiring no sophisticated equipment, but has several limitations. At low loads (e.g., 100 g), the sensitivity of hardness to local microstructural variation can result in significant scatter and poor reproducibility. This is less so at high loads (e.g., 20 kg). Strain-softening effects, if present, can cause an erroneous increase in the estimated mean service temperature. The load used for determining temperature should be equal to that used for deriving the correlation. For

nondestructive use of this method, tube access may be limited. For a detailed description of hardness techniques, the reader is referred to Ref 3.

Microstructural Cataloging and Analysis of Carbides. Similar to turbine blades, alloy steel tubes and piping microstructures can alter due to elevated-temperature exposure. Consequently, the life can also be decreased. Spheroidization, shown in Fig. 5, may be cataloged to understand how the life was influenced. Taft and Mardsen demonstrated that there are six stages of spheroidization of carbides in ferritic steels. Using a Sherby-Dorn-type parametric equation, they showed that it was possible to get a reasonable correlation of microstructure with a "weighted average" service temperature (Ref 37). Similar semiquantitative and qualitative approaches involving microstructural "cataloging" as a function of service history have been used by others (Ref 38, 39). This procedure suffers from the limitation that the current state of the microstructure depends on the starting microstructure prior to service, which is often unknown and may not necessarily correspond to the reference-starting microstructure in the laboratory samples. One way to circumvent this problem is to remove samples from the relatively colder regions of the component and assume that they represent the typical starting microstructure. If the typical starting structure is reasonably close to the starting structure in the laboratory samples, then the temperature estimate is expected to be good. In the absence of such information, a catalog of microstructural changes for different starting structures, and a scatterband of behavior, must be established.

Attempts have been made to use parameters such as matrix-solute compositions, ratios of carbide phases, and matrix-lattice parameters as quantitative indices of temperature during aging of 1Cr-$\frac{1}{2}$Mo steels (Ref 40). Nakatani and Sugita have also outlined a procedure for direct estimation of life fraction consumed based on the percent M_6C present in 2 $\frac{1}{4}$Cr-1Mo steels (Ref 41, 42). For more details, see the references. They conducted accelerated laboratory tests on simulated heat-affected zone (HAZ) samples, which had also been previously aged to produce different amounts of M_6C, and established a relation between percent M_6C and rupture-life degradation. Correlations are often made of the percent M_6C compared to the Larson-Miller curve (Fig. 30). Figure 30 shows that the formation of M_6C was accelerated by increased phosphorus (P) and unaffected by increased stress.

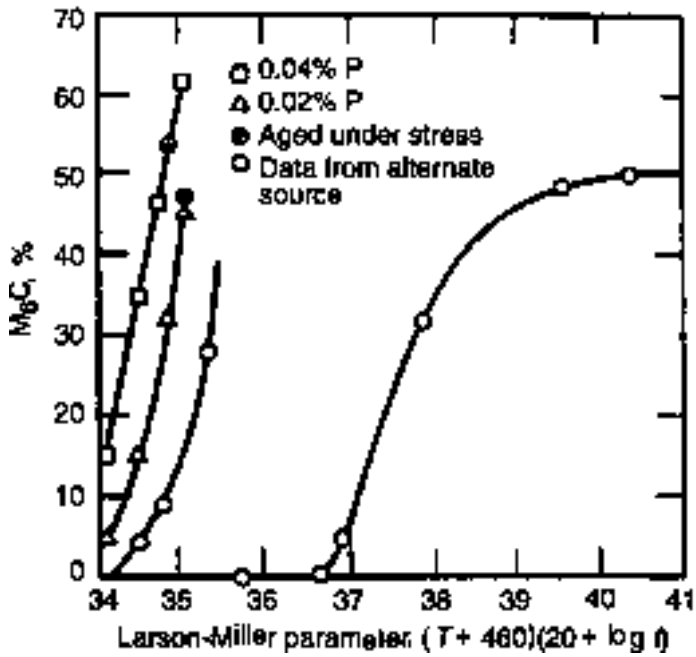


Fig. 30 Correlation between evolution of M_6C and Larson-Miller parameter for a 2$\frac{1}{4}$Cr-1Mo steel

Creep Cavitation Model. Metallographic techniques that track creep cavitation are principally applicable to the tube and piping welds and associated HAZs. Base metal failures in tubes and piping often occur due to softening, and creep cavitation appears too late in the failure cycle to be a useful monitoring index.

For example, several categories of pipes carrying high-temperature/pressure steam contain welds that may be of concern. Main steam pipes are pipes that carry steam at 540 to 565 °C (1000 to 1050 °F) to the high-pressure turbine. These pipes are small in diameter and do not contain seam welds; hence, only girth welds are of concern. The main steam pipes are, however, often connected to the steam header using thick-walled seam-welded piping. In addition, hot reheat pipes, which carry steam at 540 to 565 °C (1000 to 1050 °F) but at a lower pressure (than the main steam pipe) to the reheat intermediate-pressure turbine, are frequently made of seam-welded piping.

Replication has been widely used to characterize creep cavitation in piping, but quantitative correlations of cavitation with remaining life are sometimes lacking. Neubauer and Wedel characterized cavity evolution in steels at four stages—isolated cavities, oriented cavities, linked cavities (microcracks), and macrocracks—and formulated recommendations corresponding to the four stages of cavitation (Ref 43). Figure 31 shows the creep cavitation damage as a function of creep life. The damage classifications have therefore been correlated with life fractions, and thus a life fraction range has been established for each class. The Wedel-Neubauer classifications of material condition (undamaged, class A, class B, class C, and class D) correspond roughly to expended life fraction (t_i/t_r) values of 0.27, 0.46, 0.65, 0.84, and 1, respectively, using the conservative lower-bound curve shown in Fig. 31:

$$t_{cem} = t\frac{t_r}{t} - 1 \qquad \text{(Eq 8)}$$

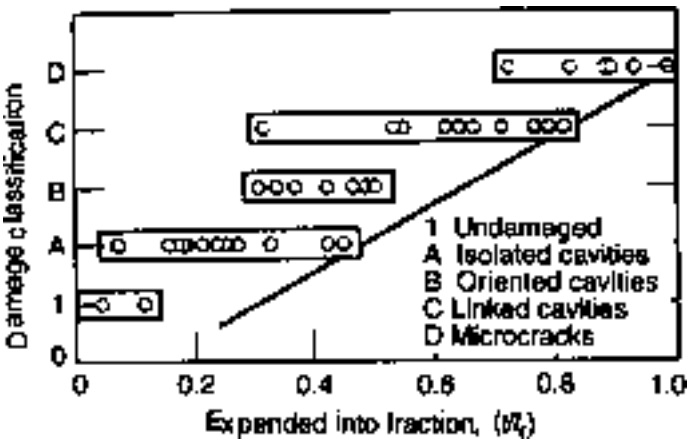


Fig. 31 Creep cavitation damage classification as a function of expended life for $1\frac{1}{4}$ Cr-$\frac{1}{2}$ Mo steels

In recent years, catastrophic failures have occurred of seam-welded hot reheat pipes in the United States (Fig. 32). These failures have been related to creep cavitation formed at elevated temperatures. The hot reheat pipe failure at the Mojave Power Station in 1986 is an example of a catastrophic failure (Ref 44). The failure resulted in the loss of life and millions of dollars as the result of downtime costs and litigation. The pipes in question are typically made of chromium-molybdenum steels and carry high-pressure steam from the boiler to the turbine at or above a temperature of 540 °C (1000 °F). Most of the failures have occurred in hot reheat pipes, because these pipes tend to be larger in diameter compared to main steam pipes and hence are more likely to be seam welded. Failures have also occurred in a few instances where short segments of seam weld pipes have been used to link the header to the main steam line.

Fig. 32 Photograph of catastrophic fishmouth rupture of seam-welded high-energy piping. These failures are typically brittle and are not predicted using simple life fraction rule calculations.

Analysis of the seam-welded pipe failures has created awareness of many issues affecting the life of weldments. Most of the field failures have occurred at calculated creep life fractions well below even 50%, sometimes as low as 15%. Some of the contributing factors to this type of premature failure have been identified as the presence of aligned nonmetallic inclusions at the fusion line, which promote creep cavitation followed by rapid linkup to form large cracks; inherently inferior rupture strength of weldments; geometric factors, such as weld bead cusps and ovality of pipe, which, in combination with the weld/base metal creep strength mismatch, create local stress concentration and through-wall stress gradients.

While replication is very useful for detecting surface damage, many types of failures, such as long seam weld and damage in the fine-grained HAZ in girth welds, originate sub-surface. In these cases, replication alone is not a reliable method to detect damage. In long seam welds in hot reheat piping and header link piping, high-sensitivity conventional or automated ultrasonic testing, focused beam ultrasonic testing, or time-of-flight diffraction ultrasonic testing methods are needed to ensure safety of the piping. In the case of girth welds, however, conventional ultrasonic testing seems to be adequate.

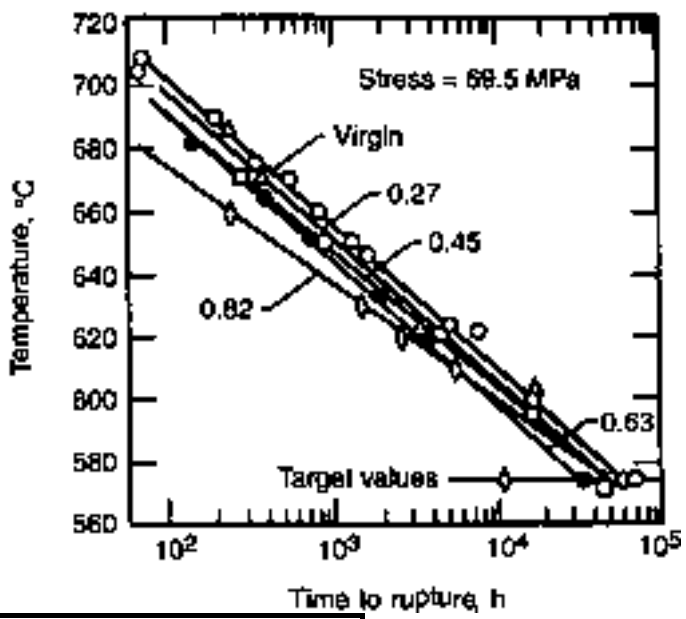
Life Fraction Testing. In the section "Gas Turbine Blade Life Assessment," the life fraction rule (LFR) was discussed. The LFR is also used for tubing and piping. The LFR states that at failure:

$$\sum t_i/t_r = 1 \qquad \text{(Eq 9)}$$

where t_i is the same time spent at a given stress and temperature, and t_r is the rupture life for the same test conditions. In the past, it has been held that the LFR is valid for temperature changes but not for stress changes. This is the basis of the use of isostress-rupture test procedures for estimating the remaining creep life of components.

Accelerated testing is often used to predict the remaining life. However, there may not always be a good match. Figure 33 shows stress-rupture tests on material in a virgin condition and the predamaged condition. The target value at 575 °C (1065 °F) is also shown in the table that accompanies Fig. 33. If the LFR is valid, the temperature versus time-to-rupture line should be parallel to the virgin material. This is not the case, however. For example, the 0.82 predamaged samples overpredicted the remaining life compared to the calculated value based on LFR. Thus, use of the LFR may need to be verified for a particular material.

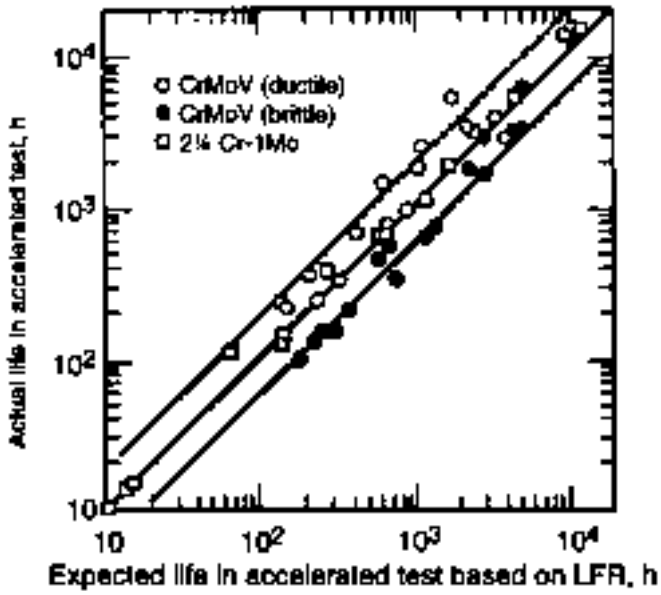
Fig. 33 Stress-rupture curves for virgin material and predamaged material showing the various life fractions. Virgin material rupture life at 575 °C (1065 °F) is 62,210 h



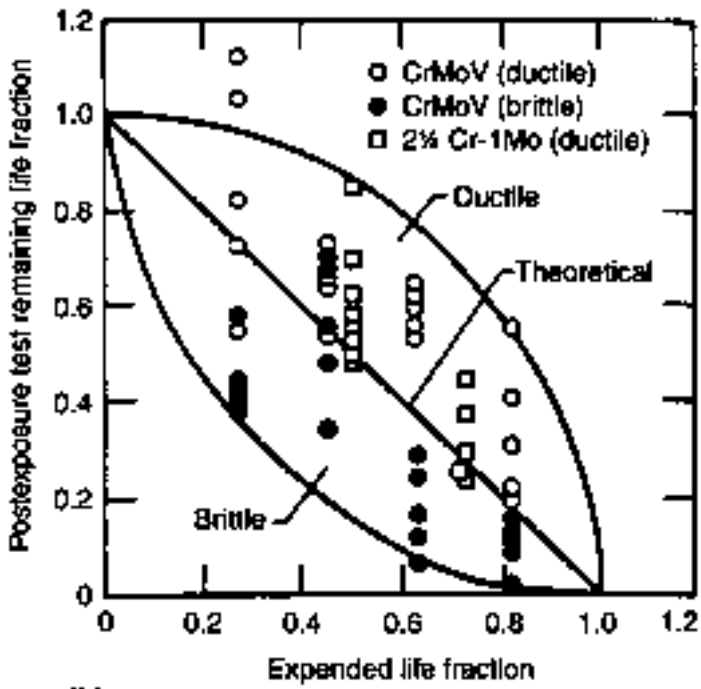
Estimated remaining life at 575 °C (1065 °F) for predamaged samples based on:			
Symbol	**Damage fraction**	**Life-fraction rule, h**	**Linear extrapolation, h**
△	0.27	45,413	49,900
□	0.45	34,215	48,300

•	0.63	23,017	33,000
◊	0.82	11,197	45,000

Some consideration should be given to the ductility of the material. Figure 34(a) shows the expected life based on LFR and the observed life for the three heats of material. For the brittle material, the actual life is shorter than the expected life. For the ductile material, the expected life agrees closely.



(a)



(b)

Fig. 34 Effect of ductility on life-fraction estimates. CrMoV (ductile) is a ferric microstructure with approximately 15% elongation (high ductility) under long-term service conditions. CrMoV (brittle) is a bainitic microstructure with 3.5% elongation (low ductility). $2\frac{1}{4}$Cr-1Mo (ductile) is a ductile tubing steel. (a) Correlation between expected life based on life fraction rule and the actually observed life in postexposure accelerated tests. (b) The expended life fraction under service conditions versus the remaining life fraction as determined from postexposure accelerated tests for three heats of steel

The correlation between prior expended life fraction and the remaining life fraction measured in the accelerated tests shown in Fig. 34(b) may be used as a modified basis for LFR. For a ductile material, the $\sum t_i/t_r = 1$ line may be used for a conservative prediction of remaining life. For a brittle material, a value of <1, as defined by the lower-bound envelope in Fig. 34(b), may be used as a conservative failure criterion. The practical implications of these results are:

- In calculating cumulative damage under changing operating conditions, the ductility of the material should be kept in mind. For instance, in superheater tubes, wherein the metal temperature is continuously increasing, $\sum t_i/t_r = 1 > 1$ or < 1 should be used for ductile and brittle conditions, respectively.
- For brittle materials (e.g., prone to type IV cavitation, fusion line cavitation, etc.), temperature-accelerated tests may cause premature failure, and isostress extrapolation of the accelerated test results may lead to pessimistic (conservative) prediction of remaining life at the operating temperature. The actual remaining life under operating conditions may well exceed the estimates from the accelerated tests.

There can be limitations to using the life fraction rule. Remaining life estimation of thick-section pipe weldments by isostress-rupture testing can be complicated by the inability to choose the correct equivalent stress in unixial tests and to reproduce the specimen size and constraint effects. There often is uncertainty in the method of calculating the stress, especially for thick-wall components. The appropriate equivalent stress calculation, whether on the basis of Von Mises, maximum shear, maximum principal, or other possible governing stress criterion, may depend on the material, temperature, stress, and stress state. Temperature gradients in thick-walled components and stress redistribution by creep render the stress estimations even more difficult. Stress redistribution in a weldment is more complex, in view of the different microstructures present with differing creep properties. Presence of material inhomogenities and inclusions, presence of weak zones, and stress concentration factors can localize the damage to specific interfaces or zones, causing crack growth rather than bulk creep damage and rupture. The manner in which these influences are felt in a component under service conditions is often different from what happens in accelerated tests on uniaxial specimens. This is the reason why, in laboratory tests, it may be difficult to simulate the field failures in terms of failure location. Cyclic service can further contribute to life reductions due to creep-fatigue interaction in a manner that cannot be simulated in accelerated uniaxial, noncyclic stress-rupture tests (Ref 45).

Failures of most of the seam-welded piping have occurred prematurely and could not be predicted based on simple LFR calculations. Life fraction calculations based on metal rupture curves in case of weldments are rendered erroneous due to inherently weaker properties of some welds and due to stress-intensification effects arising from weld-strength mismatch effects. Figure 35 is a comparison of the available cross-weld specimen rupture data on ex-service piping with the American Society of Mechanical Engineers (ASME, or ASTM) data for comparable-size specimens of $2\frac{1}{4}$Cr-1 Mo base metal, respectively (Ref 45). The as-measured, cross-weld data are consistently below the ASME average curve and often below the minimum curve. To eliminate the degradation effect of service exposure, the data should be “corrected” for service exposure by adding to the actual ex-service data an estimate of the equivalent exposure time (for the test temperature and stress) representing the life fraction consumed in service. The life fraction consumed was estimated by extrapolating the observed stress-LMP rupture data to the service stress (mean diameter hoop stress due to pressure). In most instances, this extrapolation had to be made, because isostress data were insufficient for a direct estimate. The corrected data are intended to represent an estimate of “new” weld properties extrapolated from the ex-service properties. The new cross-weld data inferred from ex-service data were also consistently below the ASME average curve and often at or below the minimum curve. While the potential inaccuracy in extrapolating from ex-service to new material rupture properties is recognized, the data nevertheless show that the performance of weldments can be inferior to that of base metal when compared on the basis of conventional estimations of stress and stress-LMP curves.

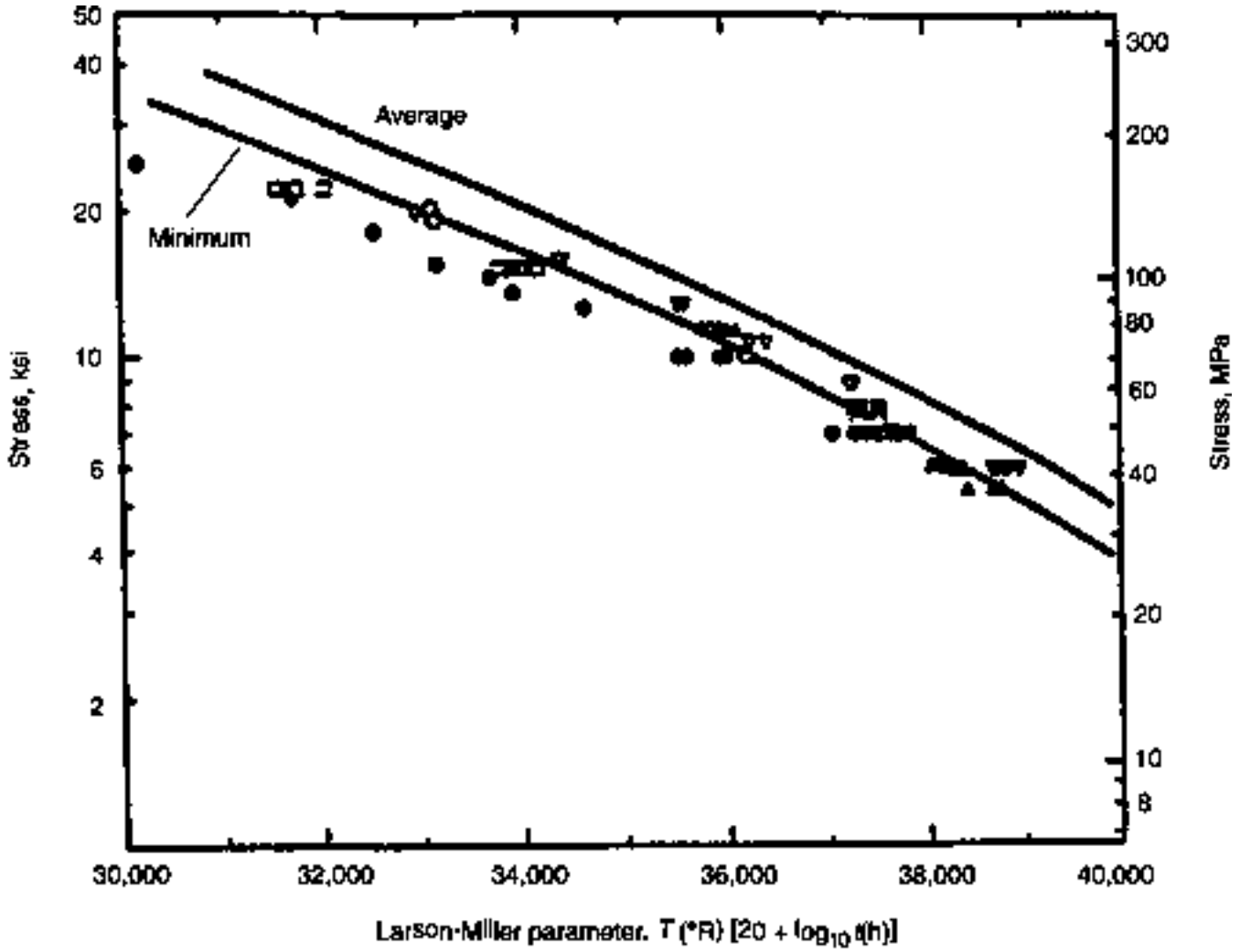


Fig. 35 Cross-weld stress-rupture data for $2\frac{1}{4}$Cr-1Mo ex-service pipe seam weldments corrected for service exposure

Oxide-Scale-Based Life Prediction. Life assessment of boiler tubes using the oxide scale thickness has proven very effective. Creep-rupture failure of superheater (SH) and reheater (RH) tubes is a major cause of forced outages of boilers. The objectives of life assessment of SH and RH tubes are to avoid forced outages and to enable scheduling of appropriate inspection intervals and eventual SH and RH replacement in a cost-effective manner. To achieve these objectives, it is important that decisions be based on the overall condition of the SH or RH, including straight sections and bends in the furnace tubes, the penthouse tubes, stub tube welds, and dissimilar-metal welds in the system. Failure at any of these locations leads to forced outage, regardless of the condition of material at the other locations.

Figure 36 depicts an overall strategy for assessing SH and RH assemblies (Ref 46). The figure reflects the point that the appearance of distress can take different forms at different locations, and hence, a different mix of evaluation tools may be applicable in each case.

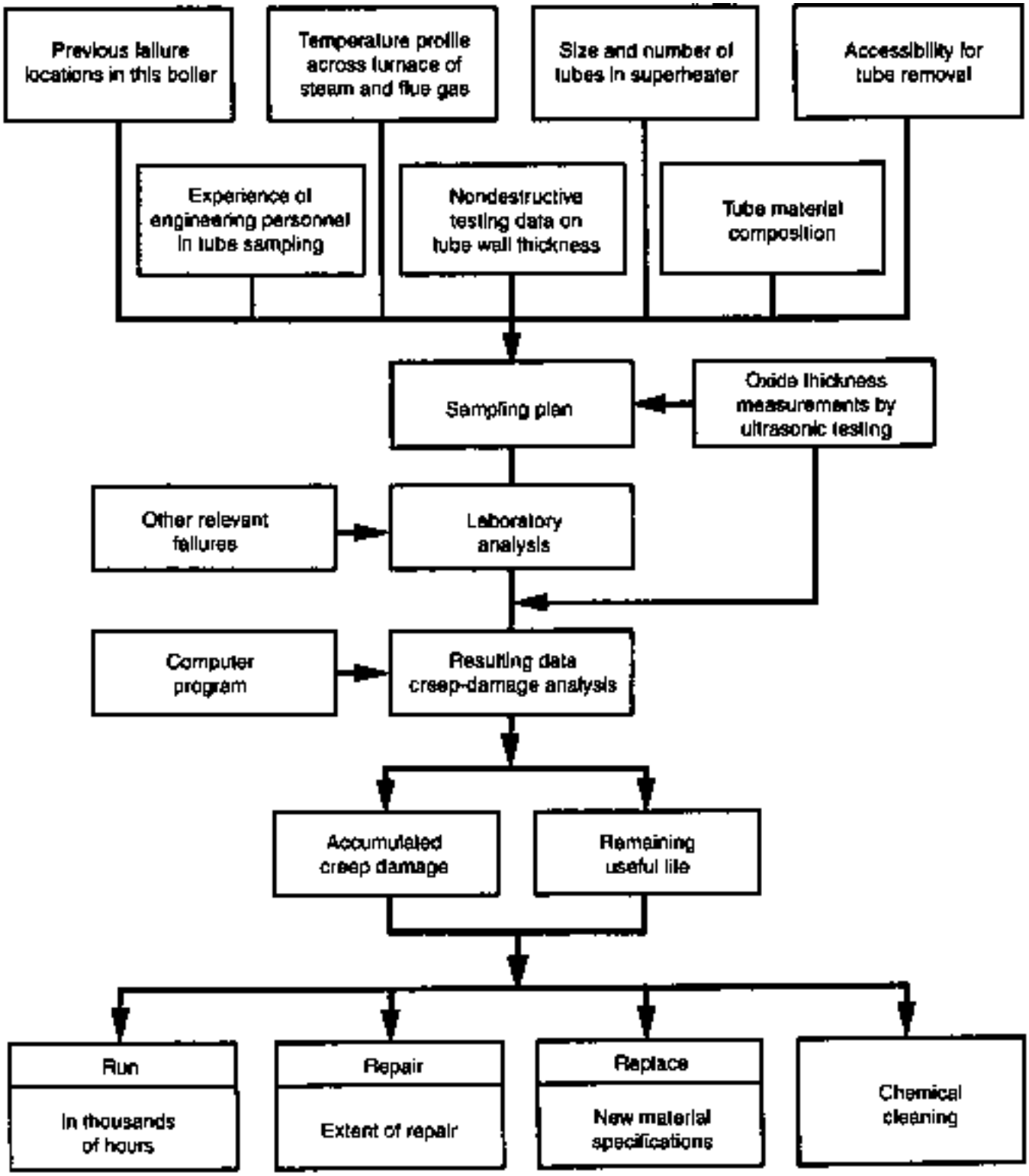


Fig. 36 Strategic approach flow diagram for general tube sampling and life-assessment plan for superheater and reheater

The earlier methods that were employed for making tube replacement decisions were based on calculations involving operating history, failure frequency, or destructive testing of tubes. A calculation procedure was developed that involves nondestructive measurement of steamside scale thickness and wall thickness.

The validity of measuring the oxide scale thickness ultrasonically has also been demonstrated by establishing a good correlation between ultrasonic measurements and direct measurements on tube samples removed from SH assemblies from several fossil boilers. Figure 37 shows the oxide scale buildup thickness versus thermocouple measurements for a $2\frac{1}{4}$Cr-1Mo tube compared to predicted values using a code. Comparison of uniaxial test results with tube burst test results has also helped identify the correct formulas for computing the reference stress in tubing, replacing a more conservative formula used in the past. The kinetics of steamside oxide scale growth in heat-flux and nonheat-flux areas

are also being measured in boilers. Cumulatively, these results have provided industry with needed confidence in its application of a more cost-effective and nondestructive method for assessing the condition of tubing.

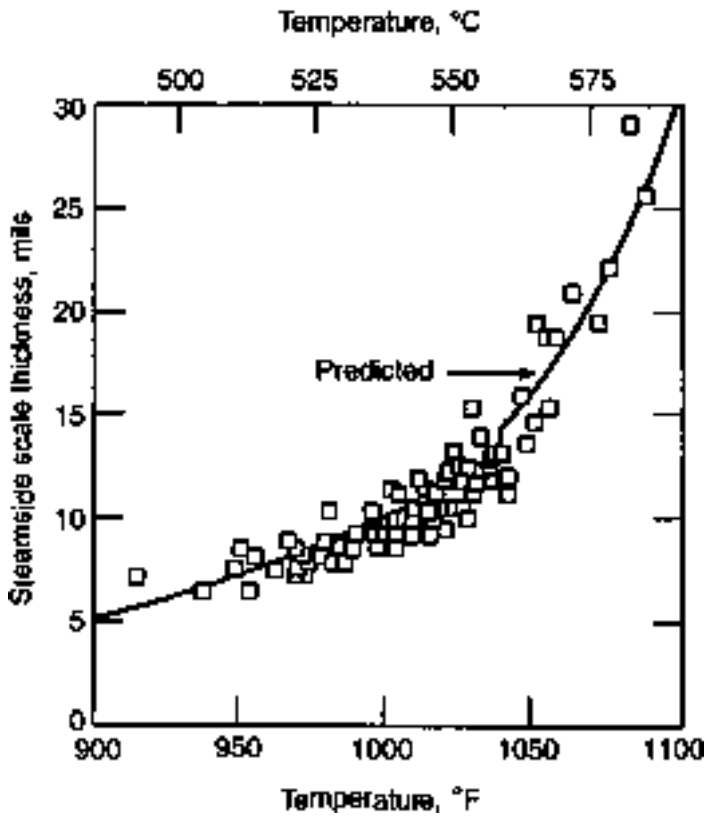


Fig. 37 Steamside oxide scale buildup thickness versus thermocouple measurements for a $2\frac{1}{4}$Cr-1Mo tube compared to predicted values using a code

Based on steamside scale thickness and tube wall thickness, a code reconstructs the thermal and stress history of a tube and calculates the creep-damage accumulation for that history. Although it is a calculation technique, the input (i.e., oxide scale and wall thickness) to it is provided from ultrasonic measurements at locations, that are selected based on historical information. Approximate oxide growth laws are used to predict the estimated thickness. A typical set of oxide curves is shown in Fig. 38.

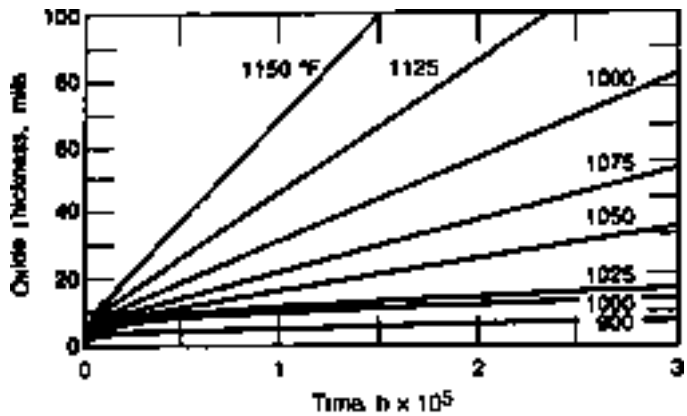


Fig. 38 Steamside oxide thickness as a function of time and temperature for $2\frac{1}{4}$Cr-1Mo steel

In a code, the current steamside scale thickness is used in conjunction with an isothermal oxide growth rate law to determine the “equivalent” temperature history of the tube. During service, SH and RH tubes oxidize according to the reactions:

$$4Fe + 3O_2 \rightarrow 2Fe_2O_3$$
(metal + oxygen) (hematite)

$$3Fe + 2O_2 \rightarrow Fe_3O_4$$
(metal + oxygen) (magnetite)

If thick, steamside scales are present and the tube is exposed to temperatures approaching or above 595 °C (1100 °F), then wustite (FeO) may also form. The tube dimensions are used to calculate the current "effective stress." The temperature and stress values are then extrapolated back to the initial conditions, assuming linear growth of steamside scale and fireside wastage and the known heat-transfer properties of the steel and the oxide scale. The corresponding equivalent temperature is calculated using the isothermal kinetics for oxide growth. By iteration, the average metal temperature and stress for each time interval are estimated, and a history of temperature and stress as a function of time is established. The life fraction expended at each interval is computed from the temperature increase caused by the steamside scale, the stress increase caused by fireside corrosion, and the stress-rupture life at each calculated temperature based on standard data.

The code uses the LFR for creep-damage summation and for establishing a failure criterion. The LFR states that, at failure:

$$\sum t_i/t_r = 1 \qquad \text{(Eq 10)}$$

where t_i is the time spent at any given stress and temperature, and t_r is the rupture life under those conditions. Typically, the scale that is formed on the steamside of SH and RH tubes can be described as multilaminated and is normally characterized by two separate oxide layers: an iron-rich inner layer (tubeside) and an oxygen-rich outer layer (steamside). Magnetite (Fe_3O_4) and Fe-Cr-Mo spinel, $(Fe, CrMo)_3O_4$, are the most abundant steamside oxide scales found on boiler tubes. A very thin layer of hematite, Fe_2O_3, is usually found at the steam/oxide interface. When the temperature exceeds approximately 560 °C (1040 °F) (or somewhat higher for chromium-molybdenum steels), and the local partial pressure of oxygen at the oxide/metal interface is sufficiently low, then wustite, FeO, forms.

When the code was initially developed, the steamside scale thickness values needed for input had to be obtained by metallographic examination of samples. Selection of sample locations was, therefore, crucial and called for considerable judgment and knowledge of prior history. With the improvement of ultrasonic techniques, these problems have been somewhat mitigated.

High-Temperature Crack Growth. Another area where major strides have been made is with respect to methodologies and data pertaining to creep and creep-fatigue crack growth (Ref 3, 44, 47). For simple creep, the crack growth rate can be expressed as a function of driving force, as:

$$\frac{da}{dt} = b(C_t)^m \qquad \text{(Eq 11)}$$

where b and m are material constants (see following). The transient creep parameter (C_t) is represented as:

$$C_t = \sigma \times \varepsilon(A, n) \times a \times H(\text{geometry}, n) \qquad \text{(Eq 12)}$$

where σ is the stress far from the crack tip obtained by stress analysis; ε is the strain rate far from the crack tip, which is a function of the constants A and n in the Norton relation for creep-strain rate ($d\varepsilon/dt = A\sigma^n$); a is the crack depth obtained from NDE measurements; and H is a tabulated function of geometry and the creep exponent, n. The values of A and n are either assumed from prior data or generated by creep testing of samples. By assembling all the constants needed, the value of C_t can be calculated.

Once C_t is known, it can be correlated to the crack growth rate through the constants b and m in Eq 10. Combining Eq 10 and 11 provides a first-order differential equation for crack depth (a) as a function of time (t). Theoretically, this equation can be solved by separating variables and integrating. However, the procedure is complicated by the time dependency of C_t and the crack-size dependence of the term H. To circumvent this, crack growth calculations are performed with the current values of da/dt (or a') to determine the time increment required for incrementing the crack size by a small amount, Δa (i.e., $\Delta t = \Delta a/a'$). This provides new values of a, t, and C_t, and the process is then repeated. When the value of a reaches the critical size, a_c, as defined by plane-strain fracture toughness (K_{Ic}) the J-integral (J_{Ic}), wall thickness, remaining ligament thickness, or any other appropriate failure parameter, failure is deemed to have occurred.

For creep-fatigue crack growth, the total crack growth rate is then given by:

$$\left(\frac{da}{dN}\right)_{Total} = \left(\frac{da}{dN}\right)_{Fatigue} + \left(\frac{da}{dN}\right)_{Hold} \qquad \text{(Eq 13)}$$

$$\left(\frac{da}{dN}\right)_{Total} = C(\Delta K)^n = C_t\left[(C_t)_{avg}\right]^g t_h \qquad \text{(Eq 14)}$$

where the first term denotes the Paris law for the fatigue crack growth component, and the second term combines the crack growth due to creep, including that due to stress relaxation.
The average da/dt and C_t are obtained as follows:

$$\left(\frac{da}{dN}\right)_{avg} = \left(\frac{1}{t_h}\right)\left(\frac{da}{dN}\right)_{hold} \qquad \text{(Eq 15)}$$

and

$$(C_t)_{avg} = \frac{1}{t_h}\int_0^{t_h} C_t\,dt \qquad \text{(Eq 16)}$$

The da/dN hold is the crack growth during the hold period and is obtained by subtracting the cycle-dependent crack growth rate from the total crack growth rate. Figure 39 shows a plot of $(da/dt)_{avg}$ with $(C_t)_{avg}$ for a 1.25 to 0.5 Mo steel at 540 °C (1000 °F). The data included test results for a 98 s hold time, 10 min hold time, and also the creep crack growth rate data. When plotted as a function of $(C_t)_{avg}$, the time rates of crack growth for these very different conditions fall on the same trend curve. In contrast, when the crack growth per cycle is plotted as a function of ΔK (Fig. 39), different curves are obtained. The significance of the previously mentioned trend with regard to predicting the hold-time effect in engineering components is obvious, because creep crack growth data can be used to estimate creep-fatigue crack growth data, and vice versa. In order to use this approach for predicting crack growth during hold time, it is necessary to estimate the magnitude of $(C_t)_{avg}$ for components.

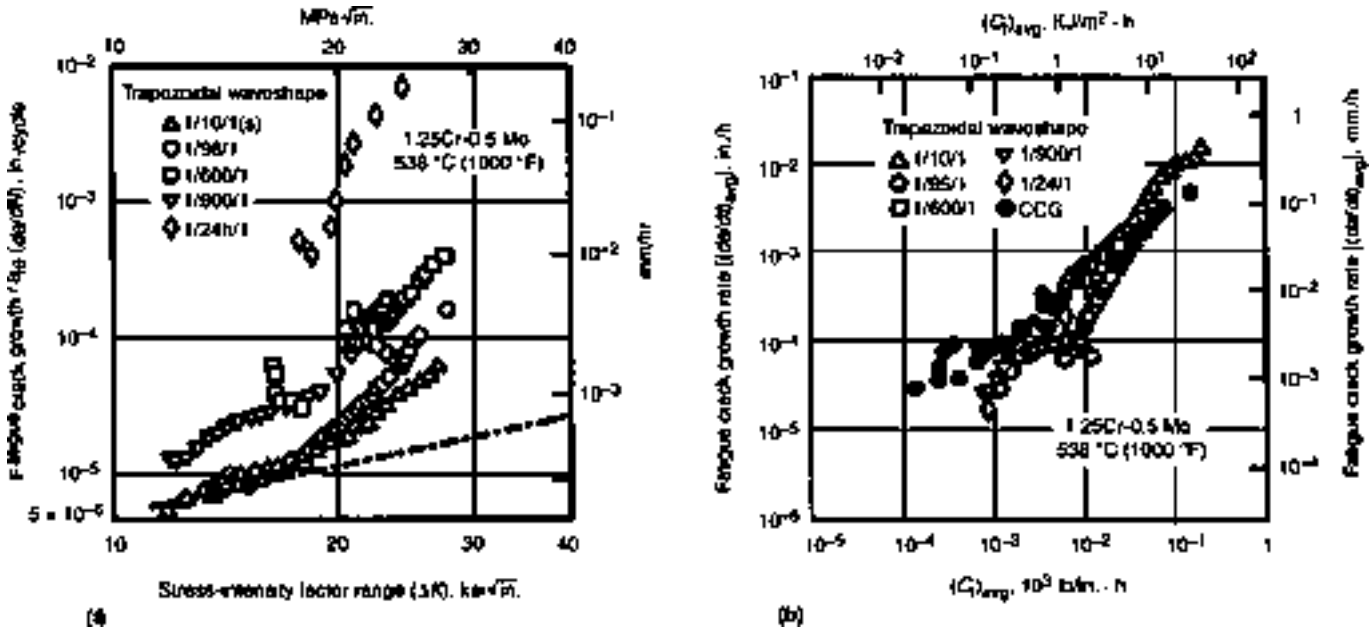


Fig. 39 (a) Comparison of creep-fatigue crack growth rates with fatigue crack growth plotted as a function of ΔK. (b) The effect of hold time estimated for engineering structures when the creep crack growth rate is plotted as a function of $(C_t)_{avg}$

References cited in this section

3. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components,* ASM International, 1989, p 188–190, 214, 228, and 369

35. R. Viswanathan, S.R. Patterson, H. Grunloh, and S. Gehl, Life Assessment of Superheater/Reheater Tubes in Fossil Boilers, *J. Pressure Vessel Technol. (Trans. ASME),* Vol 116, 1994, p 1–16

36. R. Viswanathan, J.R. Foulds, and D.I. Roberts, Method for Estimating the Temperature of Reheater and Superheater Tubes in Fossil Boilers, *Proceedings of the International Conference on Life Extension and Assessment,* (The Hague, Netherlands), 1988

37. L.H. Taft and R.A. Mardsen, The Structure and Properties of 1%Cr–0.5%Mo Steel after Service in CEGB Power Stations, *Proceedings of the Conference on Structural Processes in Creep,* (London, U.K.), JISI, JIM, 1963, p 275

38. R. Coade, "Temperature Determination Based on Microstructural Changes Occurring in 1%Cr–0.5%Mo Steel," Report SO/85/87, State Electricity Commission of Victoria, Australia, 1985

39. "Sumitomo Steel Pipes and Tubes," Section 14 on boiler tube failure and analysis, Sumitomo Metal Industries, 1984

40. M.C. Askins, "Remaining Life Estimation of Boiler Pressure Parts, Vol 3, Base Metal Mode," Technical Report CS-5588, Electric Power Research Institute, 1989

41. Y. Sugita et al., Evaluation of Creep Damage Progress by Metallurgical Examination in Aged Power Boiler Pressure Parts, *ISIJ Int.,* Vol 30 (No. 10), 1990, p 859–904

42. H. Nakatani et al., Metallurgical Damage Detection and Life Evaluation System for Boiler Pressure Parts, *Proceedings of the Conference on Predictive Maintenance of Fossil Plant Components,* Electric Power Research Institute, 1990

43. B. Neubauer, and V. Wedel, Restlife Estimation of Creeping Components by Means of Replicas, *Advances in Life Prediction Methods,* D.A. Woodford and J.R. Whitehead, Ed., American Society of Mechanical Engineers. 1983, p 307

44. R. Viswanathan and J. Stringer, Failure Mechanisms of High-Temperature Components in Power Plants, *Trans. ACME,* Vol 122, July 2000

45. J.R. Foulds, R. Viswanathan, L. Landrum, and S.L. Walker, "Guidelines for the Evaluation of Seam Welded High Energy Piping," Report TR-104631, Electric Power Research Institute, 1995

46. "Condition Assessment Guidelines for Fossil Fuel Power Plant Components," EPRI Report GS-6724, Electric Power Research Institute, 1990

47. R.H. Norris, P.S. Grooves, B.C. Hamilton, and A. Saxena, Elevated-Temperature Crack Growth, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 507–519

Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing

Daniel J. Benac, P.E., Bryant-Lee Associates; V.P. Swaminathan, Ph.D., TurboMet International

---

References

1. L.F. Coffin, Damage Evaluation and Life Prediction for High-Temperature Gas Turbine Materials, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials,* EPRI AP-4477, Project 2382-3, Electric Power Research Institute, April 1986, p 1.1–1.17

2. R. Viswanathan and R.B. Dooley, Creep Life Assessment Techniques for Fossil Power Plant Boiler Pressure Parts, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials,* Electric Power Research Institute, April 1986, p 2.1–2.28

3. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components,* ASM International, 1989, p 188–190, 214, 228, and 369

4. D.J. Benac, *Elevated-Temperature Failures,* Lesson 7, ASM Practical Failure Analysis Course, 2002

5. A.F. Liu, High Temperature Life Assessment, *Fatigue and Fracture* Vol 19, *ASM Handbook,* ASM International, p 520–526

6. N.E. Dowling, Estimating Fatigue Life, *Fatigue and Fracture.* Vol 19, *ASM Handbook,* ASM International, 1996, p 250–262

7. "Properties of Carbon and Alloy Seamless Steel Tubing for High-Temperature and High Pressure Service," Technical Bulletin 6G, The Babcock and Wilcox Company, 1955

8. R. Viswanathan and R.I. Jafee, Toughness of CrMoV Steels for Steam Turbine Rotors, *J. Eng. Mater. Techol. (Trans., ASME),* Vol 105, Oct 1983, p 286

9. V.P. Swaminathan and P. Lowden, "Gas Turbine Blade Life Assessment and Repair Guide," EPRI Report GS-6544, 1989

10. C.T. Sims, N.S. Stoloff, and W.C. Hagel, Ed., *Superalloys II,* John Wiley & Sons, 1987

11. M.J. Donachie, Jr., Ed., *Superalloys Source Book,* American Society for Metals, 1984

12. R.F. Decker, "Strengthening Mechanisms in Nickel Base Superalloys," Steel Strengthening Mechanisms Symposium, May 1969 (Zurich)

13. R.A. Stevens and P.E.J. Flewitt, The Effects of Gamma Prime Precipitate Coarsening during Isothermal Aging and Creep of the Nickel-Base Superalloy IN 738, *Mater. Sci. Eng.,* 1979

14. A.G. Pard and D. Kramer, Replication Microscopy for Estimating Mechanical Degradation of Gas Turbine Blades, *Proc. Replication Technology,* (San Antonio), 1984, p 99–101

15. A.G. Pard, "Service Embrittlement of Gas Turbine Blade Alloys Udimet 710 and IN-738," EPRI GS-6441, EPRI Project RP2775-1, Electric Power Research Institute, 1989

16. A. McMinn, "Coatings Technology for Hot Components of Industrial Combustion Turbines: A Review of the State of the Art," EPRI AP-5078, Electric Power Research Institute, Feb 1987

17. "Evaluation of Protective Coatings for Hot Components of Combustion Turbines," EPRI AP-4194, Electric Power Research Institute, Aug 1985

18. *Nondestructive Evaluation and Quality Control,* Vol 17, *ASM Handbook,* ASM International, 1989

19. Y. Lieberman, Relaxation, Tensile Strength and Failure of EI 612 and 20Khl F-L Steels, *Term. Obrabotka Met.,* Vol 4, 1962, p 6–13

20. R.V. Hart, Assessment of Remaining Creep Life Using Accelerated Stress-Rupture Tests, *Met. Technol.,* Vol 3, Jan 1976, p 1–7

21. R.V. Hart, "The Validity of Life Fraction Rule—A Review of Experimental Work," Report SSD/NE/N. 189, CEGB, 1979

22. F.R. Larson and J. Miller, A Time-Temperature Relationship for Rupture and Creep Stresses, *Trans. ASME,* July 1952, p 765–775

23. R.L. Orr, O.D. Sherby, and J.E. Dorn, Correlation of Rupture Data for Metals at Elevated Temperatures, *Trans. ASME,* Vol 46, 1954, p 113–128

24. S.S. Manson and A.M. Haferd, NASA Report TN-2890, National Aeronautics and Space Administration, March 1953

25. S.A. Manson, “Time-Temperature Parameters—A Re-evaluation and Some New Approaches,” American Society for Metals, June 1969

26. E. Campo, D. Ferrara, and S. Quaranta, Analysis of Creep Rupture Data of Nimocast Alloy 739, *High Temperature Alloys for Gas Turbines—1982,* R. Brunetaud et al., Ed., D. Reidel Publishing Co., p 575–578

27. E.A. Crombie, W. McCall, C.G. Beck, and D.M. Moon, Degradation of High-Temperature Impact Properties of Nickel Base Gas Turbine Blade Alloy, *Proc. of the 12th CIMAC Conference,* Vol C, (Tokyo), 1977

28. K. Kuwabara, A. Nitta, and T. Kitamura, Thermal-Mechanical Fatigue Life Prediction in High Temperature Component Materials for Power Plants, *Proc. of Intl. Conf. Advances in Life Prediction,* April 1983 (Albany), American Society of Mechanical Engineers, p 131–141.

29. G.T. Embley and E.S. Russell, Thermal-Mechanical Fatigue of Gas Turbine Bucket Alloys, *First Parsons International Turbine Conference,* June 1984 (Dublin, Ireland), p 157–164

30. E.S. Russell, Practical Life Prediction Methods for Thermal-Mechanical Fatigue of Gas Turbine Buckets, *Proc. Conf. on Life Prediction for High Temperature Gas Turbine Materials,* EPRI AP 4477, Project 2382-3, Electric Power Research Institute, April 1986, p 3–1 to 3–17

31. V.P. Swaminathan, J.M. Allen, and G.L. Touchton, “Temperature Estimation and Life Prediction of Turbine Blades Using Post Service Oxidation Measurements,” presented at the International Gas Turbine and Aero Engine Congress & Exhibition, June 1996 (Birmingham, U.K.), American Society of Mechanical Engineers Paper 96-GT-528

32. K.S. Chan, S.N. Cheruvu, and G.R. Leverant, “Coating Life Predication under Cyclic Oxidation Conditions,” presented at the International Gas Turbine & Aero Engine Congress & Exhibition, June 1997 (Orlando, FL)

33. K.S. Chan, S.N. Cheruvu, and G.R. Leverant, Coating Life Prediction for Combustion Turbine Blades, presented at the International Gas Turbine & Aero Engine Congress & Exhibition, June 1998 (Stockholm, Sweden)

34. S.N. Cheruvu, K.S. Chan, and G.R. Leverant, “Cyclic Oxidation Behavior of Aluminide, Platinum Modified Aluminide and MCrAlY Coatings on GTD-111,” presented at the International Gas Turbine & Aero Engine Congress & Exhibition, June 1998 (Stockholm, Sweden)

35. R. Viswanathan, S.R. Patterson, H. Grunloh, and S. Gehl, Life Assessment of Superheater/Reheater Tubes in Fossil Boilers, *J. Pressure Vessel Technol. (Trans. ASME),* Vol 116, 1994, p 1–16

36. R. Viswanathan, J.R. Foulds, and D.I. Roberts, Method for Estimating the Temperature of Reheater and Superheater Tubes in Fossil Boilers, *Proceedings of the International Conference on Life Extension and Assessment,* (The Hague, Netherlands), 1988

37. L.H. Taft and R.A. Mardsen, The Structure and Properties of 1%Cr–0.5%Mo Steel after Service in CEGB Power Stations, *Proceedings of the Conference on Structural Processes in Creep,* (London, U.K.), JISI, JIM, 1963, p 275

38. R. Coade, “Temperature Determination Based on Microstructural Changes Occurring in 1%Cr–0.5%Mo Steel,” Report SO/85/87, State Electricity Commission of Victoria, Australia, 1985

39. “Sumitomo Steel Pipes and Tubes,” Section 14 on boiler tube failure and analysis, Sumitomo Metal Industries, 1984

40. M.C. Askins, “Remaining Life Estimation of Boiler Pressure Parts, Vol 3, Base Metal Mode,” Technical Report CS-5588, Electric Power Research Institute, 1989

41. Y. Sugita et al., Evaluation of Creep Damage Progress by Metallurgical Examination in Aged Power Boiler Pressure Parts, *ISIJ Int.,* Vol 30 (No. 10), 1990, p 859–904

42. H. Nakatani et al., Metallurgical Damage Detection and Life Evaluation System for Boiler Pressure Parts, *Proceedings of the Conference on Predictive Maintenance of Fossil Plant Components,* Electric Power Research Institute, 1990

43. B. Neubauer, and V. Wedel, Restlife Estimation of Creeping Components by Means of Replicas, *Advances in Life Prediction Methods,* D.A. Woodford and J.R. Whitehead, Ed., American Society of Mechanical Engineers. 1983, p 307

44. R. Viswanathan and J. Stringer, Failure Mechanisms of High-Temperature Components in Power Plants, *Trans. ACME,* Vol 122, July 2000

45. J.R. Foulds, R. Viswanathan, L. Landrum, and S.L. Walker, "Guidelines for the Evaluation of Seam Welded High Energy Piping," Report TR-104631, Electric Power Research Institute, 1995

46. "Condition Assessment Guidelines for Fossil Fuel Power Plant Components," EPRI Report GS-6724, Electric Power Research Institute, 1990

47. R.H. Norris, P.S. Grooves, B.C. Hamilton, and A. Saxena, Elevated-Temperature Crack Growth, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 507–519

The Failure Analysis Process: An Overview

Debbie Aliya, Aliya Analytical

Introduction

FAILURE ANALYSIS is a process that is performed in order to determine the causes or factors that have led to an undesired loss of functionality. This Volume primarily addresses failures of components, assemblies, or structures, and its approach is one consistent with the knowledge base of a person trained in materials engineering. The contribution of the materials engineer to the advancement of the scientific foundation of failure analysis has been great in the last few decades. This can be shown by the fact that many people define the causes of failure in a rather binary manner: was the part defective or was it abused? Obviously, there are many types of defects, including those that come from a deficient design, poor material, or mistakes in manufacturing. Whether those "defects" exist in a given component that is being subjected to the failure analysis process can often only be determined by someone with a materials background. The reasons for this are related to the fact that many of the "defects" that people are looking for are visible only in a microscope of some sort. While microscopes may be widely available, the knowledge required to interpret the images is less widely available. The other major type of defects, those related to design issues, may also require the assessment of a materials engineer. This is because many design engineers are not very familiar with the natural variations within a material grade. Evaluation of the adequacy of a material or process specification is often best performed by a materials engineer.

Thus, materials experts have been in an excellent position to gain experience in the failure analysis process. The advent of more and more powerful and widely available scanning electron microscopes has helped provide a more fact based foundation for opinions that may have been heavily speculative in the past. Some materials engineers have become very experienced in failure analysis. As materials engineers have worked on some very spectacular failures or on failures that have caused great pain and loss, they have been led to ask deeper and broader questions about the causes that lead to failures. In many cases it becomes clear that there is no single cause and no single train of events that lead to a failure. Rather, there are factors that combine at a particular time and place to allow a failure to occur. Sometimes the absence of any single one of the factors may have been enough to prevent the failure. Sometimes, though, it is impossible to determine, at least within the resources allotted for the analysis, whether any single factor was key. If failure analysts are to perform their jobs in a professional manner, they must look beyond the simplistic list of causes of failure that some people still promote. They must keep an open mind and always be willing to get help when beyond their own experience.

Many beginning practitioners of failure analysis may have their projects defined for them when they are handed a small component to evaluate and, thus, may be able to follow an established procedure for the evaluation. This is especially true for someone working within an original equipment manufacturer. If there is someone who has a lot of experience and knowledge of the physical factors that tend to go wrong with the object and an established procedure exists, then a particular analysis may not require extensive pretesting work. However, for the practitioner who works in an independent laboratory or who is looking at a wide variety of components, following a predefined set of instructions for a failure analysis will generally prove to be an inadequate guideline for the investigation. Established "recipe type" procedures are generally inadequate for the more advanced and broad-minded practitioner as well.

Although the failure that we are investigating is that of a physical component, assembly, or structure, the failures that lead to such physical failure happen on many levels. In other words, a failure should not be viewed as a single event. It is more useful to view both the failure and the failure analysis as multilevel processes that can be explored in many useful ways. The physical failure—a fracture, an explosion, or component damaged by heat or corrosion—is the most obvious. However, there are always other levels of failures that allow the physical event to happen. For example, even a simple failure whose direct physical cause was an improper hardness value has human factors that allowed the improperly hardened component to be manufactured and used. These human factors are generally very difficult to investigate within a manufacturing organization,

because cultures that allow a particular type of failure to occur will generally not have systems in place that allow simple remedies to be enacted for the deeper level causes. For example, if someone in an organization wants to investigate causes beyond the simple fact of improper hardness, it may be discovered that the incoming (receiving) inspection clerk was not properly trained to take note of reported hardness values. Changing a corporate culture to include better training and education is generally very difficult; many corporations are structured so that the people who are responsible for training do not have an open line of communication to those doing the investigation. This only increases the difficulty of implementing change to prevent failures.

Professionally performed failure analysis is a multilevel process that includes the physical investigation itself and much more. This Section of this Volume is intended to showcase some of the latest thinking on how the different "layers" of the failure analysis process should work together, so that when the analysis or larger investigation is complete, the people involved will have useful knowledge about how to avoid future occurrences of similar problems.

Failure analysis of the physical object is often defined as a part of a larger investigation whose intent is to prevent recurrences. If we are to take the broadest view of what is required to prevent failures, there is one answer that stands out: education. Education needs to happen at multiple levels and on multiple subjects within an organization, within larger cultural groups, and within humanity in general, if we are to reduce the frequency of failures of physical objects.

Education, of which job training is a single component, is what allows people at all levels of an organization to make better decisions in time frames stretching from momentary to career long. There are many books now available that have exercises that help the reader to restructure knowledge into a more useful and accessible form (see, for example, titles in the Selected Reference list for this article). There are other books available that help the reader learn to recognize incorrect lines of reasoning; one such book is Ref 1.

Specific levels of failure causes have been defined by Failsafe Network as follows:

1. Physical
2. Human
3. Latent
4. Root

Clearly, many people involved with failure analysis today call something a root cause when what they are referring to is a simple physical cause. If failure analysis tasks are performed adequately and with luck, at the end the analyst should be able to take the causes found, show that the failure would have happened the way it did, and also show that if something different had happened at some step along the way, the failure would not have occurred or would have occurred differently. The fact that is often revealed at the end of an investigation is that this is not possible. Even a long and involved investigation leaves unknowns. The honest analyst is left to make a statement of the factors involved in allowing conditions that promoted the likelihood of failure. This is still a useful task, perhaps more useful than something that pins "blame" on a particular individual or group. Understanding the factors that promoted a failure can lead to an understanding of what is required to have a real improvement in durability of products, equipment, or structures. Understanding goes beyond knowledge of facts. Understanding requires integration of facts into the knowledge base of an individual so that the facts may be transformed into knowledge and then into product and/or process improvement.

By now it should be clear that failure analysis is a task that requires input from people with many areas of expertise. A simple physical failure of a small object may be analyzed by a single individual with basic training in visual evaluation of engineered objects. However, going to the level of using the failure analysis to improve products and processes requires expertise in the various aspects of human relations and education, at the least. Failure analysis of a complex or catastrophic failure requires much more.

People who perform failure analysis as part of their job function need to have an awareness of how their legal obligations are defined. People who perform destructive testing on a component that has failed may sometimes be held accountable for the destruction of evidence on a personal level. Company employees need to learn to protect themselves. Investigators who were "just doing the job" have been successfully sued by parties that the judicial system determined had a legitimate interest in the outcome of the failure analysis project.

The days where anyone unquestioningly agrees to destructively test a component that they know or can see "has failed" should be over. This places the destructive testing technician or engineer in a difficult position, as it

is sometimes difficult to see that something has failed. Corporate cultures that are highly structured and hierarchical can be particularly difficult environments for the failure analysis practitioner, as it may be difficult to even find out if the component has failed. Even if that information is given, relevant background details are often very difficult to obtain, even if the analyst tries. Pressure to finish the analysis in a shorter time frame than is desirable for a quality investigation is common.

Those who perform failure analysis work must realize that many people are still unaware of what failure analysts have to offer in terms of allowing clients or fellow employees to replace speculation with facts. The people who request failure analysis work may not be aware that rushing ahead into the destructive portion of an investigation may well destroy much information. The remainder of this article and the following articles in this Section of the Volume are intended to demonstrate proper approaches to failure analysis work. The goal of the proper approach is to allow the most useful and relevant information to be obtained. Readers of the various articles will see many points of view demonstrated. All the valid approaches require planning, defining of objectives, and organization prior to any destructive testing. Simultaneous preservation of evidence is also required. It should now be clear that proper failure analysis cannot be done with input from only a single individual. Even someone only participating in the “straightforward” portions of the investigation of physical failure needs to know how his or her contribution fits into a bigger picture. This is the intent of this Section, while the next Section of this Volume is intended to provide an introduction to the vast array of technical tools and information available to the failure analyst.

The competent failure analyst needs to know more than the failure analysis process and the tools used to support it. The competent failure analyst needs to understand the function of the object being analyzed and to be familiar with the characteristics of the materials and processes used to fabricate it. The failure analyst needs to understand how the product was used and the culture in which it was used. Communication skills are a must. When you ask a question, do you know for certain what the answer “yes” means? In some cultures, the word “yes” means “I heard the question” and does not imply that the answer is actually affirmative. The failure analyst must always be well versed in multiple disciplines.

The failure analysis process is something that can be approached in many different ways. Most people who do failure analysis of structural components or larger scale structures and assemblies have probably run into someone who wanted to do a failure analysis without considering a contribution from an experienced materials engineer. While the analyst may reach a conclusion in this manner, its value should be questioned. A reliable understanding of what happened and why it happened requires the input of a competent materials engineering practitioner. Every “failed” object is made of some material, and some common materials can lose more than 90% of their usual strength if they are not processed properly. Clearly, prior to reaching a conclusion as to the most significant causes of the failure, someone should determine if the correct material was used and if it was processed properly. This often requires both an investigation of documentation and a series of physical tests.

This Volume focuses on the definition of and requirements for a professionally performed failure analysis of a physical object or structure. However, many of the concepts for investigation that are described in this Section have much greater utility than for investigations of physical objects failure. The concepts in learning how to define objectives, negotiate scope of investigation, look at the physical evidence, structure both the investigation and the data that it reveals, and perform general problem solving have broad applicability in other areas of business, manufacturing, and life in general. The examples of how competent materials engineers can use these concepts in a failure analysis or failure investigation are emphasized here.

Reference cited in this section

1. D. Levy, *Tools of Critical Thinking: Metathoughts for Psychology* Allyn and Bacon, 1997

The Failure Analysis Process: An Overview

Debbie Aliya, Aliya Analytical

Principles and Approaches in Failure Analysis Work

A key principle of failure analysis is, first and foremost, to preserve evidence—the analyst must make sure that any necessary information from the subject part or assembly in the as-received condition is captured before anything is done to alter its condition. This principle can be summarized by the following guidelines:

- First, preserve evidence.
- Perform tests in order of less destructive to more destructive in nature.
- Know the limitations of one's personal knowledge.
- Know how to ask for help.
- Do not attempt a failure analysis if the basics of specimen preservation, collection, and selection have not been studied.
- Know when to say no to performing a destructive test.

Destructive testing obviously includes anything that requires cutting the part. However, even moving fragments of an explosion may cause loss of information that might have been determined from the position of the fragments. Cleaning components can also be problematic; not cleaning can lead to damage by corrosion in the case of many common materials, while cleaning may remove the substances that caused or contributed to the failure or that shed light on the nature of any physical degradation of the components. Sometimes, cleaning of dangerous or toxic substances from the debris of a failure is necessary for the safety of the investigator. The practitioner should also keep in mind that many tests described as “nondestructive” are only relatively nondestructive. There are numerous investigations during which the analysts representing different parties have spent long periods of time trying to figure out whether the dye penetrant residue that they detected is a result of the test done after the failure or before the last service period. Again, the phrase *nondestructive* should be viewed as a relative term.

It should also be recognized that the problem of failure analysis can be approached in different ways, depending on the required depth and scope of analysis. Another key principle in failure analysis work is knowing how to define the scope of the investigation at the proper time, so that the investigation has the highest chance of allowing the answers to the questions posed to become known. The circumstances of failure problems can be diverse, and even the “simple” principles of failure analysis may be subject to interpretation and examination. Even the principle of preserving evidence may involve judgments, as noted in the preceding paragraph. Even the experienced analyst can make mistakes. For example, consider a situation in which an analyst received a bearing from a regular client. The bearing had worn out prematurely in a durability test. The analyst was requested to photograph the wear marks and check the hardness and quality of heat treating of the races and balls. The bearing was covered with a black greasy substance. When analyzing wear failures, it is often the lubricant properties and the wear particles that offer the most information about the wear process. The analyst informed the client of this, and the client made the decision to sacrifice the “dirty grease” in the interest of completing the investigation in a shorter time. The client just wanted to know if the part met the specification, and whether there was evidence of gross misalignment or dimensional problems. After the investigation was finished, and the grease had been dissolved in solvent and discarded, another individual at that company requested detailed information on the actual wear mechanism. At that point, it was too late to analyze the lubricant. If either the client or the analyst had taken more time, a sample of this material could have been set aside and preserved.

Another key principle of failure analysis is that for all but the simplest routine investigations, there may be multiple, legitimate approaches. Selecting the most appropriate of these approaches to problem solving in failure investigations, some of which are described later in this article and elsewhere in this Volume, is an important skill. The classical approach is to follow a list of steps, which generally include planning the investigation, performing background research, and writing the report, as well as the actual physical tests and evaluations to which the component in question is subjected. Even though a recipe of procedures has many merits, especially for the beginners in the practice of failure analysis, performing a series of tests does not always produce results that allow the analyst to reach a clear conclusion with ease. Also, if failure analysis is presented exclusively as a series of steps or recipes, the repetitive steps may be conducive to carelessness. If the process of failure analysis is viewed as just a routine series of procedures, the practitioner may not notice the presence of new or different features. This is not to say that structured approaches are bad. However, the importance of paying attention to detail and learning to ask oneself whether each detail is consistent with the other previously noted details can hardly be overemphasized to the beginning failure analyst. Failure analysis is

an iterative and creative process, much like the design process, but with reversed roles of synthesis and analysis. See the article “Materials Selection for Failure Prevention” in this Volume

Knowing which approach to use is at least as important as knowing how to use it. This article describes some of the factors and conditions that might be considered when approaching a failure analysis problem. In any case, whichever approach is taken, it is always important to cultivate an open mind and to minimize the temptation to reach a conclusion about the cause(s) of the failure before performing the analysis and evaluation. The science of critical thinking has a principle called *the confirmation bias*, which refers to the tendency to look only for what one expects to find: that is, “Ye shall find only what ye shall seek” (Ref 1). Humans have a general tendency to see what they expect to see or to perceive things according to preconceived expectations. As Mark Twain wrote, “To the man who wants to use a hammer badly, a lot of things look like nails that need hammering.” If observation is limited to an expected outcome, helpful data may be overlooked. For example, one of the biggest mistakes that people make in failure analysis work is defining the investigation in binary terms of “was there a manufacturing defect or was the object abused?” The professional analyst should not be confined to this small set of possible causes for the failure, because it may be difficult to become aware of the situation of not finding what he or she did not set out to find.

It is also important to appreciate the value of intuition and instinct. While the importance of observation and analysis can hardly be overemphasized, sometimes intuition can provide insights and a better appreciation of the “big picture.” For years, the great 19th-century Indian mathematician Ramanujan would, immediately upon awakening, write down theorems that had come to him in his dreams. Many of these theorems remain unproven yet useful to mathematicians and physicists today (Ref 2). Another example is the discovery of the structure of the benzene ring by F.A. Kekulé. Many high-school science books report that during the period of time when he was trying to figure out how the carbon and hydrogen atoms were arranged within the molecule, he had a vision in a dream of intertwined serpents, each biting its own tail. Obviously, he and others went on to use more scientific methods to demonstrate the correctness of his theory. Likewise, while any failure investigation must stand or fall on the merits of the analytical work done, historically, too little credit has been given to the intuitive function in engineering and scientific work in general.

References cited in this section

1. D. Levy, *Tools of Critical Thinking: Metathoughts for Psychology* Allyn and Bacon, 1997

2. M. Kaku, *Hyperspace,* Oxford University Press, 1994

The Failure Analysis Process: An Overview

Debbie Aliya, Aliya Analytical

The Objectives of Failure Analysis

The objective or purpose of a failure analysis project is often described to be preventing a recurrence of the failure. However, there are many different types of failure analysis projects. Where an injury lawsuit is involved, for example, it may be important to assign responsibility for an undesired event. There are other cases when there may never be a chance for a recurrence. For example, if the item that fails is unique, there may never be a repeat incident.

Another case in which the objective of the investigation may not be the prevention of recurrences is one involving a very minor failure of a low-value component. If there is no other damage, it may be difficult to justify a prevention-oriented project. It may be more economical to live with a certain level of failure than to devote resources to prevention. The work is still worthwhile, because if certain economic situations change, there is background information available to support a broader investigation, in a more efficient manner, at a later time. Also, the understanding gained may lead to an improved product that may be appropriate for a particular market niche, for example, long-life light bulbs

It may be that there is no true product failure. In fact, one important question that may well be asked is: “Did a failure really occur?” It is possible to have an undesirable event that involves fracture, wear, deformation, or corrosion but that is not really a component failure. For example, discovering a fatigue crack in a 40-year-old structural component, in many cases, may be less of a surprise than finding one that is free from such cracks. For these and similar reasons, it is good practice to avoid the use of terms such as *failed part,* at least until the investigation has revealed strong evidence that a failure has indeed occurred. The terms *subject part, subject component,* or *physical evidence* are preferred. It is also good to have an appreciation for what a failure is not.

The objectives of failure analysis can vary, as suggested by some of the different types of objectives listed in Table 1. Early in every investigation, those with an interest should determine exactly what their objective is or, more likely, what their objectives are. If the parties have a genuine desire to prevent recurrences, even if they have no legal obligation to do so, they still need to decide how far they want to go toward this goal. Economic and timing considerations usually determine the scope of the investigation. Aside from the cost of the investigation itself, a sure answer that comes after repeat failures may be less valuable than a reasonably sound answer before repeat failures.

Table 1 Objectives in failure analysis

Types of objective	**Possible precipitating situation**	Product life cycle
Product development	Demands of the market	**Prototyping**
Product improvement	Warranty costs	**Ongoing**
Assignment of responsibility	Reparations for financial/physical damage or bodily injury or death	**After subject event**
Prevention of recurrence	**Any**	**After subject event**

However, despite these various ways of defining objectives and conclusions in failure analysis, it is still fundamentally worthwhile for the knowledge gained. Obviously, success is more satisfying than failure, but experienced practitioners of failure analysis have learned the value of taking the time to extract lessons about both the technical and people-related causes of the failure. When the challenge is to keep learning, every investigation adds to our competence. Often, problems cannot be solved from the same level of understanding in which they are created. Thus, failure analysis may often require a keen and inquisitive outlook, which is a very good and satisfying way to keep learning on a technical, professional, and personal level for an entire career.

The Failure Analysis Process: An Overview

Debbie Aliya, Aliya Analytical

Scope and Planning

The scope of a failure analysis depends on the depth and complexity of the project. Many failure analysts have experienced being told to find the root cause of a particular failure in half an hour! Usually, this is impossible or leads to superficial results. The scope of the investigation must also be targeted toward finding the real root (physical or human) cause of the failure. Root-cause failure analysis has generated much attention in recent years, and it is a good development when the term *root cause* is given to mean the particular physical or human effect that precipitated or assisted in a failure. Sometimes the term *root-cause analysis* is misconstrued to mean figuring out whether a component meets a specification. If it does not meet specification, then the lack of conformity to the specification becomes a convenient “root cause.” This use of the term is not only superficial but also invalid. Any approach that does not attempt to link the particular physical effect with the particular lack of conformity and its direct expected consequences is not a valid failure analysis approach.

During the planning stages on the possible scope of an investigation, it is sometimes useful to focus attention on the potential complexity of the problem and on when the physical cause may have occurred. Various categories of complexity are listed in Table 2. These categories of complexity are not exclusive. All failure analysis work must be founded on the physical causes. Sometimes one is asked to determine a cause based on the verbal

description of a component. In some limited cases, this may be all that is possible, if the part is gone. However, the value of the failure analysis is limited in those cases. Also note that a failure problem is defined in a broader context going from top to bottom of Table 2. This is particularly important for failure analysis within a manufacturing organization. Today, many companies manufacture components or assemblies for other companies. The individuals at the contract manufacturing companies may not have any way of learning the importance of specified or unspecified requirements of the components that they are manufacturing until they have the opportunity to learn as a result of a failure investigation. Progressive companies have been taking advantage of their human resources for years by providing more training and education, because they know that it generally has a positive effect on the profitability of the company.

Table 2 Complexity of investigation

Depth and complexity of investigation	**Example**	Allowed effects of findings
Physical cause(s) or factor(s)	Part is found to be of the wrong hardness.	**Part gets rehardened. Maybe inspection frequency is increased.**
Individual people causes	Heat treat department employee does not understand importance of hardness and does not check part.	**Heat treatment department employees are sent to a basic metallurgy and testing class.**
Latent causes or factors	Nobody tells furnace operator about the importance of hardness testing. It does not matter that she fabricated test results, until the pyrometer fails, causing incorrect temperature exposure during austenitization.	**Design and supervisory personnel are given custom training by a competent materials engineer on the subject of how to specify hardness test protocol. Maybe an investigation into the heat treat furnace uniformity is undertaken.**
Culture-based root cause(s)	**Hardness runs on low end of specified range for years but never causes a problem until someone, due to economic downturn, decides to use item for much longer than the original expected service life.**	**Marketing people are given additional training so that they do not give unreasonable expectations of product performance. (It is recognized that an optimistic outlook is part of sales and marketing. The question is, where does one draw the line as to what is reasonable and unreasonable.) Marketing people are given adequate resources to follow up with customers.**

It is also important to keep in mind when root cause(s) could have been introduced. Table 3 lists some examples. It is often difficult to assign a failure to a single phase of problem creation. Many times, there are complex interactions. Only an integrated design approach will create robust processes and thus robust parts. The designer must communicate with personnel from the sales department, as well as customers, maintenance personnel, manufacturing personnel, and material suppliers. Input from failure analysts who evaluate broken, worn, corroded, or deformed parts from durability testing is an important source of information that is often neglected by designers.

Table 3 Physical causes and time of occurrence

Physical causes	**Example**	**Likely effects of findings**	Alternative effect(s) of findings
Raw material manufacturing	Excessive microsegregation causes inability to achieve post heat treat mechanical properties.	Material supplier is blamed.	**Additional investigation is performed on how to improve heat treat line, to make process more robust.**
Design phase	Designer does not know accuracy of hardness test and specifies a range that is too narrow to achieve.	Heat treater is blamed for poor process control.	**Designer reviews specifications, and broadens specification if that is acceptable, based on**

			performance criteria, or modifies design as required.
Manufacturing phase	Part loader malfunctions, and a small fraction of the load does not experience proper thermal cycle, but none of the bad ones are in the test group	Heat treater is blamed for poor process control.	**Product design manager is inspired to perform detailed failure mode and effects analysis (FMEA) in presence of mechanical, materials, manufacturing, and maintenance personnel, and customers.**
Service phase	**Part is subject to unexpected and undetected heat, which changes the hardness**	**Hardness data can be easy to misinterpret, leading to assignment of the same likely cause listed above, if microstructure analysis is not included.**	**Product design manager is inspired to perform detailed FMEA in presence of mechanical, materials, manufacturing, and maintenance personnel, and customers.**

Avoiding Errors. The failure analyst needs to be aware that sorting out the causes of failures can cause economic and noneconomic (e.g., psychological) consequences to particular individuals or companies who are implicated for carelessness, negligence, simple ignorance, or other errors or omissions. Thus, it is important to avoid mistakes, as they could cause as much harm as, or more harm than, the original failure.

Analytical mistakes may be technical in nature, such as an incorrect measurement of a mechanical property. Analytical mistakes may also be subtle. An example may be not questioning a suspicious hardness or composition data point. Another example is an error in judgment of the significance of something that is normally a minor detail. If this causes one to overlook things that bear close scrutiny, an incorrect conclusion may be drawn. Making sure that all relevant details are examined can help point to a clear conclusion and is a key to competent failure analysis work.

In situations that involve loss of life, human injury, or large economic damage, professional analysts should be very careful to do work only within their areas of competence. It is important to know the limits of one's own knowledge and to know when help is needed. In fact, input from people from many areas will probably be involved in all but the most basic physical-cause investigations. If the failure involves complex interactions of latent factors, an interdisciplinary approach is generally required to prepare prevention strategy recommendations.

Fear of overlooking important details is probably the biggest reason that many experienced analysts refuse to perform failure analysis work unless they are given the time and budget to do a complete investigation. It is very easy to draw the wrong conclusions if one does not consider the “big picture” from multiple angles. A broad view is more likely to lead to a coherent conclusion or set of conclusions. To emphasize this important point again, a failure analysis must include an evaluation of the consistency of results from different tests or analytical methods. A single test result does not constitute a legitimate foundation for a failure analysis. Other common pitfalls in failure investigations suggested by Dennies (Ref 3) include:

- Jumping to conclusions
- Not understanding the problem
- Not understanding how the failed system is supposed to operate
- Not considering all possible failure causes
- Tearing system apart without a developed plan: “We need to tear it apart as soon as possible.”
- Failing to follow through
- Not asking for help
- Thinking it is so easy to do
- Destroying evidence due to lack of planning

Failure analysis is a profession that is rarely perfected in a given individual, and even experienced practitioners should remain aware of these potential pitfalls. It is also important to understand that failure analysis is not (Ref 3):

- “Give me your best guess”
- Not identifying root cause(s)
- Reworking or repairing: “It will be quicker to fix it.”
- Swapping parts: Remove and replace mentality
- Ignoring the problem: “Wait until the problem goes away.”
- “We were going to change it later anyway.”
- Band-aid fixes: Return to the supplier; scrap; another system, supplier, or inventory
- Never happens: “Oops”

Reference cited in this section

3. D.P. Dennies, Boeing Co., private communication

Planning and Preparation

It should be clear that the objectives and scope should be defined and understood early in every investigation. If resources for a complete and detailed investigation leading to a high degree of technical certainty are not available, the investigator is encouraged to clarify for himself or herself, as well as the others involved, what is hoped to be determined after following a particular protocol. This clarification process should be done before any destructive testing.

Even a very limited investigation is by no means useless. Often, the people involved have two or three failure scenarios in mind. It is often possible for the trained analyst to rule out some of these scenarios with a small amount of work. A case in point is when an automotive repair shop owner wanted to know if employee negligence had caused a premature fracture in an externally threaded fastener. The fracture occurred two months after the repair job. The fracture was found by a different auto repair shop. There were no indications of progressive cracking. The people from the second repair shop accused the first shop of gross negligence. In this situation, it might be reasonable to suggest that something other than the first mechanic's negligence caused the fracture. Even without revealing the whole story, the information provided with a simple fractographic evaluation was useful to those involved.

Guidelines on the preparation of a protocol for a failure analysis may vary. For a part investigation, it may be a simple checklist (Fig. 1) that is included in a client report. In larger investigations, other methods (Table 4) may be considered to help plan and identify priorities. Each has advantages, drawbacks, and limitations in any given situation. When planning the actual step-by-step activities of the investigation, one should keep in mind that the degree of comprehensiveness necessary will be determined to a large degree not only by what the involved parties want to know but also by how strong their desire is to know it. In practice, the strength of the desire is measured in practical terms by the budget and timing considerations.

Factor/Variable	Observations/Comments
Chemistry	Conforms to AISI 1030.
Microstructure	Proeutectoid ferrite and pearlite as expected in a hypoeutectoid annealed/normalized plain carbon steel.
Grain Size	Variable: coarse (1 to 2 ASTM grain size) in the heavy thickness and fine (3 to 5 grain size) in the thin web.
Knoop Hardness	Web (Thin section): 260, 266, 252, 245, 233 (HK300) Avg = 251 HK300 (20HRC) Heavy section: 275, 248, 309, 241, 273, 301 (HK300) Avg = 275 HK300 (25HRC)
Inclusions/Segregation	Appears most probably. The EDS spectra show presence of Sulfur (S), Phosphorus (P), Silicon (Si), Calcium (Ca) and other elements (Fe and Mn are expected). Some of the phosphorus may have been contributed by the laboratory cleaning procedure.
Forging Defects	No gross forging defects such as laps or folds were detected.
Overheating/Burning	Possible overheating causing grain growth and grain boundary precipitation/formation of sulfides/phosphides/oxides and other non-metallics.
Residual Stress/Tightening Overload	Unknown. Not expected to be significant.
Operating Loads	The concave side of brace is always under tension; magnitude of forces unknown.
Hydrogen Damage	Appears possible but secondary to grain growth due to overheating.
Hard spots	Higher Knoop reading near cracked edge.
Fracture Appearance	The initiation region shows fine dimpled structure which is interpreted to be due to the presence of grain boundary precipitates formed during forging processing. The fracture away from the initiation area is a mixture of ductile and intergranular modes; and has a faceted appearance. Further into the core, the fracture mode is predominantly of the cleavage (less ductile or brittle, transgranular) type.
Fatigue	No fatigue striations were detected. It is believed the crack front traveled rapidly.
Corrosion in Service	Rust only. No evidence of any other form of corrosion detected.

Fig. 1 Example of a simple checklist and report for a simple part (cracking of a brace)

Table 4 Structured decision-making and problem-solving methods for larger scale investigations

Method	Most useful in this type of analysis
FMEA	**Used during design of component and the processes that are used to manufacture it. Allows a structured approach to figuring out the consequences of failures of single features on a component or single steps in a processing sequence**
Kepner Trego	**Decision analysis method for problem solving in complex systems and prioritizing**
Fault tree analyis	**An analysis method that provides a systematic description of the combinations of possible occurrences in a system that can result in failure. It is a graphical representation of the Boolean logic that relates to the output (top) event.**
Failure wheels	**Helpful in evaluating failures from a combination of factors or damage mechanism (see the article "Determination and Classification of Damage" in this Volume).**
Root cause analysis	**There are many approaches. Most of this work has been done in the nuclear and chemical industries, in response to disastrous failures. It is also being used by medical administrators, and many other fields.**
Commercial software tools	**These exist for a variety of functions, including tracking the course of an investigation and suggesting a course of action for the background data collection.**

In many cases, particularly when litigation is or may be involved, it is preferable to define and document how the decision about the scope of the investigation is reached. Sometimes in litigation projects, there is a reluctance to create documentation due to a fear that it may be used to create a negative perception regarding the conduct of the investigation. Such fears must be balanced against the potential negative perception created if the investigator cannot later describe exactly what was done, why it was done, and what findings were determined. At the very least, for situations involving loss of life, human injury, or large economic damage, the professional analyst takes action to document the test plan and what type of information is expected from each step.

If the investigation is relevant only to those within a particular organization, such as an in-process failure, sending a simple memo summarizing the test plan to those responsible for the product may be all that is necessary. In many investigations, especially during those of parts in process at a manufacturing plant, analysis of a similar part may be useful. Often the term *control part* is used. This term implies that the similar part is free from any of the imperfections that might have contributed to the "failure." Often, however, very little is known about the reliability of the particular part supplied as a control. It may never have been in service. For this reason, it is often preferable to use a term such as *exemplar component* or *comparison specimen.*

As previously noted, a list of possible tests can be useful to ensure that nothing is unintentionally left out. Beginners looking at the finished reports of experienced analysts might think that the experienced practitioners followed a list of steps. Some of them may have followed a list, and analytical service groups may try to sell their clients a package, which includes all the tests that they can perform. However, checklists only summarize, and, like an executive summary, do not address all the detailed considerations of an investigation. One example of a relatively simple report summary is in Fig. 1. An example of a relatively limited test plan for a minor failure of a machine component due to fracture of a shaft is given in Table 5. This same table could serve for a much larger investigation. Instead of spending one or two hours on each activity, the analyst might spend three or ten or more hours on some of the activities, making sure that the locations evaluated are representative of the whole or characterizing the degree of variation from location to location.

Table 5 Example of a test protocol for a failure of minor economic importance

Activity	**Purpose/objective**	Time Frame
Visual inspection and photography of machine and surroundings, subject shaft, and mating parts	Determine general environment and condition of machine; determine whether crack was due to one or more load cycles, look for evidence of heat or corrosion damage; find crack origin(s)	**As soon as possible after failure**
Properly protect and place subject components in a	Prevent destruction of evidence	**As soon as possible after initial inspection**

protected area		
Closer inspection of broken part. Selection of test locations	Confirm that crack was due to one or more load cycles; look for more subtle evidence of heat or corrosion damage; find crack origin(s); take better photos than may be possible outside of the lab; allow the investigation to proceed	**Before subsequent destructive testing**
Ultrasonic testing	Document presence or absence and size and orientation of secondary cracks	**Before subsequent destructive testing**
Cut specimens	Provide small specimens for destructive testing	**After all nondestructive testing near crack origin is complete**
Scanning electron microscopy (SEM)	Look for evidence of usual or unusual crack paths, manufacturing imperfections, or service damage	**Before metallography**
Optical metallography	Determine whether microstructure is consistent with expectations based on crack path evidence	**After scanning electron microscopy (SEM)**
Hardness or micro-indentation hardness testing	Determine whether hardness values are consistent with apparent microstructure	**After optical metallography**
Composition verification	Confirm that proper alloy is used	**If specimen is from a different location than crack origin, any time after nondestructive testing is complete**
Tensile test	**Confirm that mechanical properties are within the expected ranges**	**If specimen is from a different location than crack origin, any time after nondestructive testing is complete**

An “Open-Mind/Open-Toolbox” Approach. In order to cover a broad perspective on legitimate approaches to failure analysis, it should also be pointed out that some people obtain useful results in an efficient manner without referring to a list of test and evaluation procedures. Rather, keeping an open and inquisitive mind may allow the analyst to formulate a series of questions. In order to answer the questions, obviously, tests must be performed, and they must be performed in an appropriate order (from least destructive to more destructive). Formulating a series of questions based on the analyst's personal knowledge and ignorance is a good way to start to try to understand how the subject of the investigation failed. One step that should never be skipped without a good reason is the written list of activities and their associated specific purposes.

This approach is termed the *open-mind/open-toolbox* approach. It can be a flexible approach, and, when used by a competent analyst, is often the most efficient approach. Legal requirements for preservation of evidence may preclude use of this flexible approach in cases where physical or financial injury results from a failure. However, it can be the most useful and appropriate approach for failures that occur during routine quality control or prototype or durability tests (where no third parties are injured), or when multiple exemplars of apparently similar failures are available. Even if the analyst is involved in a formal investigation with a protocol that is not very open to negotiation, he or she may still find the open-mind/open-toolbox approach to be a useful exercise before finalizing the investigation protocol.

A key element to understanding the open-mind/open-toolbox approach is realizing that a conceptual framework is, in an important sense, an investigative tool (Table 6). For example, understanding that the orientation of a fracture surface can indicate how the part was loaded allows one to insist that the protocol include the resources for a proper macrofractographic evaluation in a failure that was due to fracture. Fractography offers the conceptual basis for performing the visual examination. Someone familiar with the power of macrofractography would not leave this step out in any fracture-related failure. In the open-mind/open-toolbox approach, an investigative or analytical activity is selected because its cost in terms of money and time is likely to be favorably balanced by the information it provides (where information here is loosely defined as “useful data,” as distinct from just data). Information is not provided by the test alone. Information comes from the

interpretation or evaluation of data. The tools of failure analysis are not just test machines and analytical instruments. The tools of failure analysis include test machines and analytical instruments and also conceptual tools. Conceptual tools that are essential in determining the cause of any given failure may vary. They include various pattern recognition skills (in the interpretation of macrofractographs, microfractographs, and metallographic images) and engineering and scientific knowledge based on physical metallurgy, polymer physics, solid-state physics, stress analysis, chemistry, and many other fields.

Table 6 Elements of the failure analyst's toolbox

Tool	Use
Mind	**Structuring, storing and retrieving knowledge; planning investigation; conducting interviews; reviewing and integrating data and information**
Eyes	**Visual examination of subject components, the scene, the appearance of witnesses**
Communication skills	**Critical in investigation planning and interviewing of "witnesses," or even the person requesting an investigation.**
Magnification tools (magnifying glasses, binocular microscopes, close-up camera lenses, etc.)	**Detailed visual examination of subject components. Look for "witness marks" from impact events, evidence of abusive machining, burn marks, corrosion or wear evidence, anything unusual, etc.**
The science and art of macrofractography	**Provides understanding of the sequence of a fracture event. Gives information about relative stress levels and orientation of the loads and stresses.**
Visual examination of environmental damage, including corrosion	**The experienced corrosion specialist can often gain important clues about the damaging environmental factor from the appearance of the corrosion product.**
Visual examination of wear damage	**The experienced wear specialist can often gain important clues about the type of wear, from both the appearance of the worn component and any adherent material.**
Camera with macro (close-up) lenses, appropriate lighting sources	**Document the appearance of the object before destructive testing. Documenting each step of destructive testing with photographs is sometimes required for litigation support work.**
Nondestructive testing methods	**Including radiography, magnetic particle inspection, eddy current testing, dye penetrant inspection and other methods. Cracks, porosity, and unsoundness of various types may be able to be detected without cutting the part.**
Microfractrography and other techniques with SEM, including localized chemical analysis	**Microfractography may simply be used to add weight to a macro scale determination of ductile or brittle behavior. Evidence for distinguishing intergranular from cleavage fracture is most easily obtained by this method. Fine characteristics of ductile dimples can reveal additional information not available with macroanalysis. Fatigue striations can sometimes be measured with the SEM. Energy-dispersive spectroscopy (EDS) or wavelength-dispersive spectroscopy (WDS) microchemical analysis is a powerful tool for microphase evaluation and evaluation of potential corrosion accelerants or products. Microfractography gives additional information about the microstructure.**
Metallography	**Evaluate microstructure, crack path, relation of preferred corrosion attack sites to microstructural phases; check for conformance to specification for some heat treated parts, etc.**
Mechanical tests	**Get measurements of actual strength values of various types. Useful to compare with specified values. Note: specified values are often based on different test geometry than can be obtained from a fragment of a "failed" component.**
Composition analysis	**Determine if elements present in material used correspond with what is expected for the specified material. Note: May be difficult to obtain**

	accurate values for carbon, for example, on carburized steel

Understanding how the failure happened is also a big part of understanding why the failure happened. Understanding why the failure happened is often a key objective of the analysis. It can be a big problem to wait until the end of the investigation, when one has a pile of test results obtained from a predetermined set of activities, to try to formulate opinions about how or why the failure has occurred. If it is not suspected that something could have happened, one is less likely to do the tests that could show whether it did or did not happen. Something as subtle as making the measurements 0.1 mm (0.004 in.) away from the preferred site may cause a fact to remain concealed. There is no such thing as a quality investigation that is performed in ignorance of how the object is supposed to function and how it can malfunction. Knowledge of high-stress locations and operating conditions is critical to being able to draw useful and credible conclusions. While it is clearly important not to jump to conclusions before the testing is completed, it is also highly impractical in most cases to perform a competent analysis in the total absence of speculations about the manner and cause(s) of the failure.

This is an important point, because many people believe that the failure analyst should operate “blindly” in order to avoid clouding his or her judgement. However, data may be interpreted in widely varying ways by different people, or even by the same person at different times or in different circumstances. This must be recognized in the overall context of how the failure happened and how the various participants (including the analyst) are acting. Clearly, the likely and alternative findings columns of Table 3 may be viewed differently by various participants, depending on the overall situation of how a failure *might* have happened.

The Failure Analysis Process: An Overview

Debbie Aliya, Aliya Analytical

Practices and Procedures

The practices (and tools previously discussed, Table 6) that may be used by failure analysts are of many types and represent continuously developing knowledge of how to work in many diverse areas, including:

- Selecting a test
- Selecting a specimen for a test
- Setting up a photograph to highlight a particular type of fracture feature
- Conducting an interview with a friendly person who wants to know what went wrong
- Conducting an interview with a defensive person who thinks that he or she may be blamed for the failure
- Analysis of a particular alloy or stress system or component type

Like any activity requiring skills and knowledge, the “practice of practices” is also a lifetime effort, and probably most difficult, as it requires the ability to see around one's own mental filters.

The procedures of failure analysis are perhaps the easiest to define of the lists reviewed so far in this article. Procedures are relatively straightforward sets of instructions intended to provide guidance in performing a particular action. A procedure to make and polish a micromount from a specimen, which has already been selected according to the principles of failure analysis and in accordance with the approach of the investigator, is quite straightforward. A procedure to measure the thickness of a piece of sheet metal is not difficult. Procedures like these are easy to write and relatively easy to follow, after proper training. However, the importance of competent and careful technician work in failure analysis can hardly be overestimated. A sloppy technician who leaves scratches in micromounts can be excused if all that is needed is a grain size or a total case depth. If the desired datum is the crack path at the origin of the crack or the depth of a corrosion pit, careful attention to detail and pride of craftsmanship are necessities. General procedures for each of the physical analysis steps in failure analysis are given in articles in the Section on “Tools and Techniques in Failure Analysis” in this Volume. Detailed procedures regarding equipment operation may be found in the equipment manuals.

The person doing the technician work in a failure analysis must also have procedures that are specific to the components and alloys in question. It is very important to document specimen location, position, and orientation, in the absence of detailed preexisting procedures. For example, when testing an externally threaded fastener, the standard test position for a hardness test is a distance equivalent to one diameter in from the small end, at midradius. A common location for effective case depth on small gears is at the center of the gear thickness, at midheight of the tooth. The new technician may find it difficult to find terms to describe the position easily, but it is good practice to develop the analytical mind. In some cases, a sketch in the lab notebook may be an easier alternative.

Communications. The importance of developing communication skills also cannot be overestimated. The technician who does routine certification work can be taught how to do a task and go on with his or her work. With most failure analysis work, new decisions must be made each time a project is started. Some of these decisions may be as simple as figuring out how to clamp the part in the saw, without damaging the fracture surfaces, during microspecimen extraction. In this case, if the inexperienced technician knows how to listen carefully to a senior technician, things will probably work out satisfactorily.

The lead investigator in a complex investigation obviously requires much more developed communication skills, including specialized knowledge of how not to be misunderstood during a deposition or trial. Attorneys can make opposing experts look less knowledgeable than they are by asking if a certain reference is “authoritative.” If the expert answers “yes,” the attorney may proceed to give a “pop quiz” on any page in the reference. Obviously, most experts have not memorized all of the useful reference books that exist. In addition, if something in the reference is in apparent conflict with something the expert has said, the attorney will be sure to point it out. Now the expert is in a rather weak position. A better description of any reference might be “useful reference,” or “widely used reference.” It may also be useful to point out that the issue is not just whether a reference, even the present Volume, is generally authoritative, but whether it is relevant and correct when applied to a specific situation. Failures that appear to be similar may have significantly different causes. The Society for the Advancement of Science in Law has several publications on communications for expert witnesses that may be of use to those who do litigation support work. Several organizations also exist to train expert witnesses in self-defensive litigation communication skills.

In order to communicate effectively with people who request work, especially if the analyst is new to the field, there are certain things of which he or she should be aware. These include basics of organizational structures, and the perspectives of the management who will use and evaluate the information provided by the analyst. It is worth being aware of trends in the legal climate, specifically issues regarding what the company wants employees to document and to refrain from documenting. In general, documentation should be complete and correct. There is no benefit, and in fact there is actual harm in retaining incomplete or unreliable information.

These issues are particularly important to the employee who sincerely wants to prevent recurrences of failures. People who request the work may skim and file the report and may not take the action anticipated. Even errors that seem well documented to the analyst may not be convincing enough to change a mind that has decided no changes need to be made, the failure will not recur, and that in any case, it is too expensive to prevent any further failures. For example, just because the failure analyst performs an outstanding job on the latest investigation and shows that a simple change in procedure could reduce scrap by 50%, there is no guarantee that the simple change will be made. The International Organization for Standardization (ISO) 9000 registration procedures require companies to adhere to a policy of decision making for the common good of humanity, so if a direct safety issue is involved, it may be easier to make changes than otherwise. However, these guidelines are not terribly specific. Scrap reduction, while clearly beneficial to both the bottom line and the environment (and thus humanity at large), may not appear to be mandated by these ISO procedures.

Many constraints make procedural changes difficult within an organization. One of these constraints may be the ISO 9000 procedures themselves. Before changes are made, extensive testing, evaluation, and multidepartment, multilevel approvals may be required. This takes time. Because it often takes quite a long time, other priorities may push the change that was thought to be useful to a very low priority level. Other constraints may be related to budgetary considerations. There are many companies that have budgeted amounts for cost of scrap, but cannot get money shifted to a capital expenditure if a new piece of monitoring equipment is needed. The failure analyst's work often takes a long time to be seen as valuable by fellow workers.

Another important concept that affects implementation of findings is that people learn in different ways. Some people absorb information easily while reading. Some have to hear a verbal presentation. Some need flashy graphic images. Some need to take a tour of the production line and actually watch the operation. While a

written report of some sort is usually required of the failure analyst, some sort of verbal presentation of the findings is usually in order as well. Even if there is no opportunity to do a formal presentation, speaking with the people who requested the work in person or on the phone can be helpful. Especially if the written format your company prefers is quite formal, the reader may misinterpret your findings or conclusions. A short summary of the salient points in spoken form, especially useful before the written report is delivered, can be very helpful.

Knowledge Requirements for Failure Analysts. Some companies hire people with technician backgrounds to perform failure analysis work. While an engineering degree is not necessarily required to perform this type of work in a competent manner, a wide range of skills and knowledge is required. Someone without an engineering background will probably take some years to develop most of the necessary skills to a level of adequate proficiency to work on a variety of components.

For non-corrosion failures, the basic skills required include understanding fundamental concepts in stress analysis and mechanical property theory. While the analyst does not have to know how to perform complex stress analysis in a quantitative manner, he or she should be able to figure out where the high-stress areas are by looking at a part and having someone describe the function. This skill may be called *qualitative* or *heuristic stress-analysis skill,* to recognize unexpected failure locations or unexpected features on a subject component. Since the 1950s, engineers have been gathering data from mysterious failures that "should never have happened," because the operating stresses were much lower than the known strength of the materials used to make the components. The competent analyst of structural failures must be familiar with the basic concepts of fracture mechanics, fatigue crack propagation, and the causes of residual stresses.

Failure analysts who work with fractures must be familiar with macrofractography. This skill is a foundation of all fracture analysis. Without skill in fractography, the wrong test location might be selected for microfractography and metallography. Evaluating a location unrelated to the crack initiation may be worse than not evaluating the material at all with these methods, because the key evidence may be destroyed during the destructive portion of the incorrect testing.

Metallography and interpretation of microstructures are also key skills. The analyst must be able to look at a microstructure and determine whether the material is typical of its supposed composition and specified processing. This implies knowledge of how to interpret phase diagrams and isothermal and continuous cooling curves. Basic understanding of crystallography and micro- and macroscale composition effects is also required.

It is difficult for someone to independently perform failure analysis work without basic literacy. People skills and understanding of human nature are also important for the failure analyst. Failures can bring out the worst in people. During the background-information collection process of a major failure, the analyst probably needs to ask questions that make people uncomfortable and defensive. It may be difficult for the analyst to determine whether correct information is being provided. The analyst may need advice on how to encourage the people involved to tell what they know. Technical skill can help the analyst weed out some incorrect information, whether intended to mislead or confuse, or given in ignorance.

Finally, some consideration of ethics is required of the failure analyst. Ethical issues involve decisions that may be difficult to make. Most people would probably not envy the whistle blowers who decided to inform government authorities that their companies were breaking environmental regulations and who caused fellow workers to lose their jobs. Many ethical issues are not clear. Confidentiality promised to a client by an outside failure analysis service provider may conflict with the engineer's duty to protect the public, if the attitude of the client is not in line with ethical principles. (In such a case, it may be helpful to remind the client of the consequences of a repeat failure.) Anyone doing failure analysis work may wish to study some codes of ethics specifically written for engineers. There are also some interesting works on ethical systems that are conveyed by other means, including literature and the support of trusted colleagues and mentors.

The Failure Analysis Process: An Overview

Debbie Aliya, Aliya Analytical

References

1. D. Levy, *Tools of Critical Thinking: Metathoughts for Psychology* Allyn and Bacon, 1997

2. M. Kaku, *Hyperspace,* Oxford University Press, 1994

3. D.P. Dennies, Boeing Co., private communication

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

Introduction

ON 27 JANUARY 1967, the crew of Apollo I was lost when a fire broke out inside their command module during a simulation run aboard their unfueled Saturn V launch vehicle. It was a tragic event that shook a nation into reconsidering the promise of President John F. Kennedy on 25 May 1961 that "we choose to go to the moon in this decade." Weeks later, during the federal investigation of the Apollo 1 fire, astronaut Frank Borman said it was a "failure of imagination." It was just not conceivable that men at this early stage of the "space race" could actually be expected to imagine all possible problems that could arise.

Two years, 5 months, and 23 days later, on 20 July 1969, the Lunar Module *Eagle* landed on the moon. The Command Capsule, Service Module, and Lunar Module *Eagle* used for that historic space flight and landing had been improved by the rigorous and detailed failure investigation of the capsule from the Apollo 1 fire. A second capsule, the next in line, was also sacrificed to better understand the failure and to determine the root cause. During the failure investigation, many other design and manufacturing flaws were discovered, and improvements were suggested and implemented.

Failure investigation has come a long way since the Apollo 1 fire. The men and women involved in failure investigation in the new millennium have an incredible array of sophisticated and powerful tools and equipment to assist them. Unfortunately, many failure investigations performed today are not the victims of a "failure of imagination" but the victims of a "failure of organization." Too many failure investigations are started without a clear and concise goal, direction, or plan. Somewhere along the way, an investigation may get off track and never achieve its fundamental purpose of discovering root cause.

The purpose of this article is to discuss the organization required at the outset of a failure investigation and to provide a methodology with some organizational tools. The main focus is on the problem-solving tool of fault tree analysis. The main point is that time spent in preparation and planning can save a lot of time and money and even help achieve a successful conclusion of the failure investigation.

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

What Is a Failure?

A good definition of a failure is "the inability of a component, machine, or process to function properly." Failures come in all shapes and sizes. They can be individual parts, entire machines, or a process. Broken down further, failures can be physical, paper, or thinking/cultural in origin. These possibilities expand the concept of a failure, because they introduce failures that can be touched and failures that cannot be touched. A broken crankshaft is an obvious failure. However, a specification that has poorly written requirements that fail to

discern intended criteria may go unnoticed for years. In some cases, the reason things go wrong can be traced back to the culture of a company.

Always remember to look at each failure as unique and different. Considering a failure to be unique is sometimes difficult when the failure looks just like the one done before. Resist the temptation to immediately assign the same root cause to failures that look alike.

Another good point to remember is that failures are specific to an industry and the specific requirements of that industry. It is necessary to be knowledgeable about defining the requirements of the failure at hand. The saying "one man's trash is another man's treasure" applies to failures. For example, thin dense chrome plating on Unified Numbering System (UNS) 52100 martensitic steel bearings without a hydrogen bake has been used in commercial applications with great success when replacing nonplated UNS 52100 martensitic steel bearings. However, once that component moved into aerospace applications as a low-cost replacement for American Iron and Steel Institute 440C stainless steel, the requirements changed, and the lack of a hydrogen bake became a liability.

Failures in Physical Function and Expected Performance. In a very broad view, a failure can be defined by two categories. The first category is a functional failure, when a component, machine, or a process fails and everything stops. Functional failure of a component, such as a bearing or bracket, is very obvious. Functional failure of a machine can apply to something as large as an airplane or as small as a nutcracker. The difficulty in determining the root cause of the failure increases with the complexity of the component or machine, but the root cause can usually be discovered. Functional failure of processes such as heat treatment, coating, plating, welding, mail delivery, or airline scheduling can be more complex, and the determination of the root cause is also more complex. In addition, the cause-and-effect relationship for processes may not be readily apparent.

The second category is a failure of expected performance, when a component, machine, or a process fails to achieve performance criteria such as life, operating limits, and specification requirements. People have an expectation for the life of certain items, such as car tires, household appliances, car engines, biotech implants, and even pens. These life criteria usually are not written down but are debated in courtrooms. More specific criteria are operating limits, such as gas consumption, production scrap rate, gas flow in a valve, modem speed, specification turnaround time, computer-chip speed, or acquisition frequency. These items usually are well defined and, in most cases, mutually agreed upon by the parties involved. Finally, specifications requirements, such as mechanical properties, plating thickness, weight, and coating optical properties, are very well defined in written documentation. The specific definitions or limits are usually created by the customer, and tests are required to ensure compliance.

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

Why Do Failures Happen?

"Why do failures happen?" is an often-asked question, and, over the years, many categories have been devised to classify the general causes of failure. The most common reasons for failures include:

- Service or operation conditions (use and misuse)
- Improper maintenance (intentional or unintentional)
- Improper testing or inspection
- Assembly errors
- Fabrication/manufacturing errors
- Design errors (stress, materials selection, and assumed material condition or properties)

Today, service or operation is the first suspect when a failure occurs. As noted earlier, most components and machines, even processes, have a certain life expectancy. Under normal use, most items wear out in time. A failure occurs when an item wears out in a shorter time than is expected by the user. At this point, however, it is important to distinguish between normal use and misuse. Pencils and cars can be used in a normal manner and

have an expected life. Misuse, such as chewing on the pencil or constantly driving a car at high speeds or on rough roads, may reduce the life of either item.

Improper maintenance reaches all parts of our lives. The appliances in our homes, our computers, our cars, as well as the aircraft we fly, are required to be maintained. Improper maintenance results in life reduction or complete breakdown. Sometimes it is intentional, such as skipping the overhaul of a machine in order to squeeze out a few more weeks of use. Other times it is unintentional, such as the case where the person who believes that if a 50 to 50 ratio of radiator fluid to water is good, then 100% radiator fluid must be better. Another common outcome of poor maintenance is corrosion. The monetary loss due to corrosion in the United States has been estimated at $60 billion to $100 billion a year.

Improper testing can also be a root cause of failures. The cost-driven, “hire and fire” business philosophy prevalent today drives work to be performed by the lowest-cost personnel capable of performing the task. Sometimes, these personnel have just enough training, and sometimes, they have too little training. Unfortunately, this business attitude can foster a situation in which the personnel performing the task may cause a failure by selecting the wrong test, by performing the test incorrectly, or by reviewing the test incorrectly. A favorite example concerns bird-strike testing on jet engines. Bird strikes have been a problem during aircraft takeoffs and landings for many years. A bird may be ingested into an engine at takeoff or landing. The engine must be able to handle a bird strike without an uncontained failure, fire, or engine mount failure. An engine must be able to shut down in a controlled fashion to pass this U.S. Federal Aviation Administration requirement. The jet engine industry has a test to determine the effect of a bird strike. Frozen chickens are thawed out and shot from a cannon into a jet engine running on a test stand to determine if the engine stays within its nacelle. One day, a new technician was running the test, and apparently no one told him to thaw the frozen chicken before using it in the test. Needless to say, the engine failed the test as this frozen chicken hit the engine with the destructive impact of a bowling ball.

Assembly errors, fabrication/manufacturing errors, and design errors, are easier root causes to identify and are more commonly noted in the public domain through lawsuits and newspaper articles. The combination of fast and cheap with the use of low-cost personnel can be disastrous. Design and manufacturing engineers believe they can control the problem by making products and fabrication processes “idiot proof.” However, time and time again, failures occur that prove them wrong.

Benefits of Statistics. It is in the best interest of each company to keep statistics on failures. As with most databases, computers make this task easier. The type of failure, the material, the root cause, and so on, can all be easily kept in a searchable database. Because each industry, even each company, is different, they need to keep their own statistics and draw on outside sources as applicable. The reason is simple. The statistics may provide a direction as to how a company can improve profitability by indicating where the company should spend money and manpower. No company can solve every problem, so it is prudent to attack the most technically detrimental or costly problems first. The statistics can provide this direction. In addition, statistics can point to trends over time. Lastly, comparison of company statistics to industrial data indicates if a problem is company specific or industry-wide.

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

Why Is a Failure Investigation Performed?

In most instances, the purpose of a failure investigation is to determine the root cause(s). Determination of root cause is good engineering practice that crosses functional boundaries within a company and is an integral part of the quality assurance and continuous improvement programs. A proper failure investigation is not a “science project.” In addition, there is a distinct difference between a failure investigation and a metallurgical evaluation. A root-cause failure investigation determines the root cause of the failure and recommends appropriate corrective actions. In contrast, a metallurgical evaluation provides the answers to the metallurgical questions posed in the failure investigation. Therefore, a metallurgical evaluation is only one part of a failure investigation.

Why Determine Root Cause? The most public reason to discover the root cause of a failure is to determine the fault or innocence of a company or person during litigation. Once fault is established, monetary assessment follows. For industrial purposes, however, it is more common, that once the root cause is discovered, the corrective action to prevent future occurrences is implemented, thus saving the company time and money. In addition, once the root cause is discovered, it is also possible to determine if the failure is a unique situation or the symptom of a widespread problem.
The benefits of a failure investigation that discovers the root cause of the failure include being a part of quality management and continuous improvement programs, assisting in a redesign, solving a manufacturing problem, saving money, saving time, and maybe in some cases, preventing injuries and/or saving lives.

The Four-Step Problem-Solving Process

The four-step problem-solving process that most engineers are taught is a good overall view of a failure investigation:

1. What is the problem?
2. What is the root cause of the problem?
3. What are the potential solutions?
4. What is the best solution?

Therefore, a good definition of a failure investigation can be an evaluation performed to identify and/or determine the reasons for a failure that clearly identifies the root cause(s) of the failure and recommends corrective action(s). Many failure investigations end after step 2 (determine root cause). It is just as important to provide steps 3 and 4. Additionally, it would also be good to provide an evaluation of the solution. Many failure investigations discover the root cause(s) and provide a recommendation for corrective action(s), but no follow-up evaluations are performed to see if the corrective action worked.

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

Nine Steps of a Failure Investigation

There are nine steps to the organization of a good failure investigation. They are:

1. Understand and negotiate goals of the investigation
2. Obtain clear understanding of the failure
3. Objectively and clearly identify all possible root causes
4. Objectively evaluate likelihood of each root cause
5. Converge on the most likely root cause(s)
6. Objectively and clearly identify all possible corrective actions
7. Objectively evaluate each corrective action
8. Select optimal corrective action(s)
9. Evaluate effectiveness of selected corrective action(s)

Step 1: Understand and Negotiate Goals of the Investigation. Every failure investigation needs to establish four items at the onset:

- The priority of the investigation
- The resources available
- Any constraints imposed
- The goal of the investigation

There usually is not a discussion to define these four items, but there should be. These four items set the importance, direction, and expected results of the failure investigation at an early stage. Always discuss priority first, because priority sets the pace and can sometimes determine the resources and constraints. Resources and constraints may simply involve the same three items: money, personnel, and time. Constraints can also include inability to destroy the hardware, lack of other pieces to review, lack of information, and so on.

The goal of the failure investigation is the final and most important item to be understood. The customer may desire a very simple goal, one that does not require determination of the root cause of the problem. Discuss the goal, and review whether determining root cause is beneficial. For example, a customer requested the determination of a minimum property allowable for some hardware that had been received with mechanical properties below the specification limit. The goal of the customer was to provide information for a stress analysis to determine if the hardware could be deemed acceptable on a nonconformance report. However, this problem was really a failure of the heat treatment or the raw material. Because this was an approved supplier that fabricated many other pieces of hardware from a large amount of raw material from inventory stock, the internal customer was persuaded that it would be in the best interest of the company to determine the root cause of the problem. It was viewed worthwhile to investigate the heat treatment and raw material as well.

Always understand these four items at the beginning of the failure investigation so that all parties are in agreement.

Step 2: Obtain Clear Understanding of the Failure. What is the problem? This is the first step listed in the previous section, “The Four-Step Problem-Solving Process.” This should also be the first question asked in a failure investigation. What happened? Why is the investigation team here? Engineers so often want the details that they forget to hear the reason their customer wants their help. Information is crucial at this point. The investigator needs to make himself as much of an expert about the component or system or process as possible.

There are tools at one's disposal. Brainstorming is a good one. Get everyone together and brainstorm the theories as to why the failure occurred. This means everyone. The investigator may learn some very important information. For example, the drawing may call out one process specification, but the people on the shop floor may have replaced that specification a few months ago through production documentation. Just reviewing the drawing would have missed this information. Review all documentation and records available on operational history and fabrication. This process is time consuming and tedious but important. Remember to be rigorous in all aspects of the investigation. “It is not only in the finding that we learn much, but also in the looking. All things, big and little, teach us; perhaps at the end, the little things teach us the most.” This quote is by Dr. Van Helsing during his search for the elusive Count Dracula.

The failure scene is very important. Document and protect it. If the investigator is not able to visit the failure scene, he should request the drawings and/or photographs of the failure scene. Photographs of the failure scene both before and after the failure, are frequently helpful. If an investigator has the opportunity to go, he should take a field investigation kit (Fig. 1 and Table 1) and document everything. Film is cheap, and now with digital cameras, photos are free. Take hundreds or even thousands of photographs.

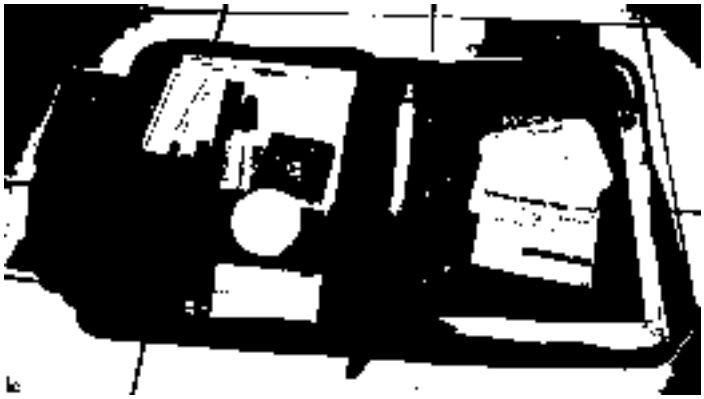


Fig. 1 Carrying case with contents of field investigation kit (Table 1) without camera, cell phone, and laptop or hand-held computer device

Table 1 Field investigation kit contents

Items	Reason/Comments
Open and questioning mind	**Be prepared. Have questions. Be ready for the unexpected.**
Good attitude	**You need the people you meet more than they need you.**
Professional demeanor	**If you look like you are organized and know what you are doing, people will consider your presence important, and you might get more help.**
Digital camera	**Take as many pictures as you can. It used to be that "film is cheap," but now "J-pegs are free." Photograph the item, the area, any supporting equipment, maybe people, anything and everything. Be careful about color.**
Known color chart, white piece of paper, etc.	**Kodak grey or color chart, Ace Hardware paint charts, red button, your tie. Anything you can use later as a standard to judge the color of your pictures.**
Ruler—steel and plastic	**1. Use in photos for scale—steel rulers are grey and good on light backgrounds. White plastic rulers are good for dark backgrounds.**
	2. Steel rulers will tell you if the something is magnetic (make sure your ruler is NOT magnetic).
	3. Plastic rulers may work better if the subject is magnetic.
Magnet—flat one and a "wand" with an extension.	**1. Identifies magnetic materials from nonmagnetic materials** **2. Can use the wand to retrieve items.** **3. Can use the wand to collect debris that is magnetic. This will immediately separate your debris into magnetic and nonmagnetic debris.**
Loop and magnifying glass	**Use to look at samples. Loop is higher magnification, 10–25×. Lower-power magnifying glass can also be helpful**
Tape measure	**Measure long distances**
Indelible ink marker (fine)	**Mark items, bags, bottles**
Flashlight	**Look at items in holes, dark areas, etc.. Good to have 90° bend attachment to look in holes and pin-light attachment to look in small crevices**
Conductivity tester	**Check conductivity or nonconductivity of surface**
Mirror	**Check around corners and under objects**
Surface-finish comparators	**Machine, cast, electrical discharge machining surface-finish standards by GAR Electroforming Division**
Microleatherman tool and pocket knife	**Scissors, screwdriver (flat and Phillips), punch, tweezers, blades, plastic toothpick (not on the plane, though). Various uses include examination tools, surface-finish profilameter, cutting, etc.**
Pen and pencil	**Just because…**
White pieces of lined	**Color standard, note paper, drawing paper, and collection funnel**

paper	
Plastic bags, 4 × 4 in.	**Sample collection**
Swabs	**Sample collection—always keep one for control sample**
Alloy reference list	**Materials, compositions, specifications, data**
Hardness Conversion Charts	**Martensitic and austenitic charts**
Addresses and phone numbers	**Assistance, information, etc.**
Technical information	**Specification lists, design criteria, drawings, notes, etc.. This is where the individual criteria come in.**
Cellular phone	**Immediate access to important contacts**
Eyewitness report forms	**Better if these are e-mailed or sent ahead to be filled out as soon as possible**
Laptop/palm pilot	**Computer provides spreadsheets, word documents, etc.**
Other	**Anything you think you will need**

Equally important are eyewitness statements or incident reports. Unfortunately, when things go wrong, eyewitnesses seem to disappear, or, at best, are reluctant to speak. Sometimes, they are shielded by management or unions. Getting information from them is a learned talent. One suggestion is to send forms ahead of time to be filled out. The number that are returned and the completeness of the forms may provide some indication of the amount of cooperation to be expected. In many cases, the investigator may never actually see the failure scene and has to depend on the photographs and eyewitness statements available.

It is advisable that the eyewitness statements be formatted in a standard form similar to a procedure. Personnel are more familiar with forms, and the document is less intimidating. It is also suggested that the eyewitness statements be electronic and in a form that is searchable. This is convenient. In addition, the information is useful later for the generation of statistics or for searching for similar failures, failures at the same plant, failures involving the same personnel, and so on.

There are other related information items that may not be obvious, such as on-the-floor “standard” process deviations, planned or unplanned plant shutdowns, time of year, weather conditions, where the failure happened (geography), and personnel losses. In fact, a good question to request on the eyewitness statement form is, “Are you aware of any recent changes in operations, personnel, equipment, and so on?” This kind of information is not apparent, so ask.

An example is the failure of a large precipitation-hardening martensitic stainless steel forging. The failures were sporadic and were discovered during acceptance testing, due to the lack of mechanical properties. It turned out that the forging company had no trouble cooling the large forgings to below -50 °C (-60 °F) during the quench cycle in the winter months when there was an abundance of snow outside to dump in the quench pit. However, during the hot summer months, the engineers had to calculate how much ice to bring in to dump in the quench tank. It became apparent that their calculations were not correct. This case is a perfect example of when geographic location and time of year are important.

Another very good question that should always be asked is, “Has this happened before?” Many times, a company simply dismisses a failure the first few times it happens and writes it off as a “unique” or “one-time” occurrence. These occurrences are then forgotten until the next one. Unfortunately, because the failure is not documented and is dismissed as unique, no one remembers the first failure, and so, the second failure may also be deemed unique. It is truly a downward spiral.

Be consistent. If the investigator always requests complete failure scene photographs, eyewitness statements, and all the information available, the customers start to collect the information before contacting the investigator. They will have been trained. If the investigator works in a large company, it would be a good idea to create the procedure “What to do when a failure occurs” and then to train company personnel on how to properly document and protect the failure scene and how to properly fill out eyewitness statements. As a start, there is the ASTM Committee E-30 on forensic sciences and ASTM Subcommittee E-30-05 on forensic engineering sciences that have created some specifications that can be reviewed and used when relevant. Some of the specifications are:

- E 620. Standard Practice for Reporting Opinions of Technical Experts
- E 678. Standard Practice for Evaluation of Technical Data

- E 860. Standard Practice for Examining and Testing Items That Are or May Become Involved in Litigation
- E 1020. Standard Practice for Reporting Incidents
- E 1188. Standard Practice for Collection and Preservation of Information and Physical Items by a Technical Investigator

Another good approach was adapted from a concept from the Failsafe Network, Inc. It is the concept of the five "Ps" that need to be documented and recorded to freeze the evidence at the failure scene:

- *Position:* fragments, equipment, parts, people (witnesses, people involved), controls, and photographs
- *People:* job descriptions, witnesses, accountabilities, information sources, experts, and those personnel that do not know
- *Paper:* drawings, design changes, manufacturing records, mill certificates, operation records, operating procedures, past failure histories, maintenance records, photographs, inspection records, stress analysis, and past nonconformances
- *Process:* design process, operational process, approved process changes, unapproved process changes (how it really operated or was made), environment, and weather
- *Part:* materials specified, materials used, mechanical and physical properties of materials for hardware or machine, fracture faces, distortion of the failed part and other hardware, remnants of failed hardware or machine, microscopy analysis, stress analysis, and metallurgical analysis

Step 3: Objectively and Clearly Identify All Possible Root Causes. The next step in the organization of a failure investigation is to objectively and clearly identify all possible root causes. There are many tools one can use. One common tool is to create a Fault Tree (Fig. 2). Computers make creation and revision of fault trees and other analytical evaluation tools easier. Once again, brainstorming is also a good tool. To discover every root cause, the investigator should ask, "Why, why, why?"

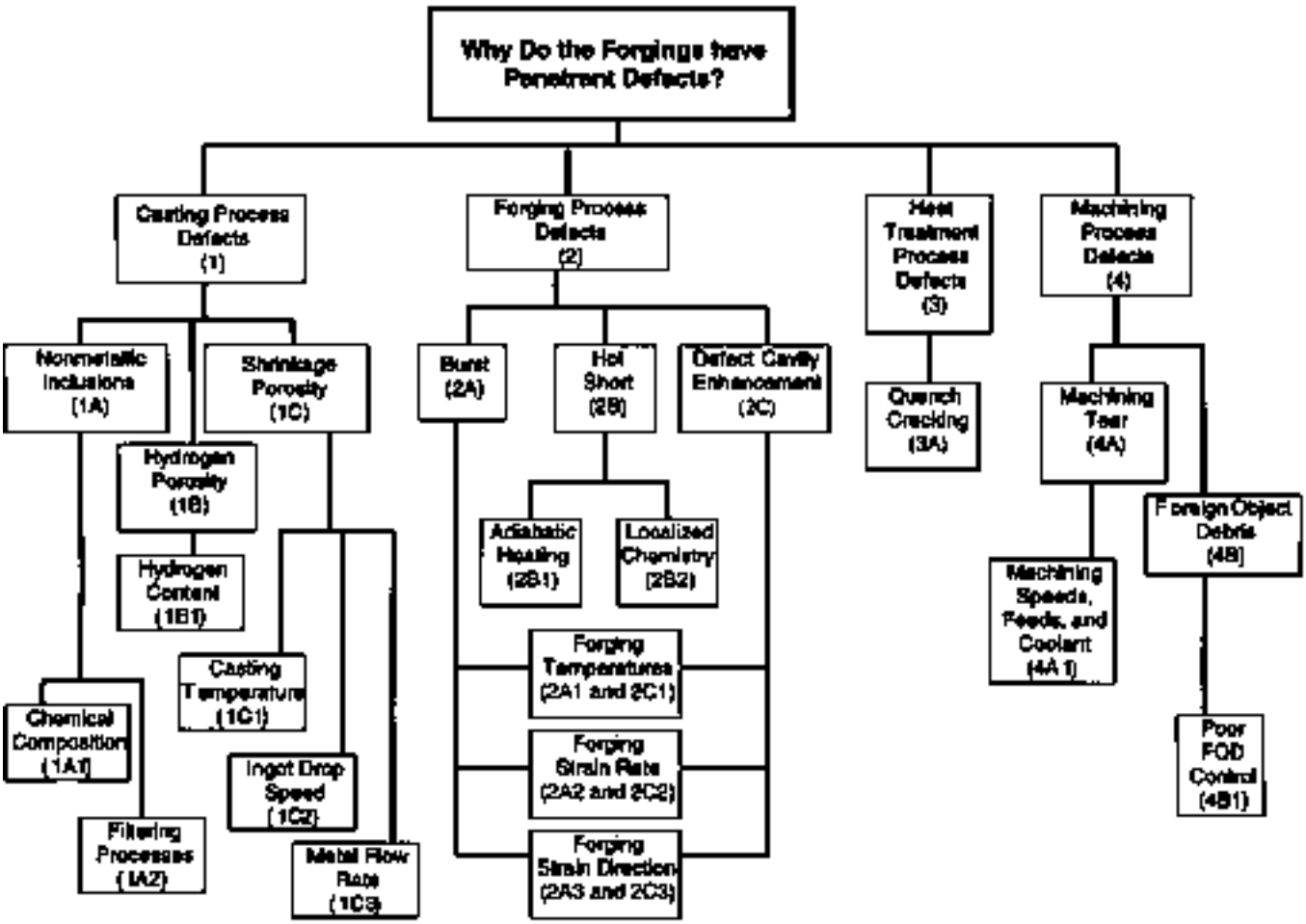


Fig. 2 Example of fault tree chart for forgings with dye-penetrant defects

First ask why the failure happened. This is the top of the fault tree. Write down the first root-cause answer, and then ask why that first root-cause answer happened. Continue asking “Why, why, why?” until each line of questioning is exhausted or a practical stopping point has been reached, but do not stop too soon. Then, go back to the top of the fault tree and start again. Keep doing this until there are no more root-cause answers to the “why” question. Be objective and list all answers; that is the key to brainstorming. If someone says that gremlins did it, write it down. The chance to be subjective comes later. It is not appropriate that any idea be dismissed. The dismissal violates the brainstorming concept, impedes the flow of ideas from other participants, and possibly ends the session.

Once the fault tree is completed, there are two other questions to ask: “What was different about this failure?” and “What are we missing?” Ask them over and over throughout the failure investigation.

There are many reasons structured root-cause analysis is important. First is to provide documentation. This is a permanent record of all possible root causes imagined by the team. No ideas are lost or forgotten. All root causes are succinctly listed and organized on one chart.

Second is to create a “living” list, which can be added to at any time. Once the failure investigation is in progress, information is generated that proves or disproves each root cause. The information may also lead to a new root cause not imagined before.

Third, these techniques can help simplify a very complicated failure. They help “divide and conquer” the problem by compartmentalizing many cause-and-effect relationships. The failure is divided into individual root causes, and each one of those is divided further. It is easier to investigate and prove or disprove each root cause than the resultant failure.

Fourth, structured problem solving may prevent the potentially misleading presumption of a “magic bullet” or “silver bullet” theory approach to failure investigation. There is always a predominant or “pet” root cause. Resist the urge to spend all the time, money, and manpower to prove this one root cause. If wrong, the investigator has wasted time and money and perhaps destroyed any evidence that would prove or disprove the other root causes in his haste to prove this one. Not all failure investigations have enough time, manpower and available test samples to test every root cause. The investigator may have to be selective in his approach. Also, the failure may be a combination of root causes, not just one. Just because the pet root cause is proved does not mean there is not another, more important root cause out there. Remember that the goal of a failure investigation is to determine the root cause(s) of a failure, not to prove one root cause. Sometimes, the failure investigation goes off track and becomes a proof of a singular theory, hence the silver or magic bullet theory.

Step 4: Objectively Evaluate Likelihood of Each Root Cause. The next step of the failure investigation is to objectively evaluate the likelihood of each root cause listed. A good tool for this is the Failure Mode Assessment (FMA) chart (Fig. 3). The FMA chart can be created in a spreadsheet.

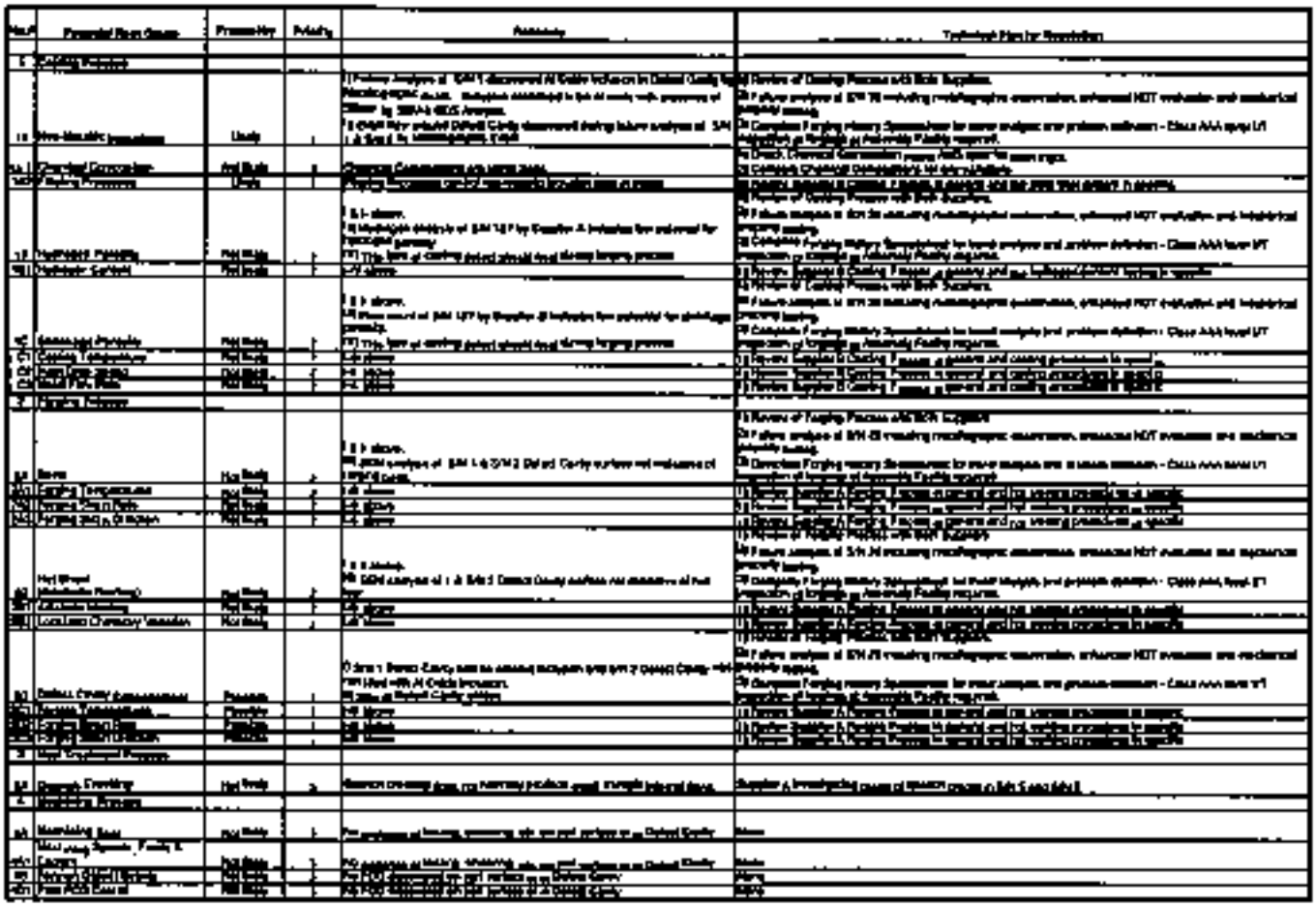


Fig. 3 Example of a failure mode assessment chart (for fault tree of forgings defects in Fig. 2)

Decision-making protocols can be as varied as the different kinds of decisions. Uncertainty and risk are part of most decisions, and so some assessment of probability is needed. As part of a failure investigation, first assess the probability of each potential root cause. How realistic is it? Each root cause is given a probability, such as likely, possible, or unlikely. Second, assign a priority to each potential root cause. This priority establishes the order in which the failure investigation will evaluate each root cause. Factors that effect this decision are hardware availability, amount of available hardware for test, cost, affect of the planned test on other tests, whether one test can be used to disprove or prove multiple root causes, and so on. Third, document the rationale used to assign a root cause its particular probability and priority. Once the failure investigation is started, no one will remember why a certain test was performed first or last if this rationale is not documented. Any action items directed toward proving or disproving the root cause should be listed in an abbreviated manner.

There are many reasons an FMA chart is important. The FMA chart is a permanent record of the assessed probability and assigned priority of each root cause, combined with the rationale for the particular assessments. The FMA chart also lists the action items to prove or disprove each root cause in an abbreviated form. Once again, this chart is excellent for briefings. The failure investigation is now documented on two easy charts, the fault tree and the FMA.

Similar to the fault tree, the FMA chart is a living list that can be changed at any time. Once the failure investigation is in progress, information is generated that proves or disproves each root cause. The information may also lead to a new probability or priority rating. A quick glance at this chart informs the investigator what happens if the probability or priority of one root cause is changed. The priority listing in the FMA provides the order in which each root cause is proved or disproved. This leads directly to the next step.

Step 5: Converge on the Most Likely Root Cause(s). Now it is time to converge on the most likely root causes(s). This would be step 2 in the process listed in the previous section, “The Four-Step Problem-Solving Process.” A good tool for this is the Technical Plan for Resolution (TPR) chart (Fig. 4), which can be created with a spreadsheet program similar to the FMA chart. In fact, a good practice is to tie the FMA and TPR charts

together on two sheets within one spreadsheet file, so that any changes made to the FMA are immediately reflected in the TPR.

Item	Potential Root Cause	Priority	Technical Approach for Resolution	What?	When?	Result?
1	Casting Process					
1A	[illegible]	1	1) Review of Casting Process with Both Suppliers.	1) Team	[illegible]	1) Casting Process, [illegible] 2) Casting Human Casting Processing - Reviewed [illegible] 3) Casting Human requires - No suggestions from Supplier 4) Process review [illegible] 5) [illegible] Casting Lean or CPI [illegible]
		1	2) Failure analysis of [illegible] including multiphase microscopy, [illegible] NDT evaluation and metallurgical property testing.	2) Engineering Center	Failure Analysis Phase I - 4/29/01 Failure Analysis Phase II - ???	
		1	3) Complete Forging History Spreadsheet for trend analysis and problem definition - Class AAA level UT inspection of forgings at Assembly Facility required.	3) Assembly Facility	Many results - [illegible]	
1A1	Chemical Composition	2	1) Check Chemical Composition results (MTR) each for each heat.	1) Supplier B	6/7/2001	
		2	2) Generate Chemical Composition for any variance.	2) Supplier B	6/7/2001	
1 B	Filtration Processes	1	1) Review Supplier B Casting Process in general and the triple filter system in specific.	1) Team	6/6/2001	1) Housekeeping Deficiencies 2) Same as 1A 3) Same as 1A 4) Same as 1A 5) Same as 1A
1B	Hydrogen Porosity	2	1) Review of Casting Process with Both Suppliers.	1) Team	6/6/2001	1) Casting process & hydrogen checks OK 2) Same as 1A 3) Same as 1A 4) Same as 1A
		2	2) Failure analysis of [illegible] including multiphase microscopy, [illegible] NDT evaluation and metallurgical property testing.	2) Engineering Center	Failure Analysis Phase I - 4/29/01 Failure Analysis Phase II - ???	
		2	3) Complete Forging History Spreadsheet for trend analysis and problem definition - Class AAA level UT inspection of forgings at Assembly Facility required.	3) Engineering Center	In Work - Constantly updated	No trends established so far
1B1	Hydrogen Content	2	1) Review Supplier B Casting Process in general and the hydrogen content process in specific.	1) Team	6/6/2001	1) Hydrogen/[illegible] Low 2) Same as 1A 3) Same as 1A 4) Same as 1A
1C	Microscopic Porosity	2	1) Review of Casting Process with Both Suppliers.	1) Team	[illegible]	1) Casting process acceptable 2) Same as 1A 3) Same as 1A 4) Same as 1A
		2	2) Failure analysis of [illegible] including multiphase microscopy, [illegible] NDT evaluation and metallurgical property testing.	2) Engineering Center	Failure Analysis Phase I - 4/29/01 Failure Analysis Phase II - ???	
		2	3) Complete Forging History Spreadsheet for trend analysis and problem definition - Class AAA level UT inspection of forgings at Assembly Facility required.	3) Engineering Center	In Work - Constantly updated	No trends established so far
1C1	Casting Temperature	2	1) Review Supplier B Casting Process in general and casting procedures in specific.	1) Team	6/6/2001	Same as 1C
1C2	Heat Over Heat	2	1) Review Supplier B Casting Process in general and casting procedures in specific.	1) Team	6/6/2001	Same as 1C
1C3	[illegible] Pour	2	1) Review Supplier B Casting Process in general and casting procedures in specific.	1) Team	[illegible]	Same as 1C
2	Forging Process					
2A	Metal	2	1) Review of Forging Process with Both Suppliers.	1) Team	6/7/2001	1) Holding of billets at high, high pressure [illegible] 2) Two Forgings with quench cracks in review 3) Forging Temperature [illegible] 4) Forging procedure review requested - Complaint issued - No Deviations from Supplier 5) Procedure [illegible] all forgings
		2	2) Failure analysis of [illegible] including multiphase microscopy, [illegible] NDT evaluation and metallurgical property testing.	2) Engineering Center	[illegible] Failure Analysis Phase I - [illegible] Failure Analysis Phase II - ???	
		2	3) Complete Forging History Spreadsheet for trend analysis and problem definition - Class AAA level UT inspection of forgings at Assembly Facility required.	3) Engineering Center	In Work - Constantly updated	[illegible]
2A1	Forging Temperature	2	1) Review Supplier A Forging Process in general and the forging procedures in specific.	1) Team	6/7/2001	Same as 2A
2A2	Forging Heat Time	2	1) Review Supplier A Forging Process in general and the forging procedures in specific.	1) Team	6/7/2001	Same as 2A
2A3	Forging Heat Duration	2	1) Review Supplier A Forging Process in general and the forging procedures in specific.	1) Team	6/7/2001	Same as 2A
2B	Hot Short Unbalanced Handling	2	1) Review of Forging Process with Both Suppliers.	1) Team	6/7/2001	Same as 2A
		2	2) Failure analysis of [illegible] including multiphase microscopy, [illegible] NDT evaluation and metallurgical property testing.	2) Engineering Center	Failure Analysis Phase I - 4/29/01 Failure Analysis Phase II - ???	
		2	3) Complete Forging History Spreadsheet for trend analysis and problem definition - Class AAA level UT inspection of forgings at Assembly Facility required.	3) Engineering Center	In Work - Constantly updated	No trends established so far
2B1	Unbalanced Handling	2	1) Review Supplier A Forging Process in general and the working procedures in specific.	1) Team	6/7/2001	Same as 2B
2B2	Unbalanced Chemistry Handling	2	1) Review Supplier A Forging Process in general and the working procedures in specific.	1) Team	6/7/2001	Same as 2B
2C	Quench Cracking Underground	1	1) Review of Forging Process with Both Suppliers.	1) Team	6/7/2001	Same as 2B
		1	2) Failure analysis of [illegible] including multiphase microscopy, [illegible] NDT evaluation and metallurgical property testing.	2) Engineering Center	Failure Analysis Phase I - 4/29/01 Failure Analysis Phase II - ???	
		1	3) Complete Forging History Spreadsheet for trend analysis and problem definition - Class AAA level UT inspection of forgings at Assembly Facility required.	3) Engineering Center	In Work - Constantly updated	No trends established so far
		1	4) Quench MR logs for other Suppliers forgings from other plants.	4) Engineering Center	6/7/2001	MR log information [illegible] quench shown type of metal, forgings [illegible]
2C1	Forging Temperature	1	1) Review Supplier A Forging Process in general and the working procedures in specific.	1) Team	6/7/2001	[illegible]
2C2	Forging Heat Time	1	1) Review Supplier A Forging Process in general and the working procedures in specific.	1) Team	6/7/2001	[illegible]
2C3	Forging Heat Duration	1	1) Review Supplier A Forging Process in general and the working procedures in specific.	1) Team	6/7/2001	Same as 2C
3	Heat Treatment Process					
3A	Quench Cracking	3	1) Supplier A investigation [illegible] of quench cracks [illegible] 3A1 and 3A2.	1) Supplier A	6/5/2001	
4	Machining Process					
4A	Machining Error	3	None	N/A	N/A	N/A
4A1	Machining Errors Cycle 1 Control	3	None	N/A	N/A	N/A
4B	Foreign Object Debris	3	None	N/A	N/A	N/A
4B1	Foreign Object Control	3	None	N/A	N/A	N/A

Fig. 4 Example of a technical plan for resolution

Previous steps (3 and 4) have provided the root causes, assessed the probability, and assigned the priority of each root cause. Now comes the time to create the technical plan to prove or disprove each root cause. The TPR

ensures that the testing or analysis is performed in a manner that should achieve maximum effect and efficiency. It is a detailed road map in which one can independently list the work needed to prove or disprove each root cause. Once the details are down on paper, a specific order of testing or analysis or a synergy of testing or analysis becomes evident. The investigator may change the priority of some testing or analysis, because it shows up under multiple root causes and is a good test to perform early. Alternatively, the investigator may discover he runs out of material by the time he gets to a certain test or analysis.
At this stage, a good tool is to ask the investigative team three questions about the failure:

- What do I know?
- What do I think?
- What can I prove?

"What do I know?" These are the hard facts and existing data. List the facts and then determine to which root cause(s) they are applicable. When the list of facts is completed, ask "What do I think?" These are the theories, ideas, and possible cause-and-effect relationships. These theories and ideas are based on history, experience, logic, and the hard facts noted previously. This is the time for personnel to ask "What if…?"
Finally, ask "What can I prove?" For each of the theories, ideas, or root causes, list the physical evidence one would expect to see if it occurred, as well as the tests, analyses, and so on required to prove or disprove each one. It is sometimes easier to disprove a theory for the root cause than to prove it. This may involve testing or analysis of the failed hardware, testing or analysis of similar hardware, exemplar testing to demonstrate if a failure mode is possible, operational data analysis, stress analysis, literature searches, a phone call to other divisions or companies that might have seen a similar problem, and so on.
It is equally important to list who has the action to get a test or analysis performed, the completion date, and the results on the TPR. It cannot be stressed enough that the person assigned must agree both to performing the test or analysis and to the completion date. Just assigning a task does not always ensure that it gets done, but if someone agrees to the task and the deadline, then the work normally gets accomplished. Documenting the results of the test plus any comments on the TPR completes the documentation.
Similar to the Fault Tree and FMA chart, there are many reasons a TPR chart is important. First and foremost, the TPR chart is a permanent record of each test or analysis that is to be performed, the person assigned, the expected completion date, and the results, with any comments. Once again, this chart is excellent for briefings. In addition, if any changes occur on the Fault Tree and FMA chart, the TPR chart will indicate whether there is a need for more testing or if the necessary testing is already planned.
Step 6: Objectively and Clearly Identify All Possible Corrective Actions. The failure investigation is not completed yet. In fact, it is just half done. After weeks or months of hard work, the failure investigation determines the root cause(s). Unfortunately, many failure investigations end here, but now the investigator needs to determine the corrective action(s) to prevent the failure from occurring again.
In this step, all possible corrective actions are identified. This is step 3 from the "Four-Step Problem-Solving Process," "What are the potential solutions?" Step 6 is similar in concept to Step 3, "Objectively and clearly identify all possible root causes." The difference is that the failure investigation team is now centered on identifying all possible corrective actions to solve the root cause identified in the failure analysis. Repeat the instructions in step 3, and create a Corrective Action Tree (Fig. 5).

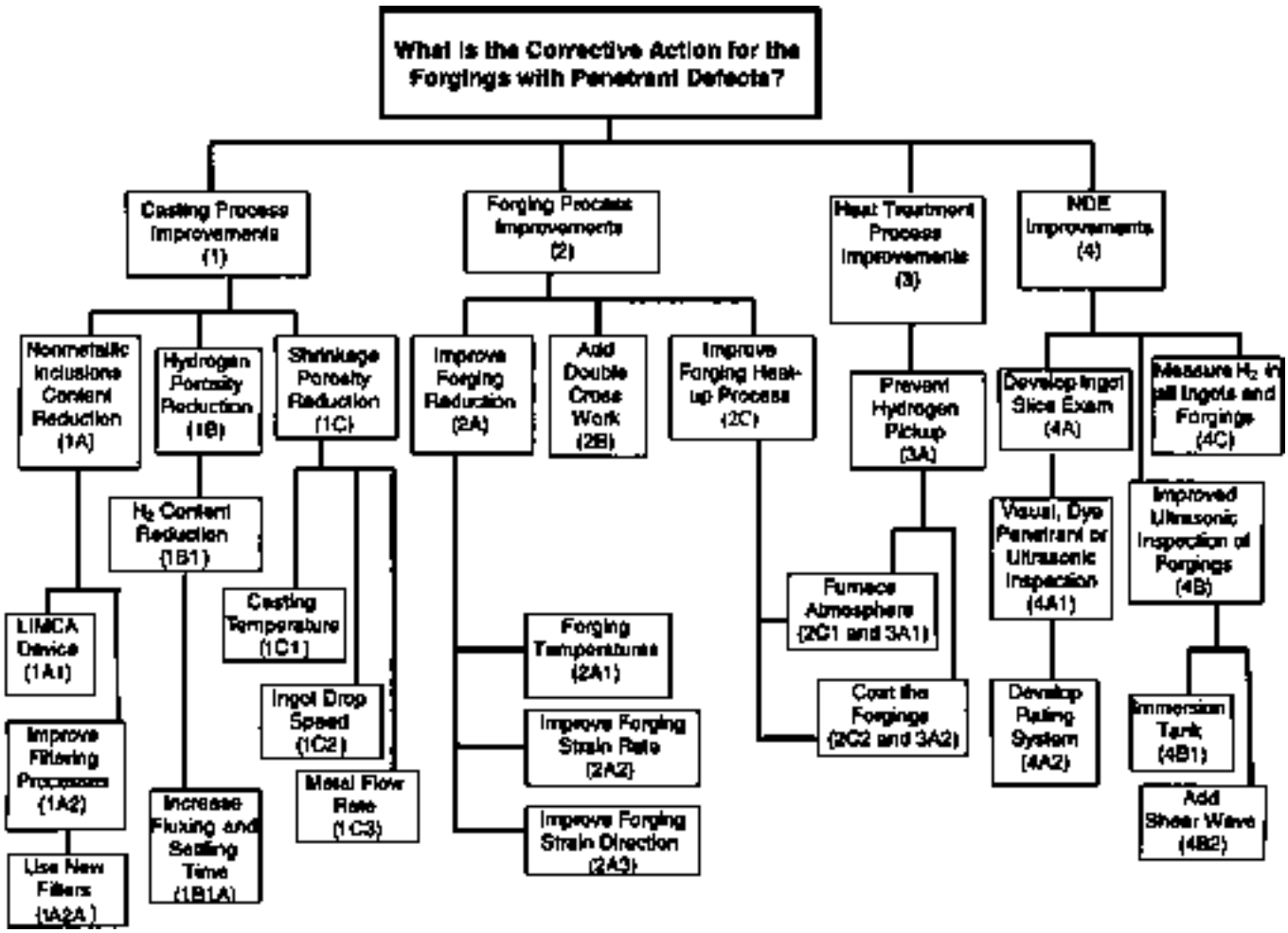


Fig. 5 Example of corrective action tree for forgings with dye-penetrant defects. LIMCA, liquid metal cleanness analyzer device

Step 7: Objectively Evaluate Each Corrective Action. The next step of the failure investigation is to objectively evaluate the likelihood of each corrective action listed in the corrective action tree. This step is a repeat of step 4, “Objectively evaluate likelihood of each root cause.” The difference is that the failure investigation team is assessing the probability and assigning priority to each corrective action, combined with the rationale. Repeat the instructions in step 4, and create a Corrective Action Assessment (CAA) chart.

Step 8: Select Optimal Corrective Action(s). Now it is time to select the optimal corrective action(s). This is step 4 from the “Four-Step Problem-Solving Process,” “What is the best solution?” This step is a repeat of step 5, “Converge on the most likely root cause(s).” At this step, the failure analysis team creates the technical plan to determine the best corrective action(s). Repeat the instructions in step 5, and create a Technical Plan for Evaluation (TPE) chart.

Step 9: Evaluate Effectiveness of Selected Corrective Action(s). The last step is to recommend an evaluation of the corrective action(s) selected. After some period of time, an evaluation of the corrective action should be performed. Whether the evaluation should be performed once or on a periodic basis is at the discretion of the failure investigation team. It is also important that the evaluation be a proper test of the corrective action.

Summary

The goal of any failure investigation is to discover the root cause(s). The first five steps of the nine-step process described here point the failure investigator toward that goal. Once the root cause(s) is determined, the second part of a failure investigation is to determine the best corrective action(s), as noted in the last four steps discussed.

The three basic tools identified in this article that are helpful in any failure investigation are a Fault Tree, an FMA chart, and a TPR chart. These three tools provide a documented, interchangeable, and concise set of

information to ensure that all possible root causes are evaluated in the most efficient manner and order. These three tools are then repeated for the corrective actions in the form of the Corrective Action Tree, the CAA chart, and the TPE chart.

Failure Investigation Pitfalls

However, there are pitfalls along the way. Common problems and mistakes are:

- The magic bullet or silver bullet theory
- Not understanding the problem
- Not considering all possible failure causes
- Jumping to conclusions
- Not identifying all the root cause(s)
- “Remove and replace” company mentality
- Not asking for help
- Failing to follow through
- Not understanding how the failed system is supposed to operate
- Tearing the system apart without a developed plan

Some companies believe there are substitutes or shortcuts for a root-cause failure investigation that can save time and money. In the end, these substitutes do neither. These substitutes include:

- “Give me your best guess”
- “Give me a five minute failure analysis”
- Rework and repair based on no failure investigation
- Return the part to the supplier and let them figure it out
- Scrapping the hardware
- Swapping out for another part
- Ignoring the problem and hoping it is a unique occurrence
- “Band-aid” fixes that do not address the root cause

In the end, these substitutes are generally neither cost-effective nor schedule friendly. Furthermore, if a failure involves personal injury, property damage, or other economic loss, the failure to find the root cause as soon as possible may expose a company to further liability.

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

Other Tools

Every failure investigation is unique. Therefore, it pays to have many tools in your arsenal. This article presents one set of tools for the organization of a failure investigation. There are other tools as well, but the author has developed this approach and finds it to be a cost-effective technique that works extremely well in most situations.

Other tools that involve failure investigation and root cause determination are processes and methods developed by Kepner-Tregoe, Inc., the Reliability Center, Inc., and The Failsafe Network, Inc.

The Kepner-Tregoe (KT) method is another problem-solving technique used in failure analysis that streamlines the process and works well in situations where industrial failures occur. The KT method is not as thorough as fault-tree analysis because potential causes get trimmed early in the process if they are found to be impossible. However, it is a useful method of problem solving and decision analysis.

The Kepner-Tregoe method is a structured problem-solving approach that first defines the problem in terms of what “is” and what “is not” as related to the questions of what, where, when, and extent. For example, gas turbine blade cracking may have been confined to one type of cracking mode, one specific row of blades, and one type of turbine blade design with the problem occurring only after a certain amount of operating hours, with the cracking observed in all plants. This shapes what “is” and what “is not” in a KT session. After the problem has been defined, possible root causes are developed and compared with the “is” and “is not” statements. Root causes are tested against the “is” and “is not” problem statements to test if they are possible. After impossible root causes are eliminated, the possible root causes are ranked in terms of probability. In real world practice, additional investigation focusing on the remaining possible causes is then performed to try to eliminate all but the one true root cause.

The Reliability Center, Inc. developed a software package called PROACT (The Reliability Center, Inc., Virginia), which is a comprehensive software tool that automates the root cause analysis (RCA) investigation process. PROACT software allows the user to concentrate on solving the problem at hand instead of having to worry about where to keep all of the data and how to present it. The PROACT software serves as a project management tool for conducting an RCA investigation.

The PROACT software proposes to provide many benefits to the user. PROACT incorporates RCA methods that have been proved to reduce production costs while increasing production output. It cuts analysis time in half by eliminating the extensive administrative work that is usually associated with conducting RCA. The software also provides a standardized reporting format so that users and decision makers can focus on the content, rather than the format, of the report. It provides a presentation module that allows the analyst to present critical information to decision makers quickly, clearly, and easily. PROACT maintains all analysis data to eliminate the need to search through volumes of computer and paper files to get needed information. Lastly, it provides an easy to use drag-and-drop “logic tree” that focuses on solving the problem and not on becoming a CAD operator.

The Failsafe Network, Inc. (Montebello, VA) has promoted various concepts in failure investigation. Operation Failsafe is a comprehensive plan for ingraining the root cause mentality within people and their organizations. The ROOTS Investigative Process is a flexible, open-ended, practical approach that helps the inquirer discover the physical, human, then latent causes of failure. This process is an evidence-driven process that proposes to drive the serious inquirer into areas most in need of clarification. The WHY Tree is the “inquiry engine.” It proposes to be a tool for keeping investigative teams on-track by documenting where the investigator is in his or her current understanding of a problem, whether it be with a team or as an individual.

The Failsafe system also includes the Combing Process, a failure mode and effects analysis (FMEA)-based tool for rapidly identifying “significant few” chronic failures in hours or days. It proposes to be an approach that goes to the hands-on operating and maintenance people for information, rather than relying on existing databases. The Situation-Filter-Outcome Model proposes to be a practical and decisive way to put oneself in someone else's shoes—to understand why people do what they do. It can be a way to surface the latent causes of a failure. The Significant 6 Barriers to True Root Cause Discovery is a series of statements, articles, and graphics intended to make the real issues visible. Lastly, the ROOT CAUSE CONFERENCE is a nonproprietary, intercompany, interindustry body of people acting as a beacon for exposing, then acting on, the real root causes of things that go wrong in business and industry. The conference is considered a “centralized mother-source” (root cause core) group.

Organization of a Failure Investigation

Daniel P. Dennies, Ph.D., The Boeing Company

Selected References

- D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 2000
- G.F. Vander Voort, Conducting the Failure Examination, *Practical Failure Analysis,* April 2001, adapted and reprinted in this Volume
- N. Schlager and H. Petroski, *When Technology Fails,* GNE Research, Inc., 1994
- C.H. Kepner and B. Tregoe, *The New Rational Manager,* Princeton Research Press, 1997
- R.J. Latino and K.C. Latino, *Root Cause Analysis,* CRC Press, 1999
- C.R. Nelms, *What You Can Learn From Things That Go Wrong, A Guide Book to the Root Causes of Failure,* Failsafe Network, Inc., 1994
- C.R. Nelms, *The Dynamics of Inculcating the Root Cause Mentality,* Failsafe Network, Inc., 1995
- C.R. Brooks and A. Choudhury, *Metallurgical Failure Analysis,* McGraw-Hill, 1993
- C.R. Brooks and A. Choudhury, *Failure Analysis of Engineering Materials,* McGraw-Hill, 2002
- V.J. Colangelo and F.A. Heiser, *Analysis of Metallurgical Failures,* Wiley Interscience, 1987
- J. Berk and S. Berk, *Managing Effectively,* Sterling Publishing Co., 1993
- D.A. Ryder, "The Elements of Fractography," AGARD-AG-155-71, Technical Editing and Reproduction Ltd, Harford House, Nov 1971
- C.R. Walker and K.K. Starr, "Failure Analysis Handbook" WRC-TR-89-4060, Pratt & Whitney, Aug 1989

Conducting a Failure Examination

George F. Vander Voort, Buehler Ltd.

Introduction

FAILURES may be caused by any of the following factors or combinations of factors:

- Design shortcomings
- Material imperfections due to faulty processing or fabrication
- Overloading and other service abuses
- Improper maintenance and repair
- Environment-material interactions

Not all failures are catastrophic. Many failures involve a gradual degradation of properties or excessive deformation or wear until the component is no longer functional long before its design life is reached. Failures due to wear or general corrosive attack usually are not spectacular failures but account for tremendous material losses and downtime every year. Of course, early failures of the spectacular catastrophic order capture the most attention—and rightly so. Nevertheless, all failures deserve the attention of the investigator, because they reduce production efficiency, waste critical materials, and increase costs. In some instances, they cause considerable damage or personal injury. Finally, failures can result in costly litigations.

The need to develop and encourage the science (and art) of performing service failure analysis and the importance of dissemination of such information is well recognized, although industry and academia may not have sufficiently emphasized this in the past, as noted in the 1975 article "Conducting the Failure Examination" (Ref 1). In some situations, investigations were carried out, but the manufacturer concerned did not wish to share the resulting information publicly. At other times, there was probably no reluctance to share such information, but reporting channels were inadequate for making this information available to those who could benefit and, in turn, provide feedback. Finally, even within a given organization, failure analysis data and conclusions were not adequately communicated to the designers so that they could implement design changes aimed at developing an improved product, or to the operations personnel so that they could eliminate processing problems and improve quality control.

Fortunately, improvements have been made since then, and examination methods have improved. This article is an update of Ref 1. Many of the fundamentals remain the same, but they warrant repeating here and in other publications or other articles in this Volume. Education and communication are the essential ingredients of failure prevention and analysis, because it is still sometimes difficult to get the failure information fed back to the design and manufacturing stages. Communication is key for any type of failure. Not all failures require a comprehensive, detailed failure examination. With many types of "routine" failures, the procedure can be simple and still be effective. Most suppliers are receptive to constructive criticism regarding their products.

Individual background, including the area of specialization and professional experience, still has a strong influence on the manner in which an examination is performed. For example, the civil or mechanical engineer may perform a very careful examination of the stresses and loads causing the failure but may not devote much attention to the metallurgical factors. By the same token, the metallurgist may carefully sort out all of the important material aspects of the failure but may examine the mechanical aspects only superficially. Similar imbalances in approach may occur when operating personnel, welding engineers, or corrosion engineers conduct the examination. Few engineers are knowledgeable in all of the possible interacting disciplines that can be involved in a given failure; thus, a team effort is often the only approach that can produce a complete picture of the failure causes. This approach has certainly enhanced the quality of the author's failure analyses work.

Reference cited in this section

1. G.F. Vander Voort, *ASM Metals Engineering Quarterly*, Vol 15(May), 1975, p 31–36; Adapted version published in *Practical Failure Analysis*, April 2001

Basic Approach to Failure Analysis

First and foremost, the investigator should not approach the examination with preconceived notions about the cause of failure. For example, if a supervisor provides a sight diagnosis to a failure task as they are assigned to a given engineer, objectivity may be lost. The common reactions may be to prove or disprove the initial cursory reaction at the expense of other hypotheses. Objectivity is lost before the job is even started.

The examination itself should be designed to test a broad range of conditions, yet not be so exhaustive as to become impractical because of excessive expense or time. Design of the investigation includes proper sequence of the selected tests; for example, a wrong sequence can introduce artifacts or mask and even destroy important evidence. Even a well-designed examination can fail of its purpose, unless it is conducted carefully and systematically.

When the series of tests has produced data that pinpoint problems, the job is still not finished. The investigator should, whenever possible, include in his report specific recommendations aimed at eliminating the problems through alternate designs, materials, processing, heat treating, or fabrication techniques. He should review these recommendations with the designer, manufacturer, stress analyst, or other specialist so that these changes produce the desired results and do not cause other problems. Potential problems in similar components should be investigated so that corrective action includes the prevention of similar failures.

Conducting a Failure Examination

George F. Vander Voort, Buehler Ltd.

Failure Analysis Procedures

The basic steps to be followed and the range of techniques available for failure analysis are outlined subsequently. Those steps requiring additional detail are expanded in subsequent sections. The basic steps listed may not all pertain to each investigation, and the order of performing these steps varies somewhat, depending on the nature of the investigation.

1. Assemble background data, including (a) all information pertaining to the failure; (b) the history of the part, that is, processing and service; and (c) pertinent codes, specifications, and standards.
2. Perform visual examination of the failure area and adjacent areas to determine, photograph, and/or schematically record: (a) origin of failure; (b) presence of stress concentrators; (c) presence of temper color or scale on fracture faces; (d) orientation and magnitude of stresses; (e) failure mode and mechanism; (f) direction of crack propagation and sequence of failure; (g) presence of contributing imperfections; and (h) sizes and other physical data.
3. Perform fractographic examination.
4. Perform chemical analyses and compare results with specification or standards. Analyze any important surface corrosion products, deposits, or coatings.
5. Determine mechanical properties and compare with specifications or standards.
6. Perform macroscopic examination to evaluate homogeneity, integrity, and quality.
7. Perform metallographic examination to evaluate microstructural features. Determine the deformation direction of wrought products and its relationship to the applied and residual stresses.
8. Perform microhardness testing to measure case depths, evaluate cold working, determine quality of weldments, and aid in identifying phases.
9. Perform high-magnification metallographic examination using the electron microscope to study phases unresolvable with the light microscope.
10. Microprobe any critical abnormalities, such as inclusions and segregations, that are too small for bulk analysis.
11. Use x-ray techniques to determine: (a) level of residual stress; and (b) the relative amounts of phases, for example, austenite or retained austenite, ferrite, delta ferrite or martensite, sigma, and carbides in steels.
12. Perform simulation tests to evaluate critical characteristics of the material (such as the stress-corrosion cracking tendency in a particular environment), to determine the degree of embrittlement, or to confirm the method of heat treatment or hardenability.
13. Summarize and analyze all pertinent data. Consult with experts when necessary.
14. Write the report and distribute it. The report should include recommendations to prevent future problems or problems in similar existing equipment.
15. Follow up on these recommendations.
16. Preserve the evidence for the benefit of other examiners.

Assembling Background Data. The investigator must develop a complete case history of the component, including details of the failure and information about the manufacturing history before he can intelligently select those tests and procedures that are best used to analyze the failure. Important items to be determined include:
Information about the failed component:

- Location, name of item, identifying numbers, owner, user, manufacturer, and fabricator
- Function of item
- Service life at the time of failure
- Rating of item, operating levels, normal and abnormal loads, frequency of loading, and environment
- Materials used
- Manufacturing and fabrication techniques used, including specifications and codes governing the manufacturing, fabrication, inspection, and operation of the component
- Normal stress orientation, operating temperature range, pressures, and speeds
- Strength and toughness
- Heat treatment, stress relief, or other thermal processing
- Fabrication procedures, such as welding, adhesive joining, coatings, bolting, and riveting

- Inspection techniques and quality control reports during manufacturing; service and maintenance records

Information about the failure:

- Date and time of failure, temperature, and environment
- Extent of damage, sequence of failure, and injuries
- Stage of operation when failure occurred
- Blueprints, photographs, or sketches of the failure and adjacent areas
- Any service deviations that might have contributed to the failure
- Opinions of operating personnel regarding the failure

The investigator should carefully study available blueprints and other graphic records to obtain a comprehensive picture of the failed component. Such information is also helpful in evaluating stresses. Whenever possible, the investigator should go to the location of the failure for first-hand examination, so that he can select critical areas for examination. The investigator should get the failed component *promptly*—before the structure and appearance of the material have changed because of handling, repair attempts, corrosion, or other factors.

While developing the case history of the component, the investigator should become familiar with the materials, the processing, and the fabrication techniques employed in manufacturing the component. Time spent researching unknowns usually pays off during the investigation.

Obtaining background information was crucial in the investigation of a Loran tower that collapsed after about 289 days service. The tower was designed considering static loads only. A pin in one of the guy wires had failed by fatigue, and fractographic analysis indicated that there were at least 250,000 stress cycles during the growth of the fatigue crack from the origin to final rupture. Figure 1 shows examples of fatigue striations seen on the pin surface. Figure 2 is a plot of the striations per micron versus distance.

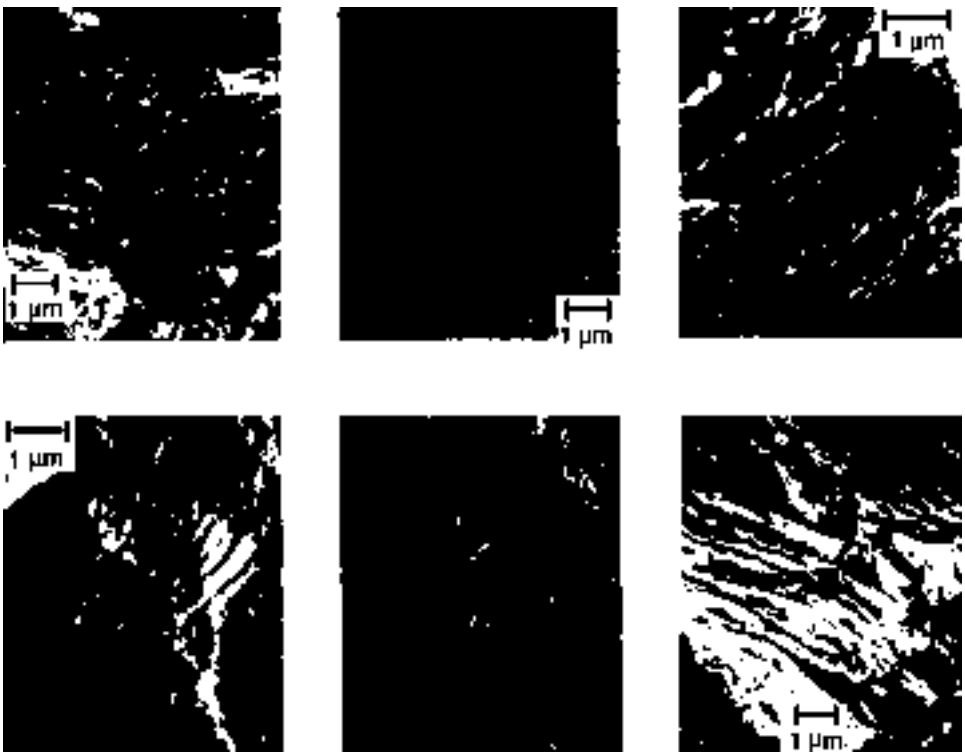


Fig. 1 Examples of fatigue striations seen on the pin surface from the Loran tower failure

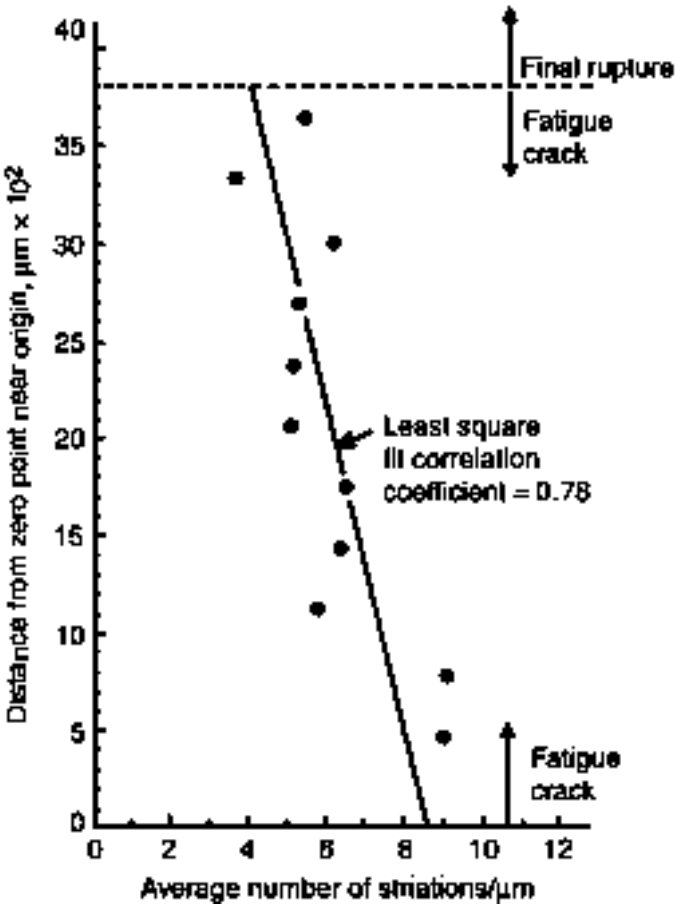


Fig. 2 Plot of striations per micron versus distance for a broken pin from the Loran tower. The original investigator cut the origin out, so this plot begins at as close as it was possible to get to the origin and extends to the final rupture. The plot was integrated graphically, revealing at least 250,000 cycles during the growth of the fatigue crack.

However, analysis of the weather conditions during the life of the tower could account for less than 10% of the cyclic loading, ignoring the number of cycles required to initiate the crack. Fortunately, a worker at the site had observed, and recorded in his notebook, strange oscillating movements in the guy wire that broke (causing the tower to collapse). He did not understand the nature of these oscillations, which looked like sine waves, but he recorded their occurrence and the date and time. What he observed were Aeolian vibrations (named after Aeolus, god of the winds) that can occur in guy wires when improperly tensioned and subject to winds of a specific direction and speed. Aeolian vibrations accounted for the missing stress cycles.

Because the guy wires were very long (the tower was about 410 m, or 1350 ft, tall), they would have interfered with the low-frequency Loran signal, so the wires were broken into smaller segments by placing a number of porcelain insulators in the strand. Porcelain, however, has limited tensile strength but good compressive strength, so the insulators must be loaded in compression with the complete strand loaded in tension.

The steel pin that broke was placed through the insulator to load it in compression. The pin was not properly aligned, however, producing a bending moment. The pin was made of 1035 carbon steel about 51 mm (2 in.) in diameter with a 63.5 mm (2.5 in.) diameter head. It was oil quenched and tempered, rough turned, pickled, and hot dip galvanized. A tool steel slip ring was placed under the head between the pin and the insulator.

Misalignment caused the tool steel ring to make marks in the galvanized coating on one side where it was in compression but not elsewhere, because there was no contact. The investigator for the plaintiff did not understand the loading and did not know how to interpret these marks, so he documented their existence but could not explain their origin. He was a metallurgist but did not consult with anyone with the right background to help interpret these observations. This resulted in an incomplete and inadequate failure analysis.

Visual Examination. Before proceeding to the visual examination, the failure analyst should ask whoever reported the failure to protect the component from further damage. Sometimes important evidence will have already been destroyed by rough handling or exposure to corrodents before the investigator sees the failure. But, the protection process should not destroy evidence. The author has received, without forewarning, wooden boxes containing broken pieces free to move within the box without any protective wrapping. This causes no end of problems for the investigator and can make correct and complete analysis impossible.

Throughout the visual examination phase of the failure study, all significant information should be properly recorded or photographed so that the information can be shown and described to all concerned parties. Color photographs are occasionally required to show contaminants, scorch, and temper color. Every possible piece of information should be squeezed out of the sample before any destructive testing is performed. Too frequently, the investigator cuts up the component before performing adequate visual examination. Location and orientation of any samples cut from the failed component should be carefully recorded. Care should be taken during sectioning, grinding, polishing, and even handling to avoid creating artifacts that could lead to false conclusions.

The investigator should preserve the fracture initiation area as it was at the time of failure. For example, any band saw cutting, abrasive-wheel cutting, or burning should be done well away from the fracture initiation area. The fracture face should not be touched by any liquid and should be kept dry and clean. If the part must be shipped, it should be carefully packed so that the fracture face does not rub against other pieces and so that it stays dry during transit. Because the original microstructure, mechanical properties, and fine fracture features are altered by tempering, a very hard piece should not be tempered to facilitate saw cutting. Rather, abrasive blade cutting or some other techniques should be used that will not alter the as-received condition. Burning (i.e., oxyacetylene or plasma arc cutting), if used, should be done far enough away from the fracture so that the heat does not alter the conditions.

If the fracture is severely altered by corrosion, fire, or rubbing, the failure analysis may be severely impaired. Mild corrosion and dirt or grease can easily be removed to allow accurate macroscopic examination and fractography. If the fracture is heated by fire or by tempering, the fine features may be destroyed, but the macroscopic appearance may still be visible. Rubbing of the fracture face obliterates the fine features and, in many instances, the macroscopic features. If fractography is to be performed, fracture faces should not be remated, because this almost invariably disturbs the fine features and could make fractography impossible.

The author has encountered cases where the fracture face was destroyed, yet a valid failure analysis study could be obtained. As an example, the author studied an Auto Train locomotive drive axle (Ref 2) that failed due to overheating of the friction bearing. 3 The train was approaching a city, and no power was being applied from the traction motor through the axle to the wheels, because they had to continuously reduce their speed as they approached the city. Lubrication was lost to the friction bearing. It got hot, and the low-melting-point Babbitt metal lining the bronze cylinder of the friction bearing melted and/or volatilized, so that the axle was rubbing against the bronze cylinder. With time, the cylinder and the axle got hot enough to melt the bronze alloy and transform the approximately eutectoidal carbon steel of the drive axle to austenite.

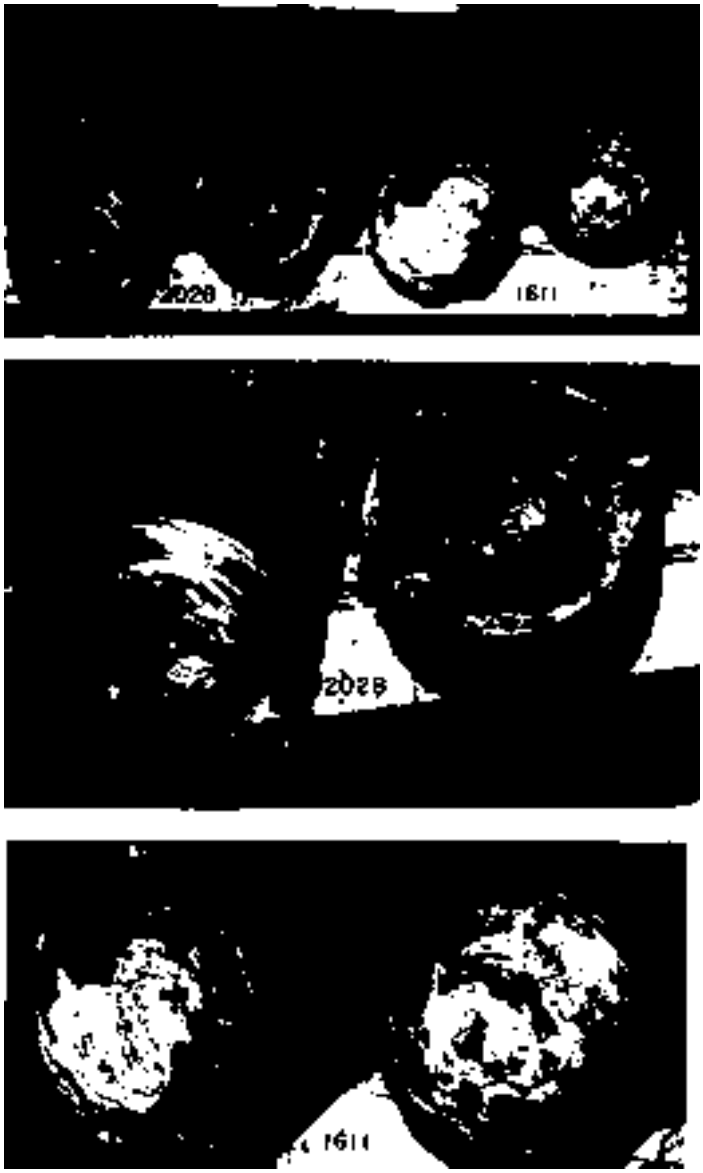


Fig. 3 Broken axles from two different Seaboard Coast Line locomotives that failed due to overheated friction-bearing failures. The pictures show both axles that were used to learn about the failure mode. The fracture faces of axle 2028 in the photo were damaged badly while those of 1611 were not.

Because there was an applied tensile stress, liquid copper started to penetrate the austenite grain boundaries (liquid metal embrittlement). Meanwhile, the axle lost much of its strength, and the partially cracked axle ended its life as a hot torsion failure. In this case, rupture occurred about 17 miles before the train derailed. The broken axle and friction bearing hung below the track and hit the ties, leaving marks that could be traced back to the geographic origin of the rupture. As the train approached the city, no power was applied to the broken axle, but it derailed when passing over a switching “frog,” because the broken axle, under the weight of the traction motor, was hanging downward at the center, and the fracture faces hit the frog.

The plaintiff's expert was asked to look for copper penetration but did not understand the failure mechanism. The axle was sectioned into nearly 60 pieces, and two were prepared metallographically (but very poorly) and examined using energy-dispersive spectroscopy (EDS) with the scanning electron microscope (SEM). No evidence of copper penetration was observed, so the expert claimed that the axle was defective, and the railroad brought suit against the axle manufacturer.

Although it was obvious that the specimen preparation was very poor and that this would make detection of copper by EDS more difficult, it was uncertain how critical this was. Further, having never seen such a failure and knowing that the fracture surfaces were badly damaged due to banging into each other for 17 miles and also from the impact with the frog, the defendant's (manufacturer's) analysts were concerned about not being able to

determine the true cause of failure. The friction bearing was not preserved after the accident and was neither examined nor documented.

The first step was to obtain examples of axles that failed due to overheating of friction bearings. Reference 2 documents the history of such failures. Two axles were obtained for study from the Seaboard Coast Line railroad (Fig. 3). Axle 1611 had failed in a yard and had minimal damage to the fracture face, whereas the other, number 2028, had failed on a passenger train and had extensive damage. Figure 4 shows a close up of the damaged commutator side fracture face of axle 2028. Examination of the microstructure along the axle outer diameter, from the fracture face moving along the portion under the friction bearing, revealed a vast partial network of prior-austenite grains filled with the bronze alloy of the friction bearing (Fig. 5). This network extended all along the surface that was in contact with the friction bearing to a depth of more than 25.4 mm (1 in.). The main rupture occurred in the part in the center of the friction bearing where the temperature was highest. Stresses were not very high but were adequate for initiation of liquid metal embrittlement.

Fig. 4 Fracture surface at the commutator side of Seaboard Coast Line axle 2028

Fig. 5 Optical micrograph of copper penetration (liquid metal embrittlement) in the broken axles. 300×

But, what could the analysts do about the expert's inability to see the copper in the axle in the lawsuit? A specimen from each of the two axles studied was ground with 120-grit SiC paper and given to the analysts' SEM operator. He was shown two color micrographs depicting the nature of the discontinuous grain-boundary

copper films in each specimen. He was asked to go through the preparation sequence, step by step, and examine the specimens with EDS after each step, in order to see when he could detect the copper.
Three days later, he showed the author a SEM micrograph and x-ray dot scans for copper. However, he said that he went all the way through the preparation sequence to a perfectly prepared condition (far better than done by the plaintiff's expert) yet could not detect the copper. So, he went to the light microscope and immediately saw the copper films, due to their natural yellowish color compared to the steel matrix. He scribed an area and photographed it with the light microscope. He located the scribed area with the SEM, took the picture, and made x-ray maps of copper.
This is a classic case of using the wrong instrument, because the SEM cannot detect the natural color of copper, which was so obvious with the light microscope. Figure 6 shows the copper penetration in the Auto Train axle. Copper and iron are atomic numbers 29 and 26, respectively, so there is little atomic number contrast between them. Faulty preparation masks the copper, because it can be easily smeared. Even when well prepared, it helped to know where the copper was located when doing x-ray maps. So, many thousands of dollars were spent defending the company in this lawsuit, simply due to poor metallographic preparation and failure to use the light microscope.

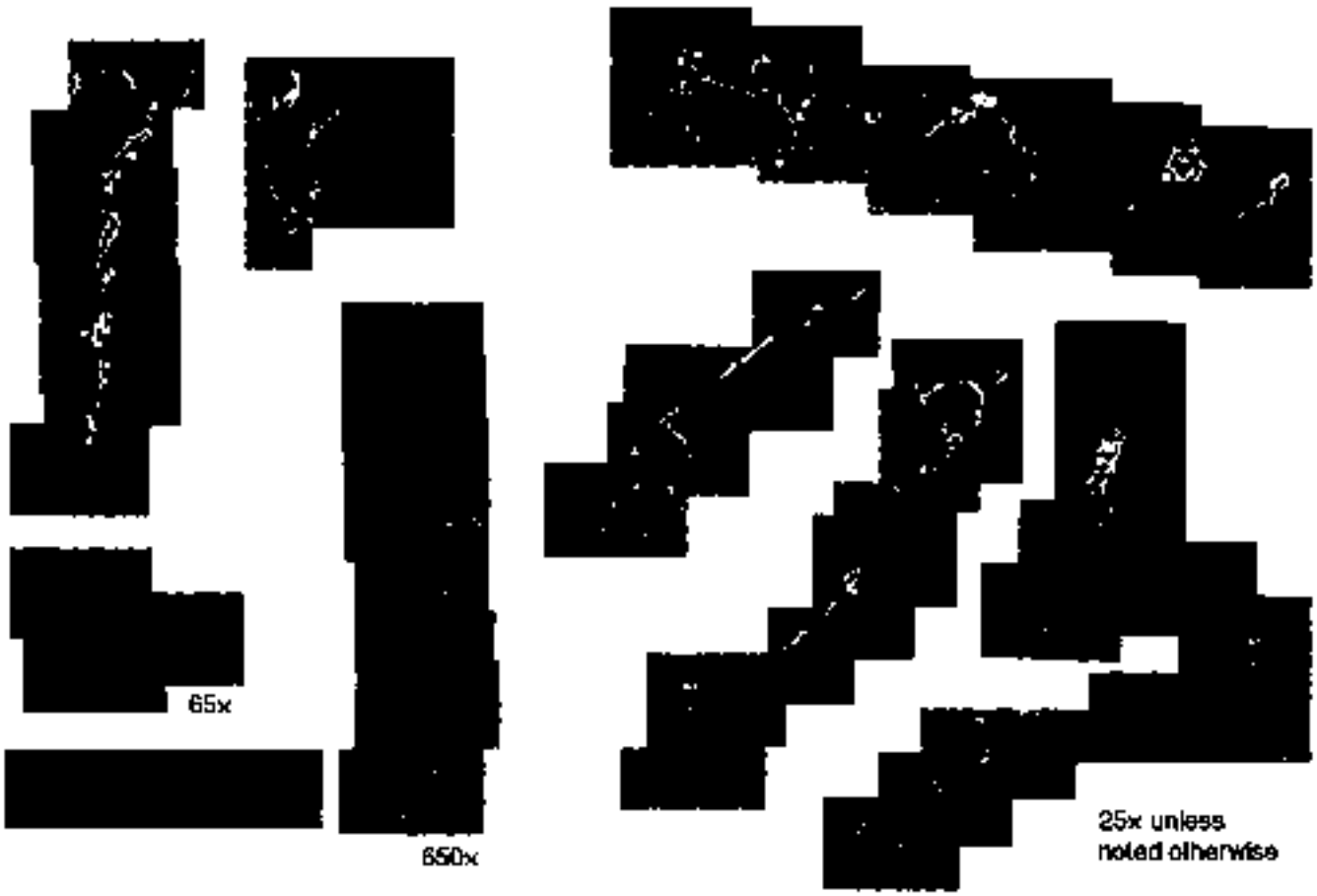


Fig. 6 Montage of the copper penetration in the Auto Train axle

Careful macroscopic examination is extremely important. Obviously, the origin or origins of failure must be determined if any conclusive results are to be obtained. When properly interpreted, fracture features describe the manner of loading, the level of the applied stress, the mechanism involved, and the relative ductility or brittleness of the material. Fracture features frequently reveal other details, such as the presence of hardened cases, apparent grain size, and internal imperfections. Presence of design-induced stress concentrators, material imperfections, fabrication imperfections, or machining imperfections that serve to locate the failure should be detected during visual examination. Techniques such as magnetic-particle inspection or dye-penetrant inspection may be required to detect or bring out tight surface defects. Ultrasonic examination or radiography may be required to detect and locate internal defects, so that they can be sectioned and examined.
Following is the usual sequence of operations in the examination of fractured components:

1. Visually survey the entire component to obtain an overall understanding of the component. Where needed, photograph all significant details.
2. Classify the failure, from a macroscopic viewpoint, as ductile, brittle, fatigue, torsion, and so on.
3. Determine the origin of failure by tracing the fracture back to its starting point or points.
4. Based on the fracture features and component configuration, estimate the manner of loading (tension, compression, bending, etc.), the relative stress level (high, medium, or low), and the stress orientation.
5. If high-magnification examination is required, use fractographic techniques to determine the fracture mode (transgranular or intergranular) and to detect other characteristic features associated with the fracture. Fractography can also be used to confirm the fracture mechanism by observation of the various dimple types, facets, striations, river marks, and so on. In a similar manner, the behavior of the material under the environmental and stress conditions can be evaluated.

Macroscopic examination of discs cut from the failed component at random locations or at failure-initiation sites can often be extremely helpful. Such examination after macroetching (per ASTM E 381) enables the investigator to evaluate or observe the following:

- Internal quality
- Hydrogen flakes
- Segregation
- Surface cases
- Hard or soft spots
- Depth of hardness penetration
- Flow lines
- Weld structure

Grinding scorch, bruises, and other surface damage can be brought out by macroetching the surface of the failed component. Special print techniques (such as sulfur, oxide, lead, and phosphorus prints) are useful for revealing the distribution of these elements in the examined section.

Not all failures involve fractures. For example, many failures occur as a result of an excessively high wear or corrosion rate. Such failures can be quite difficult to analyze, because the deficient metal is no longer present. Then, the study must be based on properties of the remaining material. However, it should be borne in mind that this material is not always representative of the missing material.

Wear and corrosion failures are difficult to analyze, because many factors, some of which may not be obvious, influence these problems. Wear failures, which can be broadly classified as adhesive or abrasive wear, are complex problems. The wear rate depends on many factors: the material used (both contact surfaces), load, temperature, sliding velocity, surface finish, and manner of lubrication, if any. Identification of the type of wear problem through visual examination is a very important first step in the failure examination.

Many corrosion failures do not involve fracture. For example, a common end result of a corrosion failure is leakage. Several different mechanisms may initiate leakage failures—for example, galvanic corrosion, crevice corrosion, pitting, and selective leaching. General corrosion may cause excessive weight loss that weakens a component until it deforms or cracks under load. Other types of corrosion mechanisms (such as intergranular corrosion, stress-corrosion cracking, corrosion fatigue, and caustic embrittlement) are generally associated with corrosion failures involving fractures. At any rate, not only careful visual examination but also other tests described later are necessary to determine the corrosion mechanism responsible for the failure.

Whenever subsequent tests are called for, visual examination remains the cornerstone of any failure examination.

Microfractography. Macroscopic fracture examination may be insufficient to properly evaluate fracture modes. Stress-corrosion cracking and fractures involving corrosion cannot be conclusively identified by macroscopic examination. Abnormalities on the fracture face may complicate macroscopic examination. Therefore, microfractographic examination of failed components is often a necessity.

Examination with the electron microscope or the SEM is, in some instances, the only effective way of defining the fracture mode and mechanism. The optical microscope does not possess sufficient depth of field to be entirely adequate for high-magnification examination of rough surfaces. The SEM permits much simpler

specimen preparation and eliminates artifacts. Furthermore, when the SEM is outfitted with energy-dispersive analysis capability, it provides the ability to analyze inclusions or corrodents observed on the fracture face.

A more recently developed tool is the scanning Auger microprobe, which is capable of analyzing surfaces on an atomic scale. This tool is very useful in analyzing surface films and in detecting impurity segregation that causes temper embrittlement.

Microfractography makes it possible to study the rate of progress of the fracture and to characterize the ductility or brittleness of the material; the fracture mechanism can usually be easily discerned. Quantitative descriptions of the fracture history can also be made. Microfractography is a very powerful tool; its application is frequently critical to the success of the failure examination.

In the case of the fractured pin in the Loran tower guy wire, microfractography was a key factor in the analysis. Only static stresses were considered in the design of the tower. The fracture revealed a well-developed fatigue fracture with classic “beach” marks. Measurement of the striations on the fatigue fracture, from initiation to final rupture, permitted an estimate to be made of the number of critical stress cycles during crack propagation (but not prior to initiation). This showed that static loading alone could not account for the failure, and an alternate source of cyclic loading—the Aeolian vibrations—was needed to explain the failure.

The analyst should realize that the lack of such sophisticated electron instruments does not mean that good failure analysis work cannot be performed. It is still possible, because good work was accomplished prior to the development of these instruments, but the investigator must be more resourceful. Fractures can be studied by light microscopy, either by direct examination of the fracture face or by the use of metallographic sections at the fracture origin. This latter technique should still be used, regardless of what electron tools are available, because it identifies possible microstructural problems associated with crack initiation and propagation.

Figure 7 illustrates the use of the light microscope to examine an intergranular fracture and compares it to SEM examination. The much greater depth of focus of the SEM is apparent, but one can still tell that the fracture is intergranular in the light micrographs. Figure 8 shows an additional example where a brittle fracture was examined using three direct approaches and three indirect approaches. All of these methods are useful and informative. It is still necessary to use replication in situations where a fracture cannot be sectioned so that it can be examined within the SEM chamber.

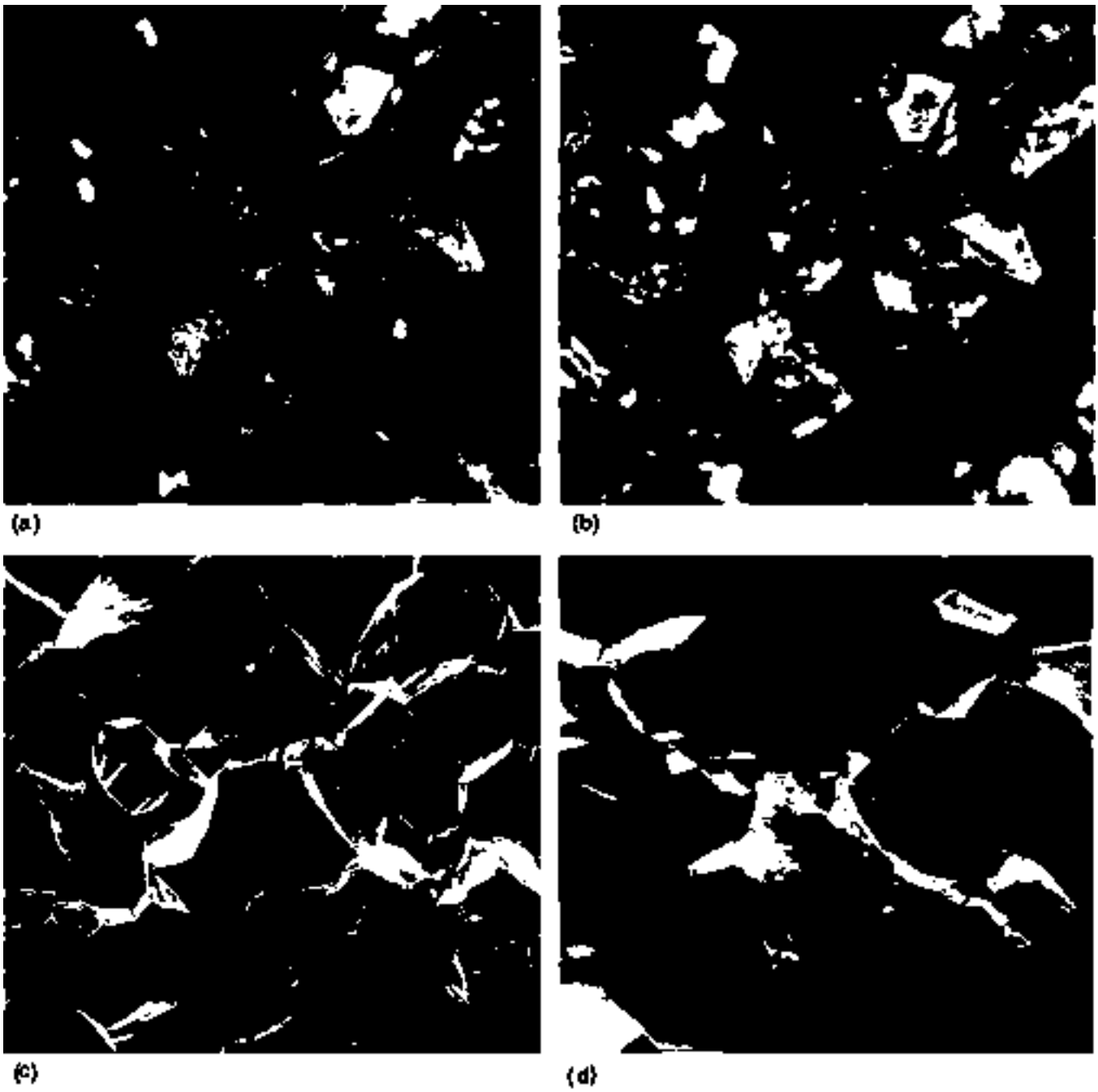


Fig. 7 Comparison of light microscope (top row) and scanning electron microscope (bottom row) fractographs showing the intergranular fracture appearance of an experimental nickel-base precipitation-hardenable alloy rising-load test specimen that was tested in pure water at 95 °C (200 °F). All shown at 50×. (a) Bright-field image. (b) Dark-field image. (c) Secondary-electron image. (d) Backscattered-electron image

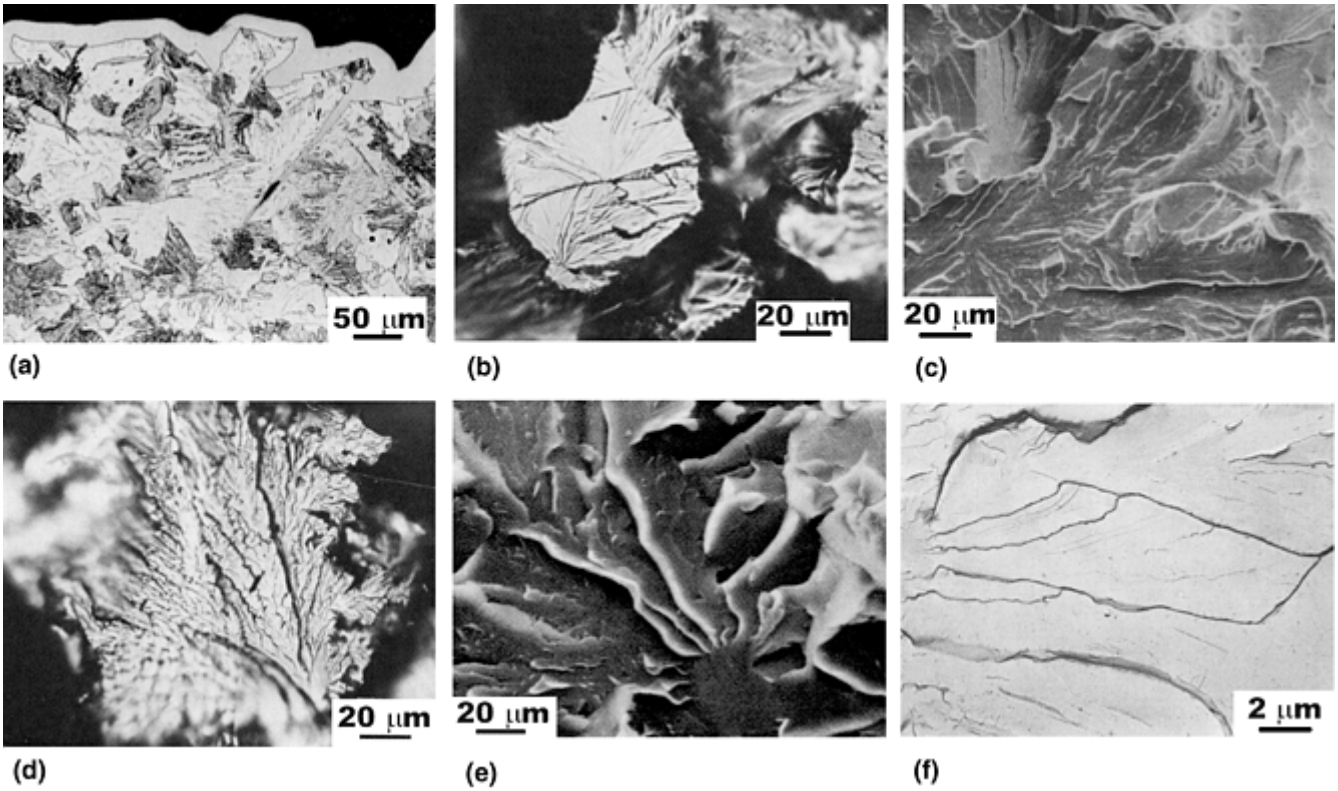


Fig. 8 Examples of three direct (a to c) and three replication (d to f) procedures for examination of a cleavage fracture in a low-carbon martensitic steel. (a) Light microscope cross section with nickel plating at top. (b) Direct light fractograph. (c) Direct scanning electron microscopy fractograph. (d) Light fractograph of replica. (e) Scanning electron microscopy fractograph of replica. (f) Transmission electron microscopy fractograph of replica

Figure 9 illustrates vertical metallographic sections through ductile and brittle fractures that were preserved by electroless nickel plating. It is clear that the cleavage cracks in the brittle specimen are crystallographic. Note the small microvoids that form around precipitates just behind the fracture path. These have sometimes been claimed erroneously to be preexisting voids.

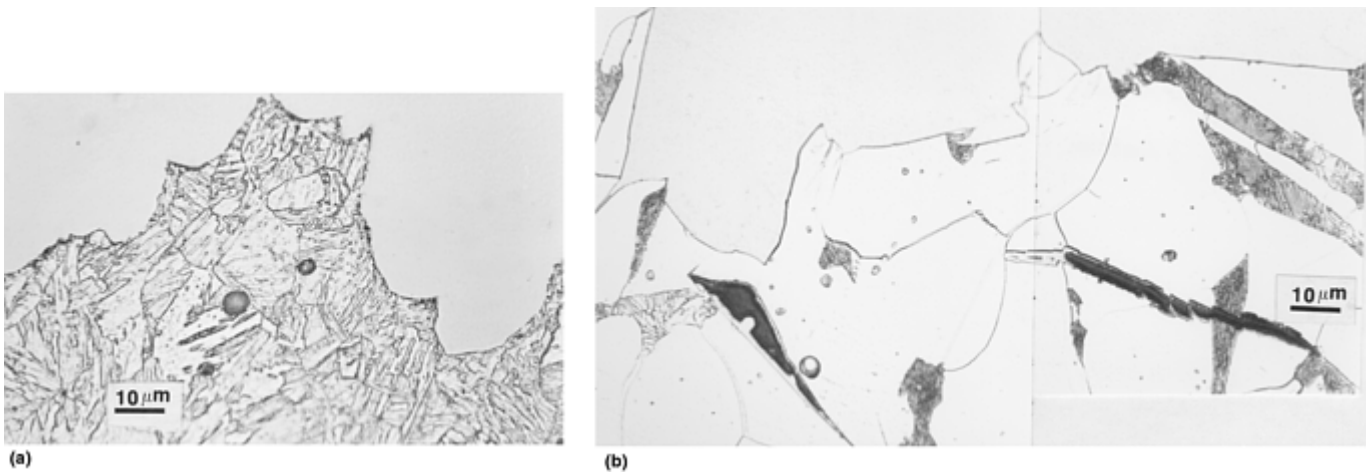


Fig. 9 Brittle (a) and ductile (b) crack paths in fractured low alloy steel specimens (both electroless nickel-plated for edge preservation and etched with 2% nital).

Chemical Analysis. A failed component should be chemically analyzed to determine whether the grade is indeed as claimed, because mixes occasionally occur at the mill, in the warehouse, or at the fabrication or manufacturing shop. A small percentage of all failures are caused by grade mixes.

In addition to analyzing the bulk composition of the material to verify the grade, the investigator may need to analyze surface contaminants by techniques such as x-ray diffraction and wet analytical procedures. The SEM and the electron microprobe may also be used. The reader is directed to *Materials Characterization*, Volume 10 of *ASM Handbook*, for more details, because a discussion of the factors governing the choice of a particular analytical technique for a given study is outside the scope of this article.

Determination of Materials Properties. Once the loads that had been applied to the failed component are known, either through estimation or from records, properties of the materials must be evaluated to determine whether the material was capable of withstanding these loads. A good case history often provides information such as whether the service load conditions were within those permitted by the materials specifications and design considerations.

In nearly all complex failure examinations, the investigator should determine the hardness and mechanical properties of the subject material and any other properties that relate to failure mechanisms, such as the toughness, endurance limit, creep strength, or stress-corrosion cracking tendency. Many test procedures have been developed to characterize the properties of materials, and the investigator should refer to appropriate materials specifications, property databases, and handbooks. The measured properties are to be compared to the specifications and to typical property data for the grade to reveal any abnormality such as might result from improper processing or from embrittlement reactions.

Metallographic Examination. Metallographic evaluation is also extremely important in determining the cause of failure. Contributing factors are frequently observable only through metallographic examination; hence, metallographic examination should always be performed.

Microscopic examination enables the investigator to evaluate the microstructural features of the failed component. The microstructure exerts a strong influence on the properties of the material and their behavior during service. Frequently, this examination can pinpoint errors made during processing, heat treatment, surface treatments, casting, and welding. Microstructural evaluation is very helpful in determining what processing methods and treatments were actually performed on the component, or whether scheduled processing steps were accidentally omitted or incorrectly performed. Microhardness testing can also be an important aid; it can be used to evaluate surface phenomena, aid in identifying microstructures, evaluate local cold working, test very thin materials, or evaluate weldments.

Electron Microscopy. The transmission electron microscope (TEM) enables the metallographer to examine fine features of the microstructure that are not resolvable with the light microscope. Thus, it is frequently used to determine pearlite interlamellar spacing or interparticle spacing, or examine the fine strengthening precipitates or dislocation substructure. The convergent-beam electron diffraction (CBED) technique can be used to identify precipitates. Interfacial relationships between the matrix and precipitate phases can also be determined. Precipitation reactions can be studied, and the extent of recovery, recrystallization, or grain growth can be evaluated. In fact, the study of such features as dislocation density or radiation damage requires the electron microscope. Microfractography can be performed using replicas, even though SEM examination is more convenient. There are times when the specimen cannot be cut, but replicas can be made and analyzed by TEM. Debris on the fracture, or corrosion deposits, can be stripped from the component by the replication method and identified using EDS or CBED methods.

Electron microprobe analysis is often required to identify unknown features observed during light microscopical examination. It is very useful in identifying complex inclusions, scale, surface layers, coatings, concentration gradients, phases, precipitates, and segregates. Minimum detectable limits are better and quantification is more precise by wavelength-dispersive analysis (WDS) than by EDS. Energy-dispersive analysis is faster, although WDS analysis has made substantial improvements in operational speed. Historically, microprobe analysis has been restricted to flat, polished specimens, and quantification is still best with flat specimens. However, curved surfaces can now be analyzed, at least semiquantitatively, using horizontal spectrometers.

X-Ray Techniques. A number of x-ray techniques are available for failure analysis. With the Debye-Scherrer camera, one can identify corrosion products and phases in refractories and minerals. The diffractometer can be applied to either qualitative or quantitative chemical analysis by indexing with the Hanawalt method or by using fluorescence analysis. Orientation of the matrix and precipitate phases can also be determined. Surface deposits can be analyzed by the reflection method or by diffraction. Chemically extracted second-phase particles can be identified by diffraction. The diffractometer is used to make residual stress measurements and to determine the amount of cold work during processes such as skin passing of sheet metal. Deformation

textures can be determined using inverse or full pole figures. With the direct-comparison method, the diffractometer can be used to determine the amount of phases present, such as retained austenite in heat treated steels or ferrite in austenitic steels.

Simulations. Occasionally, the investigator must simulate the environmental conditions encountered during service to ascertain suitability of the material to environmental conditions and to determine the effect of prior heat treatment or other processing on the service performance of the material. For example, improper heat treatment may render the material susceptible to certain types of attack. Simulation of the heat treatment as reported in the case history is valuable for both confirmation purposes as well as for further testing, particularly on a comparison basis (e.g., comparison of the toughness of the failed material as received and after various experimental heat treatments). Certain types of simulation tests require accelerated testing to obtain the desired information in a reasonable length of time. Interpretation of accelerated test data must be done with care.

When feasible, examination of similar components that did not fail in service can be most helpful in evaluating the significance of anomalies observed during the study. Simulations may also be performed on the acceptable part, and results are compared to similar test data on the failed component.

Analyzing the Data. While performing the analysis, the investigator should cross-check each of his observations against the history of the part and note any contradictions. It is frequently just such contradictions that reveal a possible cause of failure. In evaluating these factors and assessing their significance, the investigator should draw not only on his experience but also on pertinent disciplines other than his own special field.

After completion of the previously mentioned procedures, the metallurgist is ready to interpret and summarize the facts that have been gathered. The absence of certain features is often as important as the presence of these factors. Most failures are caused by more than one factor, although frequently one factor may predominate. Again, the metallurgist should make use of his own knowledge and experience as well as that of others in establishing and rating the importance of these factors. Occasionally, given the nature of the data or the theoretical state of knowledge about the particular causative mechanisms involved, the final conclusion may simply be a matter of the investigator's best judgment as to the most important factor and should be so stated in the report.

Examination of the failure and a study of the various conditions pertaining to the design, operation, and environment should raise certain questions:

- What was the sequence of the failure? The speed? The path?
- Were there one or more initiation sites?
- Did the failure initiate at or below the surface?
- Was the failure located by a stress concentrator?
- How long was the crack present?
- What was the intensity of the load?
- Was the loading static, cyclic, or intermittent?
- How were the stresses oriented?
- What was the failure mechanism?
- What was the approximate temperature at the time of failure?
- Was the temperature important?
- Was wear a factor?
- Was corrosion a factor? What type of attack?
- Was the proper material used? Is a better material required?
- Was the cross section sufficient?
- Was the material quality acceptable for the grade and specifications?
- Was the component properly heat treated?
- Was the component properly fabricated?
- Was the component properly assembled, aligned, and so on?
- Was the component repaired during service? Properly?
- Was the component properly maintained? Lubricated?
- Was the failure caused by a service abuse?
- Can the design be improved to prevent similar failures?
- Are failures likely in other similar pieces of equipment or units?
- What can be done to prevent failures in them?

In combination, the various types of investigations previously listed will, in most instances, enable the metallurgist to answer most of these questions. Because the cause or causes of failure cannot always be determined with certainty, no matter what the investigator tries, he should state what he considers to be the most probable causes or factors, together with the other legitimate possibilities. In any event, it should be made clear which findings are based on demonstrated facts and which conclusions are of a more conjectural nature.

Preparing the Report. The report analyzing the failure should be written in a clear, concise, logical manner. It should be clearly structured, with sections covering the following:

- Description of the failed item
- Conditions at the time of failure
- Background history important to the failure
- Mechanical and metallurgical study of the failure
- Evaluation of the quality of the material
- Discussion of any anomalies
- Discussion of the mechanism or possible mechanisms that caused the failure
- Recommendations for the prevention of future failures or for action to be taken with similar pieces of equipment

Extraneous data should be omitted, and, depending on the nature of the problem and the data, not every report needs full treatments for every one of the sections listed previously. Many times, the readership may include purchasing, operating, or accounting personnel who are not technically trained. If this is the situation, the report should be written so that it is comprehensible to these persons. At least, those sections of the report that bear on their decision-making or information needs should be written in language that is accessible to them. Frequently, a cover letter summarizing the most important findings and the suggested action is a good vehicle for reaching top executives who are not as interested in the technical specifics but need key findings and recommendations as a basis for decision making.

Follow-up on the recommendations is frequently a difficult task but should be undertaken for the more critical failures. Cooperation between the investigator, the designer, the manufacturer, and the user is critical in developing good, workable changes.

Preservation of Evidence. Due to the rise in the number of litigations and the strict implementation of court laws, it is a necessity to properly preserve evidence. In the past, it has been standard practice for an injured party to hire an expert to perform a failure study to determine if there are grounds for a lawsuit. If so, the experts for those sued may obtain, with the permission of the court, the specimens involved in the suit for their examination. This examination is always much more restrictive than that of the plaintiff's expert, who basically had free rein to do what he pleased. The defendant's experts must get permission even to repolish a cross section that has degraded since it was prepared. The defendant's experts can be quite hampered by the work done by the plaintiff's expert. Anyone involved in failure analysis has a catalog of horror stories regarding preservation of evidence.

The author has seen all extremes in specimen preservation. In the Loran tower failure, the U.S. Coast Guard waited 13 years before it brought suit. However, their expert did a fine job of preserving the evidence so that the fatigue fracture surface could be replicated and striations measured 13 years after he examined it. On the other hand, in the train wreck, the two specimens that the plaintiff's expert examined with the SEM were lost and not available for examination by the defendant's analysts. Further, the remaining 57 pieces of the broken wheel set (the axle broke into two pieces and the plaintiff's expert cut up the axle, producing 59 separate pieces) were stored outside in the moist Florida climate and were covered by a heavy layer of rust when the analysts obtained them. The plaintiff's expert had only a crude sketch showing how the 59 pieces related to the original two axle portions. After cleaning off the rust, the analysts were allowed to reprepare the original polished surfaces of six specimens. Fortunately, that was more than enough to reveal the copper grain-boundary penetration in the specimens.

Clearly, the expert for the railroad did a very poor job and under today's court laws would be liable for a counter suit (Ref 3). This type of mishandling of material in a litigation is grounds for dismissal of a case, because it could make it impossible for the defendant to prepare a defense. Fortunately, the analysts could work around these problems, but this may not always be the case.

This brings up a second issue that has emerged with the rise in product liability suits: the chain of evidence. It has been seen how this is important in criminal cases where the police sometimes have lost evidence, and then it cannot be admitted into the case. The same situation now applies to product liability cases. The investigator must fully document every step of the examination and keep records of everything done to the parts. The old freewheeling days of yesteryear are gone. With the creation of ASTM standards (Ref 3) covering the handling of evidence and the examination of accident scenes, one must faithfully document every step taken. To do so, the analyst must rely heavily on photography (digital imaging is being used more and more) and record keeping.

The ideal failure investigation is one where all parties have the opportunity to witness the examination and contribute to its planning and execution. If such a practice were used exclusively, the product liability caseload would drop dramatically. Further, everyone would benefit, because the actual cause of the failure would be determined more frequently, and its recognition would have more influence on preventing future failures.

Follow-Up on Recommendations. Most analysts feel that the work has been completed when the report has been written, that it is the job of other people to carry out the recommendations and determine their value. Certainly, other individuals will be involved in this work, but the analyst should check periodically to determine if the recommendations were implemented and if they were successful. If they were not implemented, the analyst should determine why they were not. There may be good reasons why they were not implemented, but a suitable alternative solution should be developed.

The author has been involved in several cases where follow-up was necessary. The first involved precipitator wires in a wet scrubbing system at a basic oxygen furnace (BOF) shop. There were six sets of wet scrubbing systems, and each had four zones. The data on the wire failure frequency are presented in Table 1. Basically, wires were hung from porcelain insulators between plates, and a charge was placed across plates and wires. The first zone of each scrubber used barbed wire, and the other three zones used cold-drawn 1008 carbon steel wire. Each was about 3 mm (0.12 in.) in diameter by about 5 m (16 ft) in length and was held taunt with a 7 kg (15 lb) bottle-shaped weight on the bottom end. The air from the BOF vessels was cooled with water and directed through the scrubbers to remove the particulate matter, which collected on the plates. Most of this debris would fall to the floor beneath the plates and be removed, but some adhered to the plates. So, the assembly was shaken periodically.

Table 1 Number of wire failures in Bethlehem basic oxygen furnace precipitators

Stage No.	Number of failures at indicated precipitator[(a)]						
	Current, mA	11	12	21	22	31	32
1	750	3	0	3	1	1	1
2	1000	16	4	10	8	10	11
3	1500	33	24	15	15	45 (27)	22
4	1500	66 (8)	15	3	9	8	3

(a) Numbers in parentheses are the number of stainless steel wires that have failed in these two selections since they were rewired late in September 1973. All other numbers refer to carbon steel wire failures. Barbed wire is used in the first stage of each precipitator (greater clearance is available).

After a year of operation, a high frequency of wire breakage was observed in two of the 18 zones strung with 1008 wire. The breakage occurred at the top of each wire where it was bent around the insulator. The short free end was held fast against the long end by a ferrule of type 430 stainless steel to form the loop that was placed on the insulator. Figure 10 shows the initial design of the loop on the top end of the precipitator wires. On the left are two loops, one with the 430 stainless steel ferrule removed. On the right of the figure is the broken wire inside the ferrule on the bottom side. Breakage occurred on the long end (5 m, or 16 ft, long with the bottle weight at the bottom) where it emerged from the ferrule.

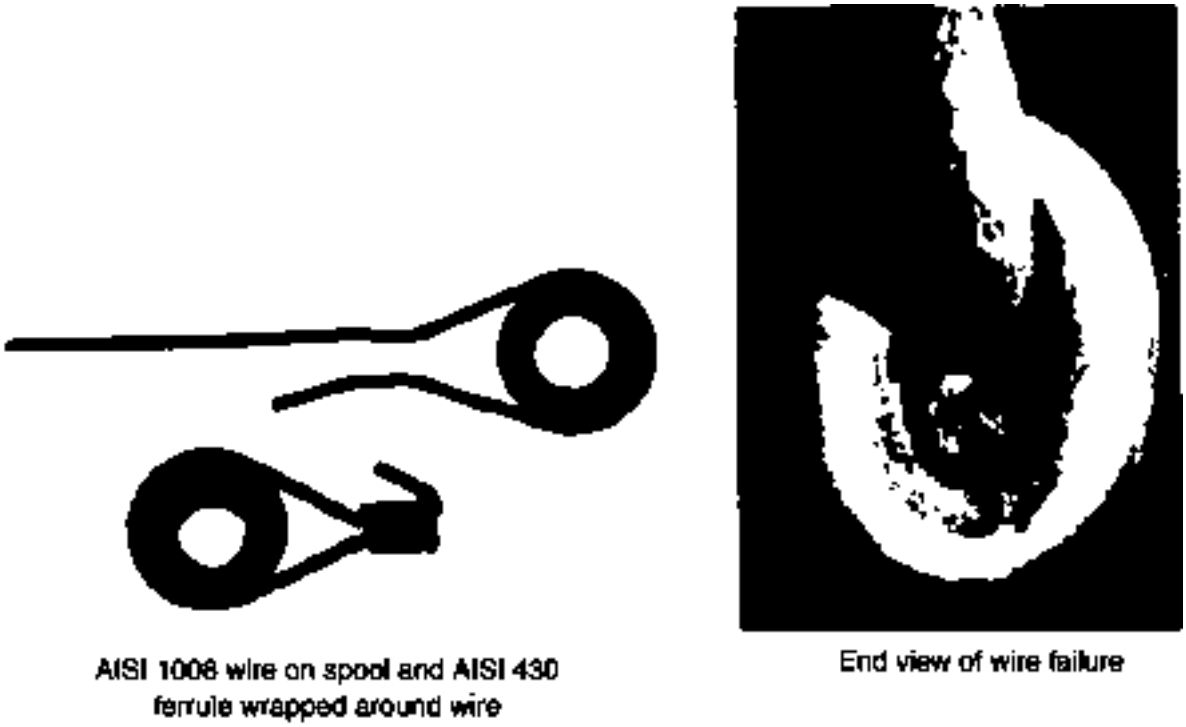


Fig. 10 Initial design of the loop on the top end of the precipitator wires. On the left are two loops, one with the 430 stainless steel ferrule removed. On the right is the broken wire inside the ferrule. 9×

The plant metallurgist diagnosed the failure as corrosion fatigue and recommended that the two zones with high wire breakage be restrung with type 304 austenitic stainless steel wires of the same diameter. Because the carbon steel wire was cold drawn, the austenitic stainless steel wire was ordered cold drawn. The method of attaching the wire around the insulator was changed. A tube of 304 stainless steel with an inner diameter slightly greater than the wire, measuring about 610 mm (24 in.) long, was bent to form a loop at one end that would fit over the insulator. The wire was placed into the hollow tube at the opposite end, and the tube was crimped in two places to hold the wire inside the tube. Figure 11 shows the original design of the loop (upper left), the new design with 304 stainless steel wire protruding from the tube (upper right), and an example of the new design: a 304 stainless steel tube with 304 stainless steel wire protruding (bottom).

Fig. 11 The picture in the upper left shows the original design of the loop. On the right is the new design, with 304 stainless steel wire protruding from the tube. On the bottom is

an example of the new design: a 304 stainless steel tube with 304 stainless steel wire protruding.

These new 304 stainless steel wires began to fail one week after installation. An example of the failed new design is shown in Fig. 12. A more-expensive material was substituted for a less-expensive material, and the time to first failure went from about seven months to seven days. Clearly, something was missed in making this recommendation. The metallurgist had not realized that austenitic grades, such as 304, are susceptible to chloride-ion stress-corrosion cracking. The operating conditions for the precipitator wires were perfect for stress-corrosion cracking. The water used to cool the hot BOF gases came from the nearby river and had been treated with chlorine to prevent algae formation as it went through the piping system. The wires were heated to about 150 °C (300 °F) by the BOF gases, and residual stress was present at the notch at the end of the hanger.

Fig. 12 An example of the failed new design for the precipitator wires. 5.5×

The author recommended using ferritic stainless steel, which is immune to chloride-ion stress-corrosion cracking, and also changed the wire diameter slightly, because its natural frequency was a multiple of the vibration frequency. No more failures of this type were observed over the next 20 years (until the shop was closed).

A colleague had a failure study of gas turbine blades that were failing by corrosion-assisted high-cycle fatigue. There were three gas turbines (one as a backup) used to drive air into each of two blast furnaces. A blade failure would cause substantial damage to the blades in the stages behind it. If two turbines were down for repair, one blast furnace could not be used, which would produce an enormous daily loss in steelmaking capacity and revenue. The fatigue failures started at pits caused by corrosion. The nearby coke ovens affected the air quality going through the turbines, causing the pitting. Stress analysis revealed that the blades were stressed in every cycle, and that the turbines operated above 3000 rpm.

First, the colleague optimized the heat treatment of the type 410 stainless steel used for the blades. This improved their life but did not solve the problem. Next, he chose a coating system that provided protection from the atmosphere; however, the coating had a finite life. So, an inspection procedure was implemented where the blades were inspected at regular intervals and recoated as needed. This prevented future failures, without major cost or redesign, but it did require regular follow-up to maintain the coatings and prevent subsequent failures.

References cited in this section

2. G.F. Vander Voort, in *Failure Analysis and Prevention*, Vol 11, *Metals Handbook*, 9th ed., American Society for Metals, Metals Park, OH, 1986, p 715–727

3. R. Dean, *Advanced Materials & Processes,* Vol 158(No. 3), 2000, p 42–45

Conclusions

Performing failure analyses is an exceptionally challenging and exciting job. The investigator must have a sound background in the areas that are central to his testing and analytical work and must also be knowledgeable enough to know when to call on other fields of specialization that interface with his work. New problems are always encountered that tax the ingenuity of the investigator, and they will sharpen and expand his skills, if he is willing to go on learning. By the same token, analysis of failures by many investigators in the ferrous and nonferrous metal industries here and abroad has, over the years, expanded the practical knowledge for processing and product developments and also contributed to basic scientific insights.

Those who frequently conduct failure examinations agree on several main points. Most failures could easily be prevented if certain well-documented precautions are taken. Many failures occur as a result of fatigue, and one or more of the following may play a decisive role: original design deficiencies, material inadequacies, machining or fabrication imperfections, abuse or neglect, improper maintenance, improper repairs, and metallurgical factors such as decarburization. Heat treatment irregularities are another major contributing factor in failures.

Conducting a Failure Examination

George F. Vander Voort, Buehler Ltd.

References

1. G.F. Vander Voort, *ASM Metals Engineering Quarterly,* Vol 15(May), 1975, p 31–36; Adapted version published in *Practical Failure Analysis,* April 2001

2. G.F. Vander Voort, in *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, Metals Park, OH, 1986, p 715–727

3. R. Dean, *Advanced Materials & Processes,* Vol 158(No. 3), 2000, p 42–45

Determination and Classification of Damage

Aaron Tanzer, Siemens Westinghouse

Introduction

THE PRINCIPAL TASK of a failure analyst during a physical cause investigation is to identify the sequence of events involved in the failure. Technical skills and tools are required for such identification, but the analyst also needs a mental organizational framework that helps evaluate the significance of observations. Like the basic process of the scientific method, failure analysis is an iterative process of narrowing down the possible explanations for failure by eliminating those explanations that do not fit the observations. The basic steps are:

1. Collect data
2. Identify damage modes present
3. Identify possible damage mechanisms
4. Test to identify actual mechanisms that occurred
5. Identify which mechanism is primary and which is/are secondary
6. Identify possible root causes
7. Test to determine actual root cause
8. Evaluate and implement corrective actions

Generally, a failure analyst will start with a broad range of possible explanations but over time will narrow and refine the existing possibilities. Often, a likely theory develops during the course of the investigation. This can be helpful, but only if the investigator does not let the theory influence his or her objectivity. Instead, such a theory has to be tested against the facts of the investigation like any other possibility.
The failure analyst must repeatedly ask the following questions as an investigation develops possible explanation(s) for actual events:

- What characteristics are present in the failed/damaged component?
- What characteristics are present or expected in an undamaged component?
- What are the possible explanations that would account for the differences between damaged and undamaged components?
- What test(s) can be performed to confirm or eliminate possible explanations and refine knowledge about the observed damage?

Asking questions, much like the Socratic method in philosophy, is an essential part of failure analysis, because a recipe approach (i.e., following a prescribed set of procedures and examples in a step-by-step fashion) often is not sufficient. A recipe (even the prescribed recipe of repeated questioning) does not necessarily answer the questions, “Why did this happen?” and “How did this happen?” The key is understanding the meaning of the observations and the hows and whys of their existence and role in a failure.
The investigator needs to understand the potential ways a component could be damaged, the clues that would differentiate between these different scenarios, and the physical meaning each of these clues would have. Comparison of observations with characteristics of expected damage and mechanisms will enable the analyst to narrow down the possible failure explanations and understand the meaning of the observations made.
Limiting conditions that refine the scope of explanations for observed damage can be defined by using the following two rules of thumb:

- *The Sherlock Holmes Rule:* When you have eliminated the impossible, whatever remains, however improbable, must be the truth (Ref 1).
- *Occam's Razor:* When two or more explanations exist for a sequence of events, the simple explanation will more likely be the correct one (Ref 2).

The Sherlock Holmes Rule is useful in keeping open possible explanations for failure that might otherwise get eliminated because they are improbable. Conversely, work in failure investigations is often performed to prove theories already found to be impossible, resulting in misdirected resources that could have been applied to testing valid explanations. Occam's Razor is a useful rule because the analyst often determines two or more possible explanations for the events that occurred. Sometimes there is no available way to verify the correct explanation, and other times it is impractical to apply the necessary resources for such a test. Occam's Razor applies order by selecting the simplest explanation for what is being observed.

References cited in this section

1. A.C. Doyle, *The Sign of Four*, 1889

2. W.M. Torburn, The Myth of Occam's Razor, *Mind*, Vol 27, 1918, p 345–353

Determination and Classification of Damage

Aaron Tanzer, Siemens Westinghouse

Characterization and Identification of Damage and Damage Mechanisms

The ultimate goals in failure investigations are to determine root cause(s) of the failure and corrective actions that could be taken to prevent future failures. Classifying and identifying the damage observed in the failure is required to achieve these goals.

Two of the critical goals in a failure investigation are to identify both the damage mechanism(s) and damage mode(s) that are present. (The terms *damage mode* and *failure mode* can be used interchangeably, as can *damage mechanism* and *failure mechanism*.) Damage mechanisms are a key component in categorizing damage and failures. The definitions are:

damage mode.

> Description of the physical characteristics of damage observed. For example, intergranular fracture, buckling, transgranular beachmarks, and pits can all be thought of as damage modes. Damage mode describes *what* damage is present.

damage mechanism.

> The specific series of events that describe both how the damage was incurred and the resulting consequences. Examples of damage mechanisms include high-temperature creep, hydrogen embrittlement, stress-corrosion cracking, and sulfidation. Damage mechanism describes *how* damage came to be present.

Much confusion has occurred because of the tendency of engineers to use the terms *mechanism* and *mode* interchangeably; in doing so, it is unclear that two distinct characteristics need to be assessed. Sometimes this occurs because, within a given system, the same wording is used to describe both the failure mode and mechanism. For example, pitting describes a damage mode because the surface of a material is pitted. In certain systems, pitting is also a possible damage mechanism. In boiler tubing, for example, a pitting damage mechanism describes a specific localized corrosion mechanism where pits form through dissolution of metal either from low pH or high oxygen conditions. The metal under the pit surfaces is unaffected (Ref 3). In this system pitting is a specific damage mechanism, but many other damage mechanisms also result in a pitting damage mode in boiler tubing, including hydrogen damage, phosphate corrosion, and caustic gouging.

It is helpful to be as specific as possible in differentiating damage mechanisms in a system. For example, fatigue is often identified as both a damage mode and a damage mechanism. A fatigue damage mode is the observable damage that occurs under fatigue loading cycles (e.g., the presence of beachmarks). Classifying fatigue as a damage mechanism is not necessarily complete because it does not point to the specific environment that results in a fatigue damage mode. Instead, specific mechanisms that can result in a fatigue damage mode have to be examined. Examples include corrosion fatigue, thermomechanical fatigue, creep-fatigue interaction, and mechanical fatigue.

Determination of damage mechanisms starts by characterizing the component(s) being examined. It is impossible to know what is different about a failure without first understanding what is expected from unfailed components. In general, the analyst should get as much information as possible about a part and its background during the course of an investigation. Some key questions worth evaluating include:

- What was the part supposed to do? How was it supposed to work?
- How was the part made? What processes were involved in its manufacture (e.g., forming, joining, and heat treatment)? What properties were expected at the time of manufacture?
- What were the specified dimensions and tolerances for the as-manufactured part?
- How was the part installed?
- To what service environment(s) was the part exposed? Typical environments to examine include operating temperatures, stresses (steady state or slowly rising and cyclic), oxidizing/corrosive environments, and wear environments. What properties were required during service? How were properties expected to change from service exposure?
- How was the part inspected during service intervals? What information was found during these inspections?
- What material characteristics were specified for the part (e.g., composition, strength, hardness, impact, and stress-rupture properties)? What specifications, industry standards, and contracts govern these properties?
- What were the various ways the part could fail?

The last item is a key question to repeatedly ask throughout a failure investigation. The list of various damage mechanisms by which a part can fail can be narrowed down through application of the Sherlock Holmes Rule because the characteristics observed will be inconsistent with the characteristics of some damage mechanisms.
Consider, for example, cracking observed in the trailing edges of several service-run gas turbine blades made from cast IN 738 nickel-base superalloy (Fig. 1). When one of the cracks was broken open and examined in a scanning electron microscope (SEM), fractographic examination revealed an intergranular crack surface at the trailing edge (Fig. 2). Comparison with an atlas of fracture features taken from earlier mechanical test specimens eliminated high-cycle fatigue (HCF) as the failure mechanism, because HCF cracking exhibits transgranular fracture features. The fracture features appeared to be consistent with both creep and low-cycle fatigue (LCF) failure mechanisms. Metallographic sectioning through the fracture at this location, however, exhibited none of the features associated with high-temperature creep damage, such as fissures and void coalescence at grain boundaries (Fig. 3). This eliminated high-temperature creep as a potential failure mechanism. The examination, through the process of elimination, determined crack initiation had occurred by LCF.

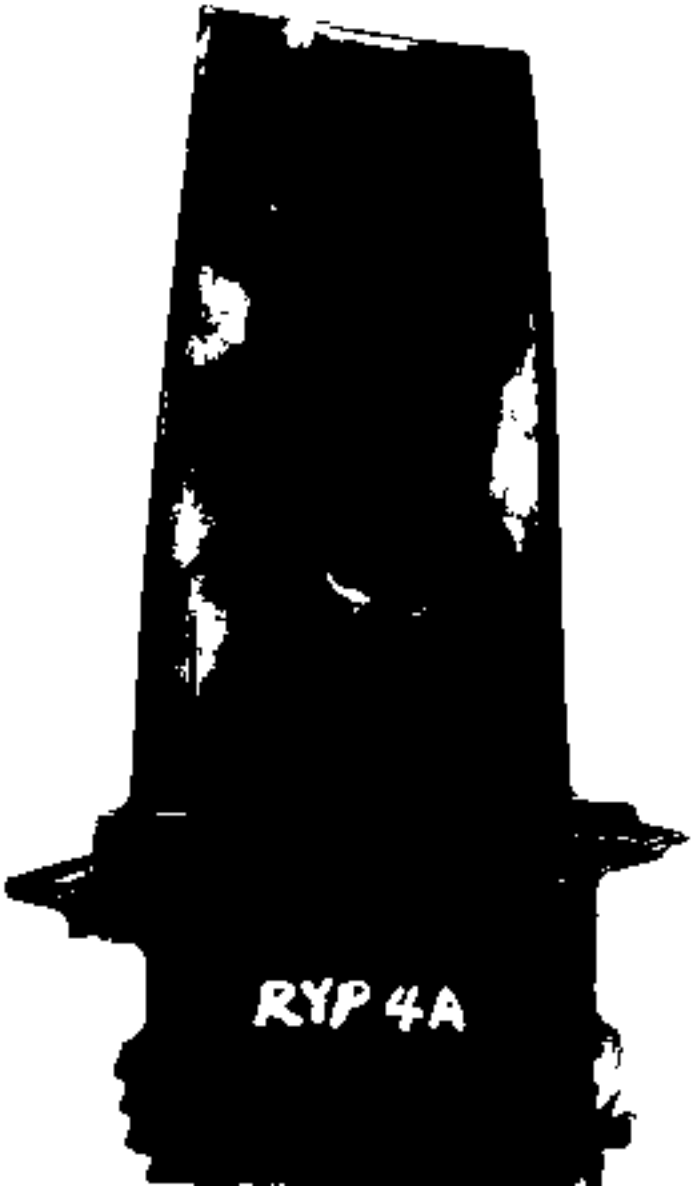


Fig. 1 Crack (arrow) observed in trailing edge near platform of service-run gas turbine blade

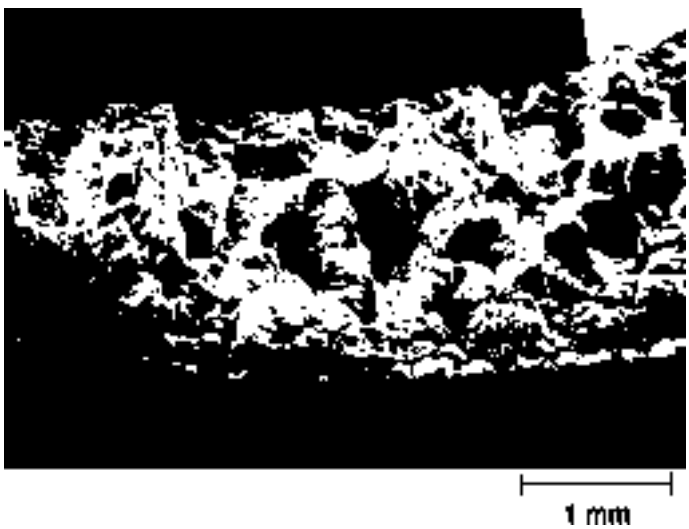


Fig. 2 Resulting fracture surface when gas turbine blade trailing edge crack is broken open in laboratory.

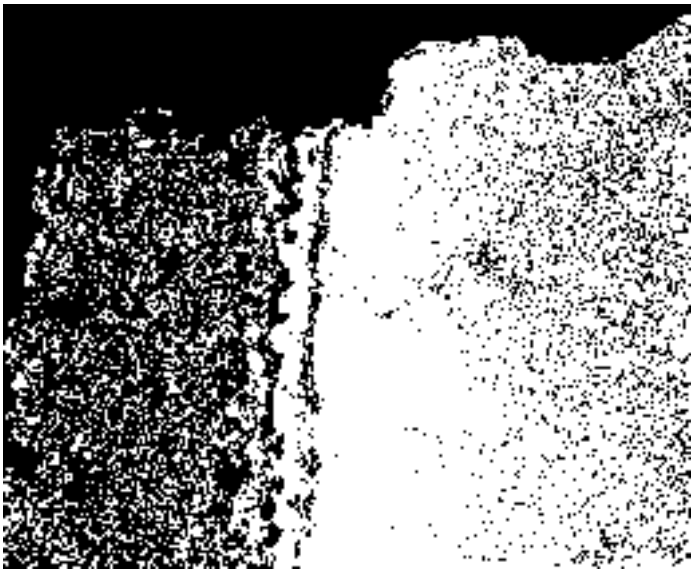


Fig. 3 Metallographic cross section through fracture surface shown in Fig. 2. Lack of fissures and voids indicates cracking was not caused by a creep damage mechanism. 15×

The relationship between specified materials properties, materials composition and microstructure, and the actual failure is frequently misunderstood. Sometimes, a part will be found to be out-of-spec in some way that tempts the analyst to attribute the failure to that deviation. This situation is commonly found in boiler tube failure investigations. Occasionally, composition of one specified element in a boiler tube deviates slightly from requirements. Yet, such minor compositional deviations have never been found to cause, or even contribute to, the observed failures. (Composition, however, may be indicative of root cause of failure when the tube is found to be made from a completely different alloy than that specified.)

Another way in which materials properties evaluation results are misapplied concerns a tendency to assume a problem is not material related because the material meets specified properties. Critical properties such as fracture toughness or corrosion resistance may never have been specified in a purchase agreement between vendor and buyer. While subsequent failure may have occurred because these properties were deficient in relation to the applied installation and/or service environments, the material was nevertheless considered in-spec. The failure analyst must, in such instances, evaluate whether the material as-specified was lacking in necessary properties and aid in application of corrective actions.

It is helpful to compare characteristics of the damaged component with those of unfailed components. If the analyst is unfamiliar with typical new and service-run components, then evaluation may include examination of such parts to isolate characteristics related to damage. However, unfailed components often are not available for examination. Review of processing, installation, and service environments can aid in understanding what properties should be typical. It is also very useful to examine the failed component in undamaged areas. For example, turbine blade roots in steam turbines are not exposed to service environment steam and temperature as airfoils are. Comparison of properties between the root and airfoil can provide useful information about service degradation of the exposed airfoils and identify damage characteristics.

One of the most important clues in identifying damage mechanisms lies in characterizing damage modes. Table 1 shows identifying characteristics for classifying six fracture and crack modes. As with other characterization techniques, identification of fracture damage mode helps narrow down the list of possible failure mechanisms and rank the possibility of those remaining. Components susceptible to progressive failure modes can often be identified and replaced or repaired prior to failure through preventative maintenance programs. For example, critical steam piping systems are now inspected at regular intervals for creep damage in the heat-affected zones of longitudinal seam welds using replication to look for void formation and cracks at grain boundaries, an early indication of impending creep failure. General observation, nondestructive evaluation (NDE) techniques (such as penetrant testing, boroscopic examination, ultrasonic testing, and radiographic techniques), and field metallography (including replication and sample removal) can identify if characteristics associated with failures are present. These techniques are discussed in detail in the Section “Tools and Techniques in Failure Analysis” in this Volume.

Table 1 Damage mode identification chart

Method	**Instantaneous failure mode**[a]		Progressive failure mode[b]			
	Ductile overload	**Brittle overload**	**Fatigue**	**Corrosion**	**Wear**	Creep
Visual, 1 to 50× (fracture surface)	• Necking or distortion in direction consistent with applied loads • Dull, fibrous fracture • Shear lips	• Little or no distortion • Flat fracture • Bright or coarse texture, crystalline, grainy • Rays or chevrons point to origin	• Flat progressive zone with beach marks • Overload zone consistent with applied loading direction • Ratchet marks where origins join	• General wastage, roughening, pitting, or trenching • Stress-corrosion and hydrogen damage may create multiple cracks that appear brittle.	• Gouging, abrasion, polishing, or erosion • Galling or storing in direction of motion • Roughened areas with compacted powdered debris (fretting) • Smooth gradual transitions in wastage	• **Multiple brittle appearing fissures** • **External surface and internal fissures contain reaction scale coatings.** • **Fracture after limited dimensional change**
Scanning electron microscopy, 20 to 10,000× (fracture surface)	• Microvoids (dimples) elongated in direction of loading • Single crack with no branching • Surface slip band emergence	• Cleavage or intergranular fracture • Origin area may contain an imperfection or stress concentrator.	• Progressive zone: worn appearance, flat, may show striations at magnifications above 500× • Overload zone: may be either ductile or brittle	• Path of penetration may be irregular, intergranular, or a selective phase attacked. • EDS[c] may help identify corrodent.	• Wear debris and/or abrasive can be characterized as to morphology and composition. • Rolling contact fatigue appears like wear in	• **Multiple intergranular fissures covered with reaction scale** • **Grain faces may show porosity**

					early stages.	
Metallographic inspection, 50 to 1000× (cross section)	• Grain distortion and flow near fracture • Irregular, transgranular fracture	• Little distortion evident • Intergranular or transgranular • May relate to notches at surface or brittle phases internally	• Progressive zone: usually transgranular with little apparent distortion • Overload zone: may be either ductile or brittle	• General or localized surface attack (pitting, cracking) • Selective phase attack • Thickness and morphology of corrosion scales	• May show localized distortion at surface consistent with direction of motion • Identify embedded particles	• **Microstructural change typical of overheating** • **Multiple intergranular cracks** • **Voids formed on grain boundaries or wedge-shaped cracks at grain triple points** • **Reaction scales or internal precipitation** • **Some cold flow in last stages of failure**
Contributing factors	• **Load exceeded the strength of the part.** • **Check for proper alloy and processing by hardness check or destructive testing, chemical analysis.** • **Loading direction may show**	• **Load exceeded the dynamic strength of the part.** • **Check for proper alloy and processing as well as proper toughness, grain size.** • **Loading direction may show failure was**	• **Cyclic stress exceeded the endurance limit of the material.** • **Check for proper strength, surface finish, assembly, and operation.** • **Prior damage by**	• **Attack morphology and alloy type must be evaluated.** • **Severity of exposure conditions may be excessive; check: pH, temperature, flow rate, dissolved oxidants,**	• **For gouging or abrasive wear: check source of abrasives.** • **Evaluate effectiveness of lubricants.** • **Seals of filters may have failed.** • **Fretting induced by slight**	• **Mild overheating and/or mild overstressing at elevated temperature** • **Unstable microstructures and small grain size increase creep rates.** • **Ruptures occur after long exposure times.** • **Verify proper alloy.**

	failure was secondary. • **Short-term, high-temperature, high-stress rupture has ductile appearance (see creep).**	**secondary or impact-induced.** • **Low temperatures**	**mechanical or corrosion modes may have initiated cracking.** • **Alignment, vibration, balance** • **High cycle low stress: large fatigue zone; low cycle high stress: small fatigue zone**	**electrical current, metal coupling, aggressive agents.** • **Check bulk composition and contaminants.**	**looseness in clamped joints subject to vibration** • **Bearing or materials engineering design may reduce or eliminate problem.** • **Water contamination** • **High velocities or uneven flow distribution, cavitation**	

(a) Failure at the time of load application without prior weakening.
(b) Failure after a period of time where the strength has degraded due to the formation of cracks, internal defects, or wastage.
(c) EDS, energy-dispersive spectroscopy.
Characteristics found on a part exhibiting a specific damage mode need to be identified to determine the damage mechanism. For example, a part may have a confirmed damage mode of fatigue. Other questions then have to be asked to characterize that fatigue: Was this high-cycle or low-cycle fatigue? Was the fatigue corrosion assisted? Was there evidence of differential thermal expansion? What cyclic loading is present that could have induced the observed fatigue damage? Did cracking only propagate by fatigue, or could fatigue be confirmed as the initiation mechanism? Evaluating questions such as these leads to specific identification of the failure mechanism and possible root causes.
It is also worth noting that not all features listed in Table 1 will always be present in a particular case for a given failure mode. Some examples are:

- Pearlitic and martensitic steel components that failed under fatigue seldom exhibit striations in SEM.
- Ratchet marks are only seen when multiple crack origins common in cyclic loading are present.
- Evidence of fracture mode is often obliterated by subsequent rubbing, erosion, or impact after cracking has initiated.

As with all other aspects of an investigation, the failure analyst should not assume the properties listed in Table 1 apply 100% of the time. Instead, understanding of *why* the characteristics listed indicate the given failure mode is key to understanding the failure itself.

Reference cited in this section

3. S.R. Paterson, T.A. Kuntz, R.S. Moser, and H. Vaillancourt, *Boiler Tube Failure Metallurgical Guide,* TR-102433, Electric Power Research Institute, 1993

Determination and Classification of Damage

Aaron Tanzer, Siemens Westinghouse

Determining Primary and Secondary Damage Mechanisms

In real world failure analyses, multiple modes and mechanisms may be present in a damaged component. A thorough investigation is critical to ensure that all damage is found and documented. Critical evidence can be missed if the analyst stops after finding a single mode. Next, it is necessary to determine which mechanism is primary and which is/are secondary. The primary mechanism is the one responsible for the failure (i.e., if the primary failure mechanism did not occur, then failure would not have occurred). Secondary failure mechanisms can be divided into three categories:

- Mechanisms that were induced because of the presence of the primary failure mechanism or came from the same root cause as the primary mechanism
- Independent mechanisms that contributed to failure. For example, such mechanisms may have contributed by accelerating the time or severity of failure.
- Mechanisms that were present but unrelated to the failure

To keep proper perspective in categorizing damage, only one primary mechanism should be identified for any particular failure, although multiple secondary mechanisms can be present. Misdiagnosis of root cause and consequent corrective actions often occur because the failure analyst has not taken the step of identifying the primary failure mechanism. Even when the primary and secondary mechanisms have been identified correctly,

the consequences of fixing only the primary mechanism need to be examined to determine if future failures could occur from the secondary mechanism(s).

This leads to a central theme of this Volume: Understanding the way in which the component functions within the overall system is critical. The analyst cannot differentiate primary and secondary causes without this understanding. Thorough materials characterization is usually the first step, but is certainly not the only step, in a failure investigation.

Example 1: Rupture of a Power Plant Waterwall Tube. A coal-fired power plant experienced a forced outage when a front waterwall tube ruptured. Examination of the fishmouth rupture confirmed a failure mechanism of short-term overheating, which occurs when either water or steam flow is inadequate in high-temperature tubing, or when tubing is exposed to excessive local combustion gas temperatures. Short-term overheating is usually classified as a secondary mechanism in tubing because flow was inhibited from an upstream source. This was found to be true in this example, where further investigation located a pinhole leak in the tube upstream from the rupture at the end of a web weld to an adjoining tube and a leak in that adjoining tube. Evaluation determined the upstream leak was the location of the primary damage mechanism, which was identified as thermomechanical fatigue (TMF) that initiated at the outside diameter surface at the web of the weld. Two secondary failure mechanisms were established: steam erosion into the next tube and downstream short-term overheat rupture. Corrective actions thus concerned eliminating TMF and did not focus on overheat, despite the problem first exhibiting itself in the form of this mechanism.

The previous example also illustrates another general rule of thumb regarding identifying where initiation associated with the primary damage mechanism first occurs: In systems involving flow and multiple failures, the initial failure associated with the primary failure mechanism will generally be the one that is farthest upstream. (Another similar rule of thumb involves initiation of a crack that propagated through a wall: Crack initiation will have occurred on the side of the wall where the crack is longest.)

Material Flaws and Welding Flaws. In differentiating between primary and secondary damage, two mechanisms deserve special consideration: material flaws and welding flaws. Material imperfections are potential failure initiation sites in any component system and pertain to deficiencies in manufacture and installation of material compared to expected and specified quality. This category can include everything from forging laps to improper heat treatment to installing a component made from the wrong material. Similarly, *welding flaws* covers welding deficiencies from expected and specified quality including lack of fusion, porosity, slag, overpenetration, and improper pre- and postweld heat treatment. These two categories are special because they usually interrelate with one or more other failure mechanisms. The failure analyst then has to establish the answer to the question: “Did the failure occur because weld or material imperfections were present, or was one of the other mechanisms the primary mechanism?”

Example 2: Cracking in Gas Turbine Blades. The following example builds on the cracked gas turbine blades mentioned previously and shown in Fig. 1, 2, 3. Cracking was found to initiate from a mechanism of LCF. Low-cycle fatigue is induced during thermal loading cycles in gas turbines (each cycle representing unit startup and shutdown). However, metallography of two cracked blades revealed a change in microstructure at as-cast surfaces for depths up to 0.41 mm (0.016 in.) (Fig. 4). Evaluation by SEM confirmed the difference in structure was associated with a lack of formation of coarse gamma prime structure in the matrix (Fig. 5). Microhardness and miniature tensile test results indicated lower strength consistent with the absence of the coarse gamma prime constituent. The blade vendor found that the lot of hot isostatically pressed (HIP) blade castings had been exposed to an improper atmosphere during the HIP process, resulting in the weakened structure.

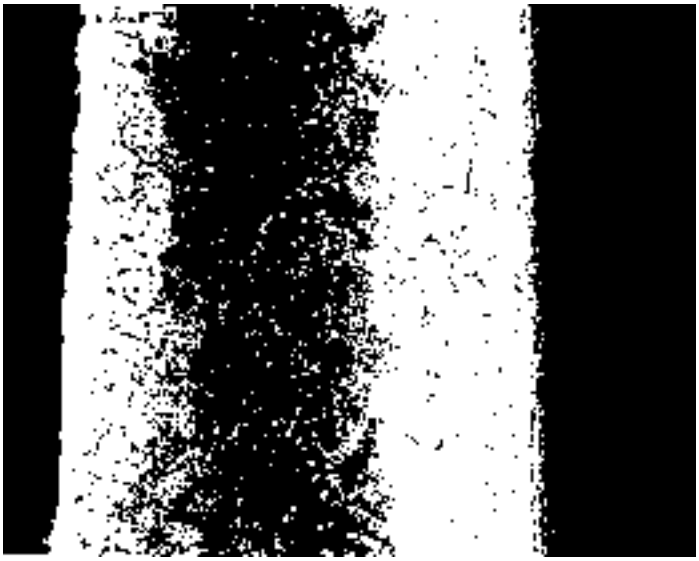


Fig. 4 Metallographic cross section through gas turbine blade. Note differences in etched structure near surfaces. Etch: electrolytic, 20% sulfuric acid in methanol. 28×

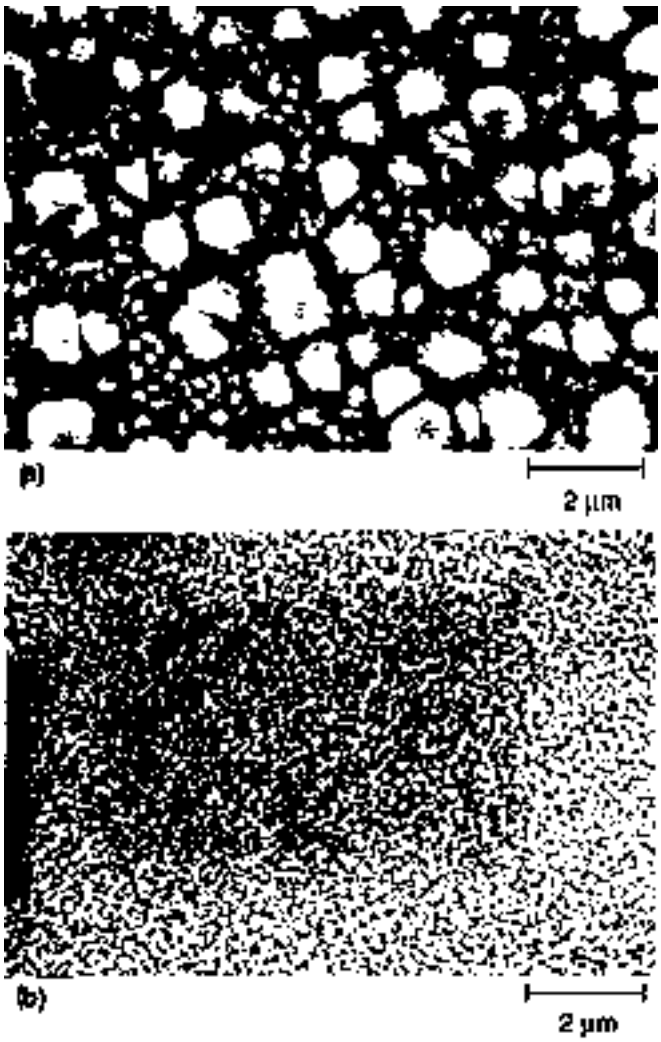


Fig. 5 Gamma prime structures from blade shown in Fig. 4 as viewed in the SEM. Note presence of coarse (primary) gamma prime particles in blade root midwall (a), which are lacking in the denuded zone of the trailing edge (b). Etch: electrolytic, 20% sulfuric acid in methanol.

Two scenarios of failure presented themselves:

- Because material strength was lowered by the resultant microstructure, the material flaw imparted from improper HIP made the turbine blades susceptible to LCF cracking in normal service where they otherwise would not be.
- All blades using the same design were susceptible to LCF cracking. The blades from the bad HIP lot were the first to fail, because reduced surface fatigue strength made them the first to crack.

Because subsequent failures were found in blades that did not come from the suspect HIP lot, the scope of the problem was considered more generic, and the primary failure mechanism was identified as LCF. Material imperfections were a secondary deficiency that had the effect of causing the blades from the bad HIP lot to crack first.

Determination and Classification of Damage

Aaron Tanzer, Siemens Westinghouse

Damage Mechanism Categorization

Identifying potential damage mechanisms in any component system is helpful in determining how a component failed. The number of potential damage mechanisms can be narrowed down during an investigation if they can be categorized further into meaningful groups that are related to observed general failure mode characteristics. Structured organization of mechanisms should be performed specifically for individual component systems because terminology and characteristics of similar damage mechanisms will vary from system to system. For example, LCF in gas and steam turbines indicates thermal cycling from unit startups and shutdowns. However, in many other systems, the presence of LCF indicates cycles related only to stress, and not temperature.
One common categorization method is to divide potential damage mechanisms among four categories of failure modes:

- Distortion
- Fracture
- Corrosion
- Wear

The appeal of this classic system is its simplicity and ease of organization. These categories serve as an organizational aid in discussing failure mechanisms (as done for this Volume, for example); however, they have limited use in actual practical diagnoses. These four broad categories of failure modes only provide the most basic forms of differentiation. Because many failure mechanisms can be categorized in more than one of these categories of failure modes, the distinctions can become less meaningful. For example, corrosion fatigue would fit in the corrosion and fracture categories. General corrosion, which eats away structural surfaces until they collapse or break, could fall under the categories of corrosion, fracture, or distortion. In practice, of course, both of these mechanisms are likely only to be considered under the corrosion category. For example, the failure analyst may see the buckled surface of a collapsed structural component exposed to a seawater environment. The Sherlock Holmes Rule says general corrosion is still a possible primary failure mechanism. In this scenario, corrosion causes wall thinning, increasing through-wall stress from the reduction in cross-sectional area. Eventually, enough wall is lost to cause buckling, which is really the secondary failure mechanism. However, the failure analyst under this “traditional” categorization might not recognize corrosion for consideration because buckling falls under the category of distortion. Corrosion is not listed as a mechanism under the distortion category. Thus (to reemphasize a point made earlier), regardless of the classification system being used, the analyst must make the effort to discover and understand all the damage mechanisms that are present.
Also, some types of damage mechanisms do not fit these four categories very well. Elevated-temperature mechanisms are often difficult to categorize. Under which category should melting and fire damage be placed? Creep is a mechanism that could result in distortion or fracture. Would either of these categories provide meaningful clues leading to evaluation of the interaction of stress, temperature, and time that produces creep?

Much of the confusion about how to apply the "traditional four" categories lies in the fact that they are not equivalent to each other. Distortion and fracture describe *how* a component physically fails, while corrosion and wear describe the physical *environment* that induces failure.
In contrast, one of the most structured and complete systems to define and categorize damage mechanisms was developed by the Electric Power Research Institute (EPRI) to characterize failures of boiler tubing. The EPRI system defines 28 separate damage mechanisms and groups them into five categories (Table 2). In this structured approach, understanding of each individual failure mechanism is broken down into:

- Metallurgical features
- Detailed damage mechanism description
- Verification of root cause
- Corrective action(s)

The detailed differentiation between various failure mechanisms enables failure analysts to identify specific root cause(s) and potential corrective actions. For example, three underdeposit damage mechanisms in waterwall tubing appear superficially to be similar. Underlying material appears gouged out or grooved. However, there are many differences in features that would enable proper identification, including the presence of intergranular tube damage and the chemistry of the deposit (Table 3). While corrective actions appear similar for caustic gouging and hydrogen damage, the former occurs from water chemistry pH being too high, while the latter occurs from a pH that is too low. Phosphate corrosion, which appears similar to caustic gouging, requires a completely different set of corrective actions compared with the other two mechanisms.

Table 2 EPRI common damage mechanisms for boiler tubing

Fluidside corrosion
Underdeposit corrosion—caustic gouging
Underdeposit corrosion—hydrogen damage
Underdeposit corrosion—phosphate corrosion
Acid cleaning corrosion
Internal deposit/corrosion product buildup
Fireside wastage
Fireside oxidation
Fireside corrosion of superheater and reheater tubing
Fly ash corrosion
Sootblower corrosion
Coal particle corrosion
Steam impingement
Fireside corrosion of waterwall tubing
Low temperature (dew point) fireside corrosion
Falling slag damage
Tube rubbing
Rupture
Short-term overheating (stress rupture)
Long-term overheating (creep rupture)
Graphitization
Cracking
Corrosion fatigue—waterside
Corrosion fatigue—fireside
Thermomechanical fatigue
Vibration fatigue
Dissimilar metal weld stress rupture
Stress-corrosion cracking
Quality control

Maintenance cleaning damage
Material imperfections/welding imperfections/discontinuities

Source: Ref 3

Table 3 Comparison between three similar boiler waterwall damage mechanisms

Comparison category	Damage mechanism		
	Caustic gouging	**Hydrogen damage**	Phosphate corrosion
Features	Rounded depressions Mottled appearance Needle Fe-rich crystals Na-rich deposits	Rounded depressions Intergranular cracking Chloride sometimes present in deposit	**Rounded depressions** **Thick grayish deposit of $NaFePO_4$ at metal surface covered by black and white deposits** **No intergranular cracks**
Causes	Localized concentrating of sodium hydroxide (NaOH) to high pH levels	Low pH conditions causing corrosion by removing steel protective oxide and producing hydrogen	**Combination of high pressure, high inside surface metal temperature, and high local concentrations of mono- or disodium phosphate**
Leak type	Pinhole leaks or thick-lipped rupture	Thick-lipped window-shaped rupture	**Pinhole leaks or thick-lipped rupture**
Locations	High heat flux regions, horizontal or inclined tubing, and locations with flow disruptions	High heat flux regions, horizontal or inclined tubing, and locations with flow disruptions	**High heat flux regions in high pressure boilers above 18,000 kPa (2600 psig)** **Only where mono- and disodium phosphate have been added**
Corrective actions	**Control water chemistry** **Minimize ingress of leakage and corrosion products** **Periodic cleaning** **Eliminate flow disruptions when possible** **Reduce local heat flux regions**	**Control water chemistry** **Remove boiler from service if pH falls below 7** **Chemical clean if pH was below 7 for more than 1 h** **Eliminate flow disruptions when possible** **Reduce local heat flux regions**	**Use equilibrium phosphate boiler water control using trisodium phosphate only (no mono- or bisodium phosphate)** **Volatile feedwater chemistry control** **Reduce drum pressure**

The methodology used by EPRI to classify boiler tube damage mechanisms is valid for application in other systems. Identification of potential damage mechanisms, acquisition of detailed information regarding material features and damage mechanisms, root cause verification, and corrective actions help prevent or minimize the scope of future failures.

Another classification methodology is a graphical system dubbed “the wheel of failure.” Failure wheel classifications are based on categorizing damage mechanisms in terms of environment. In the real world, six environments have been identified as contributors to damage mechanisms:

- Stress
- Temperature
- Corrosion

- Wear
- Radiation
- Electrical

In understanding individual damage mechanisms, it is important to also recognize that many unique mechanisms are driven by more than one environment. For example, corrosion-assisted fatigue is controlled both by exposure to stress (cyclically applied) and corrosion. In certain systems, the cyclic stress in corrosion fatigue occurs from thermal cycling. As such, temperature may also be appropriate for consideration as one of the controlling environments.

Figure 6 is a failure wheel template consisting of a circle containing three wedges surrounded by another ring. Each area on the wheel represents an environment. This particular graphical representation is limited to four environments, temperature, stress, corrosion, and wear, which are the four most commonly encountered in individual systems referenced in this Volume. The four environments selected for the wheel, however, should be germane to the particular system. For example, radiation is likely to be one of the four environments in a wheel for a nuclear reactor vessel. (Where more than four environments operate in a system, the four most relevant should be selected for the failure wheel.)

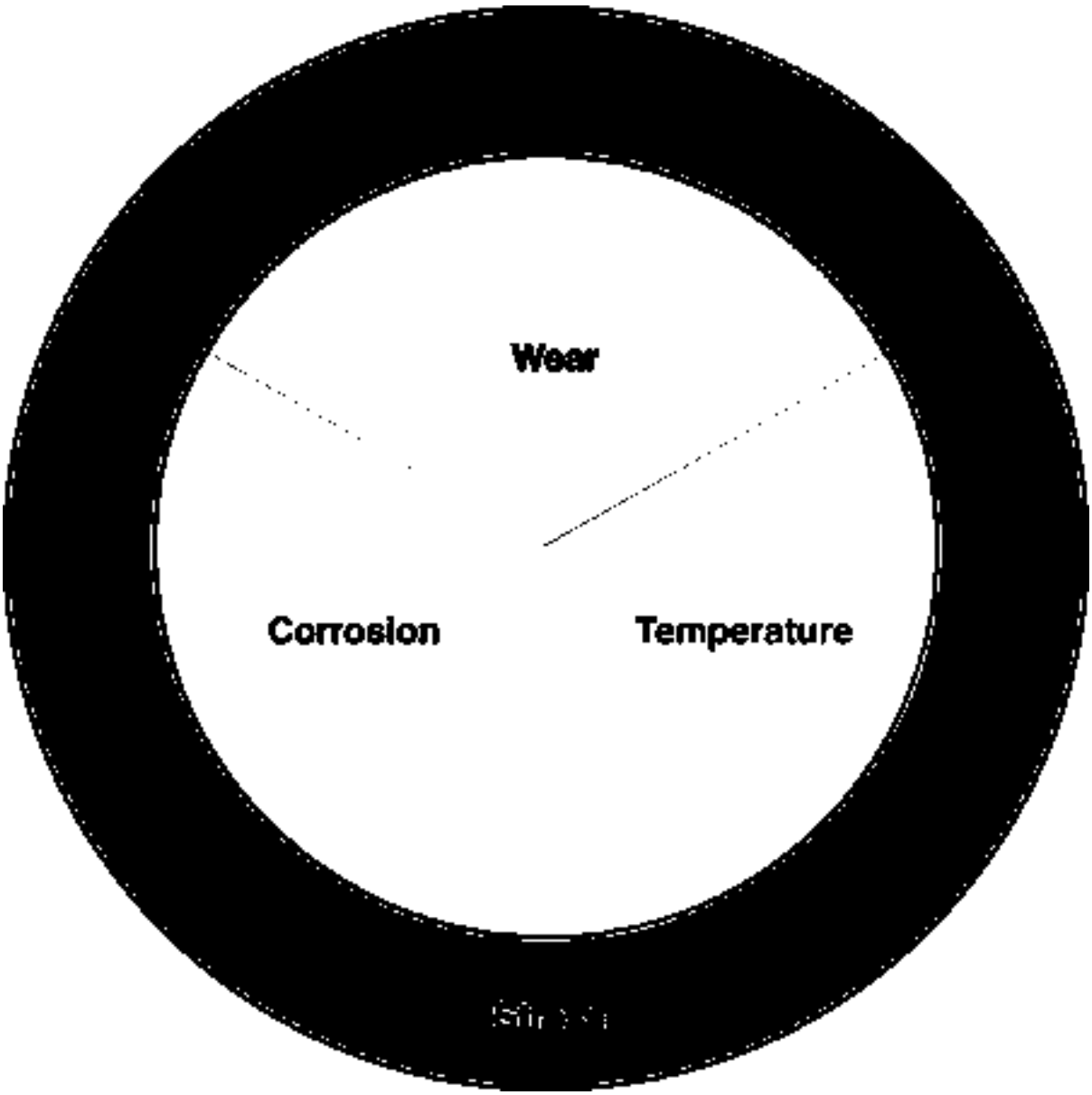


Fig. 6 Sample failure wheel template

Figure 7 shows how damage mechanisms are classified using failure wheels. In determining which environments are applicable, it is practical to apply the following test: Can the environment, as a *practical*

matter, be changed to reduce or eliminate damage occurring under the damage mechanism? For example, fracture, cracking, and distortion all indicate that stress was involved. Case in point: damage was induced on the interior of a pipe by a mechanism of flow-assisted corrosion (also commonly referred to as *erosion-corrosion*), ultimately leading to a pipe rupture from subsequent wall thinning. The damage was induced from a damage mechanism whose potential corrective actions include changing the corrosion and/or wear environments to which the pipe was exposed. Changing the stress environment, by increasing wall thickness for example, would not have been a potential corrective action in this example because wall loss from this mechanism would still cause failure at a later time. Categorizing flow-assisted corrosion as a potential damage mechanism in this system would mean locating the mechanism on the boundary between the corrosion and wear environments. The example illustrates the important point regarding categorizing damage mechanisms: The depth of understanding required to categorize a potential damage mechanism in a thorough, well-defined categorization system will enable proper identification of root cause and possible corrective actions.

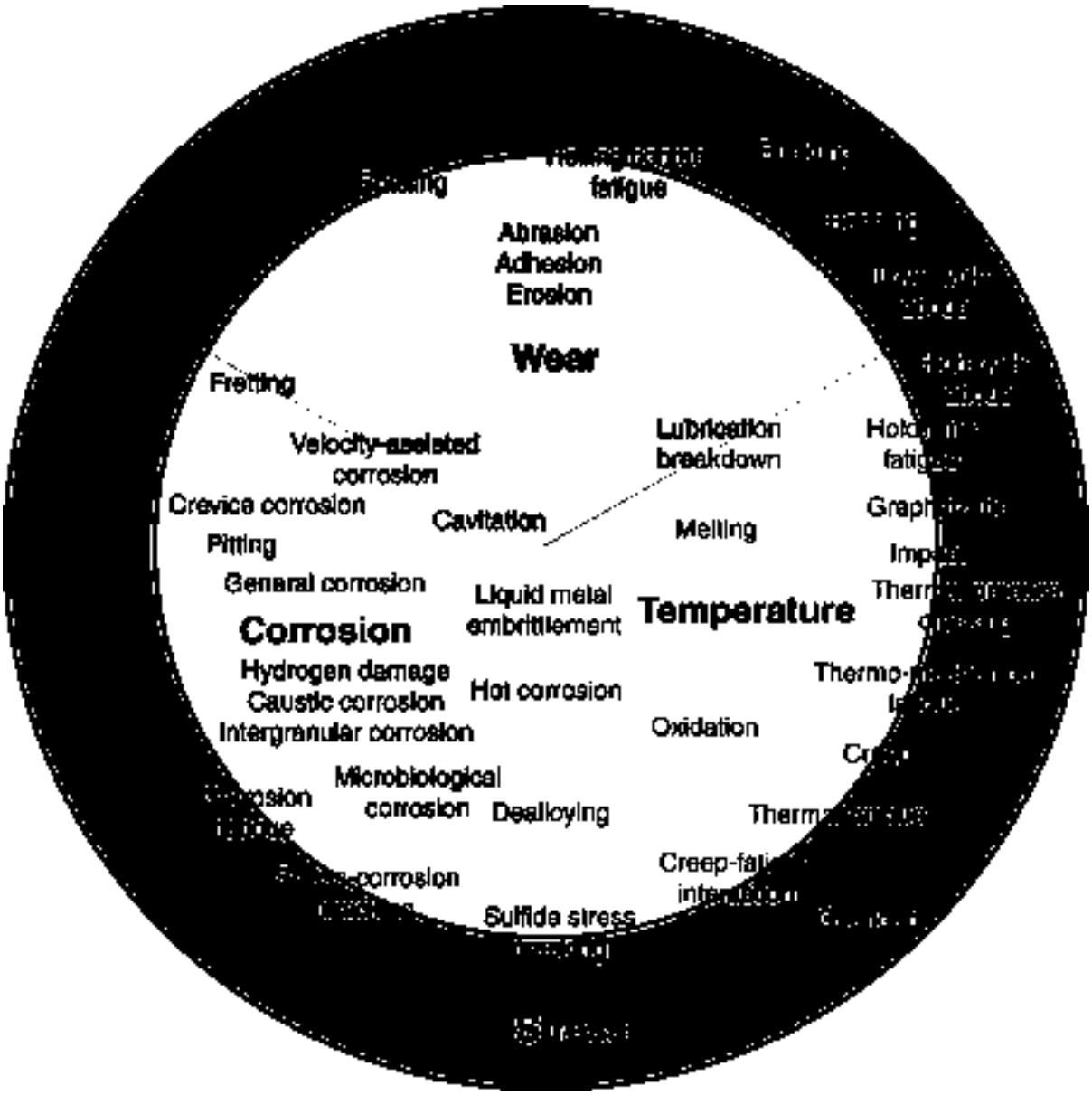


Fig. 7 Failure wheel categorization of several common damage mechanisms

The mechanisms shown in Fig. 7 are the same ones discussed in the Sections “Fracture,” “Corrosion-Related Failures,” and “Wear Failures” in this Volume. A mechanism is placed entirely in the area of an environment if only a single environment induces the damage. Otherwise, it is placed on the boundary (or, in the case of sulfide stress cracking, a triple point) between applicable environments. Determination of applicable environments is based on subjective understanding of how each individual mechanism works. Different analysts, based on their different backgrounds, will likely classify differently from those in the example shown.

When classified in this manner, two or more environments control over half of the damage mechanisms listed. Several of these are obvious. Hot corrosion is a function of both corrosion and temperature environments. Creep is induced by the combination of stress, temperature, and time. The choice for classification of other damage mechanisms may be less obvious and more subjective.

For example, impact might be considered purely a function of a stress environment. However, this may not always be the most helpful perspective. For example, consider rail car couplers (i.e., the components that join one car to the next). A series of coupler failures occurred at low temperatures in the winter. Subsequent analysis determined service temperatures were below ductile-brittle transition temperatures, resulting in brittle failures from impact between couplers during service. Heat treatment requirements were changed from normalize-and-temper heat treatment to a quench-and-temper heat treatment, lowering the ductile-brittle transition temperature. In this system, impact damage between couplers could be categorized by the temperature environment. Avoiding cold weather climates and, more significantly, changing the thermal processing environment during heat treatment could control failures. Categorizing the failure as being caused by the stress environment would not be as helpful because impact loading is a normal condition that cannot be changed. Furthermore, the subject couplers did not fail at higher temperature.

The one mechanism that transcends three environments provides another good illustration of damage mechanism classification. Sulfide-stress cracking (SSC), a damage mechanism most typically seen in drilling and oilfield applications, is a brittle failure mechanism that occurs under the combined action of tensile stress and corrosion in the presence of water and hydrogen sulfide gas. Placement of this mechanism on the corrosion-stress boundary is obvious given how this mechanism functions. It is placed on the triple point with temperature, however, because of the substantial effect temperature has on SSC resistance. (Increasing temperatures above ambient improves SSC resistance due to reduction in hydrogen permeation rate, Ref 4.)

These examples illustrate the use of a damage classification system as an aid to the analyst. Various systems are possible and different analysts will likely have different approaches. What is most important is the thought process of the analyst and achieving a correct end result. Ideally, the classification system ensures that all alternative explanations are considered.

Figure 7 is flawed in attempting to represent damage mechanisms as though the same environments drive these mechanisms in every system. For example, LCF is represented as a pure stress phenomenon, but in many instances, this is not true. In a gas turbine environment, LCF is induced by stress from thermal cycling. A failure wheel drawn for gas turbines would thus appropriately place LCF on the boundary between stress and temperature. The general purpose failure wheel shown also fails to differentiate between primary and secondary mechanisms. Consider the failure mechanism of buckling. By itself, it is purely stress driven. However, in many systems buckling occurs as a secondary failure mechanism after wall thinning has occurred. It is important in such systems to recognize that buckling was a result of the primary failure mechanism.

Classification of damage mechanisms for specific systems is much more useful in interpreting failures. Figure 8 shows the failure wheel for boiler tubes. Classification was determined based on the detailed descriptions of damage mechanisms from the *EPRI Boiler Tube Failure Metallurgical Guide.* Most of the listed mechanisms were controlled by interaction between at least two environments. Interaction between stress and temperature accounts for six mechanisms, while interaction between temperature and corrosion accounts for five. In this system, the failure mechanism that is driven by three environments is fireside corrosion fatigue. That mechanism occurs only in boilers where corrosive ash (from burning coal or oil) is present in high heat flux areas. Cracking occurs on the outside of the tubes from thermal shock exposure, most typically due to exposure to water in slag cleaning operations or splashing of water in the bottom ash hopper. The combination of thermal shock and sulfur-rich slag exposure results in corrosion fatigue initiation and propagation.

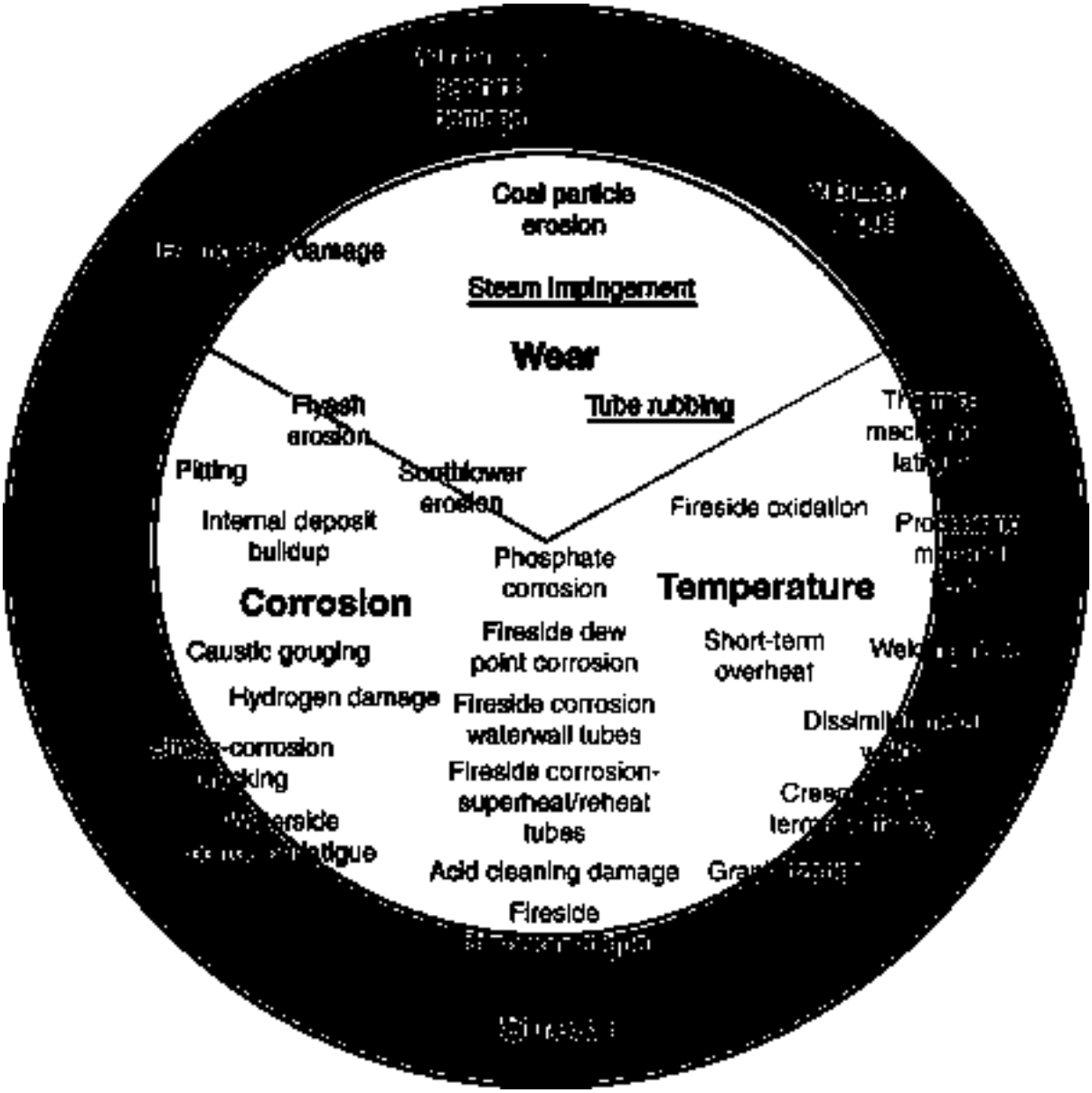


Fig. 8 Failure wheel for boiler tube damage mechanisms. Underlined mechanisms are always secondary in this system.

References cited in this section

3. S.R. Paterson, T.A. Kuntz, R.S. Moser, and H. Vaillancourt, *Boiler Tube Failure Metallurgical Guide*, TR-102433, Electric Power Research Institute, 1993

4. C.M. Hudgins and R.L. McGlasson, The Effect of Temperature (75–400 °F) on the Aqueous Sulfide Stress Cracking Behavior of an N-80 Type Steel, *H_sS Corrosion in Oil and Gas Production—A Compilation of Classic Papers*, R.N. Tuttle and R.D. Kane, Ed., National Association of Corrosion Engineers, 1982, p 90–94

Determination and Classification of Damage

Aaron Tanzer, Siemens Westinghouse

Conclusions

In a systematic approach to failure investigations, physical observations need to be compared with characteristics of the potential damage mechanisms that could occur in the system. A list of these potential mechanisms should be compiled. Included for each mechanism should be a description of how it works, what physical characteristics are exhibited by damage under it, how to verify root cause for the mechanism, and what corrective actions can be taken to reduce or eliminate the possibility of future failures. To aid in identifying the damage present, the potential list of mechanisms should be grouped in meaningful categories. Such a grouping can be unique and specific for the system being evaluated. Another way of grouping is a graphical method that involves grouping based on the environment(s) that influence the damage mechanism. Characterization will most notably identify failure mode (e.g., fatigue, overload, creep), as well as provide other clues to the damage mechanism(s) present. Comparison with the list of possible mechanisms enables identification or provides information needed to further differentiate which mechanism(s) are present. If more than one mechanism is present, the analyst also has to identify which mechanism was the primary mechanism and which is (or are) secondary. Considering the wide range of possible components, environments, mechanisms, and modes that the readers of this volume may encounter, it is not possible to present one system of damage classification that would apply equally well to all analyses. Whatever system is adopted should aid in a thorough, logical, systematic analysis, with consideration of all possible scenarios. Such an analysis provides the greatest confidence that the correct root cause will be identified and that the corrective action taken will be effective.

Of necessity, this volume is organized according to a simple system involving deformation (distortion), fracture, corrosion, and wear. Whatever classification system is employed, the analyst must ensure that all damage modes are discovered and documented. The appropriate section or sections of this Volume and other references can be used as needed to fully understand how each damage mode occurred and to prescribe corrective action.

Determination and Classification of Damage

Aaron Tanzer, Siemens Westinghouse

References

1. A.C. Doyle, *The Sign of Four*, 1889

2. W.M. Torburn, The Myth of Occam's Razor, *Mind*, Vol 27, 1918, p 345–353

3. S.R. Paterson, T.A. Kuntz, R.S. Moser, and H. Vaillancourt, *Boiler Tube Failure Metallurgical Guide*, TR-102433, Electric Power Research Institute, 1993

4. C.M. Hudgins and R.L. McGlasson, The Effect of Temperature (75–400 °F) on the Aqueous Sulfide Stress Cracking Behavior of an N-80 Type Steel, *H_sS Corrosion in Oil and Gas Production—A Compilation of Classic Papers*, R.N. Tuttle and R.D. Kane, Ed., National Association of Corrosion Engineers, 1982, p 90–94

Examination of Damage and Material Evaluation

Iain Le May, Metallurgical Consulting

Introduction

EXAMINATION of a damaged component involves a chain of activities that, first and foremost, requires good observation and documentation. If the component is being examined on site where the failure took place, for

example, its position and relationship to other components must be recorded. All interested parties should also be informed prior to handling or removal of specimens from the site. If parts have been shipped to an investigating laboratory, details of shipment must be recorded, together with the condition of all parts. The receipt of the component by the individual or individuals examining it must also be recorded, together with any observations regarding its general condition. There is always the possibility of damage having occurred during handling or shipping and of parts being mislaid, so that the investigator must ensure that he or she is not accused unjustly of having damaged or lost parts.

It is essential that the component be examined overall and that the features of interest be recorded photographically or by means of sketches or both. If possible, it is valuable to have a drawing or drawings of the components on hand to check if all parts of an assembly have been supplied and to determine whether apparent distortions are real or were part of the design. Following receipt and documentation, the features of damage can be recorded and their cause(s) investigated, as this article briefly describes, for typical types of damage experienced for metallic components. Examination and evaluation of plastics and ceramics are also briefly discussed. It should also be noted that the term “subject component” is used in text, because the determination of failure (or loss of function) may still be the objective during examination and evaluation of the subject component.

Examination of Damage and Material Evaluation

Iain Le May, Metallurgical Consulting

Visual or Macroscopic Examination of Damaged Material

Visual and macroscopic examination are performed to document the main features apparent to the eye directly or by using low-power magnification. Of interest are scores, scuffing, dents, distortions, evidence of plastic deformation, cracks, and fractures. For example, documentation of damage and cracking surrounding a fracture, including damage patterns and crack origins, is an important step in assessing relevance to final failure. This information needs to be recorded before any parts are cut from the component of interest. A considerable amount of information can be obtained from observation of the direction and sequence in which cracking has occurred. For example, where cracks meet, as illustrated in Fig. 1, it is apparent that crack “B” occurred at a later stage after crack “A” had formed, and that crack “B” was arrested at the prior crack.

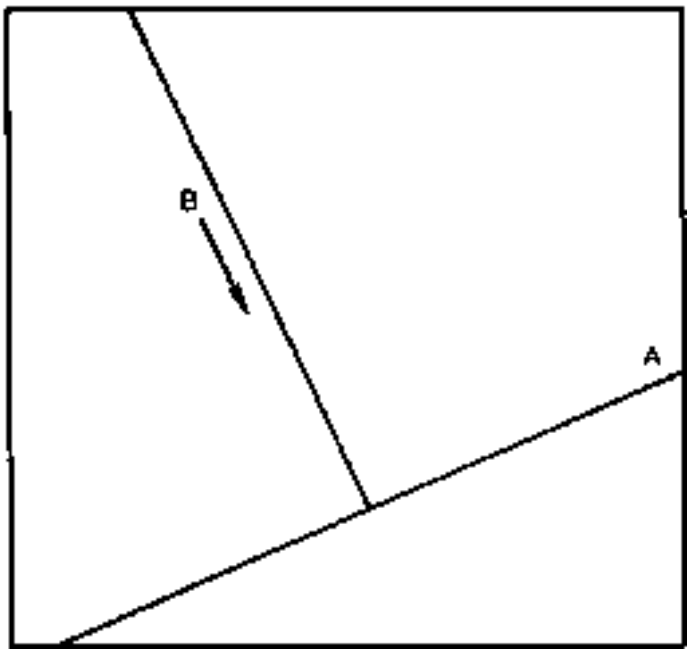


Fig. 1 Intersection of two cracks. Crack "B" has been arrested at crack "A," which occurred first.

In the case of fracture surfaces, it is important to record if these are macroscopically ductile or brittle in appearance; whether they are shiny or dull, fresh or corroded; their orientation with respect to the component; and the loading carried. In many, if not most, cases, fracture surfaces have multiple stages to their appearance. An example is a surface oriented perpendicular to the component outside surface, where a smooth thumb-nail-shaped region is seen, followed by a rough, apparently brittle area of fracture, followed in turn by a small amount of slant fracture. In such a case, the initial fracture may have been by fatigue, followed by fast fracture when the crack had reached a critical length, with a final shear lip as the crack completed its passage across the section. It is important to record the extent of each region of the fracture. Photographs of this as well as of the other features of the damage should include a scale marker of some sort, be it a scale rule, a coin, or an instrument such as a pen. For large objects such as a tank or pressure vessel, having a person standing beside them when they are photographed can serve the same purpose and allows an impression of the size to be conveyed when a report is prepared or when evidence is presented in court.

In many cases, components have corrosion or other products on the fracture surfaces that hides surface details of the fracture. However, fracture surfaces, generally speaking, should not be cleaned until after their appearance with corrosion or other contaminants present has been recorded. In some cases, the contaminating material may be of interest, and sample material may require analysis either on the fracture surface or after it has been removed but preserved. However, there may be legal issues relating to cleaning, and the potential exists for allegations of destruction of evidence if permission from all parties is not sought and obtained before any cleaning is done.

The presence of corrosion products on some or all of a fracture surface can be very important in deducing both the cause and the time at which failure occurred or initiated. For example, Fig. 2 shows the fracture surface of a bolt that failed in fatigue, but the dark nature of the oxide on part of the fracture surface indicates that this cracking occurred a long time (years, in this case) before the later cracking. If the bolt had been cleaned in an attempt to display the underlying fracture surface of the metal, this somewhat dramatic evidence would have been destroyed.

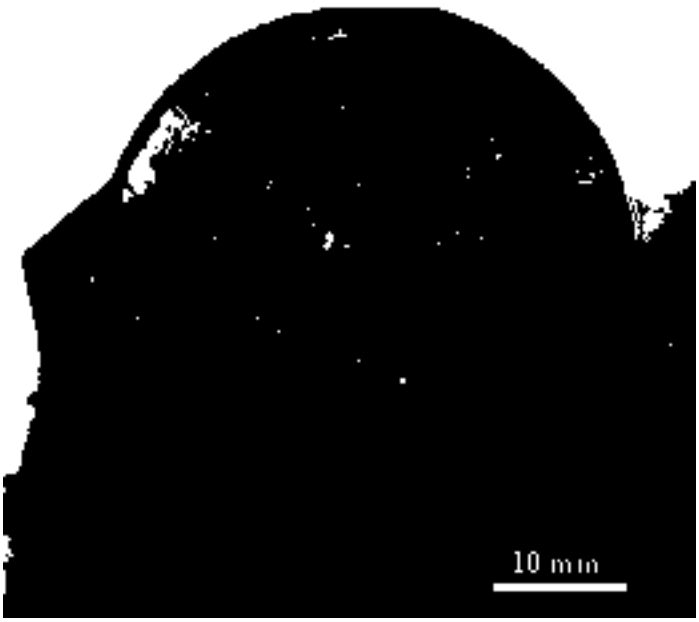


Fig. 2 Bolt that failed in fatigue. The smooth dark areas are from initial fatigue cracking, now coated with dark adherent oxide. This initial cracking took place some years prior to the later fracture, the surface of which is covered by a less adherent rust coating.

As another example, when damage occurs by electrical discharge (for example, if a crane touches a power line or if welding has been done on a component on a vehicle while the electrical ground is connected through the

unit), the damage can show up in the form of local arcing through the oil film in bearings and in local heating at contacting surfaces. Visual or macroscopic examination can reveal these features, as illustrated in Fig. 3.

Fig. 3 Electrical discharge and arcing through the bearing have produced blue-black oxide, indicating the local heating. The ball bearings also show local hot spots.

When parts have been removed from a site for examination, they must be preserved carefully and protected from damage. In some cases, it may be difficult to store them, perhaps due to space requirements, and also because there is often a very long time (ten years or more, in some cases) between examination of parts and a subsequent trial. In such cases, the investigator needs to seek permission to have the parts stored elsewhere or by another party or to dispose of the parts if all parties agree that they are no longer required. Reference should also be made to relevant ASTM standards dealing with preservation of evidence and the examination of parts that may be involved in litigation (Ref 1, 2).

References cited in this section

1. "Standard Practice for Preservation of Information and Physical Items by a Technical Investigator," E 1188–95, *Annual Book of ASTM Standards*, ASTM

2. "Standard Practice for Examining and Testing Items That Are or May Become Involved in Products Liability Litigation," E 860–97, *Annual Book of ASTM Standards*, ASTM

Iain Le May, Metallurgical Consulting

Interpretation of Damage and Fracture Features

Macroscopic Examination. The presence of distortions or plastic deformation in a component can indicate the direction of loading that was applied. The shape of indentations can also allow the nature of an impacting part to be determined. In some cases, there may be deposits of the impacting material, including coating, left on the surface of the impact area. Examples include aluminum present on the surface of a steel component that has been impacted by an aluminum component or the presence of paint on a scraped or indented part.

Figure 4 shows a bolt from a diesel engine connecting rod that had stretched after very limited service. Macroscopic examination revealed that the bolt had stretched locally in a region in which the cross section was not a minimum, indicating locally weak material. A closer examination revealed the presence of many cracks that were oxidized internally. The conclusion from the macroscopic examination was that these cracks had been produced before the bolt was placed in service, and indeed, before the steel was heat treated. This was confirmed by microscopic examination, which showed the cracks to have been produced by stress-corrosion cracking of the material. While the specific cause of the stress-corrosion cracking was not investigated, it is

suggested that there must have been a spillage of corrosive fluid on the bolt prior to heat treatment. As a result of the analysis, the damaged engine was replaced by the bolt supplier.

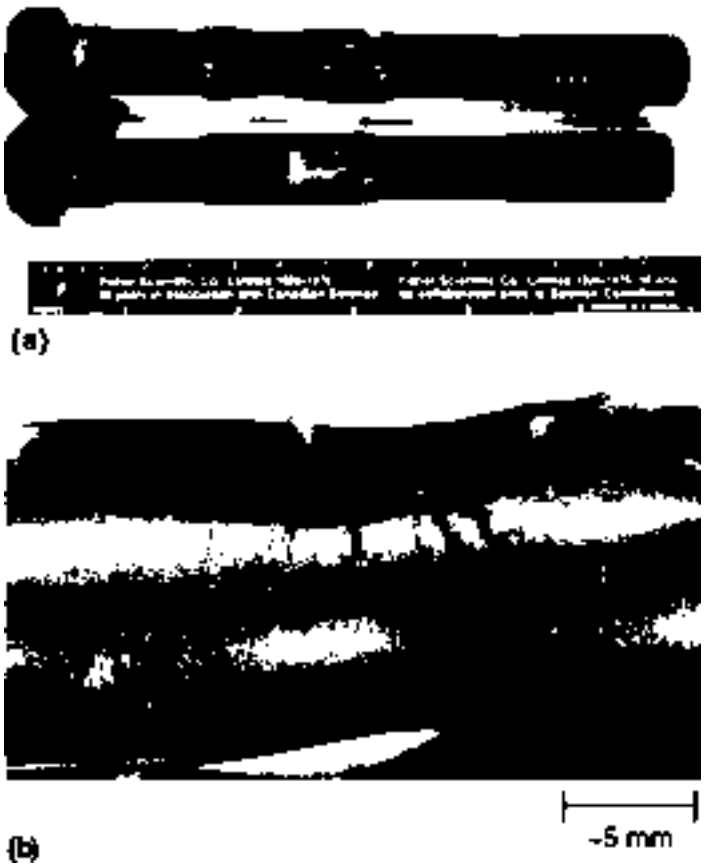


Fig. 4 Failure of a connecting rod bolt in a diesel engine. In (a), the failed bolt is the upper one, having necked down in a nominally larger cross-sectional area. The lower bolt is another removed from the engine in unstretched condition. In (b), the stretched region of the bolt is shown, with cracks that have opened up during bolt tensioning and subsequent operation.

Macroscopic examination of fracture surfaces can provide a great deal of useful information for analysis of a failure. Fatigue fracture surfaces, which are often flat in nature, frequently display “beach marks” or “clam shell” markings, indicative of interruptions in the progress of fatigue fracture through a component (Fig. 5). This can arise from shutdowns in operation of the component or changes in the level of the fluctuating load. Brittle fractures often display a chevron pattern, with the points of the chevrons pointing back toward the fracture origin, as shown in Fig. 6. Figure 7 shows the fracture origin of a brittle fracture of a shaft; the chevrons are spread out in this case, but their appearance, as in the case of a narrow or plate specimen, as shown in Fig. 6, may be seen if a slice running across from the fracture origin in Fig. 7 is considered.

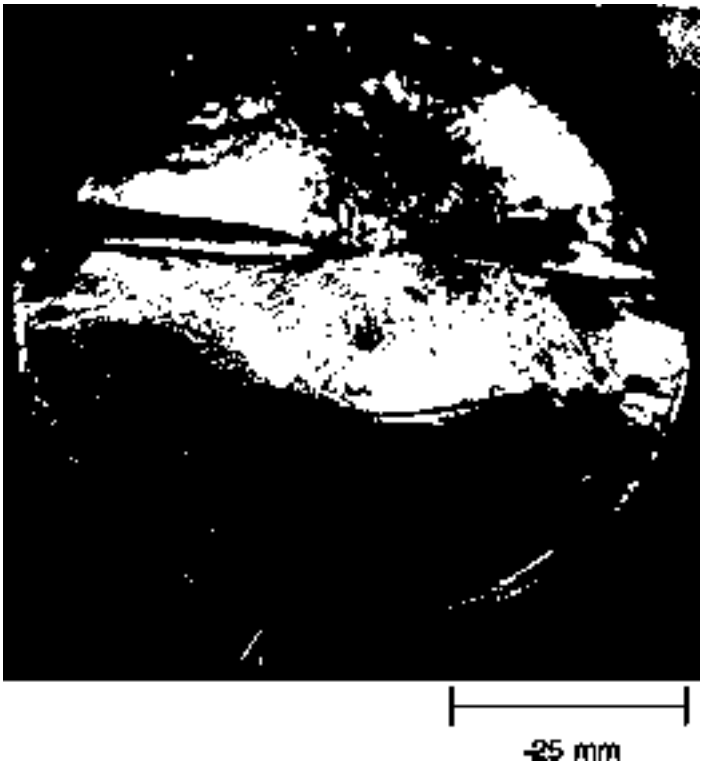


Fig. 5 Fatigue fracture surface appearance of a failed crankshaft, showing “beach marks” on the lower part. The origin of the primary fracture is indicated by the arrow.

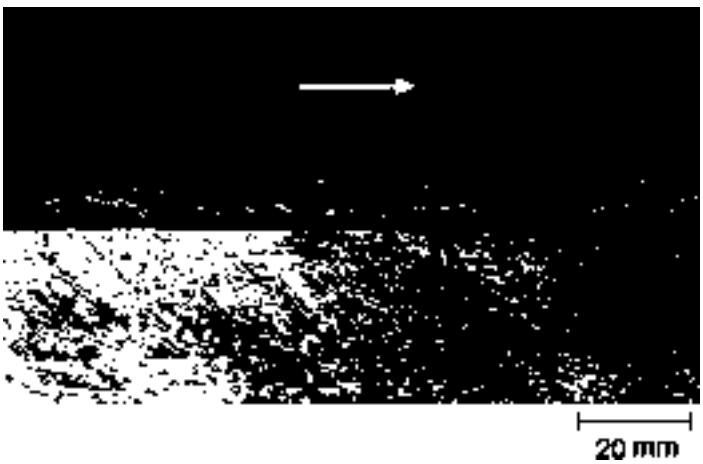


Fig. 6 Chevron marks on the fracture surface of a steel plate from an oil storage tank that fractured in a brittle manner. The arrow shows the direction of crack propagation.

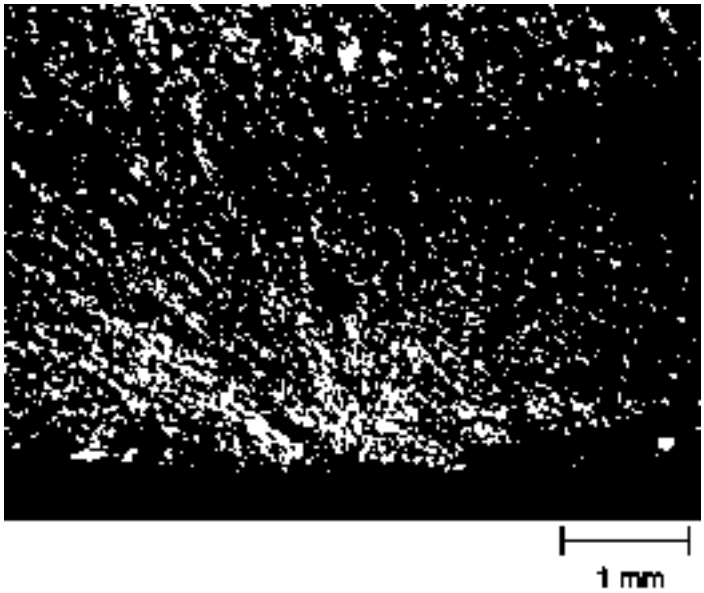


Fig. 7 Origin of a rapid brittle fracture in a steel shaft

In the case of erosion, for example, by cavitation or erosion-corrosion of a pipe in which a liquid is flowing under turbulent conditions, the characteristic teardrop shape of cavities is apparent at low magnification, as may be seen in Fig. 8.

Fig. 8 Elongated surface cavities on the inside surface of a 70-30 cupronickel tube produced by erosion-corrosion. The tube surface is clean, the attack having occurred due to brine flowing through it at 70 °C (158 °F) with turbulent flow and an excessive level of dissolved oxygen.

Oblique lighting of fracture surfaces for macroscopic examination can be used to advantage to display surface topography. This often involves arranging lights until the desired effect is shown in an exaggerated manner. The lighting from photoflood or similar lamps often results in a suppression of the surface features that are desired. Flexible fiber-optic lighting is often the most appropriate method of illumination. Figure 9 shows a view of the fracture surface of a tensile test specimen from a cast steel that has suffered embrittlement, resulting in a “rock candy” fracture.

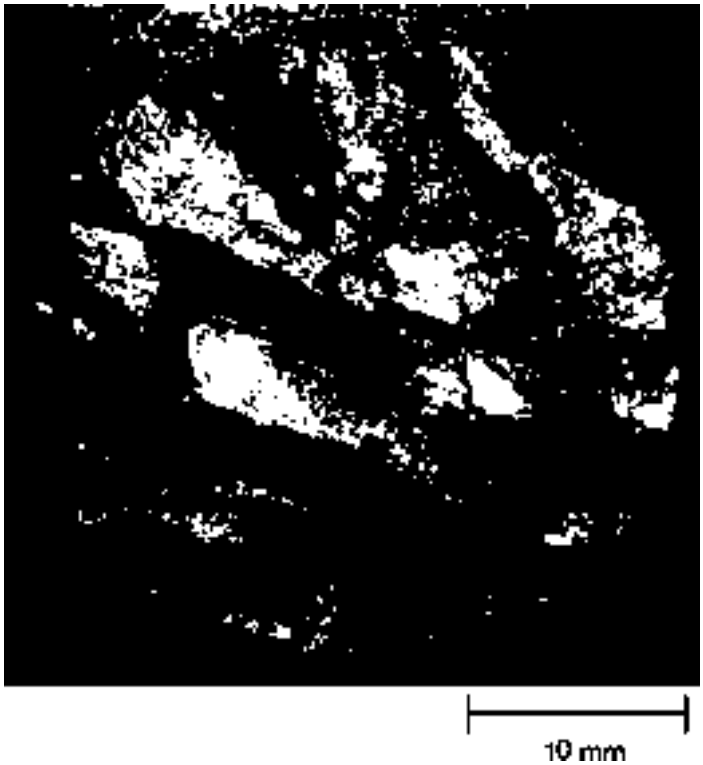


Fig. 9 Low-power light microscope view of a "rock candy" fracture in a tensile specimen taken from a cast steel that had aluminum nitrides segregated to the grain boundaries

Macroscopic examination is commonly made using a low-power binocular stereo microscope, which gives images at magnifications in the range of approximately 4 to 25×. Good macroscopic results can also be obtained on a few (only) light metallographs that are designed for the use of low-power macroscopic lenses. Figure 10 shows a copper tube with pitting that originated at the outside surface, which has been illuminated at low magnification on a metallograph with oblique lighting, together with illumination within the tube from fiber optics, to demonstrate the perforation of the wall. In general, there are considerable limitations on the size of specimens that can be examined using a metallograph. Larger specimens can be examined by using a binocular stereo microscope that can be supported by an arm from a stand that in turn can be placed on a large failed component.

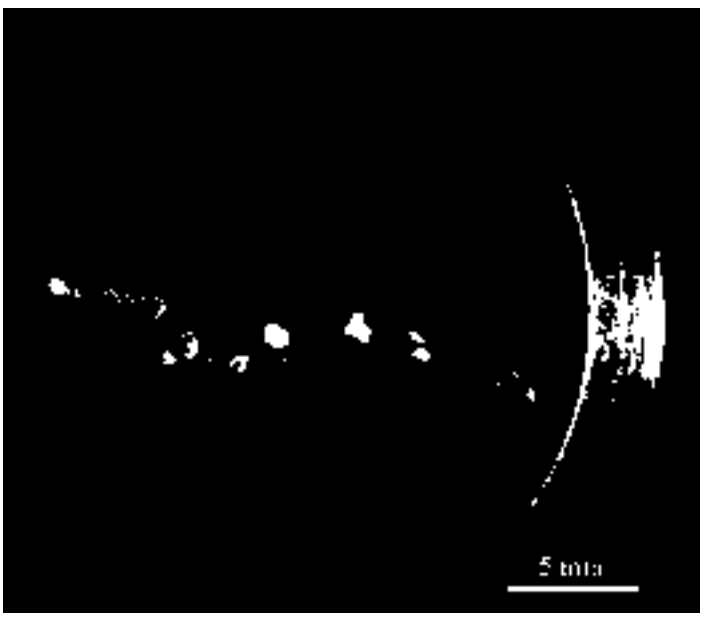


Fig. 10 Pitting from the outside of a copper tube. This is shown under oblique lighting set on the stage of a metallograph. The inside of the tube is also illuminated, using fiber optic lighting to demonstrate the perforation of the wall.

The use of the scanning electron microscope (SEM) at the lowest-magnification range can be very useful in macroscopic examination, both because of its large depth of field and its ability to provide higher-magnification views of specific areas that have also been recorded at low magnification. It is normally used after examination has been made using a light optical microscope. Size limitations of the chamber of the SEM often dictate sectioning of the subject component prior to examination.

Microfractography. Detailed examination of fracture surfaces at higher magnification can facilitate evaluation of the failure mechanisms.

Microfractographic examination can be conducted in several ways, including direct methods involving light microscopy and scanning electron microscopy and indirect methods involving replicas that can be examined by light microscopy or by scanning or transmission electron microscopy (TEM). The limited depth of field available in the light microscope is such that it is difficult to record fracture surface features for rough surfaces other than at low magnification. Hence, most microfractographic examinations today are made using the SEM, which has a great depth of field. Figure 11 shows a SEM view of a fracture surface, the excellent depth of field being apparent. Excellent results can also be obtained using replicas that are examined in the TEM. The procedure to prepare such replicas is time-consuming, however, so they are used only to a limited extent in failure analysis. Figure 12 shows examples of fracture surface replicas viewed in the TEM. These are carbon films deposited on the original plastic films used for replication of the fracture surface. Subsequently, the plastic is dissolved away, and the carbon replica, supported on a copper grid, is coated with gold or palladium applied at an angle to the surface to create contrast shadowing.

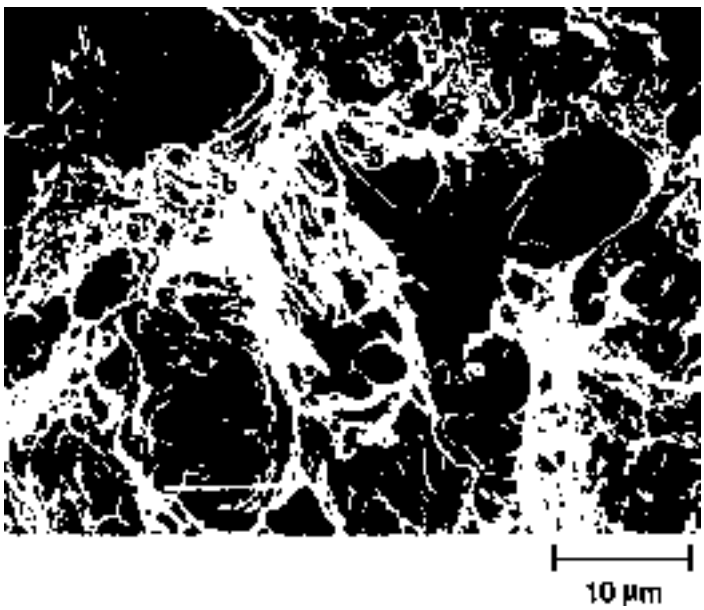


Fig. 11 SEM view of the fracture surface of a low-carbon steel specimen broken in tension, showing ductile dimples, local quasicleavage, and manganese sulfide inclusions

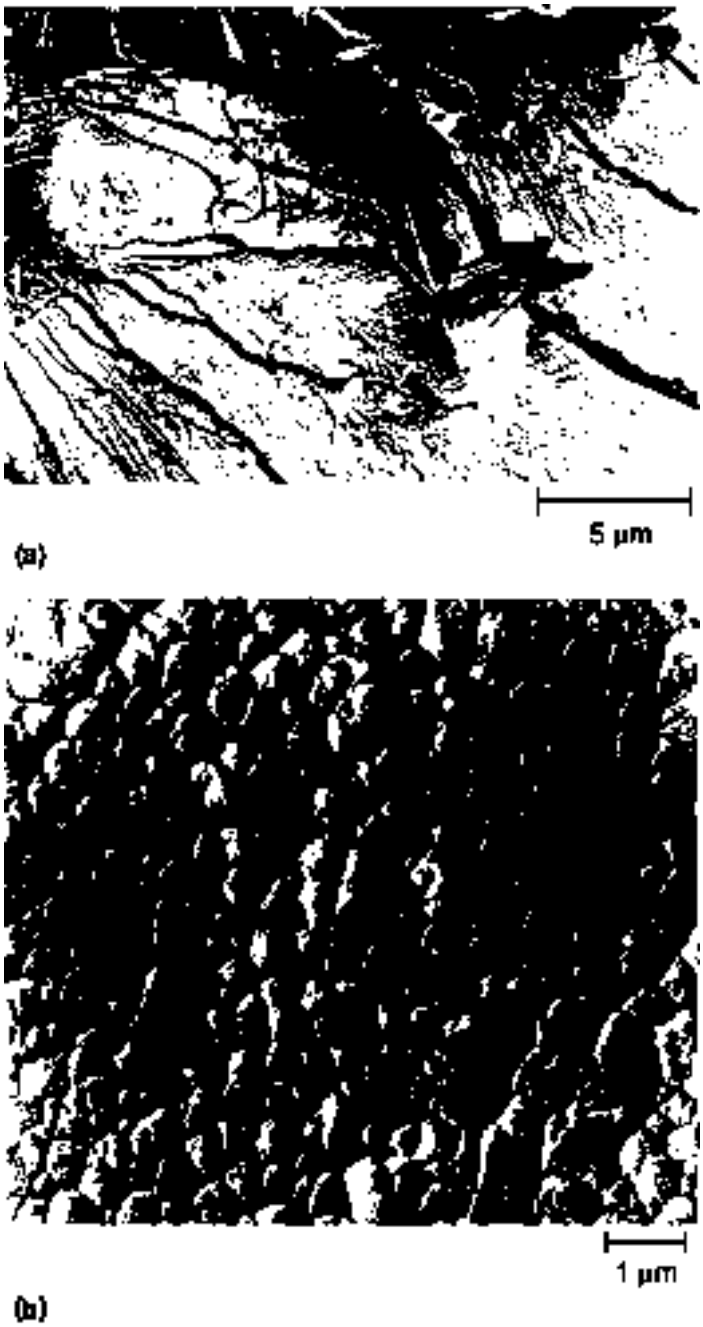


Fig. 12 Replicas prepared from fracture surfaces and viewed in the TEM. (a) River patterns of a brittle failure of a steel fastener. (b) Fine, equiaxed ductile dimples on the surface of a tensile fracture of a high-strength steel. The small carbide particles where local rupture and void formation initiated are apparent.

Cleaning of fracture surfaces in order to display the underlying fracture surface and its characteristic features is an important matter but one that is fraught with hazards for the unwary. There are dangers in that the cleaning may damage surface features, for example, by pitting attack. There are several satisfactory methods of cleaning surfaces so that they can have microfractographic examination conducted on them. If litigation may be involved, only the last method on the list is considered nondestructive. These include:

- Removal of debris with a blast of pressurized gas or air
- Brushing with a soft, nonscratching brush
- Washing with organic solvents
- Use of an ultrasonic cleaning bath and appropriate solvents. The use of an ultrasonic cleaning bath is very useful in speeding up the cleaning process.

- Use of a mild acid or alkali, depending on the metal. A solution that attacks the deposits but not the underlying metal is required.
- The use of cellulose acetate replicas to lift debris from the surface. If this method is used for litigation work, the cellulose acetate replicas should be numbered in the order that they were used, then stored for future reference.

When mild acid or alkali solutions are to be used, it is useful to try them first on a piece of similar material. This surface should be examined and recorded before the cleaning medium is applied and then again examined afterward to ensure that no attack takes place. For steel fracture surfaces, a 6 *N* hydrochloric acid inhibited with 2 g/L hexamethylene tetramine has been used successfully (Dong's solution) (Ref 3), while very good results in the removal of rust from steel can be obtained using a hydrochloric acid hexamethylene tetramine solution. For removal of rust from steel, a mixture of 10 mL HCl, 90 mL H_2O, and 3 g hexamethylene tetramine has been found to work well, removing the rust without attacking clean metal, although the exact proportions in the solution can vary, depending on the particular situation. For steel with an organic zinc coating, a solution with 40 mL HCl, 60 mL H_2O, and 5 g hexamethylene tetramine has been found to remove the coating and clean the fracture face without attack on the metal. Figure 13 shows a fracture that had taken place during a drilling operation in a water-laden soil environment. In Fig. 13(a), the fracture is initially covered with corrosion products, the surface being reddish-brown from the rust present. Figure 13(b) is a high-magnification view showing the nodules typically observed on the rusty surface, and Fig. 13(c) shows the surface after it was cleaned using Dong's solution.

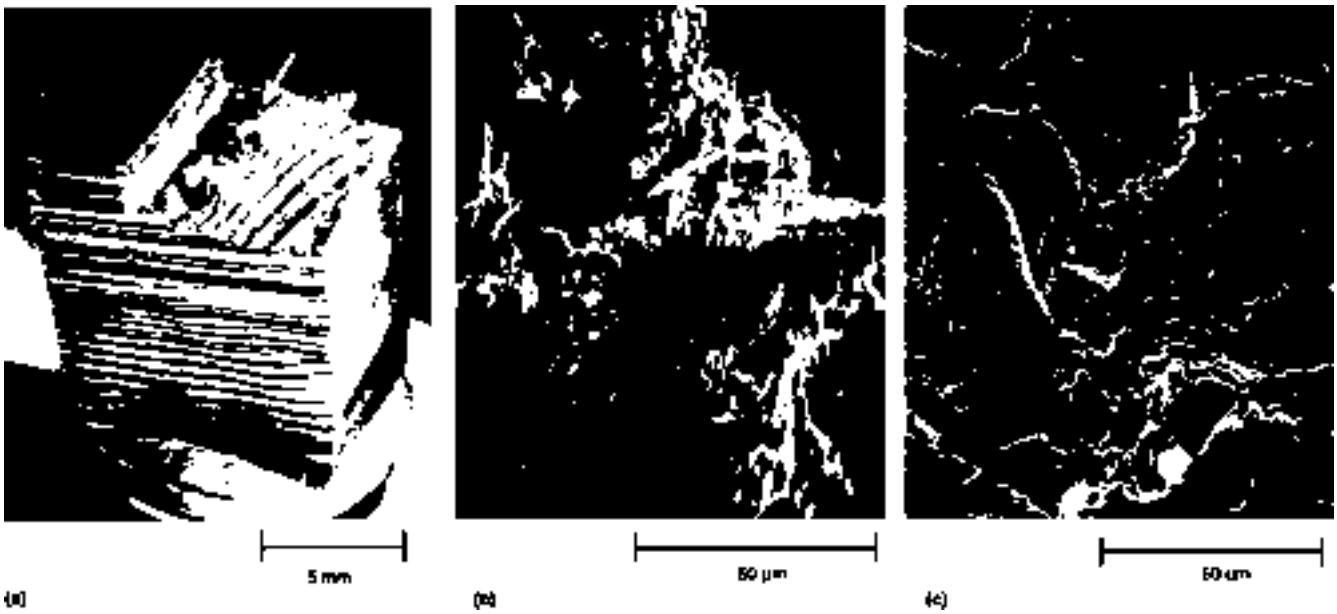


Fig. 13 Fracture surface where a welded connection failed in a drill pipe connection. (a) General view of part of the fracture surface, which was covered with rust. (b) High-magnification view (approximately 2000×) of the rust nodules on the fracture surface. (c) View of the fracture surface at the position indicated by the arrow in (a), as recorded in the SEM after being cleaned in Dong's solution, the magnification being almost the same as in (b). Cracking initiated by fatigue from an internal defect

Cellulose acetate replicas can serve a double purpose when used for cleaning fracture surfaces. In addition to the cleaning function, the debris and corrosion products removed from the fracture surface are then attached to the plastic and can be examined and analyzed subsequently. The procedure followed is to immerse in acetone a strip of acetate cut from a sheet about 1 mm (0.04 in.) thick, at most, place this on the fracture surface, and apply pressure to maintain contact over the fracture surface until the strip has dried. It is then removed and the procedure repeated until a satisfactorily clean surface is produced. The method is tedious because the drying times may be long, necessitating the use of clamps to hold the acetate in place. It is often impossible to remove

very adherent rust or oxides. Nevertheless, the method is an excellent one for removing looser debris and lightly adherent corrosion products, because it does not damage the metal surface in any way.

When the various standard references and other texts that illustrate microfractographic features are consulted, it is normal to find excellent illustrations of the features of interest. Unfortunately, the failure analyst is often faced with a very different situation, in that the characteristic features may be seen only very locally and sometimes not at all, because of damage to the fracture surface(s). The situation is particularly notable with respect to the identification of fatigue striations, which cannot normally be seen using light microscopy and require higher magnifications for them to be distinguished clearly. In most illustrations of fatigue striations, the material is ductile aluminum, in which they are very distinct. However, they are much less readily seen on the fracture surface of steel samples, even where there is little surface damage. Even when it is known that the failure was by fatigue, it may be difficult to find evidence of striations. Thus, if there is some compelling reason to find them, it is necessary for the analyst to scan carefully, concentrating on pockets where there has been less surface rubbing or contact between the two fracture surfaces. The compelling reason to find the striations may be that the analyst wants to do fracture mechanics work. Otherwise, it may be acceptable to use other criteria to assert that it appears to be a fatigue crack. This may be more intellectually honest, as far as a documentation of the crack surface appearance, than emphasizing the importance of a patch containing a few striations on a large fracture surface.

Figure 14 shows a patch of fatigue striations identified on an extruded 2024 aluminum alloy. In this case, where the crack front progressed on a number of parallel planes, the fracture surface was very rough, and there was extensive surface-to-surface contact before final failure.

Fig. 14 A local area of fatigue striations on an extruded 2024 aluminum alloy component

Another feature that is sometimes observed on fracture surfaces and is a clear indicator of fatigue fracture is the presence of “tire tracks,” caused by the entrapment of small particles between closely mating fracture surfaces. Relative motion between the surfaces produces a series of indentations, shown in Fig. 15.

Fig. 15 "Tire tracks" on a fatigue fracture surface of a 4140 steel quenched and tempered at 700 °C (1292 °F)

Nonfracture features of interest include rub marks or scores, indentations, deposits of other metals or other materials that have been deposited on surfaces by impact or rubbing, porosity, and heat damage.

Rub marks, scores, and indentations can provide important information concerning contact between the subject part and other components. The direction from which contact was made may be inferred in many cases, and this information needs to be correlated with the evidence of crack initiation and direction in a fractured component. Deposits of other metals can provide information on the contacting material, as do deposits of paint or other coatings. Evidence of wear can be very important, indicating poor lubrication, for example, or excessive loading. Evidence of surface heating, as indicated by discoloration, may relate to rubbing, service temperature, or to prior heating of the component during manufacture. In other cases, it may be indicative of electrical discharge, with arcing between surfaces separated by an air or oil gap. This can lead to the formation of bluish surface films on a steel surface such as a ball or roller bearing track or rolling member, or it may induce overall heating and softening of the metal if the current is large and sustained, which leaves a thin, dark oxide layer on the surfaces (see Fig. 3).

Porosity observed on a component that has failed or suffered damage may be an indication of poor quality control but does require investigation to determine if it was related to the failure.

In all such cases of nonfracture features being observed, they should be recorded carefully before parts are cut up, returned to the owner, or passed to another party for their investigation of the failure.

Reference cited in this section

3. H. Dong, Concurrent Technologies Corporation, Pittsburgh, PA, private communication

Corrosion and Wear Damage Features

Features of Corrosion Damage. The failure analyst is frequently faced with the problem of specifying how a component has failed by corrosion or erosion. As with so many failure investigations, the information supplied is often incomplete in terms of the service conditions that the component has been exposed to, and detailed questioning may be required, either before or following the initial investigation, in order to obtain sufficient information to enable a definite opinion to be offered. Good knowledge of the actual conditions is of particular importance when the customer wishes to have a recommendation for replacement of the corroded or eroded component by another of a different material that will last much longer and not deteriorate in the same manner.

Corrosion damage can be apparent in many different ways, including loss of material, surface pitting, and the buildup of corrosion deposits, but it is convenient to classify corrosion by visual observation of the corroded material before any cleaning is conducted. There are generally considered to be eight basic forms of corrosion (Ref 4):

- General attack (uniform corrosion)
- Galvanic corrosion
- Crevice corrosion
- Pitting
- Intergranular corrosion
- Selective leaching
- Stress corrosion
- Erosion-corrosion

This classification was initially proposed in the 1960s, and since then, terminology has been refined to more clearly distinguish the considerable overlap between these categories or types of corrosion. For example, galvanic corrosion, although sometimes considered as a form of corrosion, is more accurately considered as a type of corrosion mechanism, because galvanic action can accelerate the effects of other forms of corrosion. Uniform attack, pitting, and crevice corrosion can all be exacerbated by galvanic conditions. This distinction is important when considering the form or mode of corrosion penetration (as discussed further in the article "Forms of Corrosion" in this Volume). The form (or appearance) of attack also may not be unique, because several corrosion mechanisms may operate at the same time for a given mode or appearance of corrosion. Moreover, several additional mechanisms of attack, including hydrogen damage, "environmentally-assisted fracture," high-temperature corrosion, and biological corrosion, are normally considered today. These mechanistic-based categories may also result in one or more of the forms of corrosion (e.g., pitting, uniform attack, intergranular corrosion, and so on).

Although the distinctions between the eight basic categories of corrosive attack have become blurred, particularly when fundamental mechanisms are considered, this classification may help (at least in the first instance) to simplify the analysis. If required, fundamentals can be considered at a later stage.

A good practice is to use this basic list, together with hydrogen damage, environmentally-assisted fracture, and high-temperature corrosion, for preliminary classification of a corrosion-related failure. The general features that may be observed for general, galvanic, crevice, and pitting corrosion are:

- General attack (uniform corrosion), which is typically apparent from overall thinning and loss of metal
- Galvanic corrosion, which occurs when dissimilar metals are in electrical contact via an electrolyte (or less commonly, when similar metals are in electrical contact via either dissimilar electrolytes or even a similar electrolyte with a temperature differential)
- Crevice corrosion, which is recognized by its local nature and its attack in an area that is shielded from full exposure to the environment because of close proximity between the metal and the surface of another material
- Pitting, which usually is readily detected, although in some cases it may be obscured by the buildup of deposits and may not be fully identified until after cleaning is conducted

Intergranular corrosion, selective leaching, and stress-corrosion cracking may be more difficult to differentiate until more detailed examination is made. Erosion-corrosion can usually be identified on the basis of the locally

elongated pits that develop where a rapidly flowing liquid has been present, for example, in a pipe and particularly at a change in section or geometry (see Fig. 8). Some examples of corrosion damage follow.

Figure 16 illustrates an example of failure by galvanic corrosion. It is a connection for a water distribution system in which the contractor directly coupled a brass fitting to a steel pipe. Although an obvious recipe for failure, many such connections had been made, backfilled, and paved over before a failure occurred. The example emphasizes that materials should be specified carefully by the engineer responsible and not left to the discretion of those making the installation.

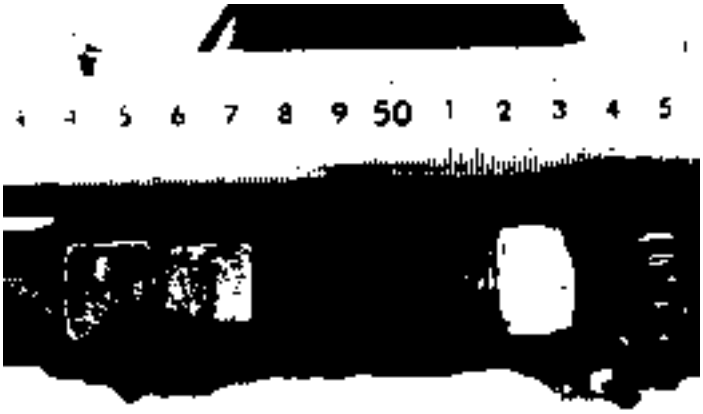


Fig. 16 Galvanic attack at a water pipe joint involving dissimilar metals

Figure 17 shows a 316 stainless tube that had carried cooling water in a heat exchanger. The outside had been in contact with oil entering the heat exchanger at approximately 80 °C (175 °F). Attack took place on the outside, as shown in Fig. 18, which is a metallographic section through the wall. In this particular case, the pit is long and has branched within the wall; it is common for pits to enlarge in a subsurface manner, often without significant evidence of this being apparent on the outer surface. The cause was contamination of the oil with city water containing in excess of 8 ppm chloride. The water side, which was not contaminated, gave no problems. It is of interest that the stainless tubing had been substituted for mild steel, which had a life of only approximately three years until perforated by general and pitting corrosion. The stainless tubes lasted only approximately two years before they perforated from pitting and subsequent stress-corrosion cracking. Determination of the direction of pitting, that is, from the outer or inner side of a wall or tube, is of importance in determining the origin of the failure and may involve metallographic sectioning.

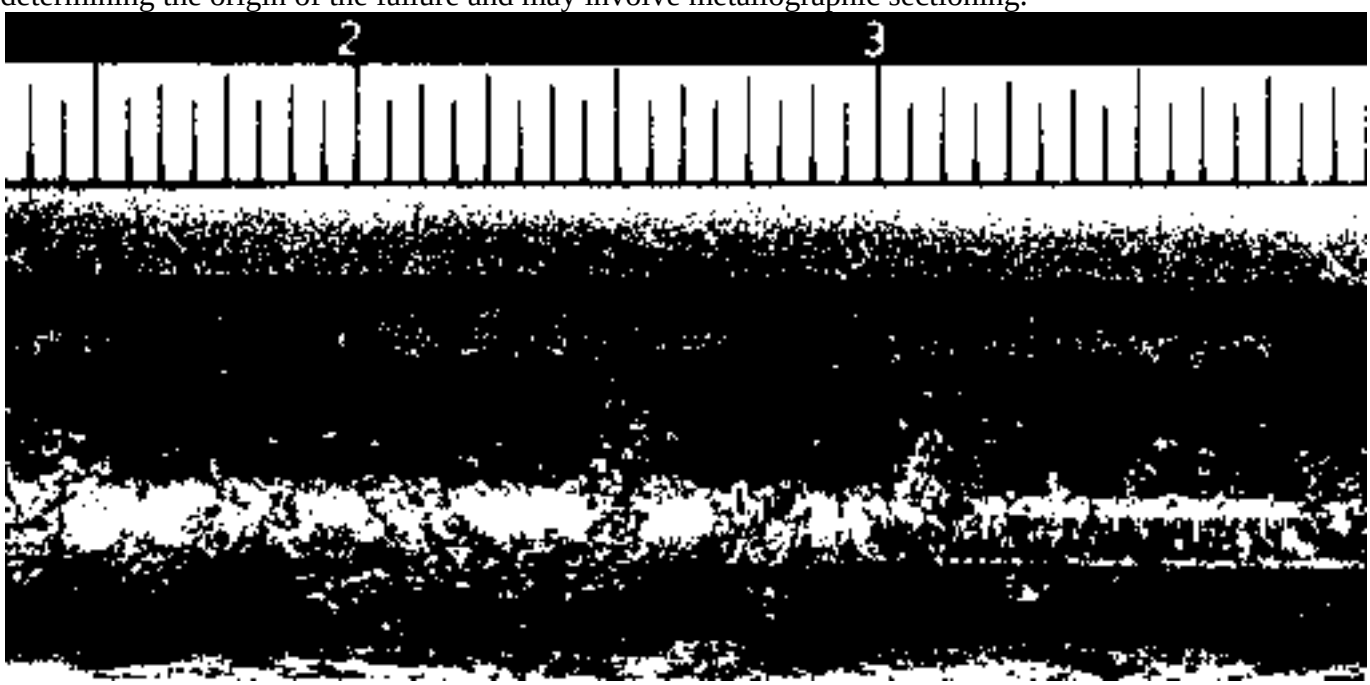


Fig. 17 A stainless steel 316 tube from a heat exchanger

Fig. 18 Attack from the outside surface of the tube shown in Fig. 17

Figure 19 shows an example of pitting attack on a pipe of 0.1% C steel that carried gasoline.

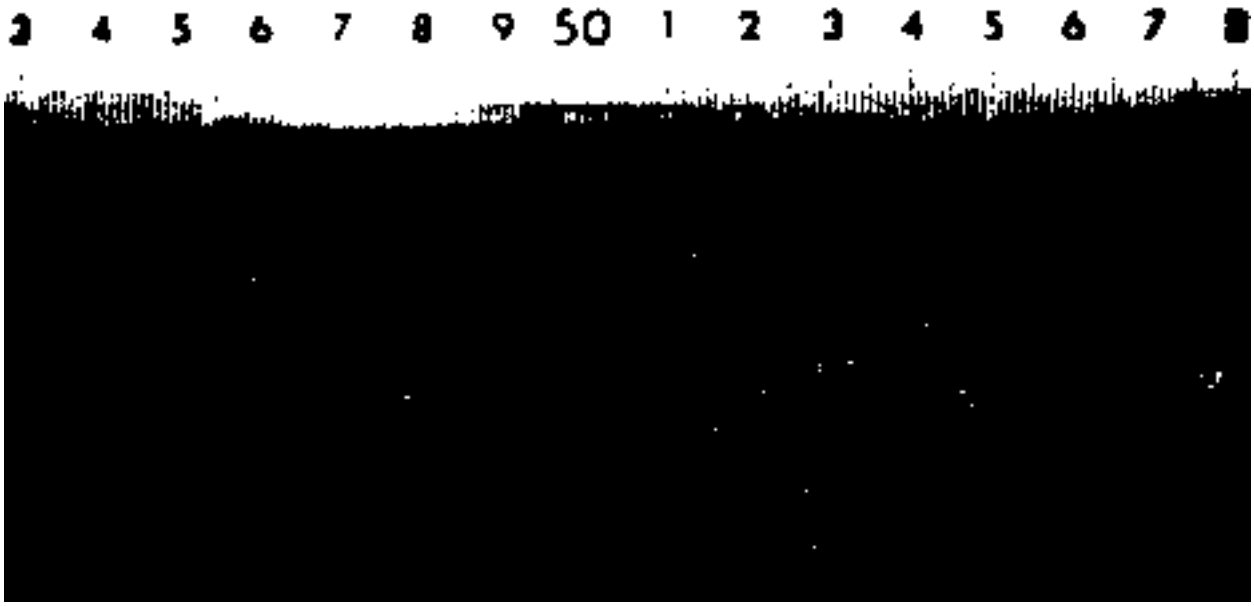


Fig. 19 Pitting and perforation on the outside of a carbon steel pipe

Corrosion was caused by inadequate drainage from the fill around the pipe and by dissimilar soil environments above and below it. The soil beneath the pipe was consolidated, while that above was loosely packed. This led to differential aeration and attack at the anodic (consolidated) region at the bottom of the pipe. This example

emphasizes the importance of soil sampling and of examining the conditions surrounding a component that has failed by corrosion.

Widespread pitting on a copper tube from a heat exchanger is shown in Fig. 20. The heat exchanger was installed in a refrigeration plant, which had some leakage of ammonia present. The tubes in the heat exchanger failed by pitting attack within approximately one year, because of the presence of the ammonia and condensed water vapor on the outside of the cooling tubes.

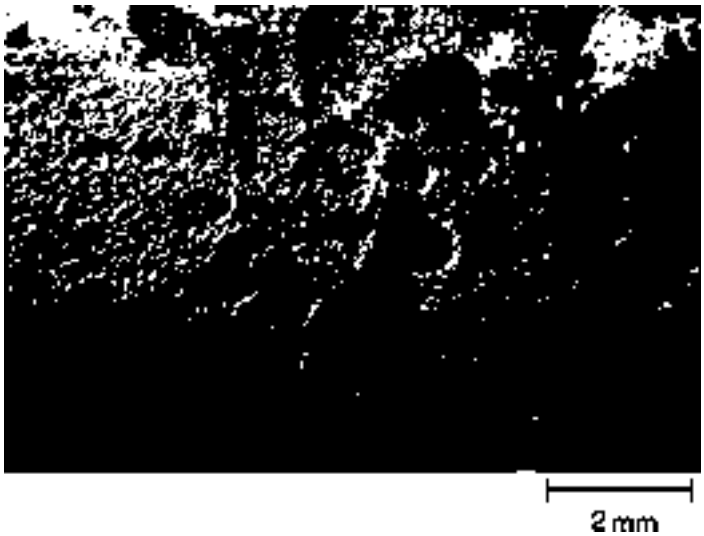


Fig. 20 Pitting on the outside of a copper heat exchanger tube

Features of Wear and Erosion Damage. Wear and erosion involve loss of material. This may be, for example, because of the absence of adequate lubrication, the rubbing together of components that are supposed to have clearance between them, or from the handling of abrasive materials that impinge on the component that continually removes surface material. In many cases, the presence of debris buildup may provide critical understanding of the specific wear/erosion mechanism involved in the failure.

A common wear problem that is encountered relates to bearings, either roller-element bearings or plain bearings. Features to be looked for include the nature of the damage, for example, overall wear or scores in the bearing surfaces. In the latter case, it is particularly important to identify any hard particles that may be embedded and trapped at the ends of the score marks.

Where unexpected wear between components in relative motion is encountered, and lubrication is observed to have been adequate, it is important to look for possible contaminants. These may arise from some other component in the system (for example, through the loss of hard chromium plating particles or through the development of corrosion products that are then transmitted through the system in circulating lubricant). Other circumstances that lead to premature wear include deliberate sabotage, where, for example, metallic particles or nonmetallics such as fine sand have been deliberately introduced into a lubrication system. These are not circumstances that the failure analyst usually considers, but they do occur from time to time. Such materials can be identified both microscopically and chemically, and careful documentation and continuity of evidence is very important in such circumstances.

A special case of wear is fretting, in which contacting surfaces are involved in vibration or very small relative motion at high frequency. Generally, corrosion is also involved, because the small particles of material that are removed expose fresh material during the small-amplitude motion oxidation. A telltale sign of fretting in materials that oxidize is the development of local discoloration at the contacting surfaces, there being a reddish color with the products of fretting on steel or cast iron surfaces and blackish stains on aluminum alloys. A major risk when fretting occurs is the development of surface cracks that might then grow by fatigue.

Energy-Dispersive Spectroscopy and Other Microchemical Analysis Procedures. In the evaluation of the causes of corrosion and wear, the determination of the chemical composition of corrosion products, debris, and contaminants is often of crucial importance. In some cases, the contaminants may be large or in bulk form, such as soil from a trench, but in many cases, they are small and may require microchemical analytical procedures. The local composition of components that have interacted with their environment may also require determination. In some cases, this may require the determination of the composition of chemically generated films and the variation in the composition of the underlying base material that has reacted with the

environment. In other cases, such as liquid metal embrittlement, it may require profiling the contamination within the bulk material or determining the composition of thin films on grain boundaries or on fracture surfaces.

Where failure has taken place in some specific direction, it may be necessary to conduct chemical analysis of specific regions. Examples are intergranular fracture, where there may be embrittling elements or phases on grain boundaries, or fracture along aligned inclusions or second-phase particles. In such cases, chemical analysis may be made using specialized techniques. A number of analytical tools are available for such evaluations and may be used by the failure analyst. The more common ones, which are described briefly in this section, are x-ray analysis by energy-dispersive spectroscopy (EDS), electron probe microanalysis (EPMA), Auger electron spectroscopy (AES), secondary ion mass spectroscopy (SIMS), and x-ray powder diffraction (XRPD).

Energy-dispersive x-ray analysis is normally conducted using a SEM fitted with an x-ray detector that can simultaneously count and determine the energy of emitted x-rays. Such an energy-dispersive detector contrasts with the wavelength-dispersive spectrometers used on the original electron probe microanalyzer developed by Castaing in 1948. The x-rays are generated when a sample of the area of interest is bombarded with a beam of electrons. Light elements to boron (atomic number 5) can be detected, and semiquantitative analysis can be made for atomic numbers of 11 (sodium) and higher. This is the most generally used microchemical analysis technique in failure investigations. A rapid qualitative analysis can be made in this way. The detection limits of elements are typically on the order of 0.1 wt%.

Figure 21 shows the analysis made of deposits within a crack in a pipe in a petrochemical plant.

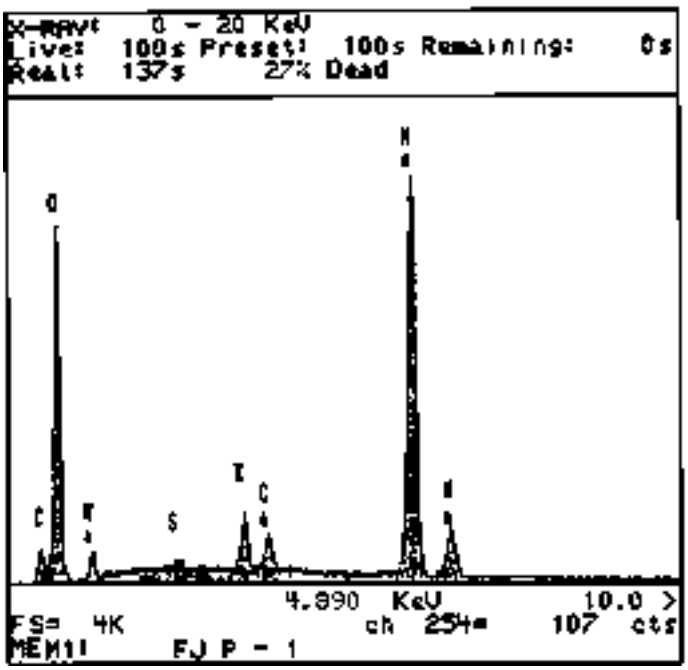


Fig. 21 Energy-dispersive spectroscopy analysis of deposit in a crack in a pipe from a petrochemical plant

Electron Probe Microanalysis. The electron probe microanalyzer is intended to combine the capabilities of the electron microscope with microanalytical capabilities. While the original instruments had an x-ray detector and single-crystal wavelength-dispersive diffractometers that moved around the sample to detect x-rays of various wavelengths, the development of energy-dispersive x-ray spectrometry detectors has permitted high-resolution detection of the composition of selected areas of the specimen. Although EDS is normally conducted using a SEM today, the electron probe microanalyzer is still widely used, and the use of wavelength-dispersive spectrometers provides much greater spectral resolution, allowing the separation of peaks that are poorly resolved with an energy-dispersive spectrometer. This is illustrated in Fig. 22.

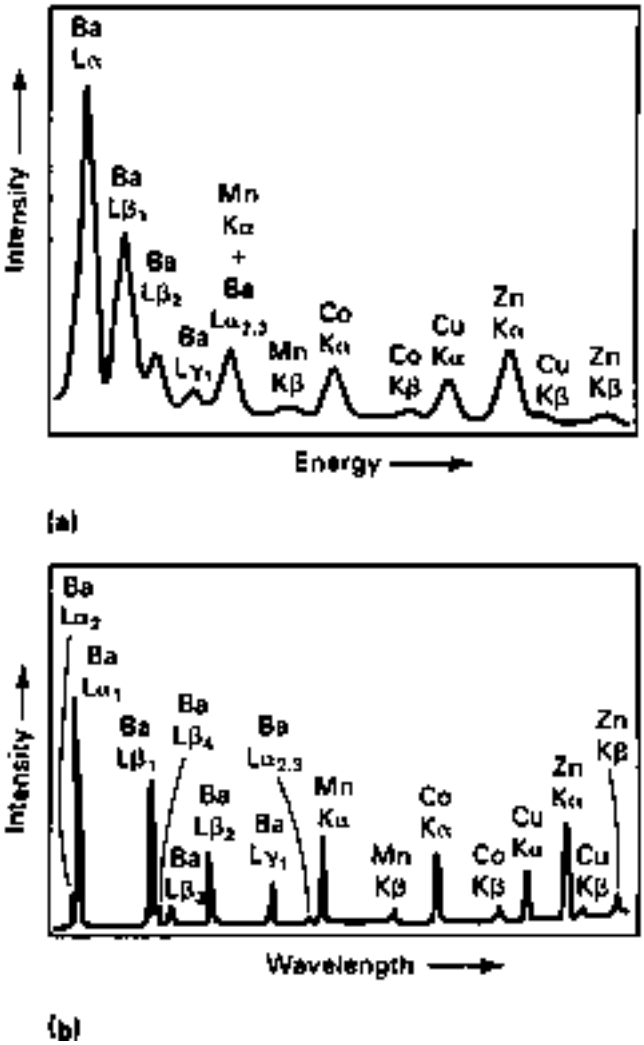


Fig. 22 (a) Energy-dispersive and (b) wavelength-dispersive x-ray spectra of a multicomponent glass. Source: Ref 5

The detection limits of EPMA using a wavelength-dispersive spectrometer are on the order of 0.01 wt%, an order of magnitude greater than is obtained using an energy-dispersive spectrometer.

Auger Electron Spectroscopy. In AES, an electron beam is focused on the area of interest on the sample being examined. Weak Auger electrons resulting from a secondary reaction of x-rays with the orbital electrons are detected. The electrons have energies that are characteristic of the elements that emit them.

The great advantage of AES over EPMA is the sensitivity of the analysis in terms of the depth of material analyzed. Using an electron probe microanalyzer, the depth of the spot analyzed can be on the order of 1 μm (0.04 mil), whereas with AES, the depth can be on the order of 3 nm, and all elements can be detected in this near-surface region, with the exception of hydrogen and helium.

Thus, the technique is ideal for evaluation of contaminated surfaces and thin coatings. Depth profiling of composition can be accomplished by step-wise progressive removal of material from a surface using an ion beam, normally of argon ions, followed by Auger analysis.

Secondary ion mass spectroscopy techniques are used to detect secondary ions created by the interaction of an ion beam with a material. Thus, when an ion beam, as used in AES analysis for depth profiling, impinges on a surface, the secondary ions can be detected using a mass spectrometer. A quadrupole mass spectrometer can be installed on an existing Auger system fitted with an ion etching gun, and simultaneous AES and SIMS analyses can be made. Stand-alone SIMS systems can provide improved performance as a result of a smaller ion beam diameter and greater sensitivity of detection.

Secondary ion mass spectroscopy analysis is capable of detecting all elements, including hydrogen, with a depth resolution of 5 to 10 nm, and the detection sensitivity of trace elements is in the parts per million to parts per billion range.

X-ray powder diffraction is an important tool in the evaluation of the composition of oxides and other contaminating crystalline material. The material must be removed and crushed to a powder, if it is not already in this condition. It is then placed in a glass capillary tube or consolidated on a sample holder and analyzed using a diffractometer, a Debye-Scherrer powder camera, or other camera and film technique to record the diffraction data produced when the crystalline material is placed in a monochromatic x-ray beam. The diffraction pattern obtained is recorded by the diffractometer or on film, and the interplanar atomic (*d*) spacings can then be calculated and compared with data for known compounds, as contained in the database maintained by the Joint Committee on Powder Diffraction Standards International Center for Diffraction Data in Swarthmore, PA.

References cited in this section

4. M.G. Fontana, *Corrosion Engineering*, 3rd ed., McGraw-Hill, 1986

5. *Materials Characterization*, Vol 10, *ASM Handbook*, ASM International, 1986, p 521

Iain Le May, Metallurgical Consulting

Analysis of Base Material Composition

Metals. In general, the composition of the base material of the subject part or component needs to be determined. This does not necessarily require chemical analysis to be made, because there may not be a question about the importance of the part conforming to the specified composition. However, it is important that the specified composition, or at least the material specification, be determined, whether or not chemical analysis is to be made. First, there is the question of whether the specification was sufficiently complete and was appropriate for the application. For example, in the case of a structural steel part, the specification would normally be for a structural-quality steel rather than a commercial-quality steel in order to meet requirements of the applicable construction code. It is not unusual to find that the specification and ordering of material has been done in a loose and unspecific manner. Other examples are the specification of “carbon steel” or “mild steel,” which allows for wide ranges of materials to be used.

Another reason for the need to determine exactly what was specified is that analysis can then be made of the key elements in the specification, rather than an unnecessarily wide spectrum of elements. Also, when a composition is determined, it is often found that this allows the material to be classified as corresponding to a range of specifications, and it is not possible to be specific about the exact grade or type of material. All specifications have a range for each element, in some cases this being given as a maximum value only. Additionally, there are variations within a product itself, particularly in the case of cast material, and the analysis has limitations on the accuracy and precision of the elements detected. Where there is a doubt about a particular analysis, it may be necessary to have a separate duplicate analysis made to see if the two are consistent. There are sometimes consistent differences in the analyses made by different laboratories, and not too much importance should be given to apparent small differences in composition or to some elements being apparently just outside the specified range. If the element is a critical one, then duplicate or multiple analyses may be justified.

Where failure has taken place in some specific direction, it may be necessary to conduct chemical analysis of specific regions. Examples are intergranular fracture, where there may be embrittling elements or phases on grain boundaries, or fracture along aligned inclusions or second-phase particles. In such cases, chemical analysis may be made using specialized techniques involving the use of an electron probe microanalyzer or energy-dispersive x-ray analysis in a scanning electron microscope, which are discussed in the section “Energy-Dispersive Spectroscopy and Other Microchemical Analysis Procedures.”

Polymers. Chemical analysis, as is normally done for metals, is not sufficient for the characterizing of polymeric materials. The main intrinsic properties that characterize all polymer molecules are their size, weight,

and weight distribution; chain stiffness or chain rigidity; the particular state, whether amorphous, semicrystalline, or crystalline; and the nature of any network structure. These elements, by means of which all polymers can be classified, are illustrated in Fig. 23 and described with examples in Table 1. Thermoplastics should, in theory, have no cross linking, but, in practice, small amounts occur as a result of thermal and oxidative degradation or through the processing of the polymeric material in fabricating the component.

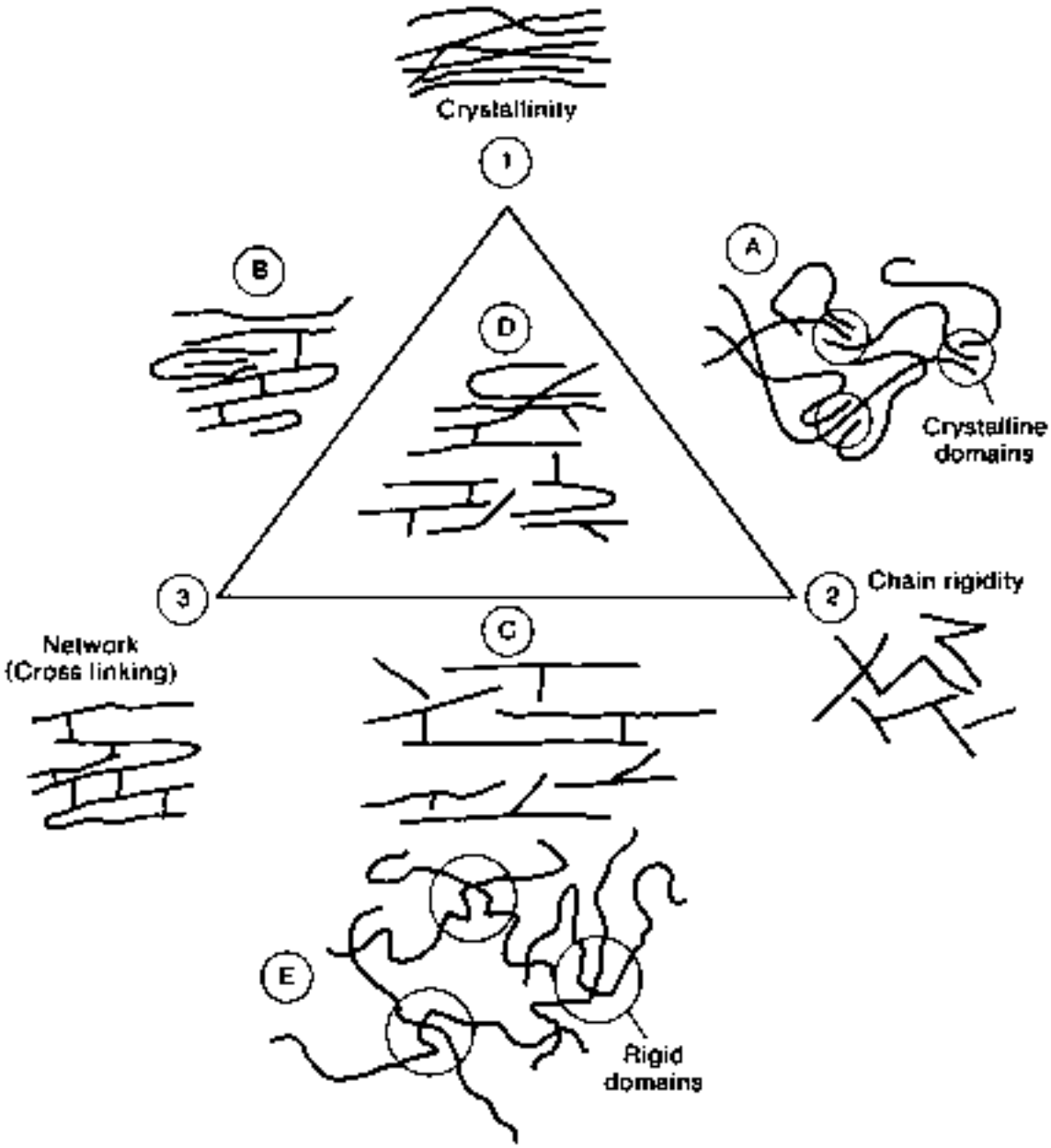


Fig. 23 Basic elements of polymeric materials. (See Table 1 for explanation of locations). Source: Ref 6

Table 1 Basic characteristics of engineering polymers

Location[(a)]	**Characteristics**	Examples
1	Flexible and crystallizable chains	**PE** **PP** **PVC** **PA**
2	Cross-linked amorphous networks of flexible chains	**Phenol-formaldehyde cured rubber**

		Styrenated polyester
3	Rigid chains	**PIs (ladder molecules)**
A	Crystalline domains in a viscous network	**PET** **Terylene (Dacron)** **Cellulose acetate**
B	Moderate cross linking, with some crystallinity	**Chloroprene rubber** **Polyisoprene**
C	Rigid chains, partly cross linked	**Heat-resistant materials**
D	Crystalline domains with rigid chains between them and cross-linking chains	**High-strength and temperature-resistant materials**
E	**Rigid-chain domains in a flexible-chain matrix**	**Styrene-butadiene-styrene, triblock polymer** **Thermoplastic elastomer**

(a) See Fig. 23. PE, polyethylene; PP, polypropylene; PVC, polyvinyl chloride; PA, polyamide; PI, polyimide; PET, polyethylene terephthalate.
Source: Ref 6

Characterization of polymers in failure analysis may involve the following methods:

- Molecular spectroscopy.
-
    1. Infrared (IR) or Fourier transform infrared (FTIR) spectroscopy
    2. Nuclear magnetic resonance (NMR) spectroscopy
- Molecular weight (MW) and molecular weight distribution (MWD) using light scattering as well as other secondary methods
- Methods of thermal analysis
-
    1. Differential scanning calorimetry (DSC) and differential thermal analysis (DTA)
    2. Thermomechanical analysis
    3. Dynamic mechanical analysis
    4. Thermogravimetric analysis
- X-ray diffraction (XRD) analysis

Further details of polymer characterization are discussed in the article “Characterization of Plastics in Failure Analysis” in this Volume.

A brief scheme of structure analysis as it relates to material failure is presented in Table 2. A typical problem a materials engineer must face is the evaluation of a piece of failed pipe. Identifying the type of molecule used to make the pipe material can be accomplished by IR spectroscopy or spectroscopic methods in general. With the chemical structure of the pipe established, the next property to address is the glass transition temperature, T_g, and the melt temperature, T_m, because these quantities characterize the useful working temperature range of the material. Either DSC or thermomechanical analysis (TMA) can be used to determine both T_g and T_m as well as other thermal events. Repeat analysis can be used on a single sample to reveal the effects of thermal or process history on the transition temperatures. The next step is to determine the MW and/or MWD of the polymer. A primary method (such as light scattering) or secondary methods (such as dilute solution viscosity, gel permeation chromatography (GPC), or melt rheology) can be used to characterize the MW and MWD. Differences in MW and MWD can have a profound effect on ultimate properties. Wide-angle x-ray diffraction as well as electron microscopy can determine the morphology or structure (amorphous or crystalline) of the material. This problem-solving approach can be extended from failed pipe to a polymeric structural member that failed under load, or a plastic gear that cracked under conditions of use, or to any other failure problem.

Table 2 Practical information derived from polymer analysis methods

Test method	Property
Gel permeation chromatography (GPC)	**Weight-average molecular weight, $\overline{M}_w$; molecular weight distribution (MWD)**
Low-angle light scattering	**$\overline{M}_w$, MWD**
Osmometry, membrane	**Number-average molecular weight, $\overline{M}_n$**
Osmometry, vapor pressure	**$\overline{M}_n$**
Dilute solution viscosity	**Inherent viscosity, reduced viscosity, intrinsic viscosity**
Quasielastic light scattering (QELS)	**Macromolecular particle diameter, diffusion coefficient**
Differential thermal analysis (DTA)	**Transition temperature, T_g; melt/crystallization temperatures, T_m**
Differential scanning calorimetry (DSC)	**Heat of polymerization, fusion, T_g, T_m**
Thermogravimetric analysis (TGA)	**Composition, weight loss with time or temperature**
Dynamic mechanical analysis (DMA)	**Elastic modulus, loss modulus, tan delta**
Thermomechanical analysis (TMA)	**Penetration temperature, expansion coefficient**
Mechanical spectroscopy	**Viscosity, normal stress difference, shear elastic and loss modulus**
Infrared (IR) spectroscopy, Fourier transform infrared spectroscopy (FTIR)	**Chemical functional groups**
Attenuated total reflectance (ATR)	**Surface functional groups**
Nuclear magnetic resonance (NMR) spectroscopy	**Chemical shift of nuclei**
Mass spectroscopy	**Mass/charge of ions produced**
X-ray diffraction analysis (XRD)	**Crystalline polymer component**
Small-angle x-ray diffraction	**X-ray scattering at low angle**
Scanning electron microscopy (SEM)	**Surface and particle morphology**
Transmission electron microscopy (TEM)	**Polymer morphology**
X-ray photoelectron spectroscopy (XPS)	**Elemental concentrations, oxidation states**
Auger electron spectroscopy (AES)	**Elemental concentrations**
Wavelength dispersive x-ray analysis	**Elements present on polymer surfaces**

Source: Ref 6

Ceramics/Intermetallics. Bulk chemical analysis of ceramic and intermetallic components that have failed may be made in several ways, including wet analytical chemistry, ultraviolet/visible absorption spectroscopy, and molecular fluorescence spectroscopy. Additionally, x-ray diffraction analysis is particularly valuable in identifying constituent compounds.

Other specialized techniques may be required to detect contamination and local compositional variations; these would use analytical techniques already described in the section "Energy-Dispersive Spectroscopy and Other Microchemical Analysis Procedures."

Reference cited in this section

6. A.T. Riga and E.A. Collins, Analysis of Structure, *Engineering Plastics*, Vol 2, *Engineered Materials Handbook*, ASM International, 1988, p 824–837

Metallurgical Samples and Hardness

Selection of Metallurgical Samples. Samples should not be taken from subject components until all information required has been obtained while still in place. For example, if there is a fracture surface this should be examined and recorded, as required before any cuts are made through it. In many cases, it is of interest to section a component through the fracture origin, but this origin is then effectively destroyed, as far as viewing the local surface of the fracture is concerned. Thus, cutting should be delayed until all required fractographic or surface chemical analysis has been made. In the case of a failure where legal action is contemplated, it is important that the other parties involved have an opportunity to examine the component before such sectioning is done. It may be necessary, in such cases, that the other parties are in agreement with a documented protocol of sectioning and examination procedures to be followed.

When samples are to be taken from a subject part or component, it is important to record exactly where they were taken from, including their orientation. Photographs of the part before and after sectioning are valuable. It can be very useful to mark the position of the cuts to be made on the part before it is sectioned or photographed. Sketches documenting cutting sequence and providing unique identification to resultant sectioned pieces are very valuable, such as the one shown in Figure 24.¨

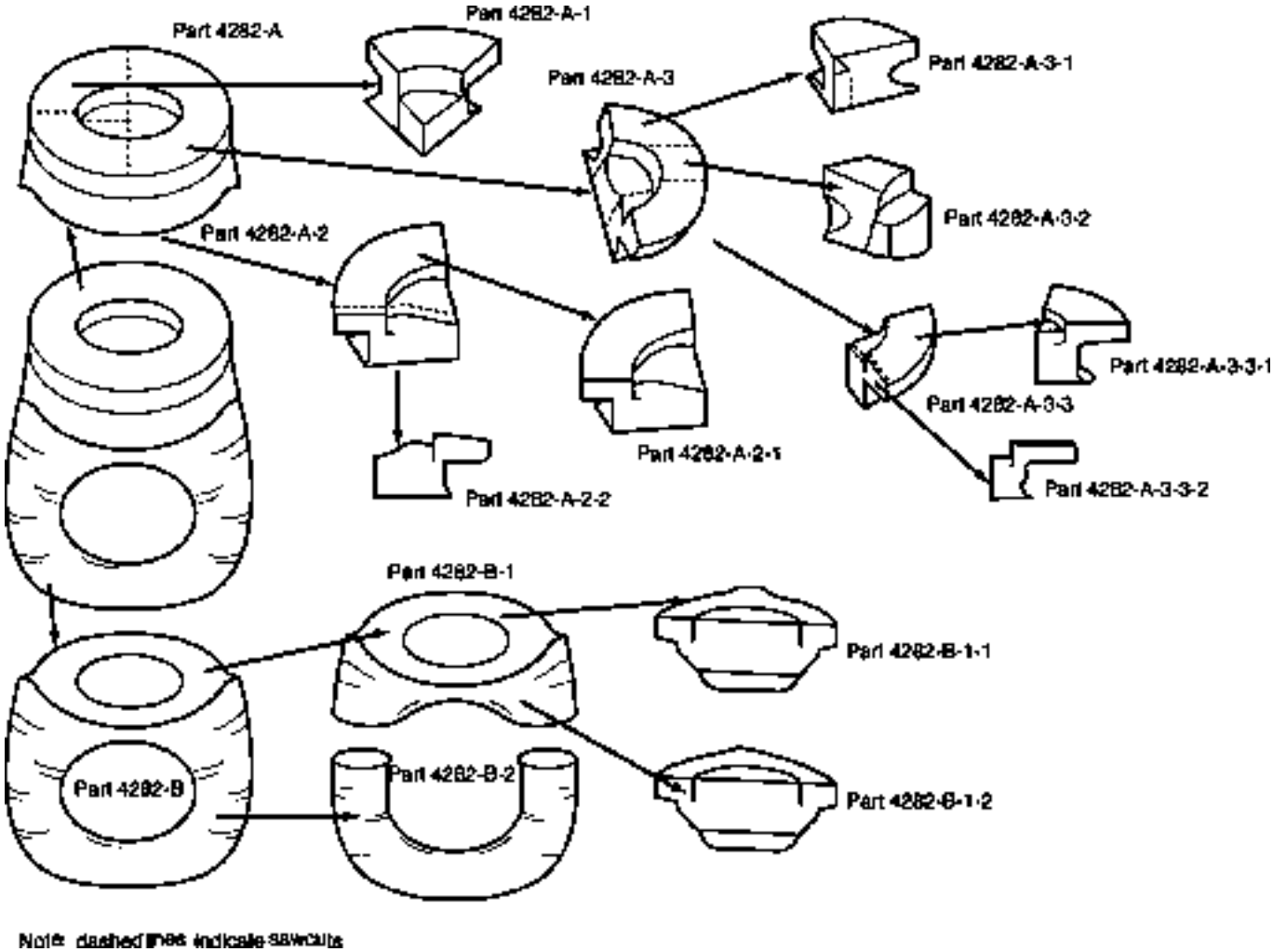


Fig. 24 Sketch showing cutting sequence for a failed trailer hitch. Parts 4282-A-2-2 and 4282-A-3-3-2 were mounted for metallographic examination.

In addition to samples taken from a subject component, it is often important to have samples taken in regions away from the area of failure or from similar components that have not failed. These serve to show the microstructure of the component in other than the damaged region and of similar components, so that any microstructural aspects that are specific to the subject component and, in particular, to the damaged region can be identified.

Hardness Determination by Microindentation and Nanoindentation Methods. Hardness measurements can provide considerable information about the mechanical properties of materials. Many specifications provide a

range of hardness, and this quantity can be correlated reasonably well with tensile strength in many cases, particularly for steels. Hence, hardness is one of the properties that is routinely measured during a failure analysis.

During the metallographic portion of a failure analysis, microindentation measurements of hardness can be informative. Examples include microhardness profiles to determine if specified effective case depth of gear teeth was achieved, and hardness in weld, base, and heat-affected zone material to identify if weld hardness properties were consistent with expected values. To answer such questions, a microindentation hardness tester with a diamond indenter is commonly used, with loads in the range of 10 g to 1 kg. Figure 25 shows a plot of hardness across the carburized surface layer of a gear tooth that had been sectioned following failure of a gearbox assembly. The effective case depth is the distance from the surface at which the case hardness curve falls to a particular predetermined value, often 50 HRC, as indicted on the figure. The indenters used in microindentation hardness testing can be of the Knoop or Vickers type. Both are diamond pyramid type. The Vickers indenter is of the same geometry as that used in conventional Vickers testing, being square in cross section, while the Knoop indenter has a rhombic base, producing a rhombic-shaped indentation with a ratio between the long and short diagonals of approximately 7 to 1. The shape of the indentations is illustrated in Figure 26.

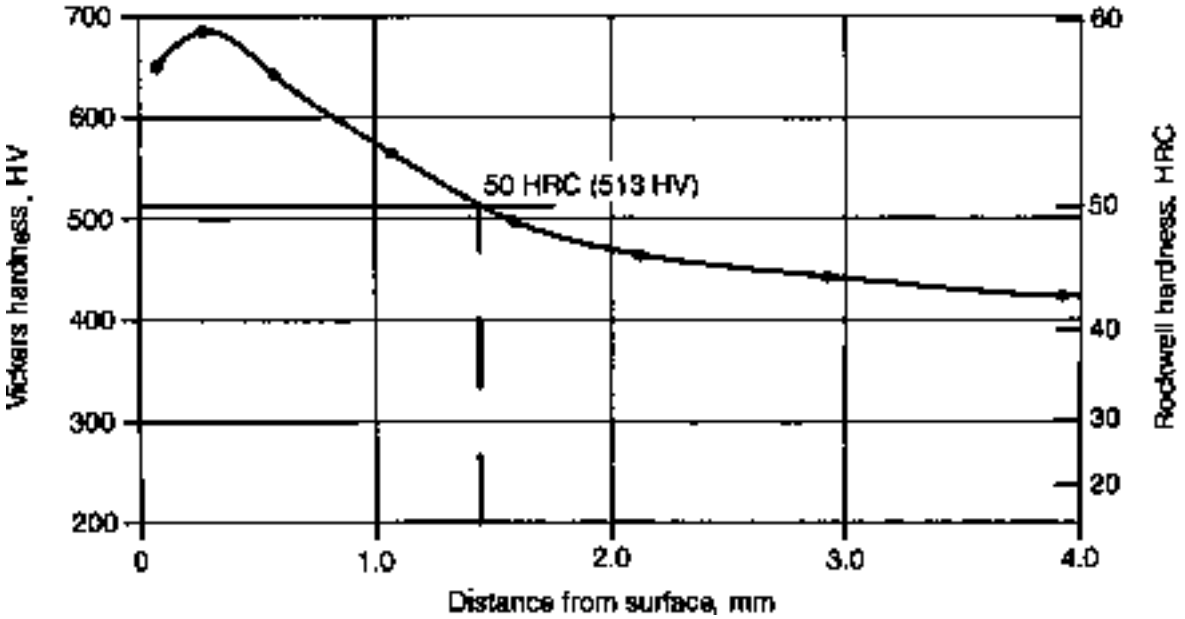


Fig. 25 Plot of hardness across the carburized layer of a gear tooth made by using a micro-indentation hardness tester with a Vickers indenter. The equivalent Rockwell C hardness is shown on the right. The effective depth of hardness is indicated by the broken line cutting the hardness plot at a level of 50 HRC.

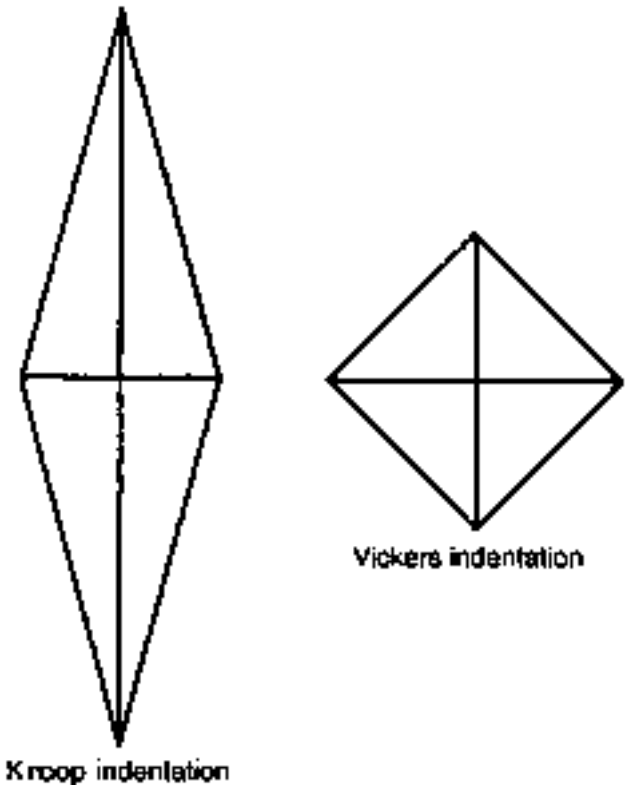


Fig. 26 Comparison of indentations made by Knoop and Vickers indenters in the same metal and at the same loads. Source: Ref 7

The choice between indenter types is often arbitrary, but the Vickers indenter is less sensitive to surface conditions, because it penetrates more deeply. With its shorter diagonals, the Vickers indenter is more sensitive to measurement errors. However, because both diagonals are measured and then averaged for the Vickers indenter, it is less subject to error when the surface is not completely level.

In recent years, the development of the ultrasonic microindentation hardness tester has provided another very useful tool for the failure analyst. This type of hardness tester uses a Vickers diamond attached to one end of a magnetostrictive metal rod. When the diamond-tipped rod is excited to its natural frequency by a piezoelectric converter and the free end with the diamond tip brought into contact with the surface of a solid body, the natural frequency of the rod changes. The area of contact between the diamond tip and the material under test can be derived, because the resonant frequency depends on this for a given modulus of elasticity. The area of contact is also inversely proportional to the hardness, provided the contact force is constant. The loads used are small, with a maximum of approximately 10 N, and the resulting indentations are small (on the order of 4 to 20 μm) so the testing is essentially nondestructive. Because the commercially available equipment is light and portable, and measurements and calculations are performed electronically when the instrument is brought into contact with the test material, ultrasonic microindentation is extremely useful in checking the hardness of failed or damaged components; it takes little time and does not involve cutting up parts in order to make the tests. One warning should be given, however: because the indentations are small, they are affected greatly by the surface condition, such as the presence of local cold working and by surface roughness. Thus, local preparation using grinding papers and, in some cases, local polishing may be necessary to achieve results that are representative of the underlying material.

Provided that the microindentation hardness tester is well set up, it is also possible to determine the hardness of second-phase particles, provided they are of sufficient size. Where this is impractical, a useful technique is to use a fine needle to scratch the surface of the metallographic specimen, passing across particles or areas where the hardness is in question. Although no absolute values of hardness can be determined in this way, the variation in the width of the scratch as it traverses different phases or second-phase particles can provide useful information as to their relative hardness.

The development of new surface-modification techniques and coatings in order to improve the service life of components in tribological applications has produced a need to measure the hardness of the surface layers, and it has become necessary to determine changes in hardness over depths on the order of 10 nm. Hardness

measurements on a nanometer scale are thus required when components that have surface-modified layers or coatings are to be assessed.

Depth-sensing nanoindentation testing can be used to characterize the mechanical properties of a surface layer. A diamond tip is pushed into the surface while the load is increased, and this is continued to some maximum depth or load, the load and displacement of the indenter being recorded. The record of load versus displacement provides information on the mechanical properties of the material over a volume whose dimensions depend on the material properties and the shape of the diamond tip. As the load increases, the volume and the depth that are being probed increase strongly. To obtain a profile of hardness versus depth, the maximum load can be increased, but the resolution of the load variation method (LVM) decreases significantly as the depth of penetration increases. This is illustrated in Fig. 27(a), from which it may be observed that the measurement of the response is for the combination of film and substrate for thin films or large loads. Although analytical methods have been developed to compensate for this, these do not help when the measurements are made on multilayers or ion-implanted surfaces in which there is a gradient of mechanical properties.

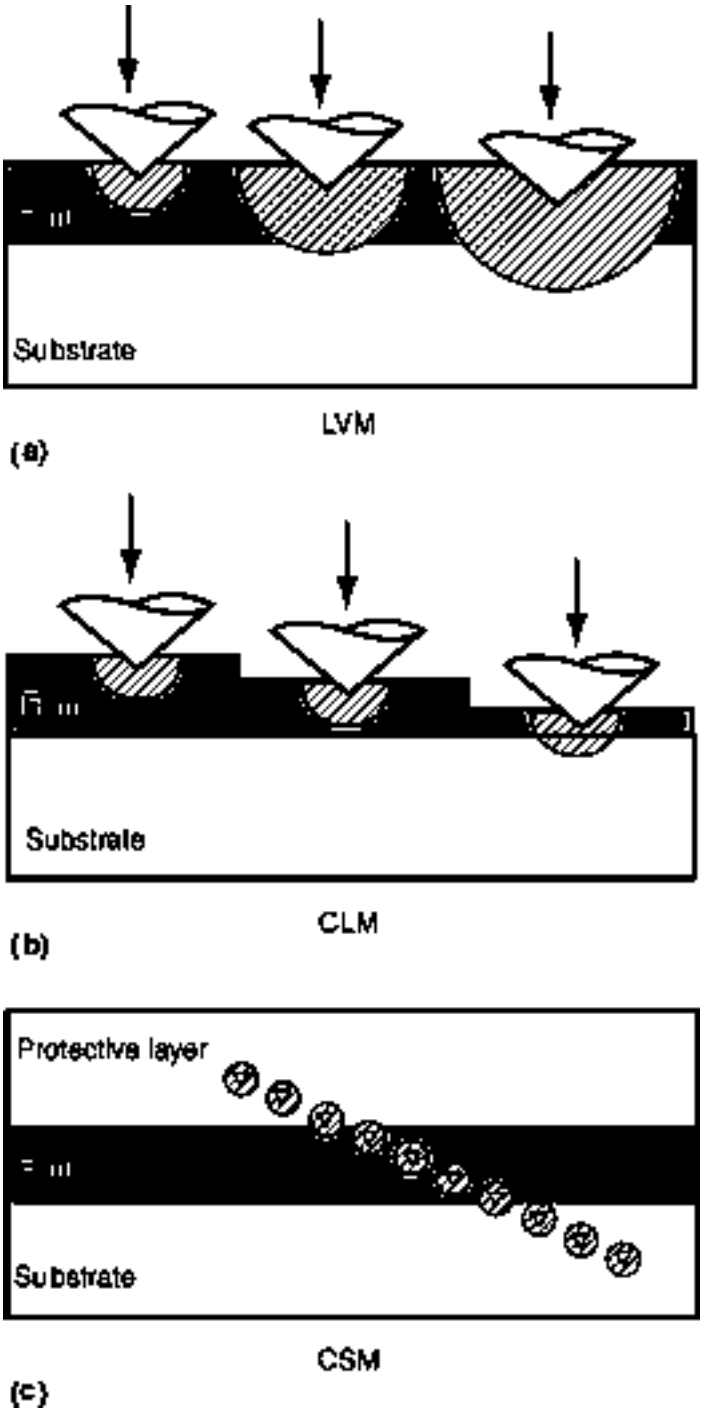


Fig. 27 Schematic presentation of different methods of hardness-depth profiling. (a) Load variation method (LVM). (b) Constant load method (CLM). (c) Cross-sectional method (CSM). The cross-hatched areas indicate the relative dimensions of the probed volumes, for which a hemispherical shape is assumed. Source: Ref 8

Alternate methods to determine hardness gradients are the constant load method (CLM), which involves removing successive thin layers of material and performing hardness indentation testing at each step (Fig. 27b), and the cross-sectional method (CSM) in which a cross section is prepared, as is common in metallurgical evaluations (Fig. 27c).
In nanoindentation hardness testing, the hardness is evaluated from the recorded load-displacement curve, the hardness being defined as the mean contact pressure under the indenter at maximum load.

References cited in this section

7. *Hardness Testing,* 2nd ed., ASM International, 1999, p 65

8. M. Kunert, B. Baretzky, S.P. Baker, and E.J. Mittemeijer, in *Metall. Mater. Trans. A,* Vol 32, 2001, p 1201–1209

Examination of Damage and Material Evaluation

Iain Le May, Metallurgical Consulting

Preparation and Examination of Metallographic Specimens in Failure Analysis

It has already been emphasized that the cutting sequence involved in the preparation of metallographic specimens must be well documented. The mounting medium used for these must be chosen carefully. When edge retention is important, as it usually is when a metallographic specimen is prepared through a fracture origin, a sample can be mounted using a hard mounting material or it can be plated electrolytically or using electroless solutions before being mounted. In general, electroless plating using nickel, copper, iron, or chromium is the preferred method, because bonding is superior, with less tendency for the plating to separate than with electrolytically plated specimens. Hard mounting materials for edge retention generally contain pelletized alumina (Al_2O_3), ground glass, or other wear-resistant filler materials in the resin.
Where specimens are to be displayed to others, for example, in court proceedings, it is useful to mount them in transparent plastic, so the complete piece of material can be seen. This allows others to see more clearly the shape of the mounted sample and to appreciate exactly the location where it was taken from a component. However, the edge retention is often inferior with transparent mounting compounds, although some relatively satisfactory proprietary materials are available.
If SEM evaluation is required for nonmetallic specimens, it is necessary to coat these with a conducting layer of carbon, gold, or gold-palladium alloy by means of vacuum deposition. In the case of metallic specimens, these may be either coated or grounded to the holder using conducting tape or metallic silver paint to prevent charging from occurring on the specimen. All specimens must be grounded to prevent this charge buildup.

Ceramic Materials*

Ceramics are inorganic compounds, such as oxides, carbides, nitrides, borides, and mixed ceramics, and are usually produced by sintering. The often-extreme hardness, brittleness, porosity, and chemical resistance of ceramic materials lead to a number of difficulties when preparing representative polished and etched samples.
Careful selection of cutting, mounting, grinding, polishing, and etching procedures is required and must be optimized for each type of ceramic. Automatic sample preparation is recommended, as are an adjustable

pressure, special grinding discs, and the use of diamond as the abrasive material. Using such methodology, a flat surface that displays an undistorted, real microstructure may be prepared in a reasonable time.

Cutting invariably introduces damage to the surface of ceramics. Grain pullouts, cracks, and cutting grooves may result, which have to be removed step-by-step during subsequent preparation (grinding and polishing). Generally, ceramics are wet-cut (water or a special cutting fluid) with a rotating diamond cutting wheel on a bench-type laboratory machine or on a precision cutting machine. The cutting speed (25 to 500 rpm for a low-speed cutting machine; 500 to 5000 rpm for a high-speed cutting machine) and the cutting pressure should be optimized, depending on the material properties.

Mounting. For automatic sample preparation, the specimens can be mounted or may be glued directly onto a sample holder. The two possibilities for metallographic mounting are hot mounting, with compression and heat, and cold mounting. It should be noted that sensitive, small, and very brittle ceramic specimens tend to become damaged and cracked when using hot mounting, because of the high pressure and temperature that are needed for this process. The mounting material should either be very hard or have good abrasion resistance. Additionally, care must be taken that the etching technique to be used is known before the sample is mounted, to enable correct selection of mounting material. For example, when using thermal etching techniques or salt melt etching techniques, the sample should be removed from the mount before etching. When etching in a boiling chemical solution, the mount should be of a material that is not chemically attacked.

Mechanical Preparation (Grinding and Polishing). The grinding and polishing procedure is best carried out on an automatic or semiautomatic machine. The structure of each and every ceramic product has been specifically adjusted to exhibit the required properties and thus shows a unique behavior when being prepared. Table 3 contains preparation standards for advanced (structural) ceramics. This table should be used as a rough guide only, and figures have to be adjusted according to the preparation requirements of specific ceramics.

Table 3 Guidelines for semiautomatic preparation of structural ceramics

Step	**Base**	**Diamond grade, µm**	**Lubricant**	**Speed, rpm**	**Pressure, N**	Time, min
Grinding	Diamond disc	65	Water	300	≈180	**Until flat**
	Diamond disc	20	Water	300	≈180	**5–10**
Fine grinding	Composite disc	6	Lubricant	300	≈100	**5–15**
Polishing	Fusible-backing nonwoven cloth	6	Lubricant	150	≈150	**15–120**
	Hard synthetic cloth	3	Lubricant	150	≈120	**15–120**
	Hard silk cloth	1	Lubricant	150	≈90	**5–10**
Fine polishing	**Short-napped fiber cloth**	**(a)**	**(a)**	**150**	**≈50**	**1–20**

(a) Silicon dioxide suspension

In general, phenolic-thermosetting-resin-bonded diamond discs are employed for grinding. The surface damage generated during grinding has to be removed during fine grinding and polishing. Fine grinding keeps the surface of the specimen flat, and no further damage is introduced. Complete removal of the damaged surface must therefore be achieved by subsequent polishing steps. Polishing should preferably be performed on hard cloths, and the highest removal rates occur during steps with the application of diamond polishing compound of 6 and 3 µm size. Polishing with 1 µm diamond removes only very little material. All breakouts and scratches are to be removed during these steps. In general, fine polishing with a suspension of SiO_2 alone is used if small and fine scratches have to be removed.

Ceramographic Etching. Microscopic examination of ceramic specimens in their as-polished condition prior to etching has proved useful. An evaluation of the number of pores, their distribution, and possible pullouts can only be assessed in this state. Evaluation of inclusions, contamination, and cracks should also be made before etching. In order to reveal grain boundaries, phases, and other microstructural details, the ceramic specimens have to be etched. The methods of etching used in ceramography are varied. All of the well-known and commonly used methods of ceramographic etching are summarized in Fig. 28 (Ref 9). Three etching methods can be distinguished: optical etching, which does not change the surface of the specimen, and electrochemical

and physical etching, both of which cause changes to the surface of the specimen. The most frequently used etching methods for advanced ceramics are shown in italics (solution etching, plasma etching, and thermal etching).

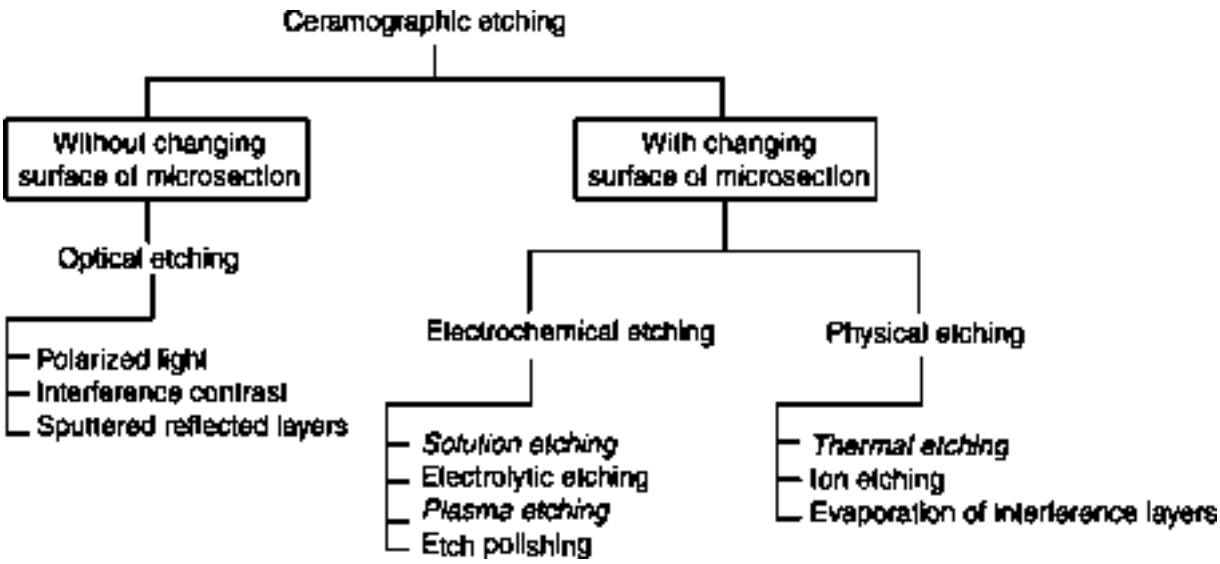


Fig. 28 Ceramographic etching procedures

Due to the chemical resistance of many ceramics, solution etching of polished specimens requires the application of very aggressive chemicals at elevated temperatures, for example, Al_2O_3 and Si_3N_4 require boiling phosphoric acid at 250 °C (480 °F).

In contrast to solution etching, plasma etching (which can be used only for SiC and Si_3N_4 ceramic materials) attacks the matrix grains of the ceramic and not the grain-boundary phase (Fig. 29) (Ref 10). Plasma etching is carried out inside a special plasma-etching apparatus, which provides a gaseous mixture of CF_4 and O_2. A radio frequency generator emits high-frequency electromagnetic oscillations and thus produces fluoride radicals inside the reaction chamber. These radicals then react with the silicon-containing matrix of the ceramics (SiC and Si_3N_4), and the intergranular grain-boundary phase remains unattacked by plasma etching.

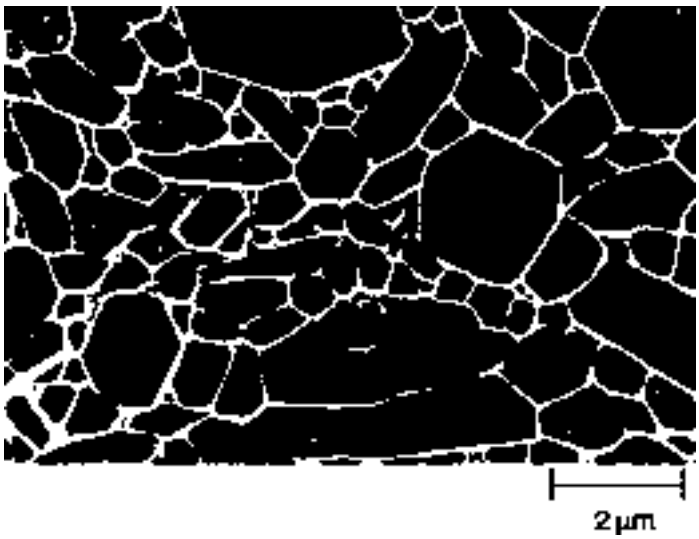


Fig. 29 Silicon nitride (Si_3N_4), plasma etched

Thermal etching in air or in vacuum is applied at temperatures approximately 150 °C (270 °F) below sintering temperature and is mainly carried out with ZrO_2 and Al_2O_3. Figure 30 shows a thermally etched Al_2O_3 ceramic. The grain boundaries are visible after thermal etching at 1500 °C (2700 °F) in air.

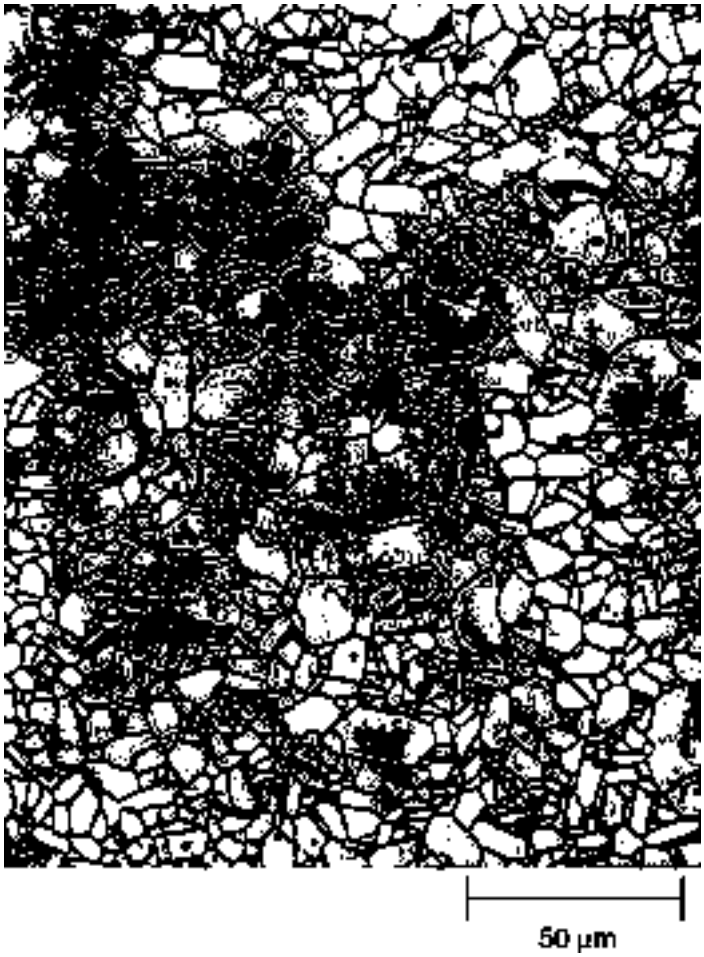


Fig. 30 Alumina (Al_2O_3) thermally etched at 1500 °C (2700 °F) in air

The choice of the etching method, of course, depends on the type of ceramics and on the structural feature of interest.

Footnote

* *This section contributed by Ulrike Täffner, Max-Planck-Institut für Metallforschung, Stuttgart, Germany.

References cited in this section

9. G. Petzow, *Metallographic Etching*, 2nd ed., ASM International, 1999

10. U. Taeffner, M.J. Hoffmann, and M. Kraemer, A Comparison of Different Physical-Chemical Methods of Etching for Silicon Nitride Ceramics, *Pract. Metallogr.*, Vol 27, 1990, p 385–390

Iain Le May, Metallurgical Consulting

Analysis and Interpretation of Microstructures

A sample of material from the subject component may be examined for several reasons. These include: determination if the microstructure is in accordance with that to be expected for the specified composition and processing of the part; evaluation of the type and extent of damage resulting from its service life, taking into account loading and environmental effects (including temperature); evaluation of welding or other joining

processes employed during fabrication; determination of the origin and the path of cracking and its possible relationship to microstructural features, such as inclusions or grain boundaries; estimation of the extent of overheating that may have occurred during service; examination of directionality in the microstructure and its possible relation to failure; and the distribution of phases within the material.

After metallographic specimens from the component of interest have been prepared by grinding and polishing, they should be examined in the unetched condition. This allows better observation of features such as inclusions, cracks, and pores and of foreign material, such as copper in steel or surface plating, that may not be observable after etching.

Metallographic examination allows the characterization of many parameters, whether or not it is being made to determine the cause of a failure. The microstructure of nonfailed components or samples of similar materials may be examined for the purposes of comparison, because the microstructural parameters may be very important in the overall analysis, even when they are not specific causes of a failure. Features that may be measured include grain size and shape, inclusion size and distribution, the distribution of second-phase particles, plating or coating thickness, and the extent and nature of a heat-affected zone produced by welding or another process. In an increasing number of cases, these parameters are determined using quantitative metallographic (or stereographic) techniques, as discussed in the article "Quantitative Metallography" in *Metallography and Microstructures,* Volume 9 of *ASM Handbook.* Recent development and cost-effectiveness of automated systems have greatly facilitated such analyses.

The microstructure of a metallic component is determined by its thermal and mechanical history. Generally speaking, cast and wrought products can be differentiated on the basis of their microstructure as well as their overall shape and external appearance prior to such examination. Segregation effects are much more likely to be present in cast products. Features such as long directional grains produced during a directional solidification or fine equiaxed grains produced at a chilled surface may be seen in cast components, and these may be relevant to the failure mechanism. In general, rolled, extruded, or otherwise mechanically worked materials and components are intended to have their maximum tensile stresses acting along the direction of deformation, because this is normally the direction in which the strength is greatest. Hence, it is not unusual to find failures that have occurred when a component is loaded in tension in another direction. An example of this is when a lug or other attachment is welded to a plate product, and then a load is applied in the through-thickness direction. A specific case involved cracking in steel beams with a minimum specified yield strength of 300 MPa (44 ksi) that had measured yield and tensile strengths in the transverse (not through-thickness) direction of 180 MPa (26 ksi) and 288 MPa (42 ksi), respectively. Cracking occurred during service by fatigue running along banded regions containing many manganese sulfide inclusions, when the material was loaded transversely in service. In this case, the fault was not with the material, which met specification requirements, but with the designer assuming that the material was isotropic. The material met all applicable specifications, tensile properties only being specified in the longitudinal direction.

Other examples where metallographic examination is of great value are in showing where decarburization has occurred at a surface during heat treatment or in illustrating the depth of a carburized layer. Such metallographic examinations are normally supplemented with hardness measurements made using a microindentation hardness tester, commonly with a hardness traverse made on the surface of a metallographic specimen, as described in the section "Hardness Determination by Microindentation and Nanoindentation Methods." Embrittlement can often be observed through the examination of metallographic sections. For example, Fig. 31 shows the microstructure of a 321 stainless steel used for the hot-gas casing of a gas turbine after 18,000 h of operation, with a gas-side temperature of 985 °C (1800 °F) and an air-side temperature of 204 °C (400 °F). Extensive carbide formation has taken place, and brittle, intergranular cracking had occurred as a consequence.

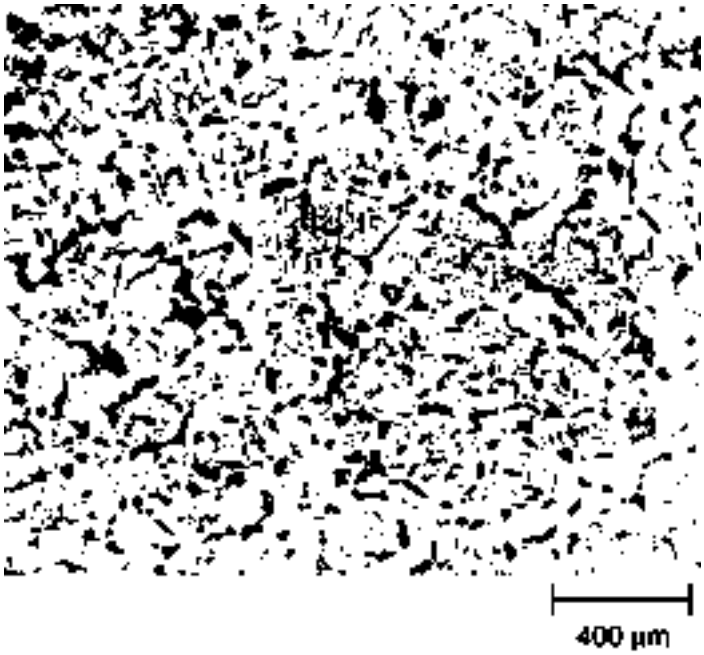


Fig. 31 Grain-boundary and intragranular precipitation at the hot side of the hot-gas casing of a gas turbine. Material is 321 stainless steel. Etched successively in Vilella's reagent, methanolic aqua regia, and Groesbeck's reagent to darken carbides

In quenched and tempered carbon and alloy steel samples, determination of the prior austenite grain size may be of importance, and this may not be readily defined from direct examination of the microstructure prepared and etched in the normal manner. However, if significant amounts of a high-temperature transformation product have formed at grain boundaries during quenching, these lie along the prior grain austenite boundaries. Also, in some cases where the tempering temperature is high, cementite particles precipitated during tempering lie along these boundaries. In these specific cases, the prior austenite boundaries may be visible following the standard metallographic preparation procedures. In the situation where the prior austenite boundaries are not readily observed, several techniques may be used to reveal them. These include the use of an etchant such as Vilella's reagent (5 mL HCl, 1 g picric acid, and 100 mL ethanol or methanol) that delineates the colonies of martensite laths or plates, differentiating the colonies formed in neighboring austenite grains. An alternate procedure is the use of a wetting agent added to an aqueous picric acid solution, which provides an etchant that delineates the prior austenite boundaries themselves. Figure 32 shows examples of its effectiveness in etching American Iron and Steel Institute (AISI) 4140 steel quenched and tempered at various temperatures.

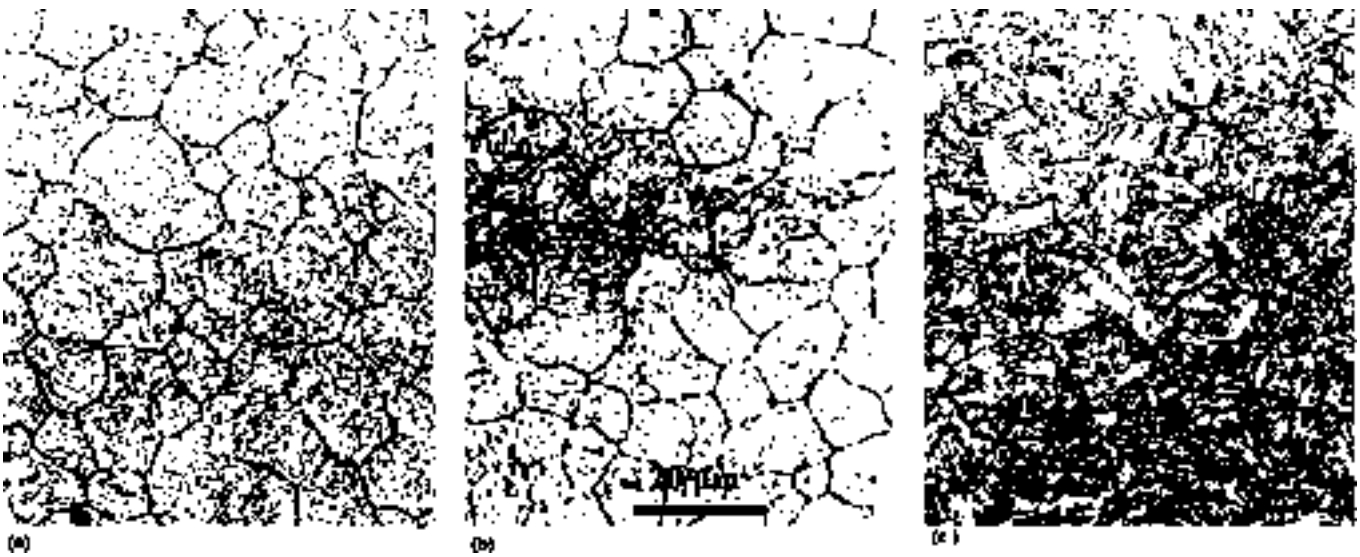


Fig. 32 An AISI 4140 steel, austenitized at 850 °C (1560 °F), oil quenched and tempered for 1 h at (a) 200 °C (390 °F), (b) 300 °C (570 °F), and (c) 550 °C (1020 °F). Etched in aqueous saturated picric acid containing sodium tridecylbenzene sulfonate (1 g per 100 mL solution)

For carbon and alloy steels with carbon contents below 0.25%, the McQuad-Ehn test may be used. In this, a specimen of the steel is carburized at 927 ± 14 °C (1700 ± 25 °F) for approximately 8 h. The specimen is furnace cooled to below the lower critical temperature at a cooling rate low enough to allow precipitation of cementite on the austenite grain boundaries in the hypereutectoid part of the case. Subsequently, the sample is sectioned and prepared metallographically to reveal the grain size in the hypereutectoid region of the case. A limitation is that the grain size of steels that have been austenitized at temperatures lower than 927 °C (1700 °F) may be finer that that indicated by the McQuad-Ehn test.

At times, it is important to determine the extent of retained austenite in a heat treated steel component, particularly when there has been brittle or intergranular fracture of the component. Austenite retained after quenching may transform isothermally to martensite at room temperature over a period of months, and this can also occur in steels that have been tempered at a low temperature. It may also transform as a result of stress during, for example, shock loading. Untempered martensite is difficult to distinguish from retained austenite; both appear white in the light microscope after the normal etching has been conducted. The importance of retained austenite is that it can lead to delayed cracking when it transforms to martensite, particularly in hypereutectoid steels. The presence of untempered martensite also makes the steel more brittle. Figure 33 indicates the extent of retained austenite in carbon steel after austenitizing and quenching in water or brine to room temperature; the amount increases with increase in carbon content. In hypereutectoid steels, etching in Vilella's reagent, rather than the more normally used picral, allows the martensite plates to be seen more clearly and the retained austenite, which remains white, to be identified more easily. This is well demonstrated in Ref 14. However, because of the difficulties in determining the extent of retained austenite using light microscopy and standard metallographic techniques, particularly when the quantity involved is small, it is usual to determine this by x-ray diffraction. The accuracy using a diffractometer is in the region of 0.5% for the range 1.5 to 38% of austenite by volume (Ref 15).

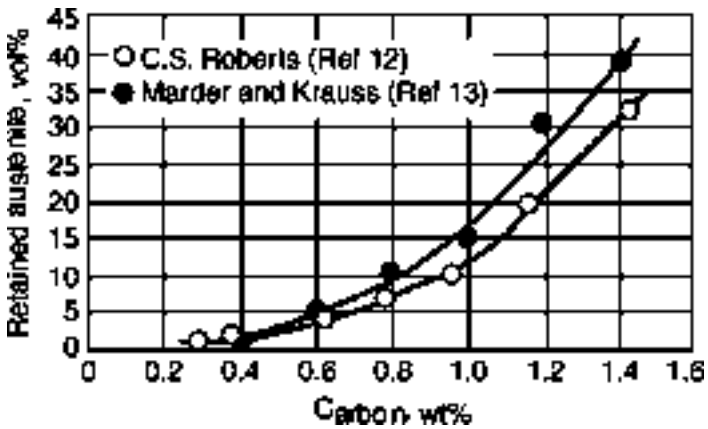


Fig. 33 Volume fraction of retained austenite in carbon steels fully austenitized and water quenched in water or brine at room temperature. Source: Ref 11

Microstructures of concern in failure analysis were formerly examined almost exclusively by means of light microscopy, with only a limited amount of examination being made in the SEM and an even more limited amount of TEM being conducted using thin foils or extraction replicas. However, it has become more common to use the higher resolution of the SEM in microstructural characterization, particularly where there are fine second-phase particles present, such as, for example, γ' in the nickel-base superalloys, as illustrated in Fig. 34.

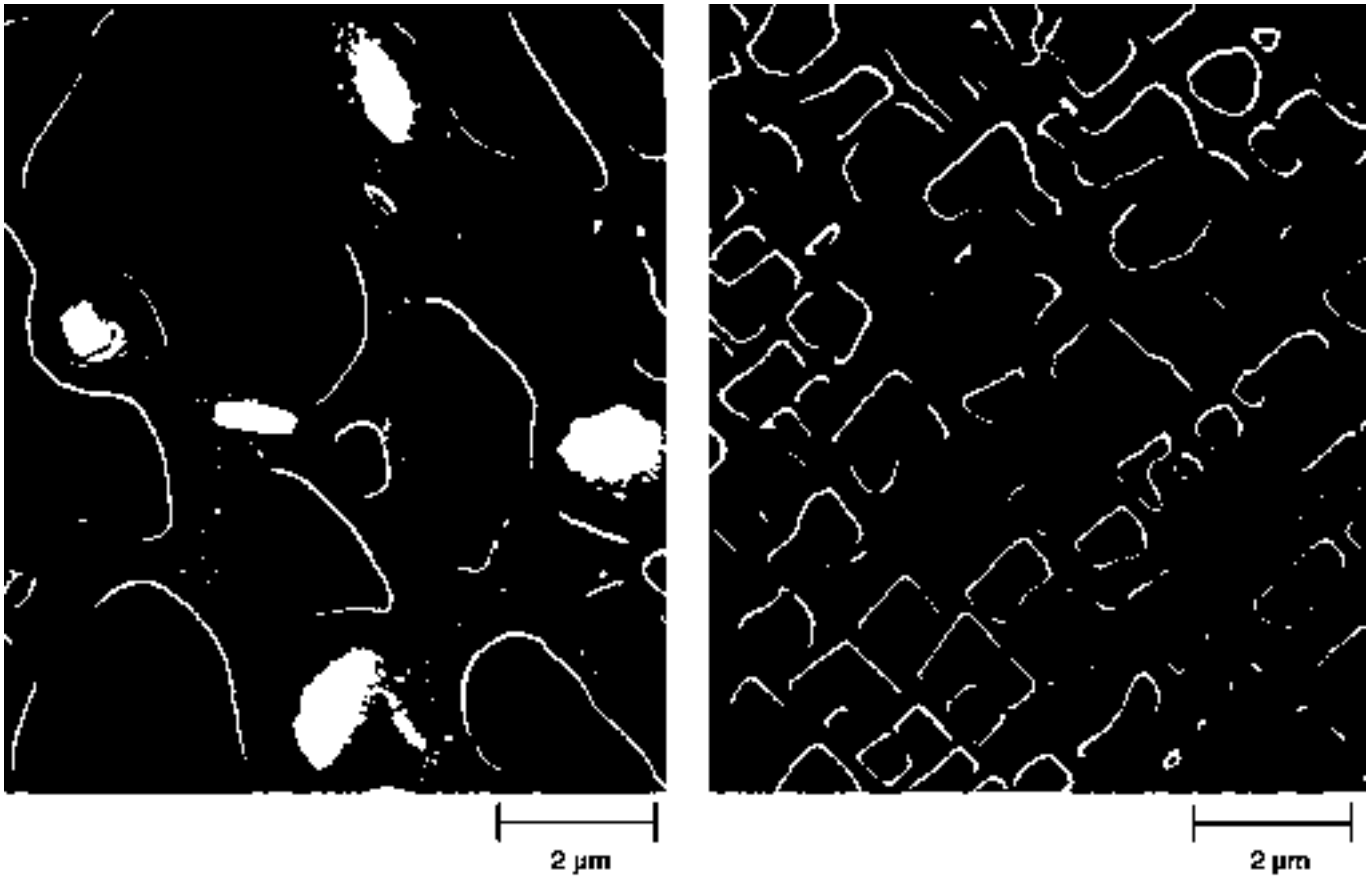


Fig. 34 Microstructure of Mar-M-247 heat treated cast alloy for gas turbine components showing different sizes of γ′ particles. Electropolished and electroetched. Courtesy of Dr. J.F. Radavich, Micro-Met Laboratories

This is also the case when components suffering from creep damage or creep failure are examined and in which there may be intergranular cavitation or decohesion at grain-boundary carbides. These features may not be readily detected using light microscopy, unless heavy etching or repeated polish-etch cycles are used; such procedures produce enlarged cavities on the specimen surface (Ref 16, 17). Figure 35 shows a region close to the fracture in a steam pipe of a CrMo steel that failed by creep. In Fig. 35(a), the specimen has been repeatedly polished and etched to open voids where carbides have precipitated on grain boundaries and are decoherent with the matrix. There is the appearance of extensive creep voids on the grain boundaries. In Fig. 35(b), the specimen has been carefully polished and etched, and grain-boundary voids are not apparent. A replica prepared from the surface of a creep-damaged CrMo steel is shown in Fig. 36 at two different magnifications. The material was repeatedly polished and etched to emphasize the damage, and the apparent voids are readily seen in the SEM. The initiation of creep voids at second-phase particles in type 316 austenitic stainless steel that was exposed to high temperature is illustrated in Fig. 37. The voids are apparent at the boundaries between ferrite and austenite and sigma phase/austenite boundaries. At lower temperatures under high stress, wedge-type cracking is observed (Fig. 38).

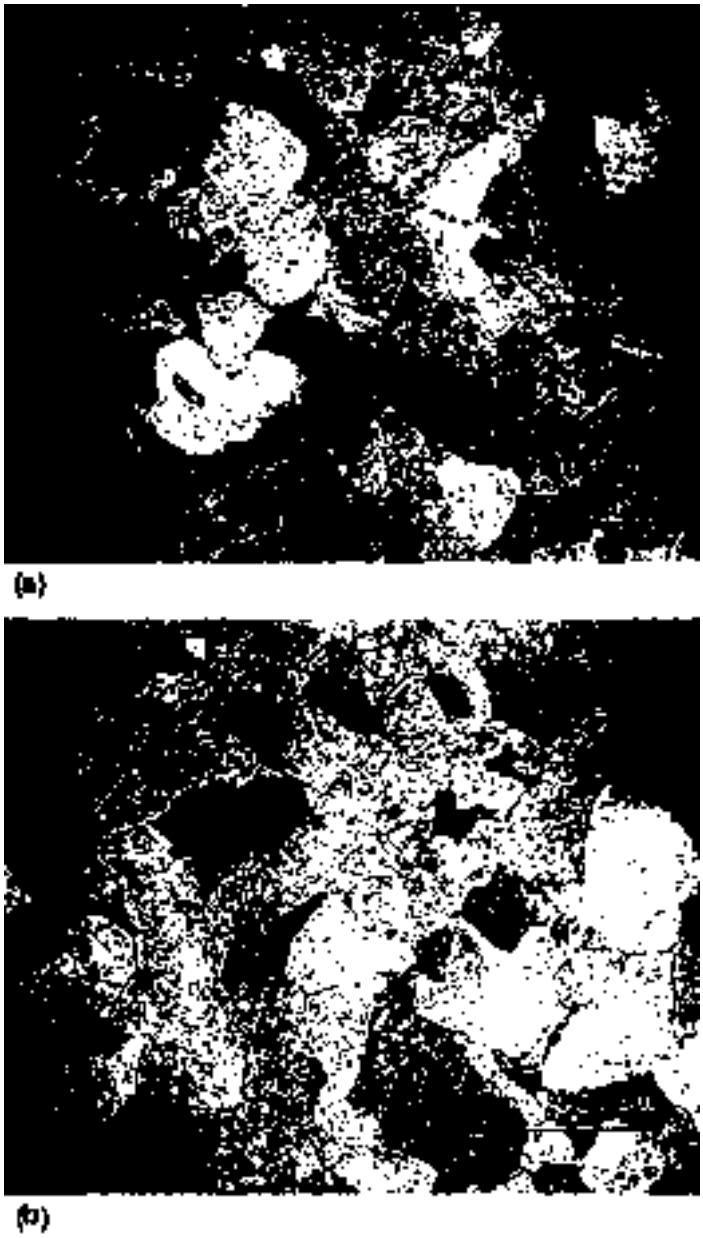


Fig. 35 Section adjacent to the fracture surface of a CrMo tube that ruptured after 70,000 h at a nominal temperature of 490 °C (915 °F). (a) Polished and etched to emphasize grain-boundary voids. (b) Polished and etched carefully to show grain boundaries. Nital etch

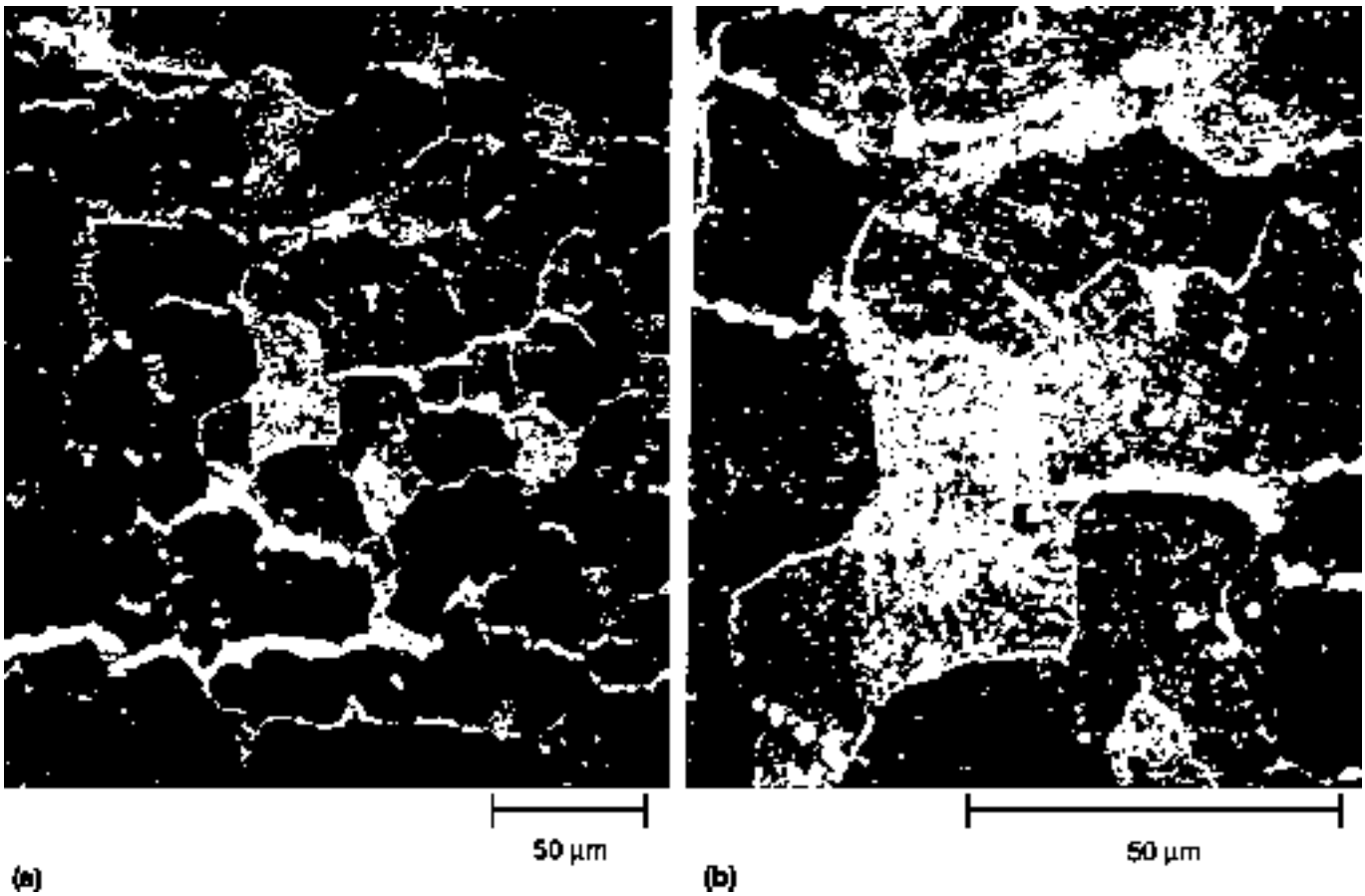


Fig. 36 SEM images of CrMo tube shown in Fig. 35, prepared from a specimen that had been polished and etched to emphasize apparent grain-boundary voids

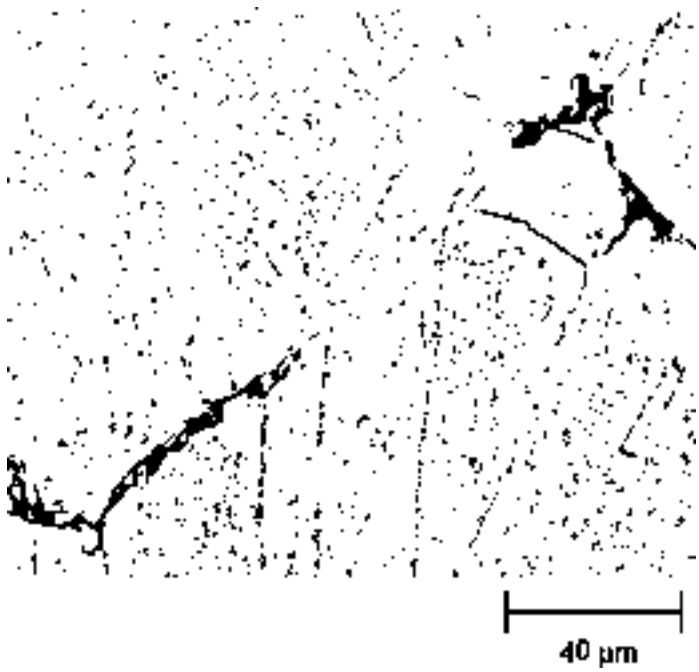


Fig. 37 Decohesion at the particle-matrix interface on grain boundaries of 316 stainless steel that failed by creep

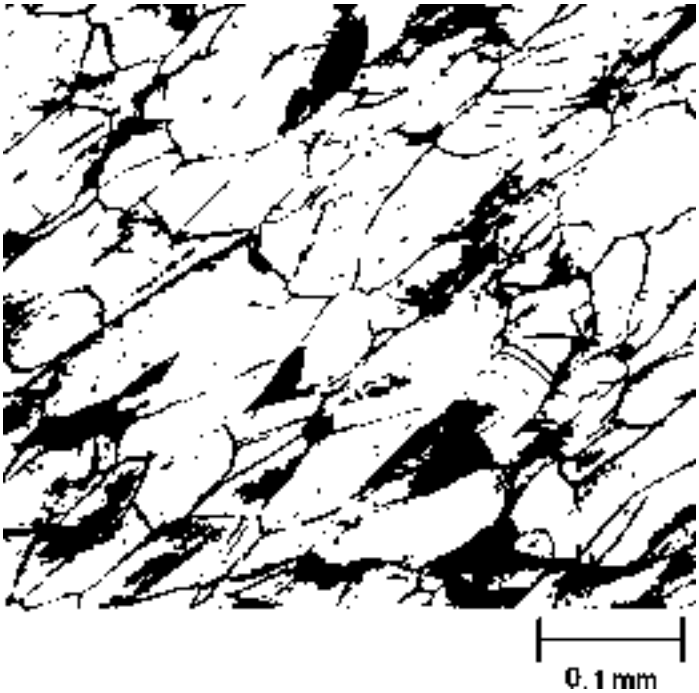


Fig. 38 Wedge cracking in 316 stainless steel that failed by creep at high stress and a temperature of approximately 700 °C (1290 °F). Oxalic acid etch

While macroscopic examination and fractography can provide much information concerning some damage and fracture mechanisms, metallographic examination can provide additional and, in many cases, more conclusive information and the identification of mechanisms that cannot be established clearly, if at all, from macroscopic and fractographic evaluation. This applies, for example, in the identification of stress-corrosion cracking (SCC), particularly when it is important to determine if the mechanism is one of intergranular or transgranular crack propagation. Figure 39 shows an example of transgranular SCC in a 316 stainless steel orthopedic implant. The cracking is typical of anodic SCC that develops from surface corrosion pits in the presence of chlorides. Crack propagation is assisted by active metal dissolution at the crack tips, as may be observed in the form of the rounded crack tips. Another example, intergranular cracking in a low-carbon steel pipeline, is shown in Fig. 40. The cracking, in this case, was caused by chemical reactions that produced hydrogen at the steel surface, which was absorbed in the steel, leading to delayed hydrogen cracking in regions of high tensile stress. Other hydrogen-induced damage is shown in Fig. 41, which illustrates voids formed by methane in a carbon steel exposed to a hydrogen-containing atmosphere at high temperature. The carbides in the original pearlite have been eliminated by reaction with the dissolved hydrogen.

Fig. 39 Stress-corrosion cracking in a 316 stainless steel orthopedic implant

Fig. 40 Hydrogen-induced cracking in pipeline steel

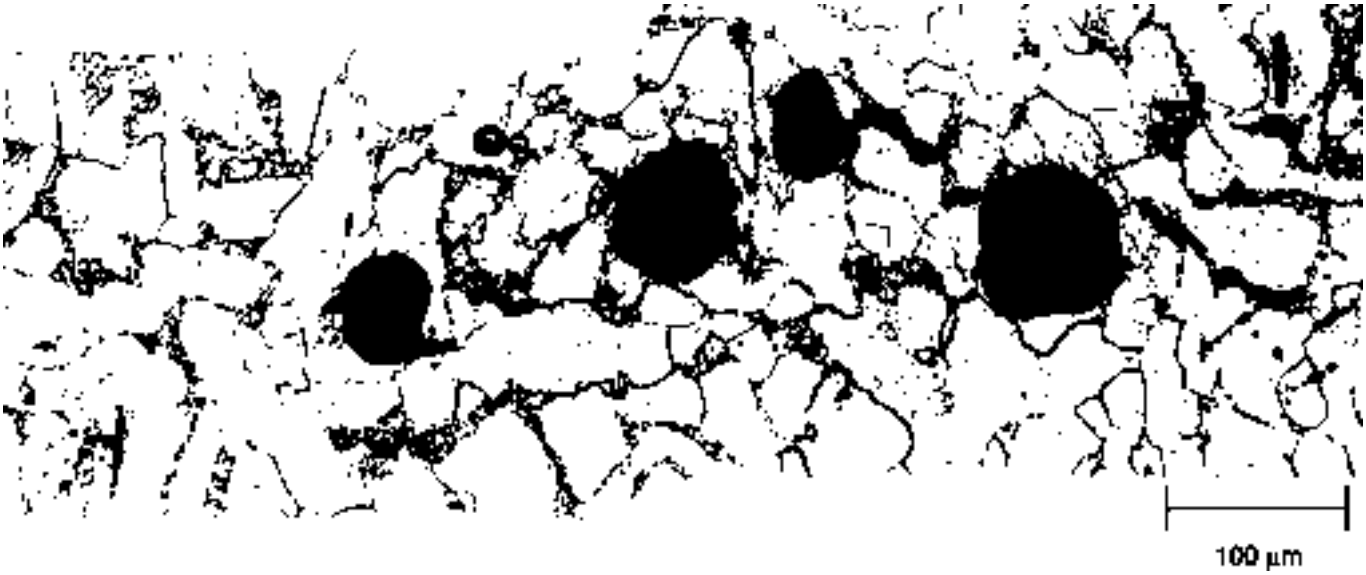


Fig. 41 Voids formed by methane in a carbon steel exposed to a hydrogen atmosphere at high temperature. The carbides in the pearlite have been eliminated by reaction with the hydrogen. Courtesy of Dr. T.L. da Silveira, Rio de Janeiro, Brazil

Metallographic examination can show the nature of subsurface fatigue fractures, as can occur in gears and bearings under high contact (Hertzian) loading. Figures 42 and 43 illustrate the subsurface nature of cracking in the carburized surface layer of a tooth in gears manufactured from Deutsche Industrie-Normen (DIN) 17CrNiMo6 steel (0.18 C, 0.38 Mn, 0.008 P, 0.020 S, 0.20 Si, 1.73 Cr, 1.58 Ni, and 0.25 Mo).

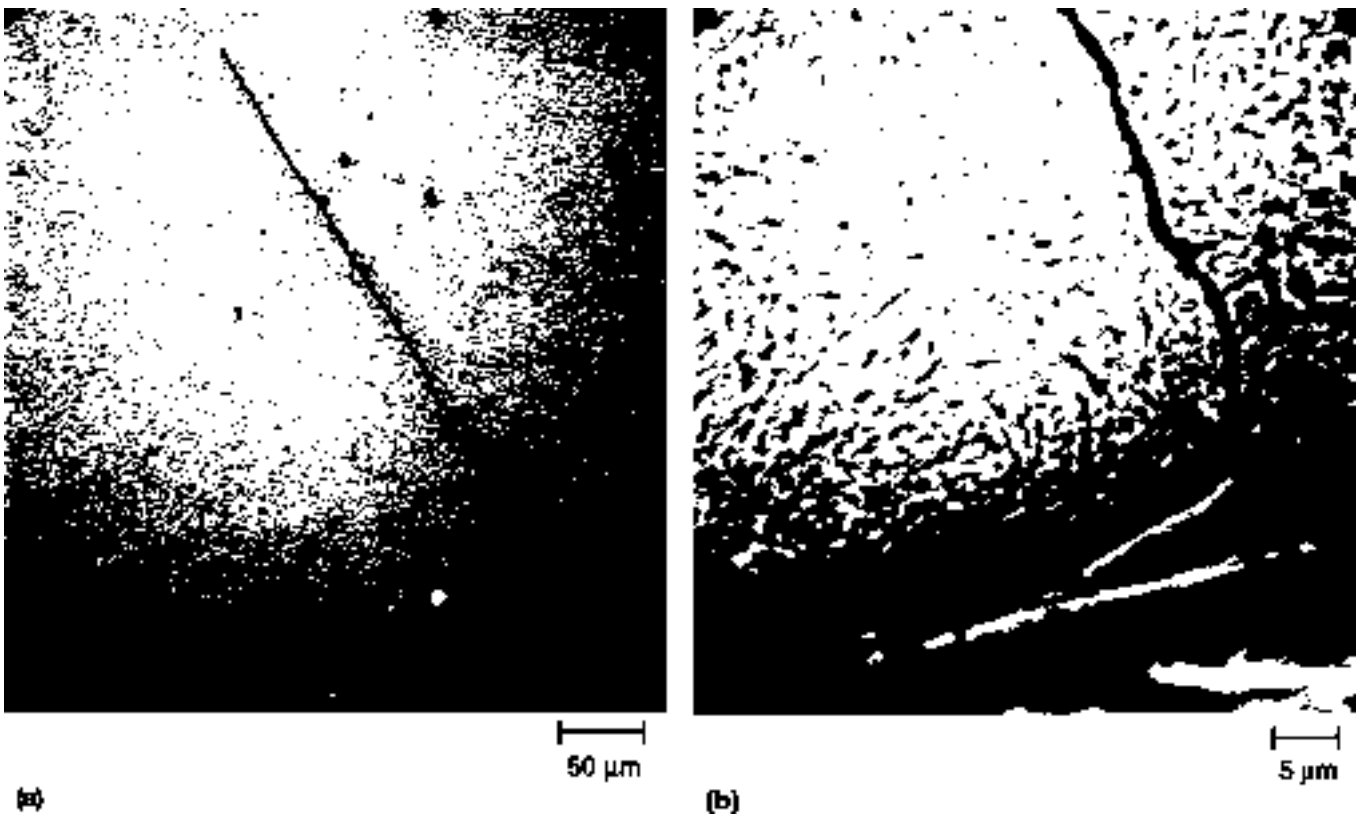


Fig. 42 (a) A crack running through the carburized surface layer on the flank of a gear tooth. (b) The region of the origin, shown at higher magnification

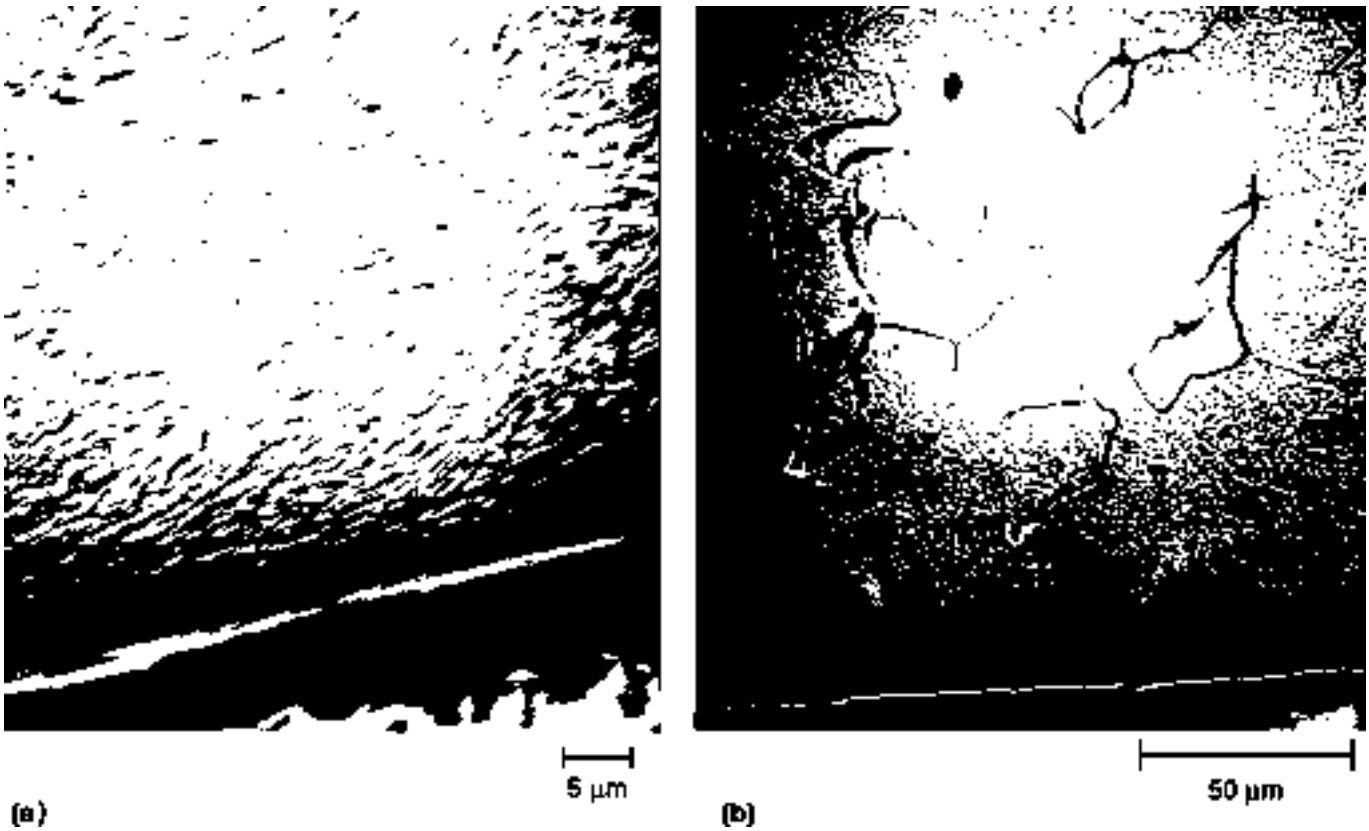


Fig. 43 (a) A fatigue crack initiating close to the flank surface of the gear tooth. (b) Multiple cracks within the carburized surface layer of the tooth in another area

In evaluating the microstructures of failed components and the mechanism of failure, the full range of analytical methods may be required. For example, it is sometimes difficult to distinguish between SCC and liquid metal embritttlement (LME). Additional analytical methods may be required to determine the chemistry within the cracks and in the adjacent matrix material. Figure 44 shows the cracking observed on a metallographic specimen cut from a locomotive axle that failed overall by fatigue. The fatigue cracking initiated from LME, which is normally attributed to embrittlement by liquid copper from melting of the backing material in the Babbitt bearing. In this case, the steel of the axle had not been austenitized. Thus, the temperature was never high enough to melt the copper alloy backing material. The material within the cracks was analyzed in a SEM. Figure 45 shows a backscattered electron image of the part of the bright-field electron image shown in Fig. 44. At several points, detailed EDS analysis was made (Fig. 46) from one of the bright spots.

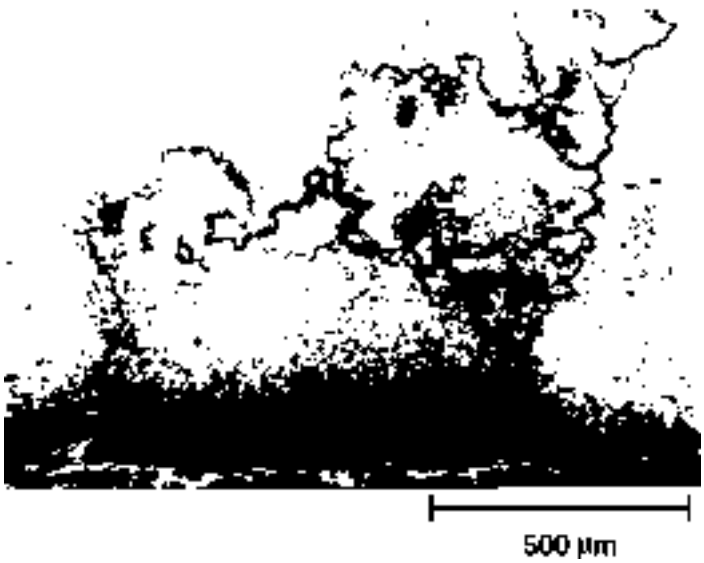


Fig. 44 SEM secondary electron image of LME in steel locomotive axle

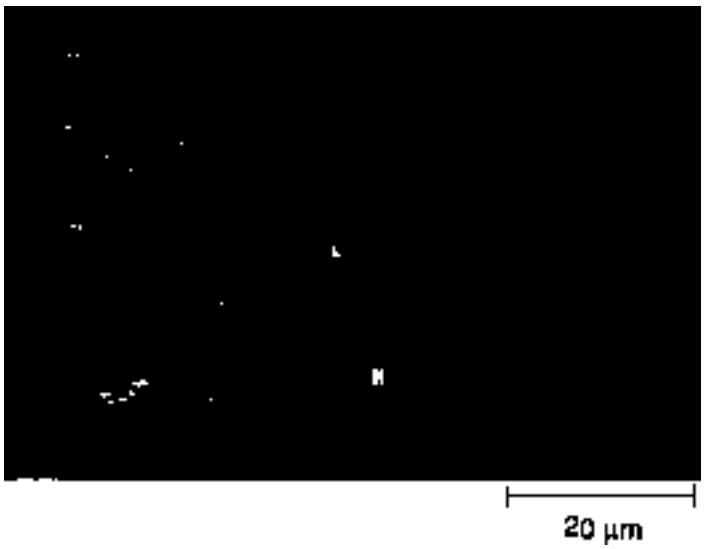


Fig. 45 SEM backscattered electron image of area within Fig. 44, showing points identified for detailed EDS analysis

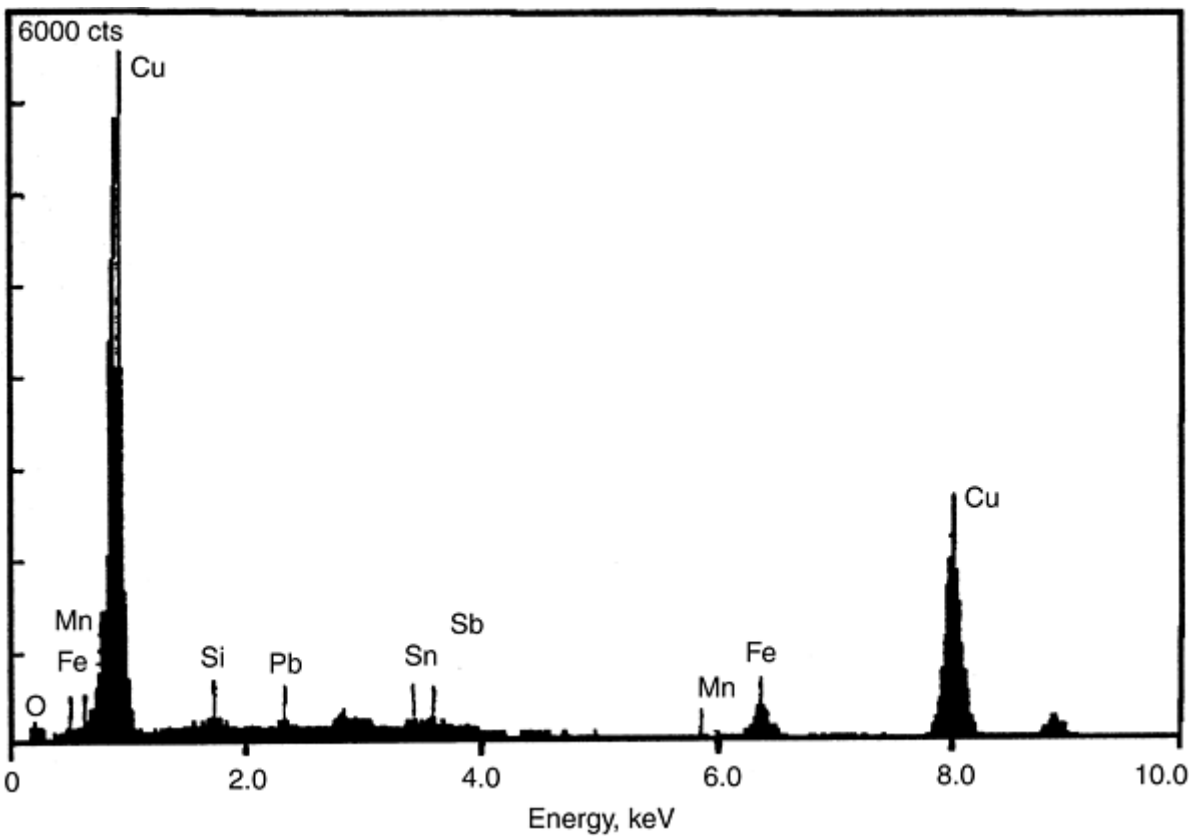


Fig. 46 Energy-dispersive spectroscopy x-ray spectrum from a shiny metallic particle in a secondary crack, as shown in Fig. 42

It is apparent that the particle is primarily copper, while small amounts of lead and tin were also recorded. The explanation is that the copper precipitated from molten Babbitt metal, a result of intermittent lubrication breakdown and subsequent overheating. The copper-lead phase diagram in Fig. 47 shows that at the solidification temperature for lead (328 °C, or 622 °F), the solubility of copper is virtually zero. At approximately 600 °C (1110 °F), it is approximately 1% by weight. Thus, when molten, the Babbitt metal in the bearing takes copper from the backing metal into solution. From the metallographic evidence, the axles were

not nearly as hot in the interior as at the surface, and copper would precipitate out from the Babbitt as temperature decreased.

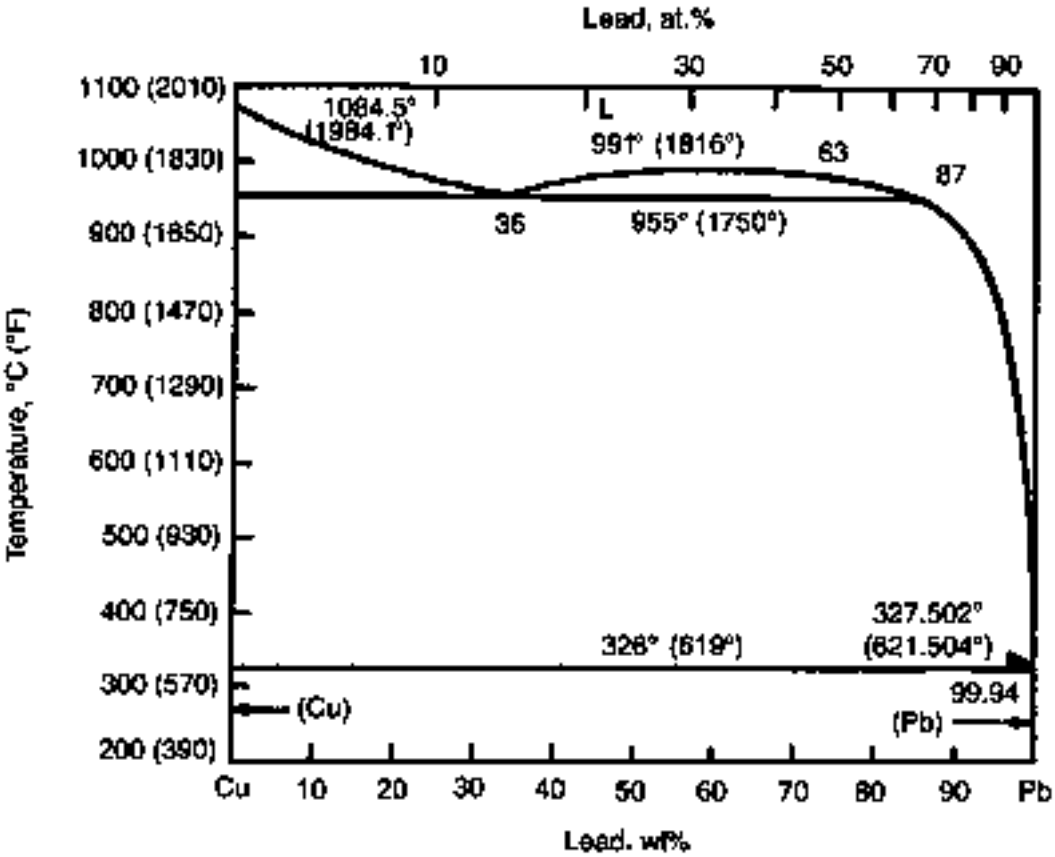


Fig. 47 Copper-lead phase diagram. Source: Ref 18

As a comparison with the foregoing case, Fig. 48 shows the type of penetration of austenite grain boundaries that occurs by copper in carbon steel in the classic copper-induced LME.

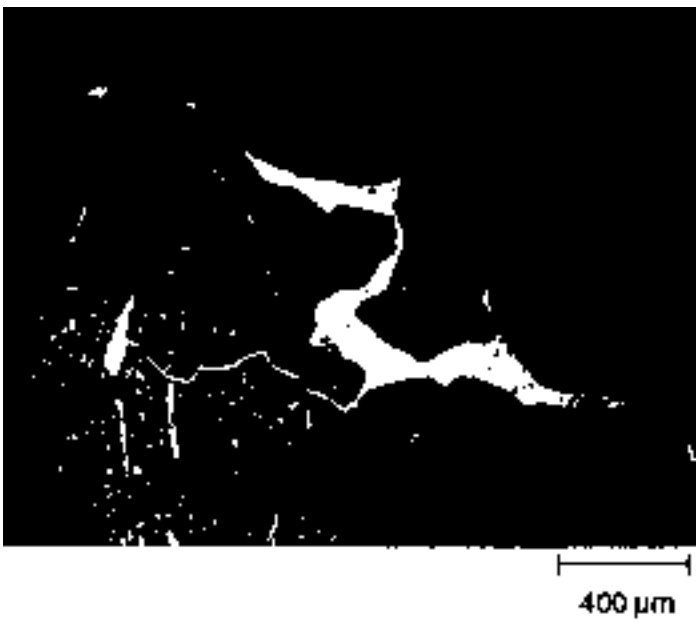


Fig. 48 Copper-induced LME in a 0.5 wt% carbon steel that occurred when the steel was in contact with liquid copper at 1100 °C (2010 °F). 2% nital etch

Figure 49 shows intergranular cracking of 321 austenitic stainless steel. While this looks like intergranular SCC, it is actually LME from liquid zinc that contacted the surface. Energy-dispersive spectroscopy was used to detect the presence of zinc at the surface where the cracks initiated. In light optical metallographic examination, some small dark spots were observed to form in the interior of the samples during final polishing, in some cases, apparently distant from observable cracks. Two such spots are visible in the lower part of Fig. 49. When viewed in the SEM, it was seen that they were produced by grains that had fallen from the surface, as shown in Fig. 50. This was caused by zinc diffusion along grain boundaries, producing decohesion.

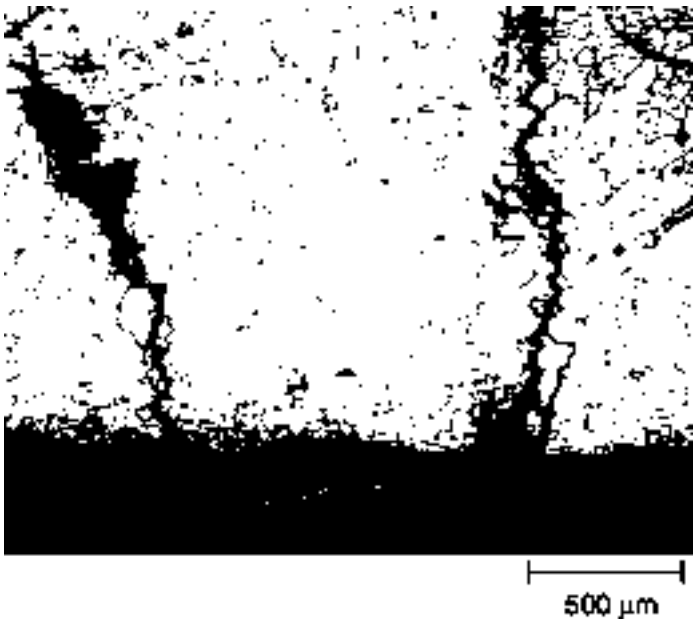


Fig. 49 Zinc-induced LME in 321 austenitic stainless steel. Etched in Vilella's reagent

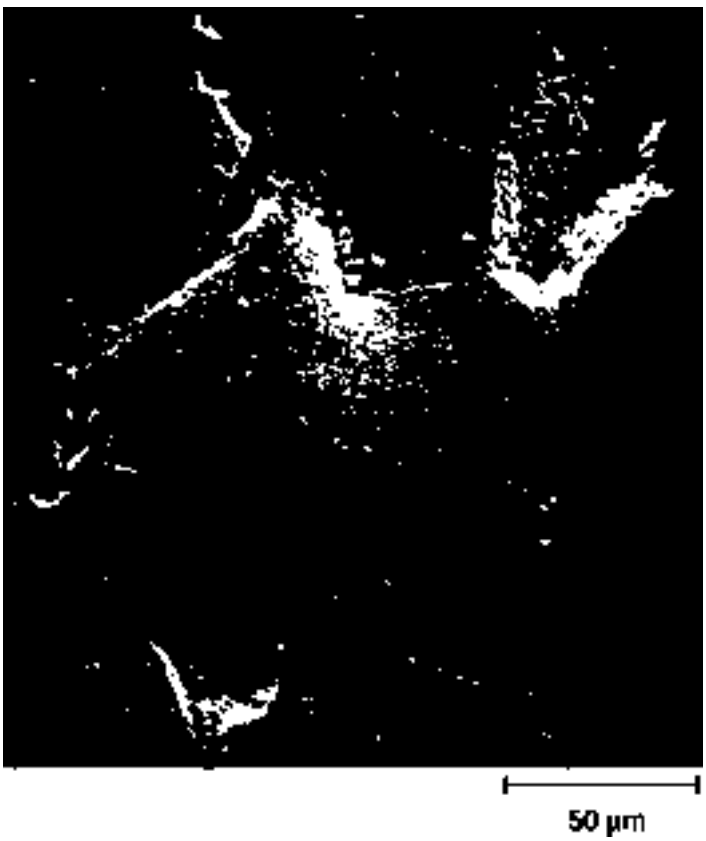


Fig. 50 SEM view of grains lost from 321 stainless steel by zinc-induced LME. Etched in Vilella's reagent

References cited in this section

11. G. Krauss, *Steels: Heat Treatment and Processing Principles,* ASM International, 1990

14. L.E. Samuels, *Optical Microscopy of Carbon Steels,* American Society for Metals, 1980

15. *Smithells Metals Reference Book,* 6th ed., E.A. Brandes, Ed., Butterworths, London, 1983, p 4–17

16. T.L. da Silveira and I. Le May, in *Mater. Charact.,* Vol 28, 1992, p 75–85

17. L.E. Samuels, R.W. Coade, and S.D. Mann, in *Mater. Charact.,* Vol 29, 1992, p 343–363

18. *Metallography, Structures and Phase Diagrams,* Vol 8, *Metals Handbook,* 8th ed., American Society for Metals, 1973, p 296

Examination of Damage and Material Evaluation

Iain Le May, Metallurgical Consulting

Evaluation of Polymers in Failure Analysis

In conducting a failure analysis of a component fabricated from polymeric material, the general methods of analysis outlined for metals are followed, including observations of deformation and damage, chemical analysis, and fractography. However, the failure mechanisms differ in a number of respects from those occurring in metallic materials, and it is appropriate to consider the mechanisms of deterioration and consequent failure in polymeric components. The various failure mechanisms are detailed in *Engineering Plastics,* Volume 2 of *Engineered Materials Handbook,* and relevant ones are reviewed briefly here.

Failure to perform their intended function may occur in a number of ways in polymeric materials. These include plastic deformation produced by creep or yielding; crazing and subsequent fracture; fatigue under fluctuating stress; thermal and residual stresses; physical aging; moisture-related degradation; photolytic degradation; degradation from organic chemicals, and environmentally-induced stress cracking.

Plastic Deformation. For an amorphous polymer, the yield stress tends toward zero when the temperature is raised to the T_g. For a crystalline polymer, the yield stress tends to zero when the polymer starts to melt. However, it should be emphasized that the onset of gross yielding is strongly dependent on strain rate and temperature. The values of T_g and T_m are important quantities, and the temperature range of operation of a polymeric material is limited. The degree of crystallinity is an important factor also, because the resistance to plastic deformation increases with the degree of crystallinity. For polymeric materials operating at temperatures above ambient, creep is a definite consideration, and stresses need to be kept to a level such that this is not an important consideration.

Crazing can occur in high-impact plastics under the combined influence of stress and a hostile environment. In a hostile environment, such as an organic solvent, if a critical strain is exceeded, cavitation (crazing) can occur, producing whitening in the plastic or opacity in clear material. The crazes degenerate into microcracks, which then grow until a critical size is reached, and failure results.

Fatigue is similar to fatigue in metals, although there are differences in the detailed mechanism of crack growth, particularly in terms of crazing and void formation ahead of the crack tip that can occur in polymers. Nevertheless, the classic appearance of beach marks and striations is observed on the fracture surfaces of polymeric materials that have failed by fatigue. For many polymers, there is no defined endurance limit below which fatigue failure does not occur, so failures may be expected where there are large numbers of repetitive load cycles. The nonlinear characteristics of the stress-strain behavior of polymers producing hysteresis losses during load cycles make them particularly subject to fatigue at high frequency.

Thermal and Residual Stress. Thermal stresses can be developed because of exposure to heat during service or may be present from the solidification or manufacturing, because the coefficient of thermal expansion is

generally high and the thermal conductivity low. As a result, considerable residual stresses may be generated after local plastic deformation takes place to accommodate the thermal stresses. The dimensional changes induced by thermal stresses may also provide sufficient change in shape, such that the component no longer serves its purpose; for example, it may lose its contact as a seal.

Physical aging occurs primarily below T_g, and there is an effect on mechanical properties, particularly on their ductile-brittle transition, with material becoming more brittle.

Moisture-Related Degradation. The effect of moisture varies widely with the particular type of polymer concerned. The general effect is that absorption of water molecules causes a reduction in T_g, an increase in ductility, and a loss in creep resistance. The difference is largely because some polymers have a very limited capacity for absorption.

Photolytic Degradation. Chemical changes in polymeric materials are induced by sunlight. Absorption of energy can cause reduction in polymer molecular weight because of rupture of main chain molecules, formation of cross links, and the formation of oxidized groups within the polymer. These effects reduce strength and increase brittleness.

Degradation from Organic Chemicals. Organic liquids, such as cleaning fluids, solvents, petroleum products, and detergents, can greatly affect the mechanical properties of polymeric materials. For the polymer under stress, the effect can be severe, with embrittlement being the usual failure mode.

Chemical attack, involving breakup of long chain molecules, is irreversible and may result in crazing, microcracking, and rupture of the material. Reversible effects, which involve absorption of fluid and swelling of the material, increase the degree of plasticization, which may allow failure to occur because of reduced mechanical strength.

Soft, plasticized polymers are susceptible to embrittlement when in contact with organic liquids, because the plasticizing additives can be leached out from the component.

Evaluation of Ceramic Materials in Failure Analysis

The principles involved in analyzing failures in ceramic materials are similar to those discussed with particular reference to metals, but a number of specific points require discussion, because of the different properties and behavior of ceramics from those of metallic components.

First, because of the brittle behavior of most ceramics, there may be many pieces resulting from a failure, and reassembly of the pieces can provide information about the form of loading and the point of fracture initiation. This is similar to what can be done for metals and is illustrated in Fig. 1. Figure 51 illustrates the types of crack pattern resulting from different loading situations in brittle materials. In practice, evaluation of a fracture in a brittle ceramic often means actually assembling the pieces together, although care must be taken not to damage or obscure features on potentially critical fracture surfaces. The situation is considerably different from that of metals, for which the failure analyst is normally very reluctant to conduct any reassembly of parts because of the much greater potential for surface damage on metallic fracture faces.

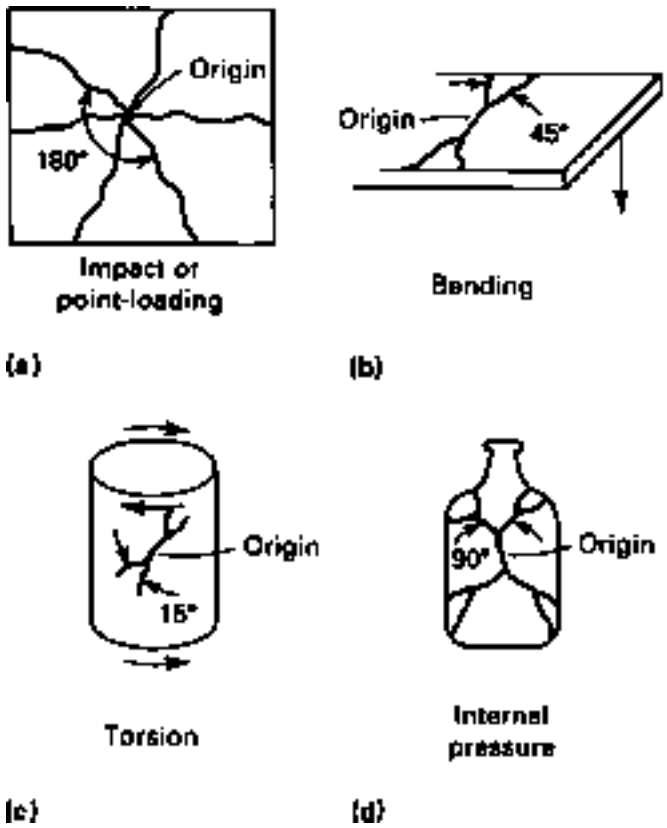


Fig. 51 Information available by examining crack direction and crack branching. Source: Ref 19

Crack branching is an important quantity in assessing fracture in ceramics. Cracks branch at a critical velocity, which is on the order of half the speed of sound in glass for the specific glass concerned. The acceleration from crack initiation to the critical velocity depends on the energy dissipation available from release of stored energy. This can arise from applied stress, prestressing, or residual stress. The last mentioned is very important; for example, tempered glass has a large stored energy, branches easily, and breaks into many small fragments. On the other hand, a ceramic component that is heated unevenly, thus developing a thermal stress, may have little crack branching, and, if the crack initiates at a small, locally heated area, there may be a single crack without any branching.

When a crack is initiated in a ceramic material and accelerates, it interacts with the microstructure, the stress field, and the acoustic vibrations that may be set up. These interactions lead to the formation of distinct fractographic features, including the fracture mirror, hackle or river patterns, and Wallner lines. Figure 52 shows schematically these features surrounding the fracture origin. In Fig. 52(a), the origin of the crack is internal, while in Fig. 52(b) it is at the surface.

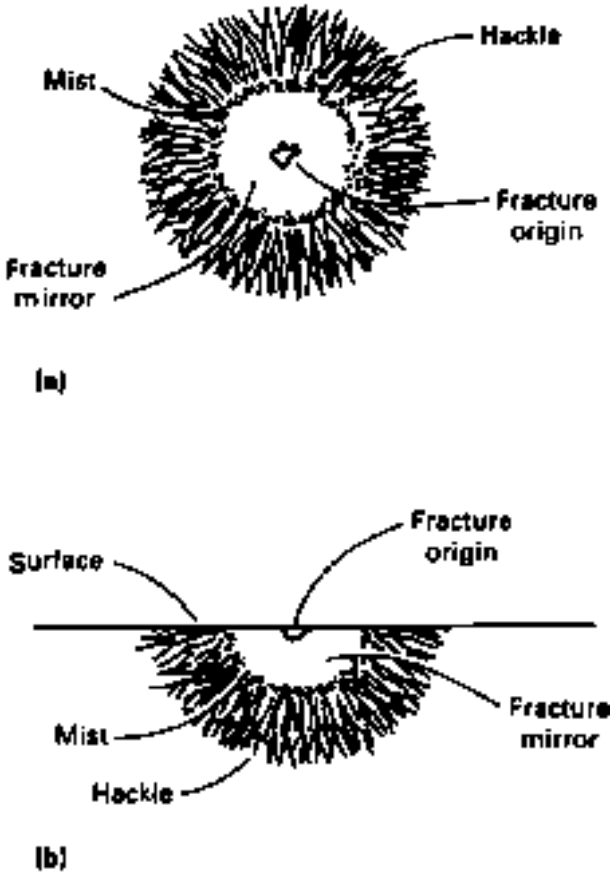


Fig. 52 Schematic showing typical fracture features surrounding the fracture origin in a ceramic material. (a) Internal crack origin. (b) Surface crack initiation.

A crack initiating at an internal flaw travels radially outward as it accelerates, forming a smooth surface termed the *fracture mirror*. It starts to deviate from its original plane when it reaches a critical velocity, intersects with microstructural features such as an inclusion of precipitate, or if the direction of the principal tensile stress starts to change. This leads to the formation of small radial ridges on the fracture plane. Initially, these may show up as “mist” on glass, but this is not necessarily seen on all ceramic materials. Subsequently, they develop into river pattern or hackle, comprised of larger ridges, and these coalesce to form a smooth surface or one having larger steps.

Wallner lines are sometimes formed on fracture surfaces, particularly on brittle ceramics. They result from sonic waves that are generated during the fracture process, and these elastic stress waves interact with the principal stress driving the propagating crack front. Thus, a series of arc-shaped steps may be left on the fracture surface. Figure 53 shows an example of these.

Fig. 53 Wallner lines (between white arrows) on the fracture surface of an aluminum alloy A356-T6 casting. The black arrow indicates the direction of crack propagation. Source: Ref 20

In conducting a failure analysis of ceramics, the fracture origins are normally determined first. The nature of these and of the cracking provides information as to the type of failure, for example, from impact; as a result of residual stresses in combination with applied load; from thermal shock; as a result of machining or other defects; or from oxidation and corrosion. Fractography (macro- or micro-) may be required, together with microstructural evaluation, as discussed in the section "Preparation and Examination of Metallographic Specimens in Failure Analysis" dealing with ceramic materials, as well as the various procedures previously outlined for chemical analysis and characterization of materials.

References cited in this section

19. V.D. Frechette, Fracture of Heliostat Facets, The ERDA Solar Thermal Projects Semiannual Review, 23–24 Aug 1977 (Washington), ERDA

20. *Fractography and Atlas of Fractographs*, Vol 9, *Metals Handbook*, 8th ed., American Society for Metals, 1974, p 83

Modeling and Accident Reconstruction

Charles R. Manning, Jr., and Thomas C. Wenzel, Accident Reconstruction Analysis Incorporated

Introduction

FAILURE ANALYSIS is generally defined as the investigation and analysis of parts or structures that have failed or appeared to have failed to perform their intended duty. The purpose is to determine how and why a part or structure has failed. In contrast, the focus of accident reconstruction is to determine how and why an accident occurred. Clearly, there is a certain degree of commonality between the terms, because most, if not all, failures can be categorized as accidents. However, not all accidents can be categorized as failures, because many accidents occur due to human error. There may be no actual physical failure of a part and/or device, although parts may "fail" because of the accident itself.

Accidents typically involve large-scale incidents such as the one shown in Fig. 1, where the sequence of events leading to the incident must be understood in order to properly focus further examination and evaluation by specialists such as materials failure analysts. Methods of field inspection and initial examination also are critical factors for both reconstruction analysts and materials failure analysts. This article focuses on the general methods and approaches from the perspective of a reconstruction analyst, although the section "Investigation of the Accident Scene" in this article includes discussions relevant to materials failure analysts at the incident scene. The elements of accident reconstruction described in this article also have conceptual similarity with the principles for failure analysis of material incidents that are less complex than a large-scale accident.

Fig. 1 Collapse of pedestrian bridge over a highway

This article provides a brief review of some general concepts on the use of modeling. A model is a physical, mathematical, or logical representation of a physical system or process. The use of modeling can be a very powerful tool for information pertaining to the reconstruction of an accident. Whether the modeling consists of simple mathematical modeling of a physical system or more complex computer modeling of a system or part, the modeling technique may directly or indirectly give insight to the cause of an accident.

Modeling and Accident Reconstruction

Charles R. Manning, Jr., and Thomas C. Wenzel, Accident Reconstruction Analysis Incorporated

Accident Reconstruction

The term “accident reconstruction” has traditionally been used to describe the investigation and analysis of motor vehicle and aircraft accidents. However, the term is also being used more often to describe the investigation and analysis of any unexpected event that causes loss or injury.

Accident reconstruction is rarely a simple endeavor, and accident reconstruction requires personnel with proper training and experience in performing investigations/reconstruction. Reconstruction also often requires the assistance of other personnel with specialized expertise to address certain aspects of the investigation. The investigation and analysis of accidents and failures must be thorough to ensure that all information pertaining to the incident has been scrutinized and that accurate conclusions have been drawn.

In order to reconstruct how an accident occurs, a number of investigative steps must be performed to obtain all the necessary information. The approach presented subsequently follows the basic approach of the scientific method as used in the physical sciences. This includes the definition of the problem, collection of data, analysis of data, development of a hypothesis, and testing and selection of the appropriate hypothesis.

Often, investigations are not able to fully answer all the questions associated with the incident. In these cases, personnel performing the investigation must present the possibilities that are scientifically plausible. Investigators may also be asked to opine on which one of the possibilities are the most probable.

Interdisciplinary Nature

Reconstruction of an accident rarely deals with only a single engineering discipline. In most cases, the issues involved in the accident are not limited to materials, mechanical, chemical, or structural engineering. The issues often extend to other engineering disciplines as well as to fields such as statistics, human factors, and other areas specific to the particular accident at issue. The person responsible for the reconstruction of an accident should have a broad range of engineering knowledge and must be able to recognize the appropriate field of expertise necessary for the investigation.

Types of Accident Reconstruction

The following are some of the more common types of accident reconstruction:

- Motor vehicle
- Aircraft
- Structural
- Electrical
- Industrial
- Mechanical
- Construction
- Fire investigations

Numerous references are available to aid the personnel responsible for the reconstruction. “The Traffic-Accident Investigation Manual” (Ref 1) is a two-volume publication by the Northwestern University Traffic Institute that deals with both the on-scene investigation and the reconstruction of motor vehicle accidents. The Society of Automotive Engineers, Inc. (SAE) publishes numerous papers on all aspects of motor vehicle accident reconstruction. Additionally, SAE regularly hosts symposiums in which new technical papers are presented. Many other sources are available for information pertaining to motor vehicle accident reconstruction.

The National Fire Protection Association (NFPA) standard 921, “Guide for Fire and Explosion Investigations” (Ref 2), gives guidelines for a systematic approach for the fire and explosion investigation process. The “Accident Prevention Manual” (Ref 3) published by the National Safety Council is an excellent general resource for workplace safety.

In addition to books/articles that specifically address investigation, reconstruction, and safety issues, codes and standards are also commonly used for reconstruction. The knowledge of codes and standards is necessary to all types of reconstruction. The following are examples of codes and standards commonly referenced by personnel involved with reconstruction:

- Occupational Safety and Health Administration (OSHA) standards for construction and industrial accidents
- American National Standards Institute (ANSI) standards for products
- Underwriters Laboratories (UL) standards for products
- Local and state building codes
- National electrical codes published by the NFPA
- National fire codes published by the NFPA
- American Society for Testing and Materials (ASTM) standards

Elements of Accident Reconstruction

Documentation and Preservation of Evidence. The top initial priorities of any accident investigation are the accumulation of information immediately available at the time of the accident and the preservation of any physical evidence so that it may be examined at a later point in time. After an accident or failure occurs, the accident scene often contains key information that may assist investigators in determining what has happened before and during the accident. In an ideal situation, the investigator arrives at the scene immediately after the accident. In this case, he/she can take charge of the investigation, document the accident scene, and preserve the evidence. However, in most cases, there is a substantial period of time before the investigator can arrive. In these cases, it is up to the personnel present at the scene to document the accident conditions and preserve physical evidence. People who are likely to be responding to accidents should be made aware of the necessity of documentation and preservation of the accident scene.
There are numerous reasons, both intentional and unintentional, why the condition of the accident scene may change after an accident:

- *Medical Attention to Injured Persons:* Persons who may be injured during the accident usually become the top priority of personnel responding to the accident. However, during this process, relocation and/or alteration of evidence can occur that may make explanation of the events surrounding the accident more difficult. For example, an issue in many motor vehicle accidents is whether vehicle occupants were wearing seat belts and their position in the vehicle after the accident.
- *Human Psychology:* Persons with knowledge of the injured person and/or accident scene may feel partially or fully responsible for the accident (whether they are or are not) and may alter the accident scene in an attempt to remove blame from themselves, the victim, or other parties. For example, someone may be motivated to remove evidence of alcohol or recreational drugs from the accident scene. Although it is difficult to ascertain how often something like this occurs, investigators must take this possibility into account during an investigation.
- *Cleanup/Restoration of Preaccident Conditions:* Personnel performing cleanup work after an accident may alter the condition of the evidence or dispose of it. This often occurs simply due to the lack of understanding of the importance of the accident scene. This also occurs when there is a rush to return the accident scene to its condition prior to the accident. Examples include a factory in which workers make an attempt to resume normal operation, or after an automobile accident in which law enforcement personnel attempt to reestablish traffic flow, as in the case of a pedestrian bridge that collapsed onto a highway (Fig. 1). In this circumstance, medical attention to the injured persons and/or urgency to restore traffic flow may cause evidence to be altered or destroyed.

If possible, the accident scene should be left untouched after the accident occurs. Accident scenes, in general, are found by investigators in the full range of possible conditions, from completely untouched to completely cleaned up. This depends on the type of accident and the knowledge of the personnel initially present at the accident scene.
If possible, personnel initially present at the accident scene should acquire photographic and/or videotape documentation of the accident scene to capture the accident scene in several levels of broad and detailed views. The personnel should document the condition of key evidence but not alter/test/disassemble any of the evidence. Further techniques for scene inspection and documentation are discussed later in this article.

Also of great importance is the saving of physical evidence from an accident scene. Examination, testing, and analysis of evidence from an accident scene may reveal important information regarding the accident. Loss and/or disposal of key evidence often occurs shortly after an accident occurs.

All of the topics discussed in this section point to the need for accident-scene control after an accident. Any changes that occur to an accident scene only make the subsequent accident investigation more difficult. Significant changes to an accident scene can make it impossible for investigators to draw conclusions regarding the cause of the accident or may lead to erroneous conclusions.

Facts Related to the Accident. The facts related to the accident should be obtained. The facts may be obtained prior to, during, or after the scene investigation and may include:

- Type of accident
- Time and date
- Location
- Weather/environmental conditions
- Eyewitnesses/personnel present in accident vicinity
- Type of equipment involved
- Equipment settings
- Equipment/scene condition
- Data from monitoring devices
- Other

The type and amount of factual information gathered by the investigator depends on the type of accident. However, the safest approach is to obtain as much information as possible. Although much of the information may not be critical or used in latter stages of the accident reconstruction, it is always better to have the information if the need for it arises. Observations that may not seem important immediately following an accident can become important as the investigation reaches its latter stages.

Witness Interview. Witnesses who actually observed the accident often have information that is important to the investigation. Witnesses generally have the best recollection of an accident immediately after an accident occurs. However, if a witness has been traumatized by viewing the accident, their recollection of what happened may be tainted. In these situations, it is often better to interview the witness at a later point in time, after they have had an opportunity to collect their thoughts. The disadvantage of waiting to take a witness statement is that witnesses may alter their recollection if given sufficient time to think of the consequences of what they say or if influenced by another person. Another disadvantage is that witnesses often tend to forget certain aspects of what they saw during the accident. A witness is much less likely to remember all the details several months or years after an accident, compared to what he/she remembers immediately after the accident.

Considerable skill is required to obtain an accurate witness statement. Several factors must be taken into account, including whether the witness has been traumatized by the accident, whether the witness was involved in or was responsible for the accident, and whether the witness has any monetary or other interests associated with the accident that might bias their statements. The interviewer should allow the witnesses to tell their own story pertaining to what they saw during the accident and not ask them leading questions that may affect their statements. The interviewer should also attempt to separate out factual statements from those containing opinions.

Often, statements made by the witnesses initially sound factual but, on further examination, are discovered to be based on an opinion. The witness may not be aware of the need to separate factual statements from opinionated statements. Although the interviewer should allow the witness to tell his story, the interviewer should also be prepared to ask further questions to uncover information that the witness either may not initially remember or may not think is important. It should also be recognized that the recollections of witnesses are often incorrect—it is very common to get different stories from different witnesses about how an accident occurred. The reconstruction investigator should take this into consideration when evaluating the witness statements.

Witnesses are not limited to those people who actually saw the accident occur. Other witnesses may include persons who were very familiar with the accident scene prior to the accident, or persons who arrived immediately after the accident who viewed the condition of the scene prior to any changes.

Witness interviews may range from recorded and/or notarized statements (by audio or videotape) taken by law enforcement personnel or professional interviewers to less formal interviews in which notes are taken by hand.
Background Information. After the initial information has been obtained, pertinent background information is collected. This type of information may be obtained prior to, during, or after the scene investigation, depending on a number of factors including when the scene investigation takes place. Much of this information is used during reconstruction and may include:

- Drawings of equipment
- Diagrams of equipment
- Equipment history
- Product recalls
- Service and maintenance history
- Test reports
- Manuals
- Warnings
- Codes and standards
- Published articles

Investigation of the Accident Scene. Examination of a nondisturbed accident scene is often key in determining what happened during an accident. Investigation should occur as soon as possible after an accident occurs. The purpose of the scene investigation is to collect data pertaining to the accident and to document the condition of the accident scene. Investigators who are not able to visit the accident scene often are limited to studying a few pictures taken by others. The following are several of the methods used to document the condition of the accident scene.
Permanent Visual Documentation. One of the first steps of any scene investigation is to obtain a permanent record of the accident scene by visual documentation, such as photography and/or videotaping. It is very important that visual documentation be performed prior to any altering of the accident scene. Although further visual documentation of the scene may occur during or after cleanup procedures, the initial documentation establishes its original condition at the time of the accident. At a later point in time, if questions arise regarding the condition of the scene at the time of the accident, these photographs/videotapes can be reviewed to attempt to answer these questions. Locations and conditions of objects, positions of knobs/switches on equipment, and so on can be permanently recorded by these techniques.
The importance of obtaining multiple angleviews, including broad overviews, narrower overviews, and several levels of detail, cannot be overemphasized. Sometimes important conditions at the accident scene may be overlooked during the initial investigation at the scene. These conditions, which may not be obvious at the time of the scene inspection, can be analyzed at a later point in time from photographs and/or videotapes. Photographs should be systematic, with some overlapping, so that later on in the investigation, perhaps years later, the context and orientation of each photograph is clear. Close-up photographs should be preceded or followed by an overall photograph, so that the location of the area in the close-up is clear. Videos are also helpful in this regard, but videos should not be relied on exclusively, because the quality of a video is usually not as good as a photograph.
It seems unnecessary to state the obvious—that the investigator should be familiar with the operation and capabilities of the camera, flash equipment, and film to be used. However, photography of the scene of an accident or other failure or suspected failure is more challenging and more critical than many other situations. Experienced investigators have often been frustrated when a critical photograph is found improperly exposed or out of focus. If an investigator is uncomfortable or unsure about personally taking photographs, then a professional should be retained. Witnesses should also be asked if they took photographs or videos, or if they know anybody else who did. Police and fire departments often document an accident or fire scene, and it is not unusual for neighbors or others to do so, too.
Notes and Measurements. Notetaking at the accident scene allows the investigator to establish a permanent record of observations. The investigator should take notes throughout the entire field investigation to establish the conditions present. Although many of these observations may also be captured by the visual documentation, written or audibly recorded notes compliment the visual documentation. The combination of photographs, video, and notes taken by the investigator at the accident scene provide the investigator with a thorough

permanent record of the accident scene conditions. Figure 2 shows a view of the remains of an industrial facility after a large fire and explosion. This type of accident scene would require substantial documentation.

Fig. 2 Remains surrounding an industrial facility after a fire and explosion

A safety evaluation is necessary prior to any investigative work being performed on the scene. Examples of potential safety hazards at accident scenes include heavy traffic at an automobile accident scene, the presence of energized/exposed conductors in a structure, unstable structures at a structural failure, and the presence of hazardous materials at an industrial accident site. The investigation that can be performed may be limited because of these hazards. If the investigator is unable to determine whether a hazardous condition exists, he/she may need to contact the appropriately qualified personnel to evaluate the condition. Investigators should be familiar with OSHA requirements that may affect their conduct on-site, including procedures for entering confined spaces and use of respirators and other protective equipment. An investigator who arrives on site without proper preparation may be limited in his access to potentially hazardous areas.

The investigator should be prepared to take basic size or length measurements. The type of accident being investigated determines the type of measuring device required. For example, inspection of an automobile accident scene would require a tape measure and roller wheel measuring device for measuring longer distances. The failure of a part in a precision device may require instruments such as calipers and micrometers. The investigator may not have the necessary equipment to perform all the measurements that are deemed necessary. The investigator often has to make the determination as to whether the measurements can wait until a later time or whether the measurement needs to be taken at the time of the initial scene investigation. This decision is based on a number of things, including the importance of the information, whether changes to the scene will cause the information to be altered or lost, and whether the equipment necessary to make the measurements can be quickly obtained during the course of the scene investigation. For example, an investigator may decide he/she needs a survey of an accident scene to document the location of debris at a large accident site. If it was known that the location of the important evidence was not going to change, then the survey could be performed at a later point in time. However, if there was a need to restore the accident to its original condition, then steps would have to be taken to have the survey done in a more timely manner.

Part of the scene investigation protocol includes preservation of evidence from the accident. The investigator must make a determination as to what evidence must be preserved. It is generally not desirable for an investigator to perform destructive testing/examination of the evidence during the scene investigation. These types of inspections are best left for a controlled laboratory setting that has better availability of disassembly

and inspection equipment. Additionally, an inspection performed at a laboratory at a later date allows the investigator to contact other parties who may have an interest in the inspection and allows the investigator to review information on the equipment and develop an inspection protocol that is acceptable to all interested parties. ASTM standards exist for the procedure of examining and testing of items that may become involved in litigation (Ref 4). Often, partial disassembly of equipment occurs (i.e., removal of access panels on equipment) either to better assess the condition of the equipment or to identify the manufacturer of the equipment. The investigator must make the determination as to whether the partial disassembly of the equipment may substantially alter the condition of the equipment. ASTM standards exist for the collection and preservation of physical evidence from an accident scene (Ref 5). Evidence should be labeled by the investigator to establish what it is, where it came from, when it was collected, and who collected it, so that no questions arise at a later point in time regarding these issues. Storage of the evidence should not cause any alterations to the evidence. This may require the advice of a corrosion specialist or chemist to ensure that deterioration of the evidence, such as corrosion of a fracture surface, does not occur while the evidence is in storage.

Occasionally, components that are too large to be effectively transported must be tested on-site. Additionally, in some instances, the transportation of evidence may alter it, thus rendering the results of any future off-site testing questionable. This situation may also require that testing be performed on-site. If on-site destructive testing is to be performed, steps should be taken to ensure that all interested parties are notified and allowed to participate in the testing. Legal implications of destructive testing are discussed in detail in the following section.

With due regard to evidentiary procedures and/or proper notification of interested parties, on-site materials evaluation might be considered by a materials failure analyst. Equipment and supplies considered for a portable field test kit might include the following:

- Cleaning supplies
- Measuring equipment (tape measure, calipers, micrometer, and angle-measuring device)
- Portable hardness tester
- Optical microscope and/or a borescope with low-to-moderate magnification
- Metallographic polishing and chemical etching kit
- Dye-penetrant kit
- Magnetic-particle inspection kit
- Simple chemical tests for plating and alloy identification

Visual inspection assisted by microscopy and some nondestructive test methods that do not affect the components as evidence may be possible. Magnetic-particle inspection and dye penetrants from liquid-penetrant inspection techniques leave residues that may be objectionable, unless all parties are in agreement as to the procedure and qualified labs/personnel to implement the procedures.

Other materials may need to be collected around the site for later analysis. Natural chemicals present in soil, cleaning and maintenance supplies, deicing salts, industrial pollutants, agricultural chemicals, and animal waste residues have all been found to contribute to environmental failures. A supply of clear plastic bottles, plastic bags, acetate replication tape, adhesive tape, and labels are practical items to have available when investigating a site and collecting samples.

Testing of Evidence in a Laboratory Setting. Often, parts/evidence from an accident scene require a more detailed inspection than can be adequately performed at the accident scene. Many benefits can be obtained by performing the inspection at a controlled location away from the accident scene. Benefits include:

- *Improved Examination Environment:* Inspections of parts/equipment at the accident scene often prove difficult due to inadequate lighting, weather conditions, and accessability. Additionally, a laboratory usually provides a cleaner environment, where it is much less likely that parts will be lost during the examination. More inspection equipment is usually available in a laboratory than at the accident scene.
- *Restoration of the Scene:* A systematic inspection of evidence may require a substantial period of time. As mentioned previously, a certain amount of documentation of evidence should occur at the accident scene. One of the main purposes of the on-scene examination of evidence is to ensure that information, which may be lost by evidence removal or scene cleanup, be thoroughly documented. By saving the

more detailed inspection of evidence for a later time in a more controlled environment, the scene may be restored to its original condition in a more timely manner.

- *Procurement of Information on Evidence:* Rarely does an investigator have complete knowledge of the evidence when he/she performs the initial scene investigation. By preserving the evidence and performing inspections at a later time, the investigator can research information on the evidence, such as facts related to the accident, witness interviews, background information, or the on-site investigation discussed in previous sections. This preparation helps in the examination of the evidence. The research of information may also include examination of exemplar parts/equipment.
- *Examination by Individuals with Specialized Expertise/Knowledge:* The investigator performing the scene investigation may not have the expertise to adequately perform the examination of the evidence. The investigator may require assistance from individuals with specialized experience or knowledge on the evidence being examined.

Nondestructive Examinations of Evidence. The evidence should be initially photographed in its original condition at the site. This establishes its condition at the beginning of the inspection and provides a comparison to its condition at the accident scene and after nondestructive/destructive examination. Throughout the nondestructive phase of the examination, photographic or videotaped documentation should be performed. Notes regarding the condition of the evidence should be made, and measurements should be taken as deemed necessary. The following list describes some of the types of nondestructive examinations that may be performed on the evidence:

- Visual examination
- Physical measurements
- Microscopic examination
- Light/stereo microscopy
- Scanning electron microscopy
- Electrical measurements
- Radiography
- Holography
- Liquid penetrant

As previously noted, judgement calls often have to be determined as to whether certain types of tests would be considered destructive or nondestructive, for example, whether residue from dye-penetrant examination is objectionable.

Destructive examination refers to examination that permanently alters the condition of the evidence, such that it cannot be returned to its condition at the time of the accident. Although this statement appears straightforward, there exists a gray area between what is considered destructive and nondestructive. Some tests, such as operational tests of equipment, may or may not be considered destructive testing, depending on the nature of the accident. The investigator must be careful in evaluating whether examinations that are performed are destructive or nondestructive. The distinction between the two is important in the United States and other countries whose legal systems require the preservation of evidence. “Spoilation of evidence” is a legal term used in the United States to describe the destruction or alteration of evidence that is or may be relevant to litigation. By law, individuals have an obligation to preserve evidence of this nature. As a practical matter, this obligation would occur as soon as the investigator has reason to believe or suspect that a particular machine or component has the potential to be critical to the reconstruction of the accident.

Any destructive testing that is ultimately performed on the evidence must be under the agreement of all parties that are or may become involved in any litigation pertaining to the accident. Therefore, prior to performing any destructive testing, all interested parties should be notified and allowed to participate in both the procedure for destructive testing and in the destructive testing itself. A mutually agreed on protocol for the destructive testing should be established. For example, Fig. 3 shows the test setup for operational testing of a high-limit thermostat from a spa. The testing of the thermostat occurred in a controlled environment rather than at the scene of the accident, where conditions were not conducive to such testing.

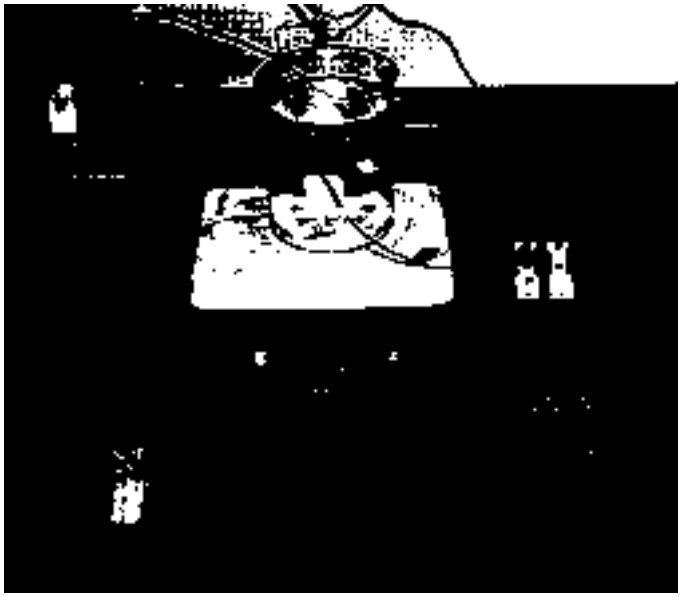


Fig. 3 Example of a photograph to document a test of a spa thermostat

For accident reconstruction being performed with no legal implications, the procedures used for preservation of evidence and destructive testing will not come under the scrutiny of the legal system. However, a systematic approach to examination of evidence should still be used.

Photographic and videotape documentation should occur throughout the destructive testing. This documentation should capture how the evidence was changed or altered and should document new information/conditions found as a result of the examination. The type of destructive examination is dependent on each individual case. The purpose of destructive examination is to further reveal information about the evidence, which may help in the overall accident reconstruction. Destructive testing is often necessary to determine:

- Identification of the evidence
- The condition of the evidence
- The construction of the evidence
- Characteristics/properties of the evidence, for example, physical, chemical, mechanical, electrical, or other
- Whether the evidence functioned properly
- Whether the evidence was altered prior to the accident
- Whether the evidence contributed to the accident

Generation of Hypotheses: Initial Stages of Reconstructing the Accident. Throughout the investigation, the investigator most probably is formulating thoughts on how and why the accident occurred. The investigator should keep an open mind throughout the information-gathering process and not jump to any premature conclusions. Prior to forming any firm hypotheses on the accident, all the information obtained should be reviewed and analyzed to form hypotheses or theories regarding the cause of the accident. The reconstruction of the accident is often analogized to fitting together the pieces of a puzzle. The reconstruction analyst should take into account all of the data generated during the investigation and determine all of the possibilities as to how and why the accident occurred that are consistent with all the collected data. The investigator must also consider the impact and effect of any missing or unobtainable data. The investigator/reconstruction analyst should be careful to consider *all possibilities* at this stage of the investigation.

The possibilities that are initially determined should be considered hypotheses or theories. The hypotheses need to be tested to determine whether they adequately explain the accident and are consistent with all the information collected. Testing of the hypothesis may be done by cognitive reasoning, modeling techniques, or actual experimentation. The need for modeling or physical testing depends on the circumstances of each individual accident. The difficulty of recreating an accident is often underestimated by investigators.

Testing of Hypotheses (Deductive Reasoning). The testing of a hypothesis or theory can be done in a number of ways, including cognitive reasoning, modeling, or physical testing.

Cognitive reasoning involves the application of physical and engineering laws/principles to determine the validity of the reconstruction hypothesis. Many investigations cease at this point, with little to no modeling or physical testing being performed. Many accidents, especially those with cause in dispute, often require further work beyond cognitive reasoning to prove that the hypotheses regarding the cause of the accident are correct and that alternate hypotheses are incorrect.

A simple example of this would be the investigation of a fire in which the investigator determines that arced wiring was the cause of the fire. The investigator may use cognitive reasoning to come to the conclusion that a defect in the wiring allowed current to flow between the conductors, which heated the conductor insulation and surrounding combustible materials to a sufficient temperature that ignition of the materials occurred. The investigator would need to be cautious when using such an approach in this type of situation. A number of questions may arise that could require testing and/or further analysis. For example, questions might include:

- Was the arcing of the subject conductors the cause of the fire?
- Could an existing fire have spread to the subject energized conductors, causing the arced condition?
- Was the current flow between the conductors sufficient to trip a circuit breaker?
- Was current flow sufficient to heat the materials to sufficient temperature to cause a fire?
- Were sufficient combustible materials present to sustain a fire?

Modeling. A model is a physical, mathematical, or logical representation of a physical system or process. The use of modeling can be a very powerful tool for information pertaining to the reconstruction of an accident. Whether the modeling consists of simple mathematical modeling of a physical system or more complex computer modeling of a system or part, the modeling technique may directly or indirectly give insight to the cause of the accident. Modeling may be used prior to or after the formulation of hypotheses. The subject of modeling is discussed further in the following section.

Testing/Experimentation. Testing of exemplar parts/equipment can be extremely beneficial in determining the validity of hypotheses regarding the accident. Tests of equipment/parts often provide conclusive evidence to prove or disprove hypotheses generated in earlier stages of the investigation. Testing can often be used to confirm results generated from theoretical modeling. The paper and/or computer-generated models rely heavily on having the correct data input. False input data result in inaccurate model results. Testing of actual components, in most cases, requires different assumptions to be made, specifically that the arrangement of the component and conditions during the test are relevant to the accident.

Testing may also reveal other occurrences that may help rule out certain possibilities. Often, testing is not feasible due to cost or other limitations. Full-scale testing to simulate large, catastrophic, or costly accidents, such as building collapses or explosions, are usually not possible due to cost and/or hazard limitations. The economics of an accident often dictate the degree of testing that can be performed. Although full-scale testing is not always feasible, smaller-scale testing/experimentation can provide information that may be vital in determining the validity of hypotheses.

Drawing of Conclusions. After testing the hypothesis, the reconstruction analyst formulates conclusions. The reconstruction analyst may or may not have sufficient information to draw conclusions pertaining to the cause of the accident. If there is not a hypothesis that is clearly correct, further testing may be necessary, or possibly, new or alternate hypotheses need be explored. The reconstruction analyst may never reach the point where he/she can fully reconstruct the accident. In these cases, more limited conclusions may be drawn that pertain to the cause of the accident. Whatever conclusions are drawn should be able to withstand an examination by deductive reasoning, and the basis for the conclusions should be clearly stated.

References cited in this section

1. J.S. Baker, and L.B. Fricke, “The Traffic-Accident Investigation Manual,” Northwestern University Traffic Institute, 1986

2. “Guide for Fire and Explosion Investigations,” NFPA 921, National Fire Protection Agency, 2001

3. “Accident Prevention Manual for Business and Industry,” National Safety Council, 1997

4. “Standard Practice for Examining and Testing Items That Are or May Become Involved in Products Liability Litigation,” ASTM E 860, *Annual Book of ASTM Standards,* ASTM

5. “Standard Practice for Collection and Preservation of Information and Physical Items by a Technical Investigator,” ASTM E 1188, *Annual Book of ASTM Standards,* ASTM

Modeling and Accident Reconstruction

Charles R. Manning, Jr., and Thomas C. Wenzel, Accident Reconstruction Analysis Incorporated

Modeling

A model is a physical, mathematical, or logical representation of a physical system or process. The use of modeling can be a very powerful tool for information pertaining to the reconstruction of an accident. Whether the modeling consists of simple mathematical modeling of a physical system or more complex computer modeling of a system or part, the modeling technique may directly or indirectly give insight to the cause of the accident. As previously noted, modeling may be used prior to or after the formulation of hypotheses.

The use of computers has greatly enhanced the ability to perform more complex calculations, modeling, and simulations quickly and thoroughly. The computer or computer program does not act as a replacement for the human analyst, but iterative calculations/procedures associated with numerical method solution techniques often performed in models can be quickly performed using a computer. Additionally, sensitivity analysis can be performed that involves changing the inputs to the model to evaluate the effect on the output of the model. This type of analysis allows the user to determine the impact that small or large changes in input have on the output. Complex models, too large to perform by hand, are now commonly evaluated using computers. The use of computers requires not only validation of the model but also verification of the computer code.

General Types of Modeling

A model can be defined as a representation of a physical system or process. Engineers and scientists commonly use models for both illustrative and analytical purposes. General types of models are illustrative or analytical.

Illustrative Modeling. Illustrative models are intended to provide a visual representation of a system in a manner intended to communicate concepts, ideas, and/or information. Illustrative models are intended to represent a relatively complex physical system in a manner that is more easily understandable. Both of these aforementioned models can be represented two dimensionally, as a drawing, or three dimensionally, using construction materials. Common examples of illustrative models used by engineering students include phase diagrams, crystal lattices, and circuit diagrams. Illustrative models may or may not have any actual visual similarities to the object or process that they represent. Whereas a scale model of an automobile looks like the real thing, a circuit diagram for the electrical system of the same automobile reveals little to no information regarding how the electrical system visually appears.

Of ever increasing importance is the generation of a geometric model using a computer. Computer-generated two-dimensional or three-dimensional models are used to generate drawings, to develop a finite element model, or to interface directly with computer-aided manufacturing to provide information for automated machining of a part. Figure 4 shows a computer-generated diagram of an accident scene in which a vehicle was driven into a guy wire from a power transmission pole.

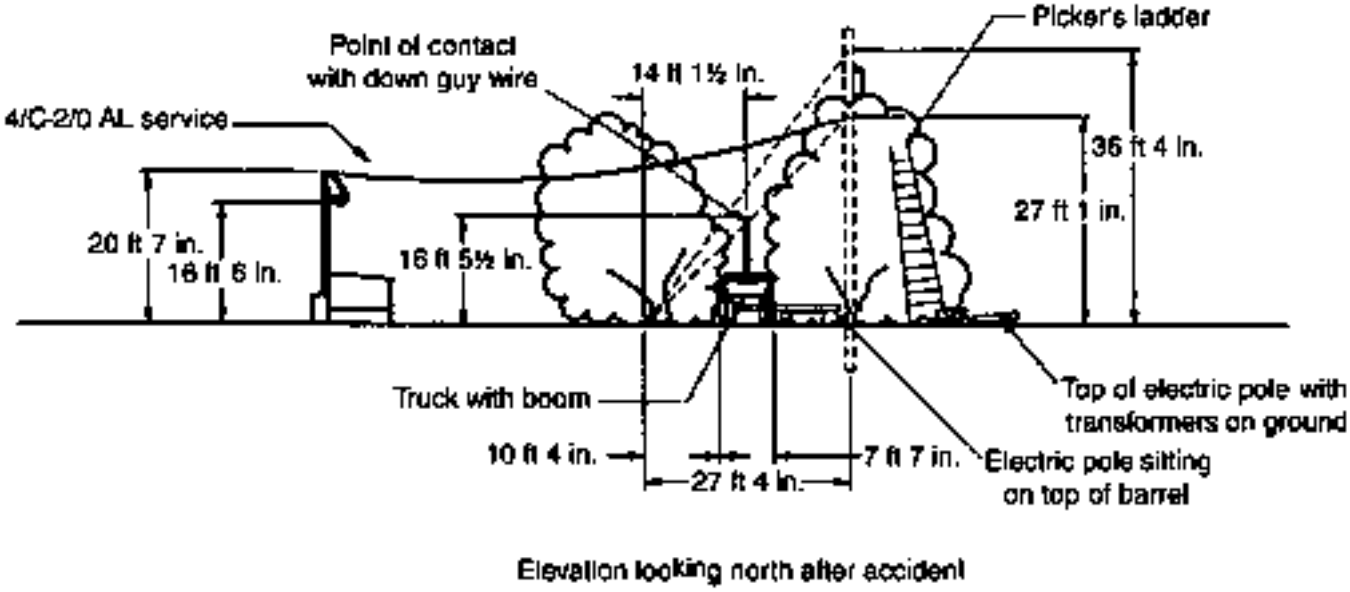


Fig. 4 Example of computer-generated diagram that illustrates an accident scene in which a vehicle was driven into a guy wire of a power transmission pole

Analytical/mathematical models give a representation of physical system or process in a different manner. These models are based on mathematical equations and are intended to be deterministic or predictive of the outcome of an event or process. In addition to providing a representation of a physical system, mathematical models are used for problem solving. Mathematical modeling is often referred to as applied mathematics.

In its simplest form, modeling is the use of mathematics to represent a real-world event, such as in the mathematical model of Newton's second law, $F = ma$, which relates force, F; mass, M; and acceleration a. This equation is a basic physical law. Use of properly selected equations allows one to predict the outcome of a physical process without performing experimentation each time a new evaluation must be made. Thousands of equations are used by the engineering disciplines to model physical systems in electrical, chemical, mechanical, and metallurgical engineering.

The more common perception of modeling is the representation of a more complex system. More complex models employ the use of multiple equations to lead to quantitative results. Examples of more complex models might include models to predict the growth of an economy, the spread of a disease in a society, or the orbit of a planet. Generally, more complex models require more assumptions and involve more variables.

Modeling of more complex systems may be classified into two different categories: deterministic and stochastic. A deterministic model is one in which there are no random elements present and all variables have known values. Only one possible output can occur from a given set of input data. A stochastic model is one that incorporates random behavior. The outputs from such a model are therefore random estimates that must be dealt with statistically. Multiple outputs may exist from a given set of input data. Of the three models listed in the previous paragraph, the growth of an economy and the spread of a disease are considered stochastic models; they both require the use of random data. The orbit of a planet requires a large number of assumptions, but the model does not normally require the use of any random data and therefore is not considered stochastic.

Simulation is a combination of illustrative and analytical techniques. Simulation involves manipulation of a model by subjecting the model to various inputs and conditions to observe how they affect the results of the model. Simulations allow for experimentation with the model to determine the effects on the behavior of the system. Most simulations are computer generated and may use a combination of equations and empirical data.

Limitations of Modeling

The cost of generating an accurate model of a complex system can be very expensive. Additionally, validating the model can be both costly and difficult.

Validation of the model to ensure that the model accurately represents the system being modeled has varying levels of importance, depending on how the results will be used. For example, assume a designer has the responsibility of designing a new product. The designer may decide to choose a modeling technique to predict

how the product ultimately performs. The goal of the designer is to model the product as accurately as possible, given cost and time limitations. Ultimately, prototypes of the product are made, and the product is experimentally tested to determine if it meets the performance goals. In this situation, the designer is not relying on the model to actually achieve the final product; he/she is relying on the model to get him as close as possible to the final design. Poor modeling assumptions may result in higher design costs, due to further refinement of the model and/or product, but should not result in a faulty design, assuming proper testing is performed. In this situation, validation of the model is not critical for the development of prototypes.

Another example is the use of a modeling technique to represent a part that has failed in service. An investigating engineer develops the model to help determine why the part has failed. The engineer decides not to perform any physical testing. In this situation, validation of the model is critical if the engineer plans on using the model to draw conclusions on the cause of the failure. Validation of a model in this situation may be difficult to impossible to perform, thus making any conclusions drawn from the model suspect.

Human error results obtained from a model are highly dependent on the input data. Use of a model (or computer program) that has been validated is not a substitute for human knowledge. Incorrect use of the program leads to results that are inaccurate. The user of the program must have a thorough understanding of what process is being used in the model to understand whether the results he/she is obtaining are reasonable and accurate. It is typical to make many simplifying assumptions, requiring the user to be fully aware of the impact of each assumption. It is often necessary for the user to perform a sensitivity study to determine the potential impact of changes to the assumptions.

Use of Modeling in Accident Reconstruction

Modeling is commonly used for accident reconstruction in the evaluation of accidents. The modeling may be both analytical and illustrative. If the reconstruction analyst chooses to use an analytical model in the evaluation of the accident, he/she must be careful in choosing how the technique is used. If he/she plans on using it by itself without any testing, the assumptions made for the model should be well known and the model should be proven. If the model is being used in combination with testing, then the accuracy of the model is not as critical, because proper testing should determine the validity of the model.

Whether modeling, testing, or both are performed for an accident reconstruction depends on a number of factors, including:

- Cost
- Availability of parts for testing
- Safety of testing

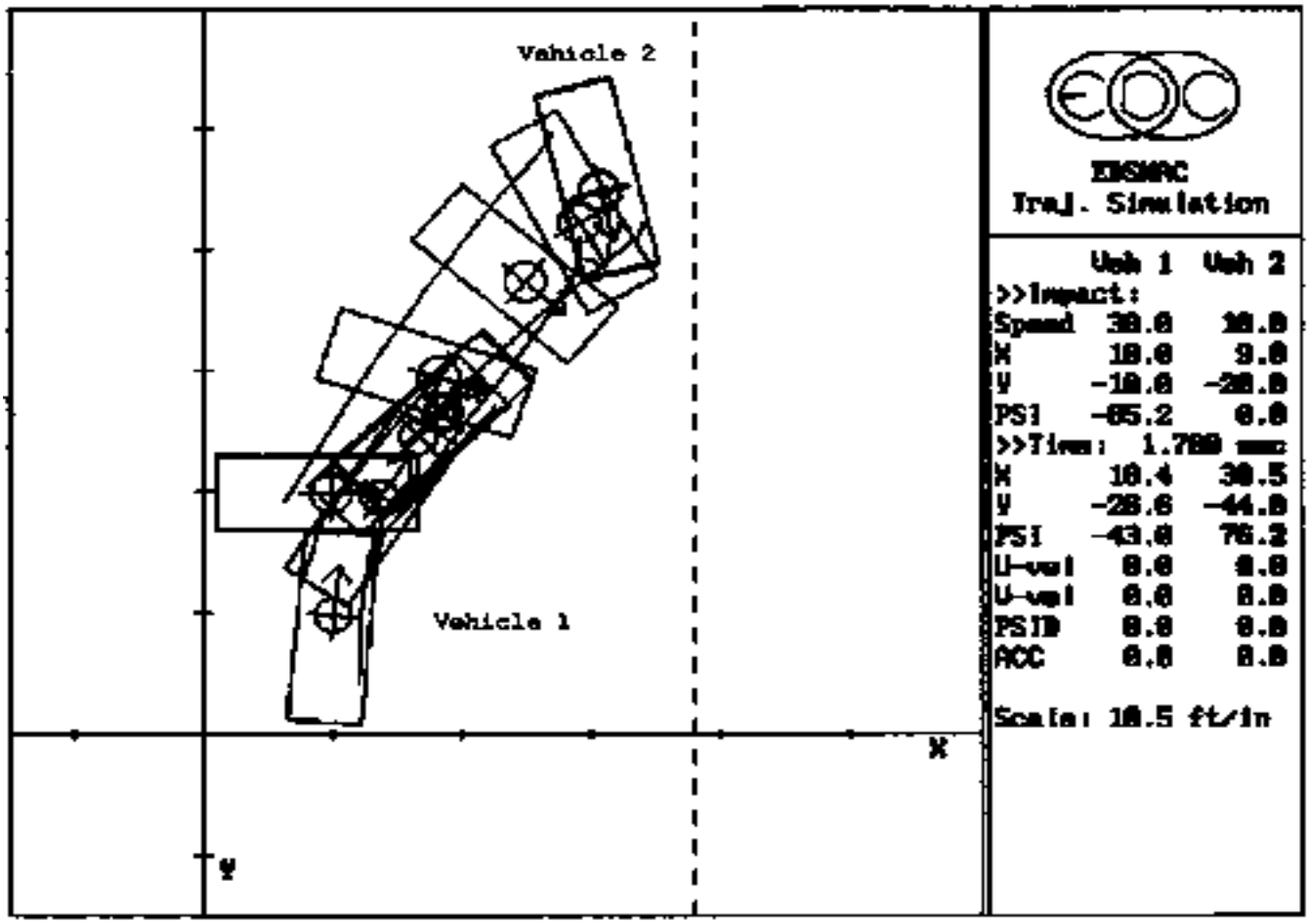
Clearly, testing to recreate a high-speed car accident is not feasible in most cases. It is very costly, and, most likely, there exists substantial unresolved safety issues. This is one of the reasons why substantial research has been performed over the years to develop reliable and accurate modeling techniques for automobile accident reconstruction. Testing is a more feasible option in other cases. An example of this is the failure of an office chair. Testing of exemplar chairs likely reveals the conditions necessary for failure. Although simple calculations may be performed to complement the testing, more complex modeling techniques for such a failure are likely more expensive and result in less conclusive results.

Significant emphasis in this article has been placed on the accuracy of modeling. This emphasis arises from the practice of many accident reconstruction experts of performing some type of modeling and then drawing conclusions from the model (simple equations or complex models) without performing any physical testing. In many cases, this has resulted in accurate conclusions. In many cases, however, this has resulted in conclusions that are inaccurate, and this often serves a specific agenda as opposed to having any real technical merit. The results obtained from physical testing are generally less susceptible to inaccuracies. Although physical testing may still not result in accurate conclusions, fewer assumptions are required, usually resulting in more accurate results. The accuracy of the results depends on whether the proper test conditions are used.

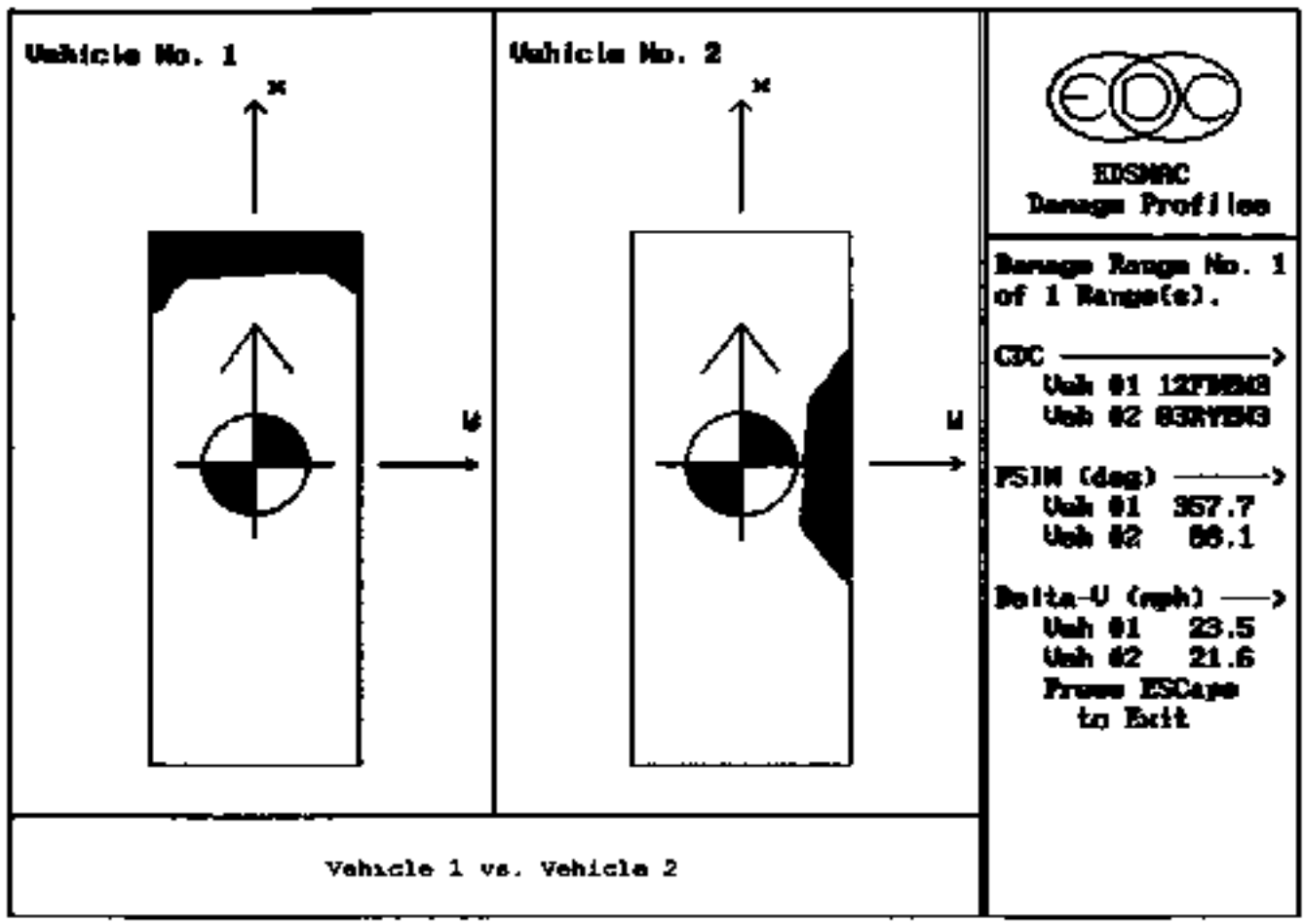
Motor Vehicle Accident Reconstruction. Mathematical modeling techniques used for automobile accident reconstruction have become very common. In the past and to this day, much of the numerical physics calculations are performed by hand (normally with the aid of a calculator). These calculations allow the user to address many aspects of an accident, including things such as braking distances and distances traveled. Many

commercially available computer programs exist for automobile accident reconstruction. They range from simple programs performing simple physics calculations that can easily be performed without the use of a computer to more complex programs using series of equations and calculations that would not be considered practical to be performed by hand. The more sophisticated programs allow for a more thorough analysis of the accident.

The National Highway Transportation Safety Administration contracted with Cornell Aeronautical Laboratories in the 1960s to develop computer programs to analyze motor vehicle accidents. A number of programs were developed, including CRASH (Calspan Corporation Reconstruction of Accident Speeds on Highways) and SMAC (Simulation Model of Automobile Collisions). The CRASH program is used to calculate impact speeds and severity of impacts based on information collected from both the vehicles and the accident scene. The SMAC is used to simulate two-car collisions. The user inputs vehicle speeds and positions, and the program simulates the collision and predicts both the movements of the vehicles and damage to each vehicle. Figure 5 shows a printout of the results of a trajectory simulation and vehicle damage profile from a multivehicle accident using a commercially available version of SMAC called EDSMAC (Engineering Dynamics Corporation). These are only two examples of a number of commercially available programs that pertain to automobile accident reconstruction.


Vehicle 2
Vehicle 1
X
Y
EDSMAC
Traj. Simulation
Veh 1 Veh 2
>>Impact:
Speed 30.0 10.0
X 10.0 9.0
Y -10.0 -20.0
PSI -65.2 0.0
>>Time: 1.700 sec
X 10.4 30.5
Y -28.6 -44.0
PSI -43.0 76.2
U-vel 0.0 0.0
V-vel 0.0 0.0
PSID 0.0 0.0
ACC 0.0 0.0
Scale: 10.5 ft/in

(a)


Vehicle No. 1
Vehicle No. 2
X
U
EDSMAC
Damage Profiles
Damage Range No. 1
of 1 Range(s).
CDC
Veh #1 12FBEW3
Veh #2 03RYEW3
PSIN (deg)
Veh #1 357.7
Veh #2 86.1
Delta-V (mph)
Veh #1 23.5
Veh #2 21.6
Press ESCape
to Exit
Vehicle 1 vs. Vehicle 2

(b)

Fig. 5 Printout of a simulation model of an automobile collision. (a) Trajectory simulation. (b) Vehicle damage profile

Vehicles also are crash tested by manufacturers to demonstrate conformance with various federal motor vehicle safety standards. Other organizations also conduct crash tests to study how vehicles perform in various types of accidents. The empirical data generated from some of the organizations/manufacturers is available and can be used in the estimation of such things as, for example, the impact speed based on deformation of the vehicle.

Aircraft Flight Simulators. The use of flight simulation software allows an investigator to evaluate a number of different variations in flight conditions in the evaluation of an aircraft accident. The use of flight simulation software is an extremely valuable tool for the investigator, because experimental flight testing to determine such variations most likely cannot be done safely. Additionally, the cost of performing actual flight testing is not feasible in many cases. Information from flight conditions can be obtained from flight recorder and radar data.

As with any other model, the simulation obtained from the computer is dependent on the model developed and the user inputs. The user of the program should be fully aware of the operation and limitations of the program, including what conditions (such as the normal flight envelope) the simulator can accurately simulate. Simulation programs often are limited to flight within the flight envelop and cannot accurately simulate abnormal conditions. The user should be aware of what aspects of the simulation have been verified by actual flight testing and what aspects are based on mathematical modeling only.

Finite-element analysis (FEA) is a computer-based numerical method used for a number of different analysis problems, including stress/strain analysis, vibration, deflection, stability, fluid flow analysis, heat transfer analysis, civil engineering structural analysis, and others. In stress/strain analysis of a solid object, the object is modeled with a mesh of separate elements connected together at their corners by points called nodes. The idea of dividing a domain into smaller subdomains is one of the basic principles of FEA, as described in more detail in the following article, “Finite Element Modeling in Failure Analysis” in this Volume.

A number of different shapes are commonly used for the elements that comprise the mesh. Because both the size and the arrangements of the elements can be selected in an infinite number of ways, very complex shapes can be analyzed using this method. Stress and strain equations are used to compute the deflection of each element. The external loads are transmitted through the structure by the nodal points connecting each element. The analysis is complex, because the elements are acting similar to a set of springs—the deflection of each spring is dependent on the forces acting on its nodal points.

A complex system of equations is required for the analysis related to the interaction of the elements. The larger the number of elements in the mesh, the larger the number of equations necessary to achieve a solution. Usually, it is desirable to use the smallest number of elements to achieve the needed accuracy. Element density may be increased in localized areas of the part that are of greater concern while maintaining a smaller element density in less critical areas. Larger numbers of elements achieve greater accuracy but at the cost of a more complex, expensive, and time-consuming analysis. The use of computers is essential for the efficient solution to the series of equations. While it is possible to perform FEA by hand, it is not considered practical, because FEA can require hundreds of simultaneous equations with as many unknown terms.

Originally, the most common use for FEA was design. However, with the increasing speed of computers and the number of qualified personnel capable of performing FEA, it is more commonly being used in the field of accident reconstruction/failure analysis. Finite-element analysis may be used to determine the stress concentrations in a part that has failed or to determine whether a part was resonating at its natural frequency at the time of failure. As with other modeling techniques discussed, the user must have thorough knowledge of the technique and know the limitations of the technique. Figure 6 shows the results of a finite-stress analysis performed on the strut area of a mechanical heart valve. The heart valve had experienced in-service failures.

Fig. 6 Black-and-white reproduction of color output from a finite-element analysis program

Fracture Mechanics Software. Several different software packages are commercially available that perform fracture mechanics analysis. The software can be used to predict the growth of preexisting flaws and cracks in stressed components. The software can also be used to determine the damage tolerance of a given structural design. The software uses numerical methods and equations related to the calculation of stress-intensity factors. Some of the programs use FEA for the calculation of stress intensities and evaluation of crack propagation rates.

The use of FEA and fracture mechanics analysis as well as other similar techniques can provide valuable confirmation that the failure analysis is correct and complete. Specifically, the loads that the analyst believes existed in service should be used to compute stresses and, if appropriate, stress intensities, and then these values can be compared with the material properties to determine if, in fact, a fracture would be expected in the location observed. If an inconsistency is found, then further work is necessary to determine if there is an unknown loading in service, if the analysis is not representative of the part, or if some other factor has been overlooked. For example, it has been known to happen that a part fractures at a location other than the maximum stress predicted by the stress analysis. Such situations require further investigation.

Commercially Available Software. The previous examples touch on some of the modeling/simulation techniques being used for accident reconstruction. Commercial software is available that pertains to some aspect of practically every type of accident reconstruction. The modeling software may be both illustrative and analytical. The majority of the analytical software does not necessarily attempt to generate a model and simulate the entire sequence of the accident. Rather, the software is designed to solve aspects of the accident not easily determined by hand calculations. Examples include software to determine groundwater activity, human movement in a motor vehicle, smoke and fire propagation, rates of corrosion, and fluid flow characteristics.

Computer-generated animation is a technique used by reconstruction analysts to dynamically illustrate the results of their reconstruction. Animation programs are not analytical tools; the objects in an animation can be programmed to perform almost any act in defiance of the physical laws. The reconstruction analyst must be careful in generating an animation to ensure that the animation is an accurate depiction of the reconstruction of the accident. The use of animation has become very popular, because it often provides a very realistic-appearing visual representation of an accident or process.

Charles R. Manning, Jr., and Thomas C. Wenzel, Accident Reconstruction Analysis Incorporated

References

1. J.S. Baker, and L.B. Fricke, “The Traffic-Accident Investigation Manual,” Northwestern University Traffic Institute, 1986

2. “Guide for Fire and Explosion Investigations,” NFPA 921, National Fire Protection Agency, 2001

3. “Accident Prevention Manual for Business and Industry,” National Safety Council, 1997

4. “Standard Practice for Examining and Testing Items That Are or May Become Involved in Products Liability Litigation,” ASTM E 860, *Annual Book of ASTM Standards,* ASTM

5. “Standard Practice for Collection and Preservation of Information and Physical Items by a Technical Investigator,” ASTM E 1188, *Annual Book of ASTM Standards,* ASTM

Modeling and Accident Reconstruction

Charles R. Manning, Jr., and Thomas C. Wenzel, Accident Reconstruction Analysis Incorporated

Selected References

- S.J. Brown, Brittle Fracture Explosive Failure of a Pressurized Railroad Tank Car, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, 240–248

Finite Element Modeling in Failure Analysis

Thomas Service, Altran Corporation

Introduction

FAILURE of a structural or mechanical component usually can be associated with materials-related problems and/or design-related problems (which may include, depending on the definition of design, unexpected service environment). Materials failure analysis investigations are usually carried out first, and the objective is to determine what the material can reveal about the cause of failure. This type of failure analysis, as discussed in this Volume, consists of first evaluating the failed material for evidence of the failure mechanism such as fatigue, overload, corrosion, environmentally assisted cracking, and so forth. The material is then examined in relation to the design specifications for noncompliance in chemistry, composition, or structure. It is not uncommon for the wrong material to have been used in an application that results in failure. Finally, the material is examined for inherent defects in manufacture, processing, or assembly. Microstructural defects, subsurface laminations, and welding defects are common examples of materials defects that cause failure.

In many cases when there is clear evidence of a materials-related cause of failure, simply changing the material to one that is more resistant to the specific cause of failure will eliminate the problem. However, after a materials failure analysis is complete, often there is no material factor associated directly with the failure. Moreover, in some cases the material may indeed have an imperfection or discontinuity, but it is not clear whether the imperfection is a defect (i.e., the cause of failure). For example, in some overload or fatigue failures the material performed to expectations, but it simply could not withstand the loads and stresses experienced during use. In some applications it is also not feasible to change the material.

At this point a design-based failure analysis may be performed to understand the specific cause of failure. This consists of an engineering review of the component design and service application to determine the loads, displacements, temperatures, vibrations, and other service-related factors that the part experiences. It may also include numerous analytical, classical, and computer techniques that are available to assist in a structural design failure analysis. A traditional closed-form analytical technique such as beam analysis or plate theory provides a simple and quick way to estimate stresses and deformations in structures and components. It is usually the first technique in both the design of a component and in the analysis of a failed component. However it has its limitations in that only relatively simple and idealized structures can be analyzed using simplified loading and materials property assumptions. When complex designs, transient loadings, and nonlinear material behavior need to be evaluated, computer-based techniques are used. This is where finite element analysis (FEA) is most applicable and provides considerable assistance in design analysis as well as failure analysis.

Finite element analysis is one of the most common tools used by design engineers. It has application in the structural/mechanical fields to determine the stress, strain, and displacements of structures subjected to different types of loading. It also has applications in the heat transfer and thermal fluid areas and, most recently, in the electromagnetic area. This article is not meant to review the current state of FEA development but rather to highlight its uses in the area of failure analysis and design with emphasis on structural analysis.

Finite Element Modeling in Failure Analysis

Thomas Service, Altran Corporation

General Development of FEA

Before the development of FEA, engineering structural analysis and design was carried out using classical techniques developed and refined since the mid-nineteenth century. Any analysis of a structure to determine the response to loads requires that the basic physical relationships of equilibrium/strain compatibility and material constitutive behavior/boundary conditions be satisfied. Classical theory of elasticity (Ref 1, 2, 3) involves developing complex differential equations to satisfy these requirements, but such an approach is practical only for situations where closed-form solutions are available. Simplifying many engineering structures to more basic components allowed the development of beam and plate theories; these are described in many standard engineering texts (Ref 4, 5, 6).

Finite element analysis is a computer-based analytical technique that allows numerical solutions to be obtained for complex mathematical and engineering problems. It is a technique that relies on creating a geometric mathematical model of the structure out of discrete or finite numbers of individual nodes and elements. Displacement functions are assumed that allow displacements throughout the elements to be evaluated. Once a discrete-element model has been created, mathematical techniques are used to obtain a set of equilibrium equations for each element and the entire model. By applying various boundary conditions and loads to the model, the solution of the simultaneous set of equations provides the resulting displacements anywhere in the model while still providing continuity and equilibrium. Assuming various constitutive relationships allows the results to be expressed in strain or stress rather than displacement (Ref 7, 8, 9).

While FEA is a powerful tool in structural analysis, its usefulness and accuracy depend on the ability to create accurate, realistic models of a component that cannot be analyzed using traditional techniques. Such a model usually consists of large complex geometry with hundreds if not thousands of nodes and elements. This complexity was a limiting factor preventing the widespread use of FEA before the use of interactive design. Models had to be created by hand, requiring the input of the geometry of each individual node and element. The analyst needed to determine the geometric coordinates of each individual node and then connect them correctly to form the elements—a time-consuming task. These models then had to be solved, and the limiting size of models was the memory and computing speed of the computers available.

Because computing power has been (and still is) the limiting factor in the generic use of FEA, models and analyses tend to be simplified as much as possible through the use of two-dimensional (2-D) analyses, symmetry, linearity, and steady-state conditions. As computing power increased, so did the complexity of finite element modeling and analysis techniques. The first analyses and elements available were limited to 2-D beams

and spars. The progression was then towards three-dimensional (3-D) spars and beams, 2-D solid and axisymmetric elements, and the full 3-D elements (see Fig. 1). Along with the progression in element capability was the advancement toward nonlinear and transient analyses. Contact elements, elastic/plastic properties, large deflection, and viscoplastic and dynamic impact analyses are now readily available.

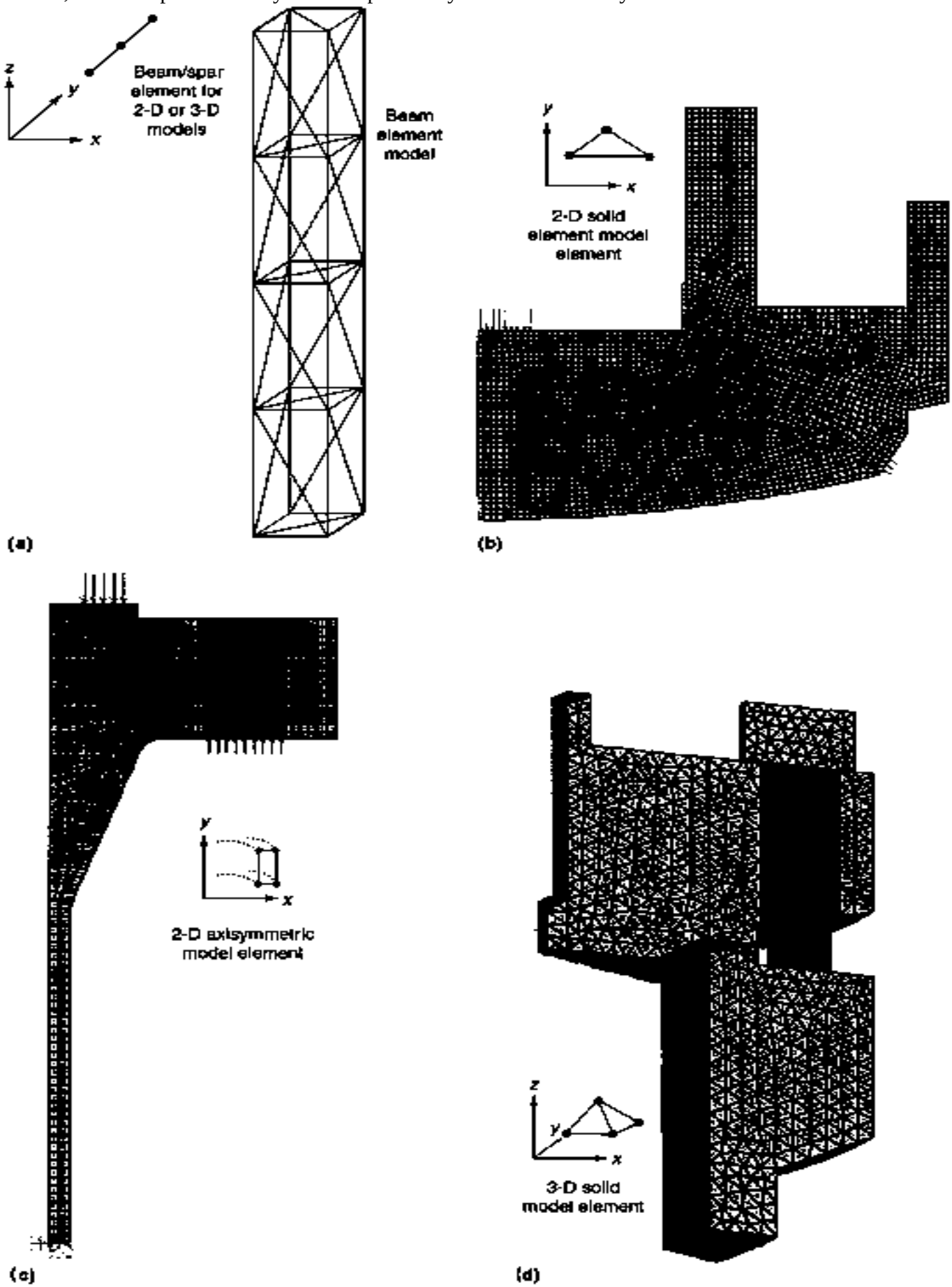


Fig. 1 Common types of finite elements used in modeling with examples. (a) Beam spar elements used to construct, for example, a beam element model. (b) Two-dimensional solid-model element with example. (c) Two-dimensional axisymmetric solid-model element with example. (d) Three-dimensional solid-model elements with simplified example

The other major advancement in FEA is in the area of interactive model generation, solid modeling, and the interface with other design software. It is now possible to quickly create 3-D solid models with automatic mesh generation rather than having to input each individual node location. It is common for desktop computer and workstations to be able to handle complex 3-D FEA linear and steady-state models and simpler nonlinear and transient analyses. The largest nonlinear, transient, and dynamic models are still best solved on the fastest mainframes. However, as computing power continues to increase, so will the ability to solve even more complex problems.

References cited in this section

1. E.H. Love, *A Treatise on the Mathematical Theory of Elasticity,* 4th ed., Dover Publications, New York, 1944

2. Y.C. Fung, *Foundations of Solid Mechanics,* Prentice-Hall, Englewood, NJ, 1965

3. L.E. Malvern, *Introduction to the Mechanics of a Continuous Medium,* Prentice-Hall, Englewood, NJ, 1969

4. W.C. Young and R.G. Budynas, *Roark's Formulas for Stress and Strain,* 7th ed., McGraw-Hill Publishing Co., 2001

5. J.M. Gere and S.P. Timoshenko, *Mechanics of Materials,* 4th ed., PWS Publishers, Boston, 1997

6. F.P. Beer et al, *Mechanics of Materials,* 3rd ed., McGraw-Hill Publishing Co., 2001

7. O.C. Zienkewicz, *The Finite Element Method in Engineering Science,* McGraw-Hill Publishing Co., 1971

8. R.H. Gallagher, *Finite Element Analysis: Fundamentals,* Prentice-Hall, Englewood, NJ, 1975

9. R. MacNeal, *Finite Elements: Their Design and Performance,* Marcel Dekker, Inc., 1994

Finite Element Modeling in Failure Analysis

Thomas Service, Altran Corporation

General-Purpose Applications

General-purpose FEA software programs can perform a variety of analyses including structural and thermal, static and transient, linear and nonlinear, and more. The programs are thorough, detailed, and relatively complex and rely heavily on the analyst to accurately create models, understand the solution process, and interpret the results. Currently there are numerous commercial software packages available with a variety of options designed for use on all types of personal computer (PC) and mainframe platforms.

Generating models through interactive design and importing model geometry from computer-aided design (CAD) software is making FEA analysis much easier. However, with this increased ease comes the increased

need to correctly understand and interpret FEA results. Care must be taken so that models are developed and analyses are performed by trained personnel who understand the technology.

Structural Analysis. The most common application of FEA is in the determination of displacement and stress profiles in structures. Bending beams, rotating disks, pressure vessels, buildings and bridges, airplanes and automobiles, and heart valves and hip joints are all analyzed using FEA structural analysis techniques. This structural analysis area of FEA provides results analogous to those obtained through classical beam and plate technique with the advantage that much more complex geometries and loadings can be evaluated. More important is that the FEA provides results that are immediately available for the whole structure rather than one location that is identified for analysis using classical techniques. This enables graphical representations of the results that are more easily visualized and interpreted than analytical results. The graphical representations can clearly highlight areas of stress concentration and excessive deformation relative to the entire structure. This graphical capability makes FEA particularly useful when performing failure analysis investigations.

Consider the case of a simple uniformly loaded cantilever beam shown in Fig. 2(a). The stress and deflection profiles for this structural member are derivable from beam theory and are readily available in many engineering tests. Closed-form equations provide the results for any location desired in the beam, provided it meets the criteria assumed in beam theory. For example, if the beam length is reduced to a value similar to the beam height, then the solutions are not valid and other techniques must be employed to determine the stress and deflection. Now consider the same beam problem solved using finite element techniques. Figure 2(b) shows a simple finite element model and loading and boundary conditions for the same problem. The solution provides not only a graphical depiction of the deflection profile but also the stress contours throughout the beam. Further, unlike beam theory, the solution is valid for all beam length and height dimensions.

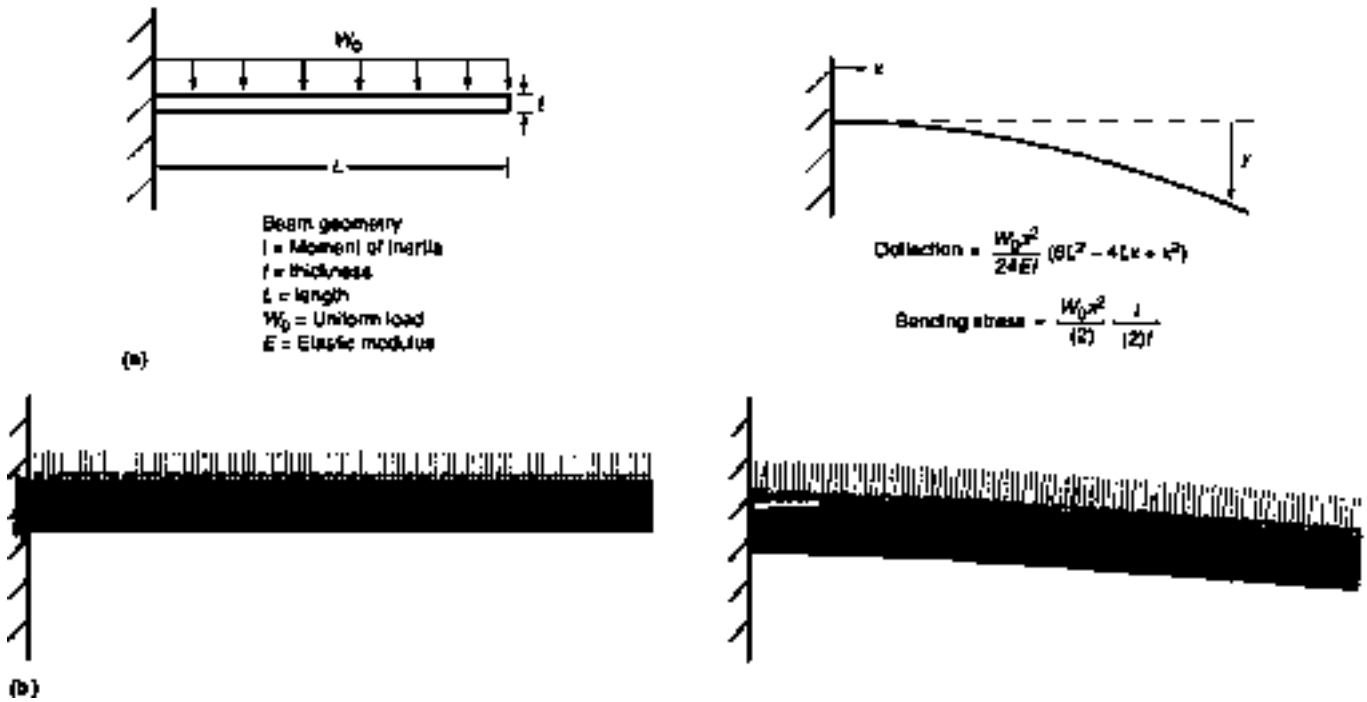


Fig. 2 Simple cantilever beam analysis. (a) Classical beam theory showing key elements (top left) and stress and deflection (top right). (b) Finite element analysis (FEA) cantilever beam elements (bottom left) with FEA cantilever beam results (bottom right)

Now consider the same cantilever problem but with the addition of two simple supports under the beam (Fig. 3). This simple change makes the problem statically indeterminate and thus adds an order of magnitude to the complexity of solution using beam theory. Techniques such as the moment-area method, double integration, and superposition have been developed to provide analytical solutions to these types of problems. Using FEA, no such complexity exists. Simply adding deflection boundary constraints at the appropriate locations will provide the correct solution. Again, the complete deflection and stress profiles are quickly determined, providing the necessary information at any location in the beam. Further, the localized stress profile in the area of the support points is also known.

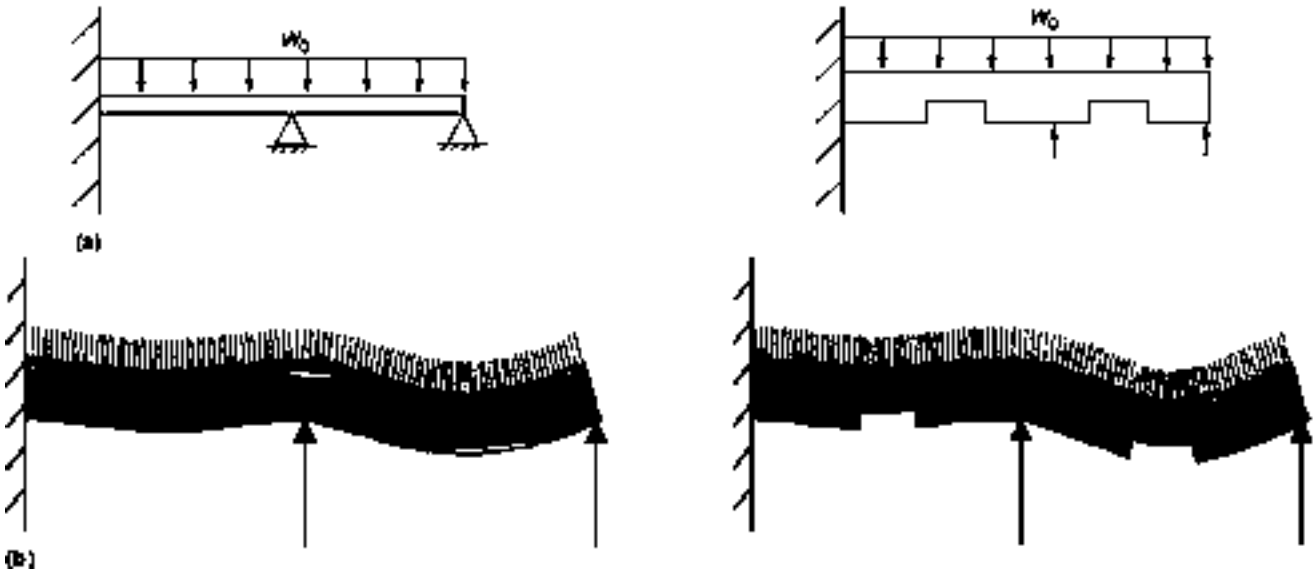


Fig. 3 Example of a statically indeterminate structure. Unlike the simple cantilever in Fig. 2, there is no closed formed solution in classical beam theory for a supported cantilever as shown in (a) for deflection of a beam (top left) or a notched beam (top right). Results from finite element analysis (FEA) are shown in (b) for the supported beam (bottom left) and the notched beam (bottom right).

Localized stress concentrations resulting from support locations and, in particular, changes in beam cross-sectional geometry are always a concern in design and often act as sites for crack initiation and failure. While traditional beam theory can account for changes in beam section through changes in the moment of inertia, it does not account for the localized stress concentration resulting from geometric concentrations. Handbooks of stress concentration factors for most any type beam or plate with geometric changes that are under stress (Ref 10) have been developed to account for these localized changes. These handbooks cover designs containing notches and grooves, fillets and radii, and through holes. They also cover beams and plates in tension, torsion, and bending. The design engineer must recognize the specific situation in the component being analyzed using beam theory and then apply the appropriate stress concentration.

Analyzing the same design using FEA requires only that the local geometric changes be incorporated into the original model. The simple beam model can be easily modified with local changes in the cross section of the beam. The localized effects of the reduced section and the stress concentration are then easily analyzed and the results evident. This is how FEA can be used to more easily analyze specific items that complicate analyses using traditional methods. This ability to analyze complex components is the reason for the proliferation of FEA as a design tool. The size and complexity of models that are routinely analyzed is increasing as fast as the computing speed and data storage capacity of computer workstations. Today it is not uncommon for large 3-D solid models to be solved on PC-based computing systems. Figure 4 shows an example of just such a 3-D model that can be analyzed on a desktop computer.

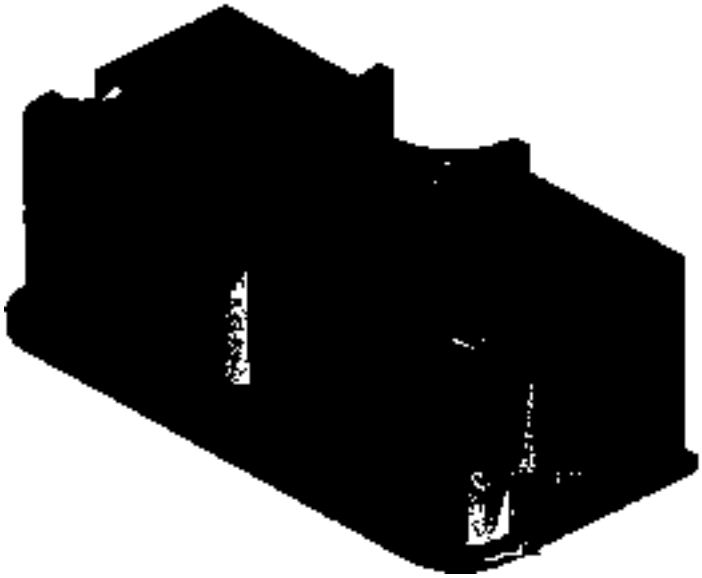


Fig. 4 Example of a three-dimensional half-symmetry finite element model of a housing

Thermal Analysis. General-purpose FEA software programs also are commonly used for thermal analysis to determine heat transfer and temperature profiles. Because the focus of this article is on FEA and structural failure analysis, discussion is limited only to those thermal areas that affect structural integrity. As engineered components are subjected to external or internal temperature changes, several things happen. First, as expected, the temperature throughout the component changes. The temperature profile will be dependent on such variables as the component geometry, material thermal properties, and how the temperature is applied to the component. Second, as the temperature of the component changes it will either expand or contract depending on the temperature profile and the thermal expansion behavior of the material. Finally, the mechanical properties of the material, such as the elastic modulus or fracture toughness, can change with temperature. Common design applications for FEA thermal analysis are power piping, pressure vessels and reactors, and turbine and heat engine components.

A thermal analysis begins the same as a structural analysis, with the creation of a model of the component to be analyzed including the material thermal properties. The boundary conditions and applied thermal loads are then applied to simulate the conditions expected. An analysis is then performed to determine the temperature profile in the structure. If all that is needed is this temperature profile or the corresponding heat flow, then the analysis is complete. However, in most cases, the effects of this change in temperature on the structural response are important. If an understanding of these effects is needed, then the results of the thermal analysis are used in a structural analysis to evaluate how the temperature profile creates a thermal deformation profile. These thermal deformations can be incorporated with other loads in a structural analysis to determine the overall response. Some general-purpose FEA programs have special coupled-field elements that allow the thermal and structural analyses to be performed together.

Thermal loads and the resulting stresses can be very significant. If they are not accounted for in the original design, extremely high stresses can occur causing distortion, cracking, or failure.

Nonlinear Analysis. As the computing capability of computer systems increases, so does the ability to analyze larger and more complex FEA models including nonlinear behavior. Most general-purpose FEA programs can solve a variety of nonlinear problems including elastic/plastic material behavior, nonlinear thermal properties, viscoelastic behavior, large deflection analysis, and contact analysis. These analyses are usually solved using an iterative procedure performed automatically by the FEA software and requiring numerous steps to achieve convergence on a solution. As a result, nonlinear analyses can take considerably longer than linear analyses if convergence is a problem or if the original model is excessively large. Nonlinear models usually are simplified as much as possible at the start to obtain a solution in a reasonable time. The model is then refined as needed to achieve the desired results.

Contact analysis is one of the most common and useful nonlinear analysis capabilities of FEA. It is an advanced capability that is difficult to solve using classical techniques but is becoming more automated in FEA software programs. In many cases, as an entire structure responds to mechanical or thermal loads, there are localized

points of contact or separation within individual elements of the structure or with other structures. Since these are usually localized points of contact or separation, the result can be very localized stress concentrations. These are also the areas that are difficult to analyze in original design analyses and often the sites of damage, cracking, or failure initiation. Examples of structures or components that can benefit from contact analysis are those bolted or otherwise connected; components with mating surfaces; vessels with caps, lids, or sealed openings; thermal expansion supports; or components simply contacting nearby components.

Figure 5 shows a model of an axisymmetric threaded connection subjected to internal pressure and external loading. The threaded connection is designed to maintain a pressure boundary under these loads. Using contact elements in the threaded region between the two pieces allows the localized effects of the contact on the individual threads to be determined.

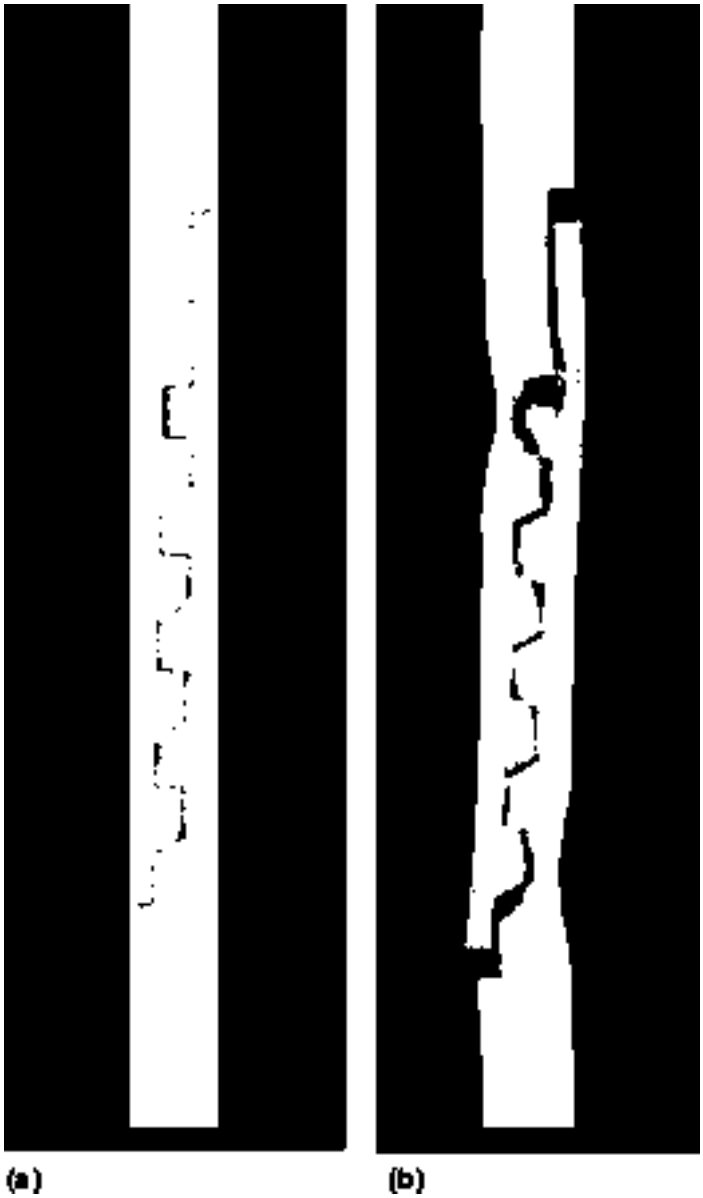


Fig. 5 Axisymmetric finite element model of a threaded connection showing (a) thread contact and (b) distortion and stress state

Elastic/plastic material behavior analysis is another often-used capability of general-purpose FEA software programs. It is also very useful when performing failure analyses, because failure and fracture of materials often involve some form of plastic deformation before failure. Most analyses start out as linear elastic problems. The results of a linear analysis can indicate whether stresses in the component exceed the yield strength of the material and give an indication of the size of the high stress plastic zone. In many cases, particularly in the design phase, an elastic analysis is sufficient because it reveals areas of excessive stress, thus prompting a redesign to reduce the stress. However, there are certain design cases where high plastic stresses

are unavoidable and need to be analyzed. This is also the case in failure analysis where an existing component design has failed and it is important to determine the magnitudes of the underlying stress and displacement profiles. Elastic/plastic analyses can determine not only the magnitude of stress, but also energy absorption and dissipation, as well as predict permanent deformation.

Figure 6 shows an example of an elastic/plastic analysis to predict the deformed shape of a transformer housing that was involved in an internal overpressurization. Shown compared to a photograph of the distorted housing are the FEA results detailing the stress state and permanent distorted shape.

Fig. 6 Example of an elastic/plastic finite element analysis. (a) Photograph showing distorted transformer housing from internal overpressurization. (b) Finite element results showing permanently distorted shape and stress contours

Transient and Dynamic Analysis. All of the analyses previously described have been either static or steady state. However, analysis capabilities have advanced so that they also can address transient and dynamic conditions. Whenever there is a change from one type of loading to another or from one temperature state to another, a structure experiences transient or dynamic conditions to some degree. These conditions may not be a concern if the change is slow or controlled, but in many cases such changes can lead to failure if not accurately understood and accounted for in the design. Thermal transients and vibration are common problems that can be modeled using general-purpose FEA.

Thermal Transients. Temperature changes are major source of problems in the design of engineered components. Earlier it was discussed how temperature changes could introduce steady-state thermal expansion stresses. In some cases stresses generated during temperature transients are more severe than in steady-state conditions, and some materials can easily fracture from thermal shock if the temperature transient is too severe.

Consider the case of the design of gas and steam turbines where the components are subjected to severe thermal and rotational loads; aerospace gas turbine engines have stringent weight restraints that limit the options of designers and thus make careful and thorough analysis crucial for their safe and reliable operation. Finite element analysis is the tool of choice in the analysis of these very complex, severely loaded parts. Steady-state analyses of the blades and rotor of a turbine can show what the stresses are once the engine reaches the operating temperature and design rotational speed. Before reaching these steady-state conditions, the various components experience temperature changes and gradients that can significantly affect the state of stress. At startup the entire engine is at ambient temperature. High rotational stresses are then imposed as the engine spins and the temperature of the turbine blades increases. With time the temperature of the rotor also increases until equilibrium is reached. Because there are thermal gradients and transients through the blades and rotor, there are corresponding stress transients and gradients that must be quantified. Finite element analysis is used to determine the worst case conditions, thus allowing performance limits to be placed on the components. The analyses are also used to determine heat up and cool down conditions and procedures and to predict the thermomechanical fatigue behavior of the most highly loaded components.

Vibration and dynamic response is another area where general-purpose FEA software analysis provides useful information to design engineers and failure analysis investigations. Modal analyses to determine natural and harmonic frequencies of structures and components are often performed to assure that resonance problems are not a concern. In other cases, if the magnitude of a vibration spectrum is known, then the specific response of the structure can be quantified to determine actual displacement and stress fields.

Large rotating equipment and machinery are routinely subjected to FEA vibration modal analysis to determine the natural frequencies and the resulting mode shapes. Equipment such as turbines, compressors, pumps, and fans all are subject to vibration-related problems. Components that are connected to this equipment, such as shafts, drives, couplings, inlet and outlet manifolds, and ducting and pipework, are also influenced by any vibrations generated. It is thus very important in this type of equipment that the rotational speeds of the shafts, gears, impellers, and vanes do not coincide with any of the natural frequencies of the component and their harmonics. For example, there are many types of large fans (some up to 3 m, or 10 ft, in diameter) that are used to move large volumes of air in traffic tunnels, mines, power plants, and large buildings. These types of fans can operate with tip speeds of over 800 km (500 miles) per hour; if vibration problems occur, they can result in cracking and fatigue crack propagation in very short times due to the high rotational speeds. Modal analysis by FEA allows the tuning of the fan such that these critical speeds and frequencies are avoided during startup and use.

In addition, seismic finite element dynamic analysis is used to determine the response of structures to various seismic vibration conditions. Knowing the magnitude of typical seismic vibration spectra allows design engineers to evaluate the effects of these input vibrations on the behavior of structures. These vibration power-density spectra are used to determine actual stresses and displacements at locations and frequencies of concern. The analysis is performed on large structures such as buildings and bridges. These analyses are also required as part of the design qualification of nuclear power plants. Similar types of dynamic vibration analyses are performed for aircraft components, high-speed transport systems, and spacecraft. All these critical applications are subject to vibrations that can cause severe structural problems if they are not accounted for in the design.

Reference cited in this section

10. W.D. Pilkey, *Peterson's Stress Concentration Factors*, 2nd ed., John Wiley & Sons, Inc., 1997

Finite Element Modeling in Failure Analysis

Thomas Service, Altran Corporation

General-Purpose Applications

General-purpose FEA software programs can perform a variety of analyses including structural and thermal, static and transient, linear and nonlinear, and more. The programs are thorough, detailed, and relatively complex and rely heavily on the analyst to accurately create models, understand the solution process, and interpret the results. Currently there are numerous commercial software packages available with a variety of options designed for use on all types of personal computer (PC) and mainframe platforms.

Generating models through interactive design and importing model geometry from computer-aided design (CAD) software is making FEA analysis much easier. However, with this increased ease comes the increased need to correctly understand and interpret FEA results. Care must be taken so that models are developed and analyses are performed by trained personnel who understand the technology.

Structural Analysis. The most common application of FEA is in the determination of displacement and stress profiles in structures. Bending beams, rotating disks, pressure vessels, buildings and bridges, airplanes and automobiles, and heart valves and hip joints are all analyzed using FEA structural analysis techniques. This structural analysis area of FEA provides results analogous to those obtained through classical beam and plate technique with the advantage that much more complex geometries and loadings can be evaluated. More important is that the FEA provides results that are immediately available for the whole structure rather than one location that is identified for analysis using classical techniques. This enables graphical representations of the results that are more easily visualized and interpreted than analytical results. The graphical representations can clearly highlight areas of stress concentration and excessive deformation relative to the entire structure. This graphical capability makes FEA particularly useful when performing failure analysis investigations.

Consider the case of a simple uniformly loaded cantilever beam shown in Fig. 2(a). The stress and deflection profiles for this structural member are derivable from beam theory and are readily available in many engineering tests. Closed-form equations provide the results for any location desired in the beam, provided it meets the criteria assumed in beam theory. For example, if the beam length is reduced to a value similar to the beam height, then the solutions are not valid and other techniques must be employed to determine the stress and deflection. Now consider the same beam problem solved using finite element techniques. Figure 2(b) shows a simple finite element model and loading and boundary conditions for the same problem. The solution provides not only a graphical depiction of the deflection profile but also the stress contours throughout the beam. Further, unlike beam theory, the solution is valid for all beam length and height dimensions.

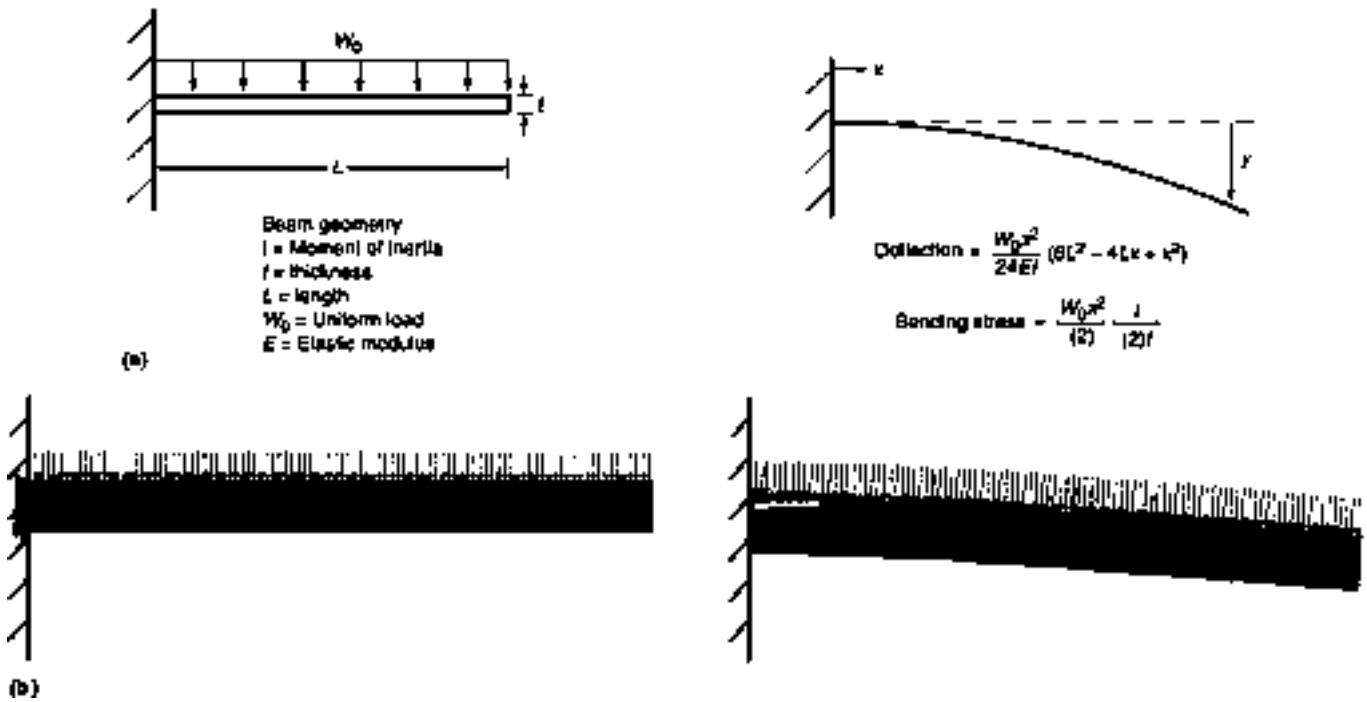


Fig. 2 Simple cantilever beam analysis. (a) Classical beam theory showing key elements (top left) and stress and deflection (top right). (b) Finite element analysis (FEA) cantilever beam elements (bottom left) with FEA cantilever beam results (bottom right)

Now consider the same cantilever problem but with the addition of two simple supports under the beam (Fig. 3). This simple change makes the problem statically indeterminate and thus adds an order of magnitude to the complexity of solution using beam theory. Techniques such as the moment-area method, double integration, and superposition have been developed to provide analytical solutions to these types of problems. Using FEA, no such complexity exists. Simply adding deflection boundary constraints at the appropriate locations will provide the correct solution. Again, the complete deflection and stress profiles are quickly determined, providing the necessary information at any location in the beam. Further, the localized stress profile in the area of the support points is also known.

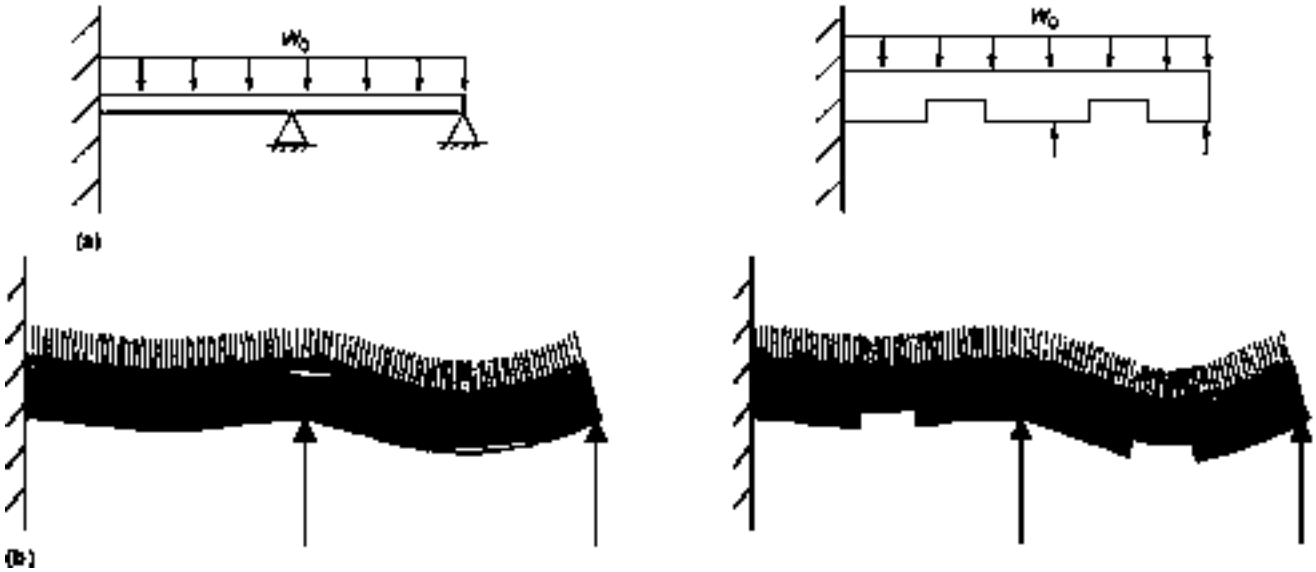


Fig. 3 Example of a statically indeterminate structure. Unlike the simple cantilever in Fig. 2, there is no closed formed solution in classical beam theory for a supported cantilever as shown in (a) for deflection of a beam (top left) or a notched beam (top right). Results from finite element analysis (FEA) are shown in (b) for the supported beam (bottom left) and the notched beam (bottom right).

Localized stress concentrations resulting from support locations and, in particular, changes in beam cross-sectional geometry are always a concern in design and often act as sites for crack initiation and failure. While traditional beam theory can account for changes in beam section through changes in the moment of inertia, it does not account for the localized stress concentration resulting from geometric concentrations. Handbooks of stress concentration factors for most any type beam or plate with geometric changes that are under stress (Ref 10) have been developed to account for these localized changes. These handbooks cover designs containing notches and grooves, fillets and radii, and through holes. They also cover beams and plates in tension, torsion, and bending. The design engineer must recognize the specific situation in the component being analyzed using beam theory and then apply the appropriate stress concentration.

Analyzing the same design using FEA requires only that the local geometric changes be incorporated into the original model. The simple beam model can be easily modified with local changes in the cross section of the beam. The localized effects of the reduced section and the stress concentration are then easily analyzed and the results evident. This is how FEA can be used to more easily analyze specific items that complicate analyses using traditional methods. This ability to analyze complex components is the reason for the proliferation of FEA as a design tool. The size and complexity of models that are routinely analyzed is increasing as fast as the computing speed and data storage capacity of computer workstations. Today it is not uncommon for large 3-D

solid models to be solved on PC-based computing systems. Figure 4 shows an example of just such a 3-D model that can be analyzed on a desktop computer.

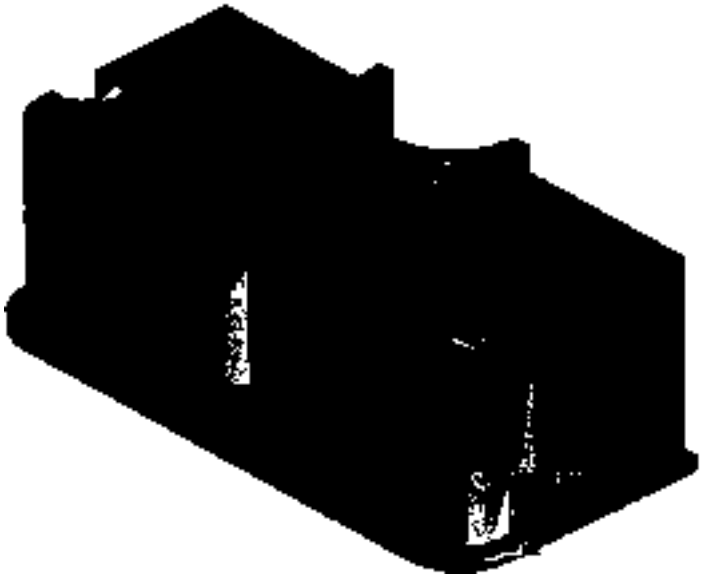


Fig. 4 Example of a three-dimensional half-symmetry finite element model of a housing

Thermal Analysis. General-purpose FEA software programs also are commonly used for thermal analysis to determine heat transfer and temperature profiles. Because the focus of this article is on FEA and structural failure analysis, discussion is limited only to those thermal areas that affect structural integrity. As engineered components are subjected to external or internal temperature changes, several things happen. First, as expected, the temperature throughout the component changes. The temperature profile will be dependent on such variables as the component geometry, material thermal properties, and how the temperature is applied to the component. Second, as the temperature of the component changes it will either expand or contract depending on the temperature profile and the thermal expansion behavior of the material. Finally, the mechanical properties of the material, such as the elastic modulus or fracture toughness, can change with temperature. Common design applications for FEA thermal analysis are power piping, pressure vessels and reactors, and turbine and heat engine components.

A thermal analysis begins the same as a structural analysis, with the creation of a model of the component to be analyzed including the material thermal properties. The boundary conditions and applied thermal loads are then applied to simulate the conditions expected. An analysis is then performed to determine the temperature profile in the structure. If all that is needed is this temperature profile or the corresponding heat flow, then the analysis is complete. However, in most cases, the effects of this change in temperature on the structural response are important. If an understanding of these effects is needed, then the results of the thermal analysis are used in a structural analysis to evaluate how the temperature profile creates a thermal deformation profile. These thermal deformations can be incorporated with other loads in a structural analysis to determine the overall response. Some general-purpose FEA programs have special coupled-field elements that allow the thermal and structural analyses to be performed together.

Thermal loads and the resulting stresses can be very significant. If they are not accounted for in the original design, extremely high stresses can occur causing distortion, cracking, or failure.

Nonlinear Analysis. As the computing capability of computer systems increases, so does the ability to analyze larger and more complex FEA models including nonlinear behavior. Most general-purpose FEA programs can solve a variety of nonlinear problems including elastic/plastic material behavior, nonlinear thermal properties, viscoelastic behavior, large deflection analysis, and contact analysis. These analyses are usually solved using an iterative procedure performed automatically by the FEA software and requiring numerous steps to achieve convergence on a solution. As a result, nonlinear analyses can take considerably longer than linear analyses if convergence is a problem or if the original model is excessively large. Nonlinear models usually are simplified as much as possible at the start to obtain a solution in a reasonable time. The model is then refined as needed to achieve the desired results.

Contact analysis is one of the most common and useful nonlinear analysis capabilities of FEA. It is an advanced capability that is difficult to solve using classical techniques but is becoming more automated in FEA software programs. In many cases, as an entire structure responds to mechanical or thermal loads, there are localized points of contact or separation within individual elements of the structure or with other structures. Since these are usually localized points of contact or separation, the result can be very localized stress concentrations. These are also the areas that are difficult to analyze in original design analyses and often the sites of damage, cracking, or failure initiation. Examples of structures or components that can benefit from contact analysis are those bolted or otherwise connected; components with mating surfaces; vessels with caps, lids, or sealed openings; thermal expansion supports; or components simply contacting nearby components.
Figure 5 shows a model of an axisymmetric threaded connection subjected to internal pressure and external loading. The threaded connection is designed to maintain a pressure boundary under these loads. Using contact elements in the threaded region between the two pieces allows the localized effects of the contact on the individual threads to be determined.

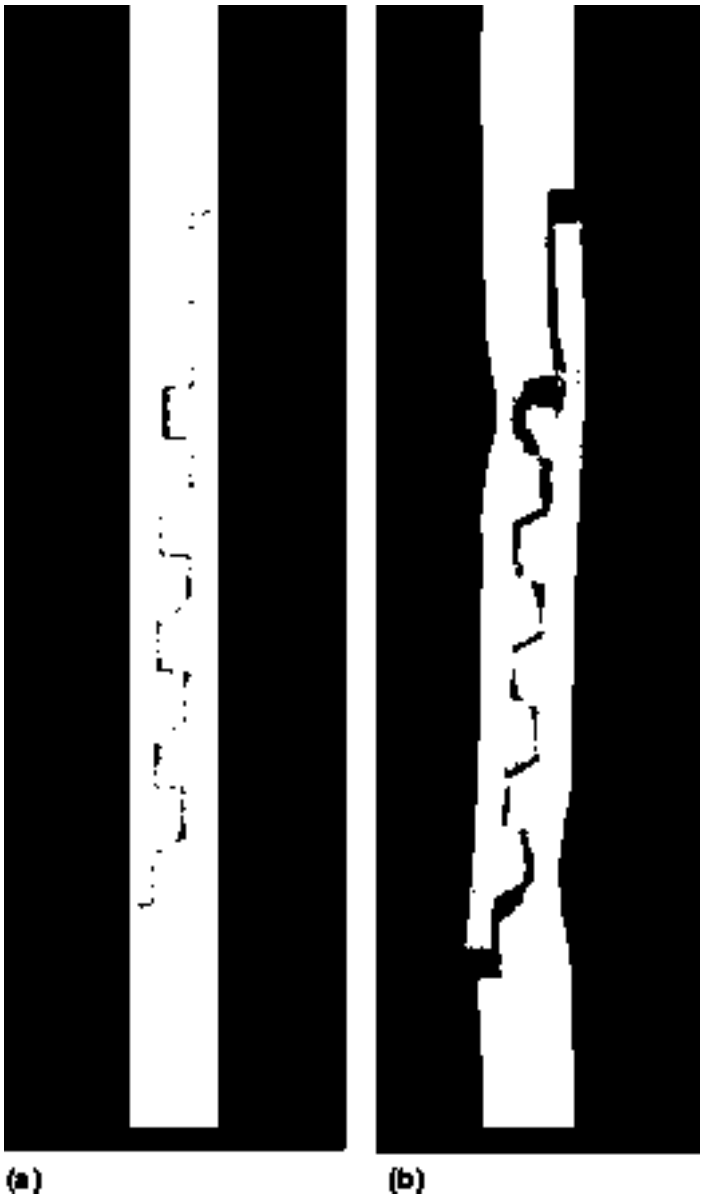


Fig. 5 Axisymmetric finite element model of a threaded connection showing (a) thread contact and (b) distortion and stress state

Elastic/plastic material behavior analysis is another often-used capability of general-purpose FEA software programs. It is also very useful when performing failure analyses, because failure and fracture of materials often involve some form of plastic deformation before failure. Most analyses start out as linear elastic problems. The results of a linear analysis can indicate whether stresses in the component exceed the yield

strength of the material and give an indication of the size of the high stress plastic zone. In many cases, particularly in the design phase, an elastic analysis is sufficient because it reveals areas of excessive stress, thus prompting a redesign to reduce the stress. However, there are certain design cases where high plastic stresses are unavoidable and need to be analyzed. This is also the case in failure analysis where an existing component design has failed and it is important to determine the magnitudes of the underlying stress and displacement profiles. Elastic/plastic analyses can determine not only the magnitude of stress, but also energy absorption and dissipation, as well as predict permanent deformation.

Figure 6 shows an example of an elastic/plastic analysis to predict the deformed shape of a transformer housing that was involved in an internal overpressurization. Shown compared to a photograph of the distorted housing are the FEA results detailing the stress state and permanent distorted shape.

Fig. 6 Example of an elastic/plastic finite element analysis. (a) Photograph showing distorted transformer housing from internal overpressurization. (b) Finite element results showing permanently distorted shape and stress contours

Transient and Dynamic Analysis. All of the analyses previously described have been either static or steady state. However, analysis capabilities have advanced so that they also can address transient and dynamic conditions. Whenever there is a change from one type of loading to another or from one temperature state to another, a structure experiences transient or dynamic conditions to some degree. These conditions may not be a concern if the change is slow or controlled, but in many cases such changes can lead to failure if not accurately

understood and accounted for in the design. Thermal transients and vibration are common problems that can be modeled using general-purpose FEA.

Thermal Transients. Temperature changes are major source of problems in the design of engineered components. Earlier it was discussed how temperature changes could introduce steady-state thermal expansion stresses. In some cases stresses generated during temperature transients are more severe than in steady-state conditions, and some materials can easily fracture from thermal shock if the temperature transient is too severe.

Consider the case of the design of gas and steam turbines where the components are subjected to severe thermal and rotational loads; aerospace gas turbine engines have stringent weight restraints that limit the options of designers and thus make careful and thorough analysis crucial for their safe and reliable operation. Finite element analysis is the tool of choice in the analysis of these very complex, severely loaded parts. Steady-state analyses of the blades and rotor of a turbine can show what the stresses are once the engine reaches the operating temperature and design rotational speed. Before reaching these steady-state conditions, the various components experience temperature changes and gradients that can significantly affect the state of stress. At startup the entire engine is at ambient temperature. High rotational stresses are then imposed as the engine spins and the temperature of the turbine blades increases. With time the temperature of the rotor also increases until equilibrium is reached. Because there are thermal gradients and transients through the blades and rotor, there are corresponding stress transients and gradients that must be quantified. Finite element analysis is used to determine the worst case conditions, thus allowing performance limits to be placed on the components. The analyses are also used to determine heat up and cool down conditions and procedures and to predict the thermomechanical fatigue behavior of the most highly loaded components.

Vibration and dynamic response is another area where general-purpose FEA software analysis provides useful information to design engineers and failure analysis investigations. Modal analyses to determine natural and harmonic frequencies of structures and components are often performed to assure that resonance problems are not a concern. In other cases, if the magnitude of a vibration spectrum is known, then the specific response of the structure can be quantified to determine actual displacement and stress fields.

Large rotating equipment and machinery are routinely subjected to FEA vibration modal analysis to determine the natural frequencies and the resulting mode shapes. Equipment such as turbines, compressors, pumps, and fans all are subject to vibration-related problems. Components that are connected to this equipment, such as shafts, drives, couplings, inlet and outlet manifolds, and ducting and pipework, are also influenced by any vibrations generated. It is thus very important in this type of equipment that the rotational speeds of the shafts, gears, impellers, and vanes do not coincide with any of the natural frequencies of the component and their harmonics. For example, there are many types of large fans (some up to 3 m, or 10 ft, in diameter) that are used to move large volumes of air in traffic tunnels, mines, power plants, and large buildings. These types of fans can operate with tip speeds of over 800 km (500 miles) per hour; if vibration problems occur, they can result in cracking and fatigue crack propagation in very short times due to the high rotational speeds. Modal analysis by FEA allows the tuning of the fan such that these critical speeds and frequencies are avoided during startup and use.

In addition, seismic finite element dynamic analysis is used to determine the response of structures to various seismic vibration conditions. Knowing the magnitude of typical seismic vibration spectra allows design engineers to evaluate the effects of these input vibrations on the behavior of structures. These vibration power-density spectra are used to determine actual stresses and displacements at locations and frequencies of concern. The analysis is performed on large structures such as buildings and bridges. These analyses are also required as part of the design qualification of nuclear power plants. Similar types of dynamic vibration analyses are performed for aircraft components, high-speed transport systems, and spacecraft. All these critical applications are subject to vibrations that can cause severe structural problems if they are not accounted for in the design.

Reference cited in this section

10. W.D. Pilkey, *Peterson's Stress Concentration Factors*, 2nd ed., John Wiley & Sons, Inc., 1997

Thomas Service, Altran Corporation

Design Review in Failure Analysis

Engineering design consists of a process that starts with the identification of a specific need and ends with the preparation of detailed drawings describing the fabrication and assembly of the final component that satisfies the original need. The steps involved in the design process may include some or all of the following:

- Identify needs.
- Establish goals.
- Perform any research needed.
- Establish overall design specifications and requirements.
- Create ideas and concepts.
- Develop detailed analysis design concepts.
- Build and test prototypes.
- Create and analyze final detailed design.
- Initiate manufacture, fabrication assembly, and production.
- Perform model testing and failure analysis on the pilot models (iterative design).

The overall design specifications provide requirements for use and functionality such as: a steam turbine needs to produce a certain amount of power; an off-road vehicle needs to carry 20% more cargo; a railroad bridge must span a river; a replacement knee joint or a handheld computer are needed. Inherent in any design are other limitations and considerations such as cost, ease of fabrication, weight, size, safety, and reliability, all of which factor into the final design.

Achieving the overall design requirements involves the design and specification of many individual parts and components. The exact specifications for these components can be dictated by a variety of codes and standards. There are domestic mandatory standards such as those in the Code of Federal Regulations (CFR) and the Occupational Safety and Health Administration (OSHA). There are also domestic standards and codes such as those of the American National Standards Institute (ANSI), the American Society for Testing and Materials (ASTM), Underwriters Laboratories (UL), the American Society of Mechanical Engineers (ASME), which issues the Boiler and Pressure Vessel Code, and more. In addition to all these external regulations, standards, and codes, most individual manufacturers have internal standards, design guidelines, and requirements specific to their products and developed from years of experience and expertise. These codes and standards can be general in nature or detailed and specific to include precise materials applications and properties, specific loadings that must be addressed, and the maximum allowable stresses under these different loads.

Although FEA is used in the research and conceptualization phases, it is used most often in the detailed analysis phase. Finite element analysis and other analyses are performed to verify the mechanical and structural design requirements. These tend to be very specific and detailed analyses and calculations with the purpose of demonstrating compliance with the design requirements and all applicable codes and standards.

A failure analysis investigation can be broken into two components—a materials or metallurgical investigation and a design and analysis review investigation. Whenever a structural or mechanical failure occurs and the decision is made to conduct a failure analysis investigation, the first objective should not necessarily be to perform a finite element analysis. As previously noted at the beginning of this article, the first objective is to determine the failure mechanism, typically by materials/metallurgical failure analysis. The results of a metallurgical failure analysis provide some indication of potential failure mechanisms or root causes and whether the appropriate material and manufacturing specifications were met. Failure may occur by one or more mechanisms associated with the four general categories of failure by distortion, fracture, corrosion, and/or wear. For example, the mechanism of failure might be a fatigue crack that initiates from a surface pit caused by corrosion, where the crack propagates initially by fatigue until final fracture (i.e., overload fracture) occurs

from crack propagation by ductile and/or brittle fracture mechanisms. The failure mechanism also may be may be a combination of design and/or manufacturing imperfections acting to cause crack initiation.
In many cases a metallurgical failure analysis will provide sufficient information to correct any problems and prevent future failure. For example, if the failure is attributed to use of the wrong material, manufacturing defects, or material imperfections, the use of the correct material in the proper condition may remedy the problem. On the other hand, the metallurgical failure analysis may provide the failure mechanism but give no information regarding the underlying cause of the mechanism. For example, the results of the analysis may identify the mechanism of failure to be some type of fatigue or overload fracture. In these cases the metallurgical analysis may not provide sufficient information to make informed recommendations, because crack initiation may be associated with mechanical conditions or factors of part geometry. What is needed is a review of the design to determine what conditions led to the fatigue or overload conditions that caused failure. This is where FEA plays an important role in failure analysis investigations.
The major steps involved in a failure analysis design review include:

- Collection of data and background information regarding overall design criteria and applicable regulations, codes, and standards
- Review of analyses and design calculations performed and assumptions used to determine if they were correct and applicable
- Development of any necessary analytical models (including FEA) that can be used to simulate all possible loading conditions the failed structure may have experienced (i.e., thermal, structural, vibration, or impact conditions)
- Performance of analyses to determine the loading that results in the conditions necessary to cause the type of failure mechanism identified in the metallurgical failure analysis
- Evaluation of results to form conclusions

A failure analysis design review provides an understanding of the structural loading conditions that led to the failure mechanism and the observed fracture morphology that are consistent with the data and information obtained during the investigation. Design review is discussed in more detail in the article “Design Review for Failure Analysis and Prevention” in this Volume. Possible conclusions from a failure analysis design review may include:

- No original analyses were performed, and the failed component was simply not adequate for the intended use.
- The original analyses were incomplete or in error.
- The original analyses underestimated the magnitude of the actual loads.
- The original analyses were correct and accurate as far as they went, but unanticipated loading conditions arose during operation.
- The original analyses were correct and accurate, but the loads placed on the component were the result of misuse or abuse.

When structural failure of components occurs from lack of design analysis or errors in analysis, the solution that prevents future problems is in the redesign and analysis. Underestimating the loads or the effects of loads on a component often leads to time-dependent failure mechanisms such a fatigue and creep. Finite element analysis is very useful in quantifying subtle loading changes or cyclic loads that lead to these failure mechanisms. A common cause of structural failure is when unanticipated loads occur during use. These often occur when acceptable proven designs are modified slightly, have material changes, or where size or operating speed is increased. Sometimes small changes in design that are not anticipated to result in significant loading changes produce unexpected results; these could be in the form of higher stresses, temperatures, or vibration. Finite element analysis allows for the repeated analysis of a variety of load cases that can influence the structure in the manner that caused failure. It can also clearly demonstrate the effects of overload or abuse.
Finite element analysis is used in the design of components or structures to verify criteria specified in design requirements, codes, and standards. Once these requirements have been fulfilled, the analysis is complete. When using FEA in a failure analysis, the primary purpose is to determine the most likely set of loading conditions that produce results that are consistent with the evidence and the observed failure. The former is

analogous to determining the stresses resulting from known loading conditions, and the latter is analogous to determining the loading conditions that cause known stresses.

Case Studies of FEA in Failure Analysis

This section provides two case histories that involved the use of FEA in failure analysis. Additional examples are provided in the preceding article "Modeling and Accident Reconstruction" in this Volume.

Example 1: Cracking in a Yankee Dryer Shell (Ref 11). A Yankee dryer is a critical mechanical component in the manufacture of tissue paper. It is a large, cylindrical, rotating, pressurized, high-temperature, cast iron pressure vessel that has evolved in design since the early 1900s. A wet sheet of paper is placed on the rotating face of the dryer, which is located toward the back end of a crepe papermaking machine. As the dryer rotates, contact with the high-temperature face removes considerable moisture from the paper by the time it is removed from the dryer. The drying process only occurs while the paper is in contact with the high-temperature face during approximately one dryer rotation.

The demand for more efficient papermaking has led to larger and faster machines. Current Yankee dryers can be 5.5 m (18 ft) in diameter with a 6.1 m (20 ft) face length, operating at 175 °C (350 °F), 830 kPa (120 psi) internal pressure and rotating at 110 revolutions per minute. This translates to making a 5.5 m (18 ft) wide sheet of tissue paper at over a mile a minute. The extremely high temperature and pressure of these machines means that there is an enormous amount of stored energy that must be safely contained through proper design and analysis. The case study, as documented in Ref 11, is summarized here on the use of FEA in the original design of the dryers and in a failure analysis investigation to understand a cracking problem that developed.

The typical components of a Yankee dryer consist of a cast iron shell, two cast iron concave heads, and a large cast iron internal center stay attached to journals. The heads are attached to the shell and center stay with high-strength bolts. The primary design requirements for this type of equipment are governed by the ASME Boiler and Pressure Vessel Code Section VIII, "Rules for Construction of Pressure Vessels." Individual dryer manufacturers also have internal design requirements based on their own experience and expertise. Although these are voluntary codes and standards, owners of the machines and insurance companies always require that the dryer manufacturers demonstrate that the equipment meets these design safety requirements.

The complexity in design detail of Yankee dryers has increased significantly from when the first 3.0 to 3.7 m (10 to 12 ft) diameter machines were used a hundred years ago, even though the basic design has not. Past design analyses tended to be hand calculations based on the rules specified in the ASME design requirements. These rules specify minimum material thickness to account for pressure, thermal, and rotational loads. The relatively low pressures and operating speeds of early dryers combined with the high factors of safety required by the code resulted in robust designs.

Over the years the basic design has remained essentially the same with a general evolution toward larger size, higher temperature and pressure, and faster machines. Current machines are so large and the operating conditions so aggressive that the only way to accurately determine the stresses in the dryer and the corresponding design safety factor is through the use of FEA. There are considerable thermal, pressure, and rotational loads on an operating dryer that must be quantified. Figure 7 shows a typical axisymmetric finite element model that is used to model a Yankee dryer. The model accounts for the overall thermal expansions and localized thermal gradient through the shell due to the heat removal of the paper, the internal pressure, and rotational loads. It also evaluates the effects of the head/shell interface and the attachment bolts. Combining all these loads results in a FEA stress and displacement profile that the design engineer compares to the code requirements. The results of these analyses show that an area with the most complex stress state is at the head/shell interface. Analyses similar to this have shown that the dryer design meets all the necessary code and standard requirements. This could not be done easily without the use of FEA.

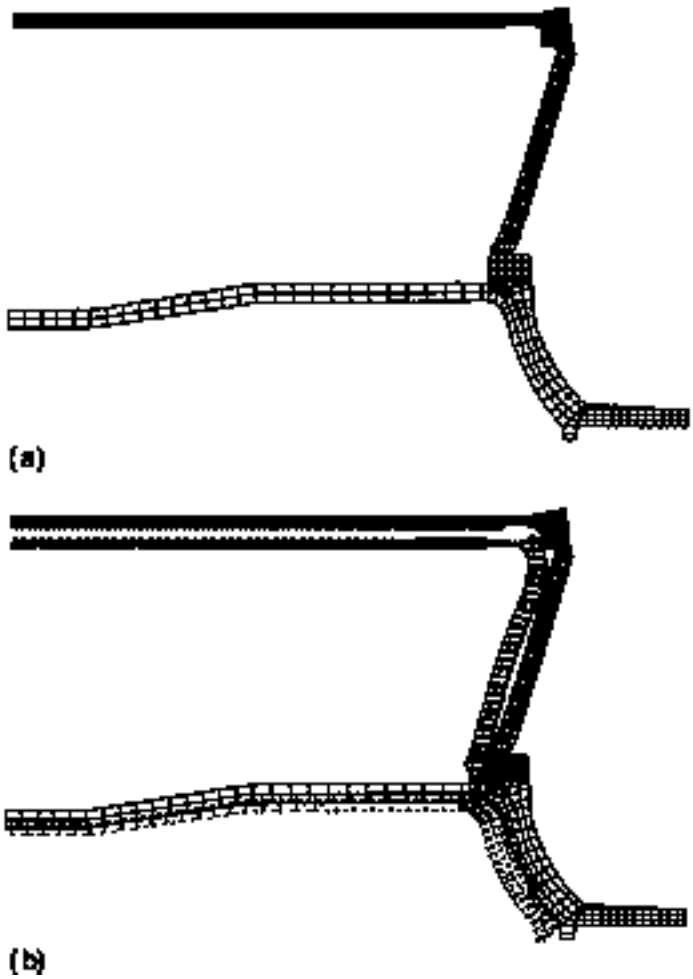


Fig. 7 Finite element model used in the design of a Yankee dryer. (a) Finite element analysis (FEA) model. (b) Deformed shape showing stress state resulting from thermal, pressure, and rotational loads. Source: Ref 11

Several years after the introduction of the larger generation of Yankee dryers, routine inspections uncovered a cracking problem in the heads of some machines. These were circumferential cracks located in the bolt-hole region of the head. The discovery of the cracking prompted an investigation by the industry to understand the causes of the problem and propose changes to eliminate the problem. Metallurgical failure analyses were performed as well as reviews of the design. The results showed that there were no material or manufacturing-related causes. The investigation also showed that the design operational stresses in the region were not high enough to initiate or propagate any form of cracking. What was discovered was a layer of corrosion inside the head/shell interface that initiated at the periphery and progressed radially in the joint interface. Given this information, analyses were performed to determine if this could be the cause of the cracking.

Figure 8 shows the proposed mechanism that includes corrosion product buildup at the head/shell interface causing the joint to displace open. This “corrosion jacking” was modeled using FEA by prescribing a displacement of the joint interface equal to those measured on actual machines. The results showed that relatively small displacements of the interface at the region where corrosion was observed could cause stresses high enough to cause the observed cracking. The results also showed that the compressive bolting loads could slightly open the head/shell interface at the periphery. Based on the results of these analyses, design changes in the head/shell joint were made and retrofitted on machines in service. Detailed preventive maintenance inspection procedures were also initiated.

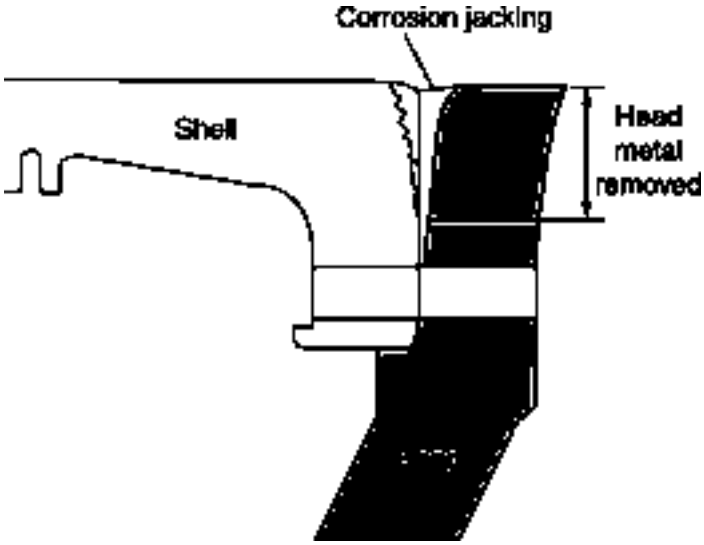


Fig. 8 Finite element model showing analysis of corrosion jacking in the head/shell interface. Source: Ref 11

The cause of this problem was not inadequate design analysis of the original design, rather, it was the result of an unexpected type of load placed on the machine that would have been difficult to anticipate. The problem also did not appear immediately but progressed with time as the corrosion product accumulated. It is also an example of how the design evolution process of scaling up smaller designs to meet the needs of larger designs has limits before problems arise. Smaller Yankee dryers with similar head/shell designs did not have this problem. It was only when the machine size and operating conditions reached current levels that the problem was manifest. This illustrates how codes and standards set only minimum requirements and that it is up to the design engineer to be careful with any new design and to be as thorough and diligent in the analysis as possible. Even then unexpected conditions can arise.

Example 2: Cracking in a Steam Generator U-Tube. A steam generator is a key component used in the generation of steam in a closed-cycle pressurized-water nuclear power plant. It is essentially a large heat exchanger that circulates high-temperature pressurized reactor coolant through a large bundle of U-tubes. Heat is exchanged to feed water surrounding the tubes that is heated and flows up through the tube bundle. Steam is eventually extracted at the top and directed to drive a turbine. A typical steam generator tube bundle can be over 9 m (30 ft) tall and 3 m (10 ft) in diameter with over three thousand 22 mm (7/8 in.) diameter Inconel Alloy 600 tubes. There are tube support plates (TSP) to support and separate the tubes and allow flow of the heating water/steam. The normal design conditions for the tubes are an approximate pressure differential of 10.7 MPa (1550 psi) at a temperature of 315 °C (600 °F). There can be up to four steam generators for a given reactor. Figure 9 shows a schematic cross section of a typical steam generator.

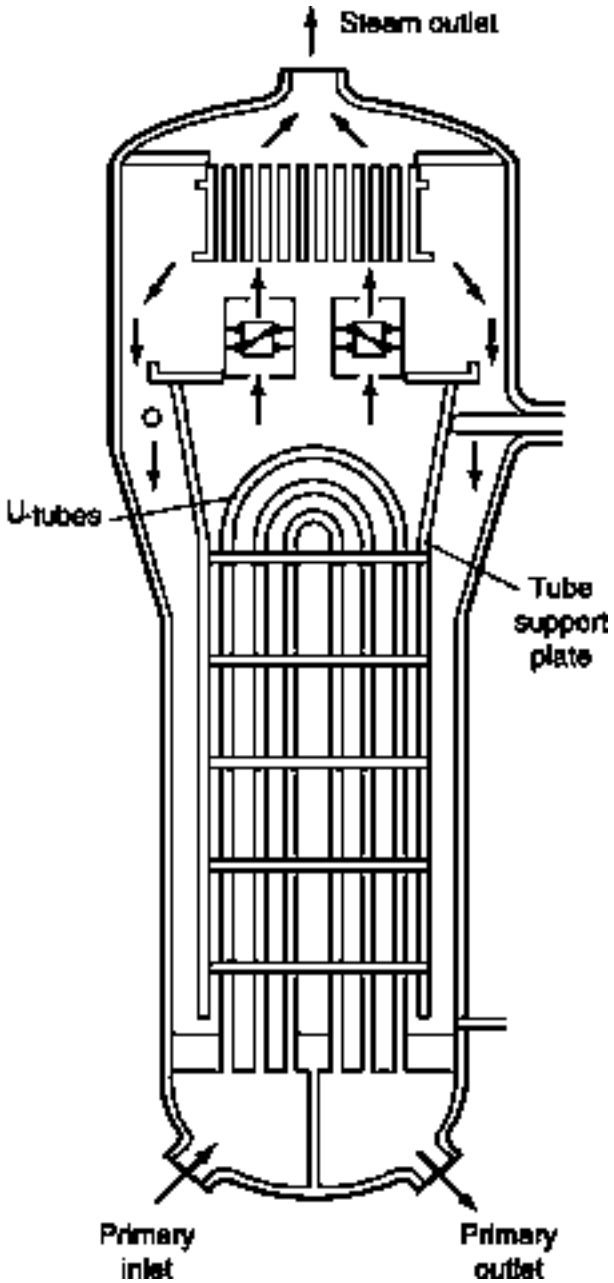


Fig. 9 Cross-section schematic drawing of a steam generator showing U-tube bundle

The original design of many of these steam generator components was conducted over thirty years ago under the requirements of the ASME code applicable at the time. At that time it was common for the appropriate code analyses involving the U-tubes to be performed using classical techniques and code calculations. These calculations were designed to provide adequate factors of safety under the expected operating conditions by determining minimum tube wall thickness. It is unlikely that FEA, which was in its infancy at that time, was utilized in U-tube analyses. It was, however, starting to be used in the more complex nozzle geometry analyses as well as in seismic analyses.

During the years that the steam generators have been in service they have been subject to extensive inspection, analysis, research, and upgrades to assure safe and reliable operation. One concern that has since been discovered is that the material used in the U-tubes, Inconel 600, is susceptible to intergranular stress-corrosion cracking (IGSCC). Intergranular stress-corrosion cracking is a degradation mechanism that requires the combination of an aggressive environment, susceptible material, and a sufficient applied stress (see the article "Stress-Corrosion Cracking" in this Volume for more details on the precondition required for stress-corrosion cracking). It starts out as an intergranular surface attack that progresses to linkage and the formation of cracks. These cracks can then propagate with time under constant stress until a critical size is reached, at which point rupture occurs. Considerable research has since shown that Inconel 600 can be susceptible to IGSCC from the

environment in a steam generator. It is therefore important to assure that the stress in the tubes is below that required to initiate IGSCC such that this degradation mechanism is not a safety concern.

A rupture of a thirty-year-old U-tube occurred that resulted in the limited release of reactor water. The safety considerations of this serious incident required that a root cause failure analysis be performed to identify the cause of the problem, fix the problem, determine the overall reliability of this and similar components, and take measures to assure that it does not reoccur. This consisted of extensive inspections and review of operational records, maintenance history, and procedures, and considerable analysis. The failure analysis investigation determined that a small longitudinal through-wall crack occurred along the flank of a row 2 U-tube just below the apex of the U-bend. The failure mechanism was attributed to IGSCC. Because stress is the crack driving force in IGSCC, it was important to analyze the observed in situ conditions of these components to quantify the stress state at the failure location. This could only be done using FEA.

Three-dimensional solid models of the U-tubes were created and analyzed. Initial complexities of the analyses included thermal gradients, nonlinear material behavior, and residual stress resulting from the forming of the tubes. Another complexity analyzed was the change from a circular tube cross section to a slightly oval cross-section during the forming of a U-bend and the resulting changes in wall thickness. All of these behaviors were analyzed and the influences on tube stress were quantified, with the results showing that the stress in the region where the cracking occurred was not sufficient to initiate cracking. Internal inspections of the steam generator indicated that there were distortions of the tube support plate in the region of the ruptured tube that could result in additional movement of the U-tubes. This movement tended to “pinch” the legs of the U-tube together, causing additional bending of the tube. Additional finite element analyses were performed to determine the effects of the movement of the uppermost TSPs and the influence on the displacement of the U-tube legs. Figure 10 shows a FEA model of the U-tube.

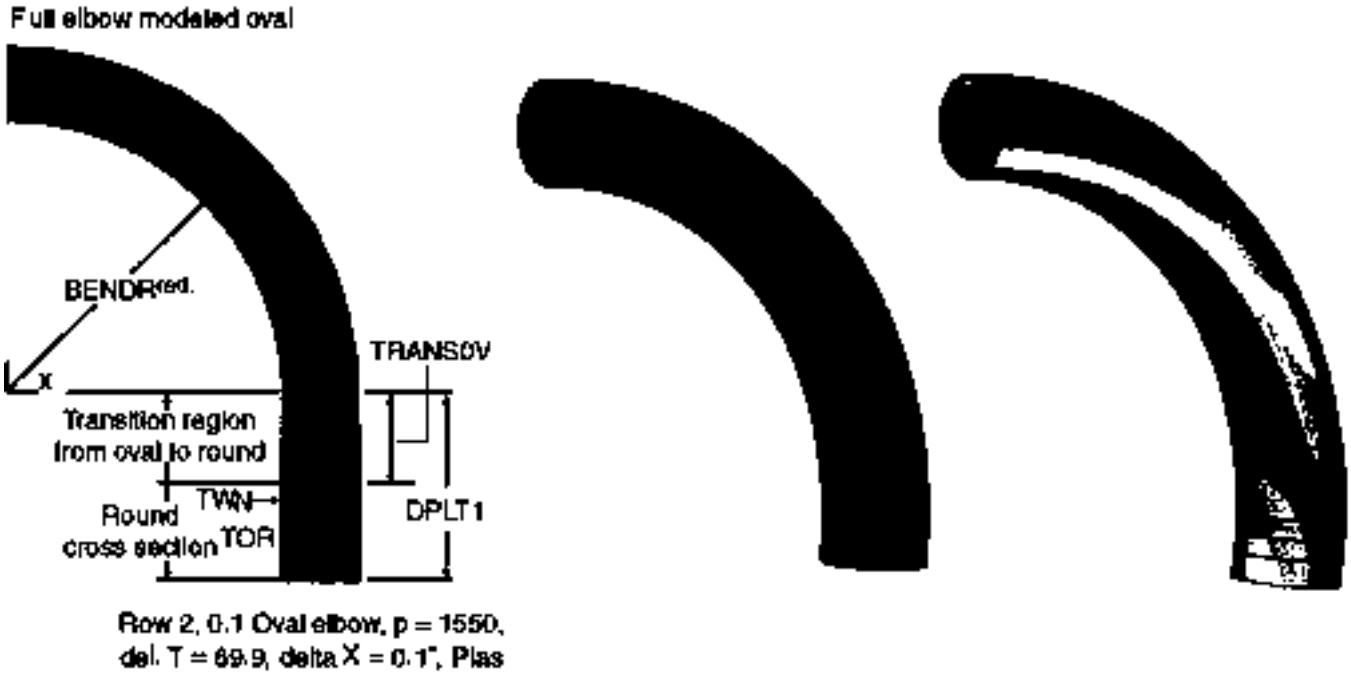


Fig. 10 Three-dimensional finite element analysis (FEA) model of U-tube showing effects of thermal gradients, internal pressure, and tube leg displacement

The results of these analyses clearly showed that the distortion of the TSPs and resulting “pinching” of the U-tubes, combined with the operational stresses, caused high stresses at the location where the tube cracked. The stress state was also consistent with that required to initiate and propagate a longitudinal crack. This resulted in the plugging of all tubes that were observed to have a similar amount of distortion from the tube support plate. The results also showed that this TSP distortion effect was very localized and had minimal effect on the majority of the tubes in the steam generator.

This again demonstrates the power of FEA in a failure analysis investigation. Most of the complexities evaluated simply could not have been analyzed using classical techniques. This example also shows how a problem or degradation that occurs during the life of a component can change the loading conditions. This

problem could not have been foreseen in the original design analysis. It was only after the failure occurred that conditions were observed that could be modeled and analyzed. The FEA performed in the failure analysis then confirmed the adverse effect on the stress state creating conditions conducive for failure.

Reference cited in this section

11. *Tappi J.*, Vol 78 (No. 7), July 1998, p 101

References

1. E.H. Love, *A Treatise on the Mathematical Theory of Elasticity,* 4th ed., Dover Publications, New York, 1944

2. Y.C. Fung, *Foundations of Solid Mechanics,* Prentice-Hall, Englewood, NJ, 1965

3. L.E. Malvern, *Introduction to the Mechanics of a Continuous Medium,* Prentice-Hall, Englewood, NJ, 1969

4. W.C. Young and R.G. Budynas, *Roark's Formulas for Stress and Strain,* 7th ed., McGraw-Hill Publishing Co., 2001

5. J.M. Gere and S.P. Timoshenko, *Mechanics of Materials,* 4th ed., PWS Publishers, Boston, 1997

6. F.P. Beer et al, *Mechanics of Materials,* 3rd ed., McGraw-Hill Publishing Co., 2001

7. O.C. Zienkewicz, *The Finite Element Method in Engineering Science,* McGraw-Hill Publishing Co., 1971

8. R.H. Gallagher, *Finite Element Analysis: Fundamentals,* Prentice-Hall, Englewood, NJ, 1975

9. R. MacNeal, *Finite Elements: Their Design and Performance,* Marcel Dekker, Inc., 1994

10. W.D. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, Inc., 1997

11. *Tappi J.*, Vol 78 (No. 7), July 1998, p 101

Finite Element Modeling in Failure Analysis

Thomas Service, Altran Corporation

Selected References

- K.J. Bathe, *Finite Element Procedures in Engineering Analysis,* Prentice-Hall, 1995
- R.D. Cook, D.S. Malkus, M.E. Plesha, and R.J. Witt, *Concepts and Applications of Finite Element Analysis,* 4th ed., John Wiley & Sons, 2001

- D.L. Dewhirst, Finite Element Analysis, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 176–185
- K.H. Huebner, E.A. Thornton, and T.G. Byrom, *The Finite Element Method for Engineers,* 3rd ed., John Wiley & Sons, 1995
- B. Szabo, I. Babuska, and B. Szabi, *Finite Element Analysis,* John Wiley & Sons, 1991
- O.C. Zienkiewicz, *The Finite Element Method in Engineering Science,* 4th ed., McGraw-Hill, 1987

Practices in Failure Analysis

Introduction

ANALYZING the inevitable failures that occur during testing, manufacturing, and service is an essential engineering process for continual improvement in product reliability. This article describes the general procedures, techniques, and precautions employed in the investigation and analysis of metallurgical failures that occur in service. The stages of investigation are discussed, and the various features of the more common causes of failure characteristics are described for fracture, corrosion, and wear failures. The mechanisms of fracture, corrosion, and wear failures are explained in more detail in other Sections of this Volume.

Stages of a Failure Analysis

Although the sequence is subject to variation depending on the nature of the failure and the availability of physical evidence or background information, there are stages that are common to all successful failure analyses. The combination of these stages comprises the total investigation and analysis. The following list includes many of the commonly used stages. The sequence in which these stages are used is not necessarily critical, and not all of the stages will or can be used in every failure analysis. However, a key principle is to not unnecessarily disrupt conditions that may require closer examination at a later date. Moreover, an additional constraint is to follow the Federal Rules of Evidence during investigation of a failure that may be destined for court.
The stages discussed in this article begin first with the preliminary steps of information gathering such as:

- Collection of background data and selection of samples
- Preliminary examination of the failed part (visual examination and record keeping)
- Nondestructive testing

These preliminary steps may then be followed by assessment of the damage and conditions leading to failure. These stages may differ depending on whether fracture, corrosion, and/or wear conditions are being investigated. In an analysis of a fracture, the following steps are described:

- Selection, identification, preservation, and/or cleaning of critical specimens
- Macroscopic examination and analysis (fracture surfaces, secondary cracks, and other surface phenomena)
- Microscopic examination and analysis of fracture surfaces
- Stress analysis to determine the actual stress state of the failed component
- Fracture mechanics
- Determination of the fracture mode

Following these topics on the analysis of fractures, separate sections also briefly describe factors and methods in the analysis of corrosion and wear failures.
In addition, investigations of a failure may utilize various techniques to characterize the condition of material. These include:

- Metallography or microstructural analysis
- Mechanical testing

- Chemical analyses (bulk, local, surface corrosion products, and deposits or coatings).
- Testing under simulated service conditions

Finally, the investigation concludes with a synthesis and interpretation of results. This step may actually require reiteration of previous steps or the introduction of new steps. Similar to design, failure analysis can be an iterative process of discovery and reexamination. Failure analysis can also be a multidisciplinary process and that may require consulting with experts in other disciplines throughout the investigation. Once all information has been assembled, then the final step is to synthesize all the evidence and formulate conclusions. This requires writing a report with follow-up recommendations on preventing future failures. The goal of every failure analyst is to determine not only the failure mechanism but also the root cause, which may be related to misuse, poor maintenance practices, or improper application, or related to the material properties, design, or manufacture of the product.

In cases that involve personal injury or will most likely involve legal pursuit of compensation from another company, care must be taken in preserving the scene and physical evidence. Accidental or deliberate destruction of evidence can result in diverting the legal liability of a failure to the person or company destroying the evidence, even though they may not have caused the original failure.

Practices in Failure Analysis

Collection of Background Data and Samples

The failure investigation should include gaining an acquaintance with all pertinent details relating to the failure, collecting the available information regarding the design, manufacture, processing, and service histories of the failed component or structure, and reconstructing, insofar as possible, the sequence of events leading to the failure. Collection of background data on the manufacturing and fabricating history of a component should begin with obtaining specifications and drawings and should encompass all the design aspects of the failed part as well as all manufacturing and fabrication details—machining, welding, heat treating, coating, quality-control records, and pertinent purchase specifications. Additional information on upfront planning of investigations is also described in the article “Organization of a Failure Investigation” in this Volume.

Collecting Data and Samples

On-Site Investigation. In the investigation of failures, it is also often desirable for the analyst to visit the scene, but for the analysis of some components it may be impractical or impossible for the failure analyst to visit the failure site. Under these circumstances, data and samples may be collected at the site by field engineers or by other personnel under the direction of the failure analyst. A field failure report sheet or checklist can be used to ensure that all pertinent information regarding the failure is recorded.

There also are situations where it is essential to perform failure analyses on the site. While it is recommended that examination be done in a laboratory, the requirements for on-site testing may involve the use of portable laboratories with metallographic equipment for grinding, mechanical polishing, and etching. Small specimens can be cut from a part on the site for preparation, examination, and photography immediately or upon return to a fully equipped laboratory. Photography is, of course, essential; it should be performed by the analyst or perhaps a professional photographer in the case of a large-scale accident scene. Other considerations for on-site examination at an accident scene are also discussed in more detail in the article “Modeling and Accident Reconstruction” in this Volume.

It is also frequently desirable to make acetate tape replicas or room-temperature-vulcanized (RTV) rubber replicas of fracture surfaces or of wear patterns of large parts during an on-site failure analysis. Several replicas should be made of the fracture-origin region using acetate tape softened in acetone, dried, then carefully stripped from the fracture surface. Upon return to the laboratory, the replicas may be gold coated and examined with a scanning electron microscope (SEM). Foreign particles removed from the fracture surface also may be analyzed.

The RTV rubber replica can be applied over a rather large area with less chance of missing a critical spot. A combination of acetate tape and RTV rubber replicas can assure the investigator of better coverage of the area in question. Room-temperature-vulcanized rubber does not provide the sensitivity of an acetate replica, and a setup time of several hours is required. However, the added area can be very important in an investigation.

Hardness testing with a portable hardness testing instrument also may be performed during on-site failure analysis. Several different types of testers are available and in general are either electronic or mechanical in principle. Obviously, small size and light weight are advantages in portable testers.
The major components of the portable laboratory may include:

- A custom-made machine, plus auxiliary materials, for grinding and polishing small, mounted or unmounted metal specimens
- A right-angle head, electric drill motor with attachments and materials for grinding and polishing selected spots on large parts or assemblies. It is also used for driving the grinding and polishing machine described in the previous item
- A portable microscope, with camera attachment and film for use in photographing metallographic specimens
- Equipment and materials for mounting and etching specimens
- A handheld single-lens reflex 35-mm camera, with macrolenses and film
- A pocket-size magnifier, and a ruler or scale
- A hacksaw and blades for cutting specimens
- Portable hardness tester
- Acetate tape, acetone, and containers
- RTV rubber for replicas

Service History. The availability of a complete service history depends on how detailed and thorough the record keeping was prior to the failure. A complete service record greatly simplifies the assignment of the failure analyst. In collecting service histories, special attention should be given to environmental details such as normal and abnormal loading, accidental overloads, cyclic loads, temperature variations, temperature gradients, and operation in a corrosive environment. In most instances, however, complete service records are not available, forcing the analyst to work from fragmentary service information. When service data are sparse, the analyst must, to the best of his or her ability, deduce the service conditions. Much depends on the analyst's skill and judgment, because a misleading deduction can be more harmful than the absence of information.
Photographic Records. Photographs of the failed component or structure are oftentimes critical to an accurate analysis. A detail that appears almost inconsequential in a preliminary investigation may later be found to have serious consequences; thus, a complete, detailed photographic record of the scene and failed component can be essential.
Photographs should be of professional quality, but this is not always possible. For the analyst who does his own photography, a single-lens reflex 35-mm or larger camera with a macrolens, extension bellows, and battery-flash unit is capable of producing excellent results. It may be desirable to supplement the 35-mm equipment with an instant camera and close-up lenses. Techniques and lighting are discussed in more detail in the article “Photography in Failure Analysis” in this Volume.
When accurate color rendition is required, the subject should be photographed with a color chart, which should be sent to the photographic studio for use as a guide in developing and printing. Some indication of size, such as a scale, coin, hand, and so forth, should be included in the photograph.
Samples should be selected judiciously before starting the examination, especially if the investigation is to be lengthy or involved. As with photographs, the analyst is responsible for ensuring that the samples will be suitable for the intended purpose and that they adequately represent the characteristics of the failure. It is advisable to look for additional evidence of damage beyond that which is immediately apparent. For failures involving large structures or key machinery, there is often a financially urgent need to remove the damaged structure or repair the machine to return to production. This is a valid reason to move evidence, but a reasonable attempt must be made to allow other parties, who may become involved in a potential legal case, to inspect the site. All concerned parties then can agree on the critical samples and the best way to remove them. If all parties are not available, care must be taken not to damage or alter critical elements to avoid spoiling evidence. Guidelines governing sample collection are covered in ASTM E 620, E 860, E 1020, and especially E 678. It is also recommended that samples be taken from other parts of the failed equipment as they may display supportive damage.
It is often necessary to compare failed components with similar components that did not fail to determine whether the failure was brought about by service conditions or was the result of an error in manufacture. For example, if a boiler tube fails and overheating is suspected to be the cause, and if investigation reveals a spheroidized structure in the boiler tube at the failure site (which may be indicative of overheating in service), then comparison with an unexposed tube will determine if the tubes were supplied in the spheroidized condition.
As another example, in the case of a bolt failure it is desirable to examine the nuts and other associated parts that may have contributed to the failures. Also, in failures involving corrosion, stress-corrosion, or corrosion fatigue, a sample of the fluid that has been in contact with the metal, or of any deposits that have formed, will often be required for analysis.
Abnormal Conditions and Wreckage Analysis. In addition to developing a history of the failed part it is also advisable to determine if any abnormal conditions prevailed. Determine also whether events—such as an accident—occurred in

service that may have initiated the failure, or if any recent repairs or overhauls had been carried out and why. In addition, it is also necessary to inquire whether or not the failure was an isolated example or if others have occurred, either in the component under consideration or in another of a similar design. In the routine examination of a brittle fracture, it is important to know if, at the time of the accident or failure, the prevailing temperature was low, and/or if some measure of shock loading was involved. When dealing with failures of crankshafts or other shafts, it is generally desirable to ascertain the conditions of the bearings and whether any misalignment existed, either within the machine concerned or between the driving and driven components.

In an analysis where multiple components and structures are involved, it is essential that the position of each piece be documented before any of the pieces are touched or moved. Such recording usually requires extensive photography, the preparation of suitable sketches, and the taking and tabulation of appropriate measurements of the pieces.

Next, it may be necessary to take an inventory to determine if all of the pieces or fragments are present at the site of the accident. For example, an investigation of an aircraft accident involves the development of a considerable inventory, including listing the number of engines, flaps, landing gear, and the various parts of the fuselage and wings. It is essential to establish whether all the primary parts of the aircraft were aboard at the time that it crashed. Providing an inventory, although painstaking, is often invaluable. An experienced investigator determined the cause of a complex aircraft accident when he observed that a portion of one wing tip was missing from the main impact site. This fragment was subsequently located several miles back on the flight path of the aircraft. The fragment provided evidence of a fatigue failure and was the first component separated from the aircraft, thus accounting for the crash.

The most common problem encountered in examining wreckage involves the establishment of the sequence of fractures to determine the origin of the initial failure. Usually, the direction of crack growth can be detected from marks on a fracture surface, such as V-shaped chevron marks. The typical sequence of fractures is shown in Fig. 1(b), where A and B represent fractures that intersect at about 90°. Here the sequence of fractures is clearly discernible from crack branching. Obviously, fracture A must have occurred prior to fracture B because the presence of fracture A served to arrest cracking at fracture B. This method of sequencing is called the *T-junction procedure* and is an important technique in wreckage analysis.

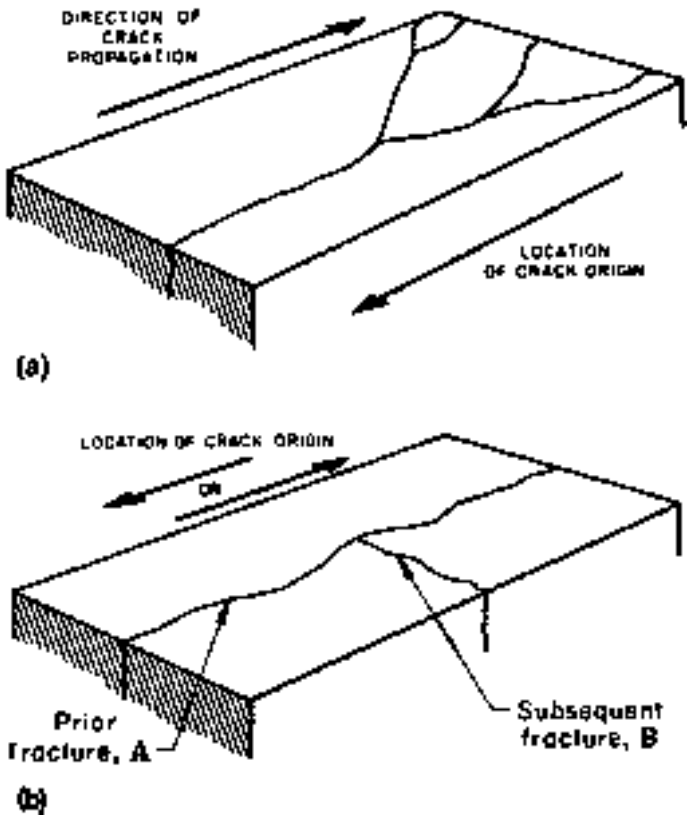


Fig. 1 General features to locate origin from crack paths (a) branching and (b) sequencing of cracking by the T-junction procedure, where fracture A precedes and arrests fracture B

Provided the fragments are not permitted to contact each other, it is also helpful to carefully fit together the fragments of broken components which, when assembled and photographed, may indicate the sequence in which fractures occurred. Figure 2 shows a lug that was part of a pin-joint assembly; failure occurred when the pin broke out of the lug. With the broken pieces of the lug fitted together, it is apparent from the deformation that fracture A must have preceded fractures B

and C. However, parts will not fit well together because of plastic deformation that occurred before or during the fracture process.

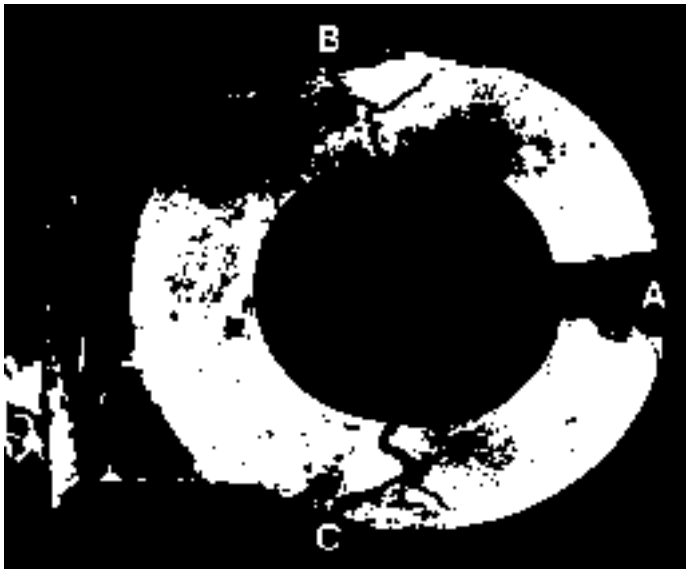


Fig. 2 Fractured lug, part of a pin-joint assembly, showing sequence of fracture. Fracture A preceded fractures B and C.

Preliminary Examination

The failed part, including all its fragments, should be subjected to a thorough visual examination before any cleaning is undertaken. Often, soils and debris found on the part provide useful evidence in establishing the cause of failure or in determining a sequence of events leading to the failure. For example, traces of paint or corrosion found on a portion of a fracture surface may provide evidence that the crack was present in the surface for some time before complete fracture occurred. Such evidence should be recorded photographically.

Visual Inspection. The preliminary examination should begin with unaided visual inspection. The unaided eye has exceptional depth of focus, the ability to examine large areas rapidly and to detect changes of color and texture. Some of these advantages are lost when any optical or electron-optical device is used. Particular attention should be given to the surfaces of fractures and to the paths of cracks. The significance of any indications of abnormal conditions or abuse in service should be observed and assessed, and a general assessment of the basic design and workmanship of the part should also be made. Each important feature, including dimensions, should be recorded, either in writing or by sketches or photographs.

It cannot be emphasized too strongly that the examination should be performed as carefully as possible, because clues to the cause of breakdown often are present, but may be missed if the observer is not vigilant. Inspection of the topographic features of the failed component should start with an unaided visual examination and proceed to higher and higher magnification. A magnifying glass followed by a low-power microscope is an invaluable aid in detection of small details of the failed part.

Examination and Photography of the Damaged/Failed Part or Sample. The next step should be preliminary examination and general photography of the entire part and damaged or failed regions. Where fractures are involved, the entire fractured part, including broken pieces, should be examined and photographed to record their size and condition and to show how the fracture is related to the components. This should be followed by careful examination of the fracture. The examination should begin with the use of direct lighting and proceed at various angles of oblique lighting to delineate and emphasize fracture characteristics. This should also assist in determining which areas of the fracture are of prime interest and which magnifications will be possible (for a given picture size) to bring out fine details. When this evaluation has been completed, it is appropriate to proceed with photography of the fracture, recording what each photograph shows, its magnification, and how it relates to the other photographs. For information on photographic equipment, materials, and techniques, see the article “Photography in Failure Analysis” in this Volume and the article entitled “Photography of Fractured Parts and Fracture Surfaces,” on pages 78 to 90 in *Fractography*, Volume 12 of the *Metals Handbook*, 9th edition (now Volume 12 of the *ASM Handbook*).

Nondestructive Inspection

Although often used as quality-control tools, several nondestructive tests are useful in failure investigation and analysis: magnetic-particle inspection of ferrous metals, liquid-penetrant inspection, ultrasonic inspection, and sometimes eddy-current inspection. All these tests are used to detect surface cracks and discontinuities. Radiography is used mainly for internal examination. A photographic record of the results of nondestructive inspection is a necessary part of record keeping in the investigation.

Magnetic-particle inspection utilizes magnetic fields to locate surface and subsurface discontinuities in ferromagnetic materials. When the material or part to be tested is magnetized, discontinuities that generally lie transverse to the direction of the magnetic field will cause a leakage field to be formed at and above the surface of the part. This leakage field, and therefore the presence of the discontinuity, is detected by means of fine ferromagnetic particles applied over the surface, some of which are gathered and held by the leakage field. The magnetically held collection of particles forms an outline of the discontinuity and indicates its size, shape, and extent. Frequently, a fluorescent material is combined with the particles so that discontinuities can be detected visually under ultraviolet light. This method reveals surface cracks that are not visible to the naked eye.

Liquid-penetrant inspection is used to detect surface flaws in materials. It is used mainly, but not exclusively, with nonmagnetic materials, on which magnetic-particle inspection cannot be used. This technique involves the spreading of a liquid penetrant on the sample. Liquid penetrants can seep into small cracks and flaws (as fine as 1 μm) in the surface of the sample by capillary action. The excess liquid is wiped from the surface, and a developer is applied that causes the liquid to be drawn from the cracks or flaws that are open at the surface. The liquid itself is usually a very bright color or contains fluorescent particles that, under ultraviolet light, cause discontinuities in the material to stand out.

The main advantages of the liquid-penetrant method are its ability to be used on nonmagnetic materials, its low cost, its portability, and the ease with which results can be interpreted. The principal limitations of the liquid-penetrant method are:

- Discontinuities must be open to the surface.
- Testpieces must be cleaned before and after testing because the liquid penetrant may corrode the metal.
- Surface films may prevent detection of discontinuities.
- Penetrant may be a source of contamination that masks results in subsequent chemical analysis of fracture surfaces.
- The process is generally not suited to inspection of low-density powder-metallurgy parts or other porous materials.

Ultrasonic inspection methods depend on sound waves of very high frequency being transmitted through metal and reflected at any boundary such as a metal/air boundary at the surface of the metal, or a metal/crack boundary at a discontinuity within the part or component. High-frequency sound waves can detect small irregularities, but they are easily absorbed, particularly by coarse-grained materials.

The application of ultrasonic testing is limited in failure analysis because accurate interpretations depend on reference standards to isolate the variables. In some instances, ultrasonic testing has proved to be a useful tool in failure analysis, particularly in the investigation of large castings and forgings. Cracks, laminations, shrinkage cavities, bursts, flakes, pores, disbonds, and other discontinuities that produce reflective interfaces can be easily detected. Inclusions and other inhomogeneities can also be detected by causing partial reflection or scattering of the ultrasonic waves or by producing some other detectable effect on the ultrasonic waves.

The disadvantages of ultrasonic inspection include:

- Manual operation requires careful attention by experienced technicians.
- Extensive technical knowledge is required for the development of inspection procedures.
- Parts that are rough, irregular in shape, very small or thin, or not homogeneous are difficult to inspect.
- Discontinuities that are present in a shallow layer immediately beneath the surface may not be detectable.
- Couplants are needed to provide effective transfer of ultrasonic wave energy between transducers and parts being inspected.
- Reference standards are needed, both for calibrating the equipment and for characterizing flaws.

Radiography uses x-rays or gamma rays, which are directed through the sample to a photographic film. After the film has been developed, it can be examined by placing it in front of a light source. The intensity of the light passing through the film will be proportional to the density of the sample and the path length of the radiation. Thus, lighter areas on the plate correspond to the denser areas of the sample, whereas darker areas indicate a crack or defect running in the direction of the incident beam.

The main advantages of radiography are its ability to detect internal discontinuities and to provide permanent photographic records. However, certain types of flaws are difficult to detect by radiography. Laminar defects, such as cracks, present problems unless they are essentially parallel to the radiation beam. Tight, meandering cracks in thick

sections usually cannot be detected even when properly oriented. Minute discontinuities, such as inclusions in wrought material, flakes, microporosity, and microfissures, cannot be detected unless they are sufficiently segregated to yield a detectable gross effect. Laminations normally are not detectable by radiography because of their unfavorable orientation, usually parallel to the surface. Laminations seldom yield differences in absorption that enable laminated areas to be distinguished from lamination-free areas.

Eddy-current inspection can be used on all materials that conduct electricity. If a coil conducting an alternating current is placed around or near the surface of the sample, it will set up eddy currents within the material by electromagnetic induction. These eddy currents affect the impedance in the exciting coil or any other pickup coil that is nearby. Cracks or flaws within the sample will cause distortions in the eddy current, which in turn cause distortion in the impedance of the coil. The resulting change in impedance can be detected by attaching the appropriate electrical circuits and a meter. Flaws or cracks will show up as some deflection or fluctuation on the meter.

The advantages of electromagnetic inspection are:

- Both surface and subsurface defects are detectable.
- No special operator skills are required.
- The process is adaptable to continuous monitoring.
- The process may be substantially automated and is capable of high speeds.
- No probe contact is needed.

Limitations of electromagnetic inspection include:

- Depth of penetration is shallow.
- Materials to be inspected must be electrically conductive.
- Indications are influenced by more than one variable.
- Reference standards are required.

Residual Stress Analysis. X-ray diffraction is the most common method for direct, nondestructive measurement of residual (internal) stresses in metals. Stresses are determined by measuring the submicroscopic distortion of crystalline lattice structures by tensile or compressive residual stresses. However, it should be pointed out that measurement of residual stresses near fractures or cracks may be erroneous because the residual stresses have already been relieved by the fracture and cracks. Testing of undamaged similar, or exemplar, parts is frequently used as the only alternative in order to understand the residual stress system in the failed part prior to failure. More information is in the article “X-Ray Diffraction Residual Stress Measurement in Failure Analysis” in this Volume.

Acoustic-emission inspection detects and analyzes minute acoustic-emission signals generated by discontinuities in materials under applied stress. Proper analysis of these signals can provide information concerning the location and structural significance of the detected discontinuities.

Some of the significant applications of acoustic-emission inspection are:

- Continuous surveillance of pressure vessels and nuclear primary-pressure boundaries for the detection and location of active flaws
- Detection of incipient fatigue fracture in aircraft structures
- Monitoring of both fusion and resistance weldments during welding and cooling
- Determination of the onset of stress-corrosion cracking (SCC) and hydrogen damage in susceptible structures
- Use as a study tool for the investigation of fracture mechanisms and of behavior of materials
- Periodic inspection of tanks and aerial-device booms made of composite materials

This type of data may be useful background information in a failure analysis, or the technique might be used in evaluation of stress effects. Sources of acoustic emission that generate stress waves in material include local dynamic movements, such as the initiation and propagation of cracks, twinning, slip, sudden reorientation of grain boundaries, and bubble formation during boiling. This energy may originate from stored elastic energy, as in crack propagation, or from stored chemical-free energy, as in phase transformation.

Experimental stress analysis can be done by several methods, all of which may be valuable in determining machine loads and component stresses that can cause failures. Stress coating can be used effectively for locating small areas of high strains, determining the directions of the principal strains, and measuring the approximate magnitude of tensile and compressive strains. Gages can then be placed at the high-strain areas and in the principal-strain directions to measure the strain accurately on gage lengths 0.5 to 150 mm (0.02 to 6 in.). Although there are many mechanical, optical, and electrical devices capable of accurate strain measurements, the bonded electrical-resistance strain gage has become the standard tool for general laboratory and field use.

Photoelastic coatings have also been used for laboratory stress measurements. For this technique, a birefringent coating of controlled thickness is bonded to the testpiece with a reflective cement. Optical analysis is similar to conventional analysis but requires special equipment. The analysis may be recorded on color film with single-frame or movie camera.

Fractures

Establishing the origin of a fracture is essential in failure analysis, and the location of the origin determines which measures should be taken to prevent a repetition of the fracture. The fracture-surface characteristics that show the direction of crack propagation (and conversely, the direction toward the origin) include features such as chevron marks, crack branching, and river patterns. Features that help identify the crack origin include concentric fibrous marks, radial marks, and beach marks. By a study of these features, crack progress can be traced back to the point of origin, and then it can be ascertained whether the crack was initiated by an inclusion, a porous region, a segregated phase, a corrosion pit, a machined notch, a forging lap, a nick, a mar, or another type of discontinuity, or was simply the result of overloading.
However, time employed in ascertaining all the circumstances of a failure is extremely important. When a broken component is received for examination, the investigator is sometimes inclined to prepare specimens immediately without devising an investigative procedure. To proceed without forethought may destroy important evidence and waste time. Some of the questions that should be raised concerning the nature, history, functions, and properties of the fractured part, and the manner in which it interacts with other parts, are:

- *Loading.* Were the nature, rate, and magnitude of the applied load correctly anticipated in the design of the part? Were repeated or cyclic loadings involved? What was the direction of the principal stress relative to the shape of the part? Were residual stresses present to an undesirable degree?
- *Material.* Was the recommended alloy used? Were its mechanical properties at the level expected? Were surface or internal discontinuities present that could have contributed to failure? Did the microstructure conform to that prescribed?
- *Shape.* Did the part comply with all pertinent dimensional requirements of the specification? Did the part have sufficient section thickness to prevent local overloading? Were fillets formed with sufficiently large radii? Were there adequate clearances between interacting parts? Were any of the contours deformed during service? Was there evidence of mechanical surface damage?
- *Environment.* Was the part exposed to a corrosive environment or to excessively high or low temperatures? Was the surface of the part suitably protected? Were the properties of the part altered by the exposure? Was there interaction (for example, galvanic) between the material of the part and that of adjacent components?

Examination of a fracture begins with visual scrutiny, which establishes:

- Whether there is gross evidence of mechanical abuse
- Whether there are indications of excessive corrosion
- Whether the part is deformed
- Whether there are obvious secondary fractures
- Whether the origin of the crack can be readily identified
- Whether the direction of crack propagation can be easily recognized

Often it is helpful to have an undamaged part of the same design as the fractured part available during this portion of the examination. The findings of this scrutiny will permit many deductions to be drawn concerning the service conditions existing at and prior to the time of fracture. These findings can then be extended by an examination of the fracture surface at low magnification with a stereomicroscope and then at high magnification by electron microscopy, metallography (occasionally), or some combination of these examination techniques. A survey at low magnification is important for identification of those areas that need further inspection at high magnification. The salient features are recorded in fractographs of appropriate magnification for report purposes and for future reference should subsequent handling or sectioning destroy evidence needed for failure analysis.
Fracture of a part in service is often intimately associated with the type of environment to which the part was exposed. Active chemical environments include water, salt air, salt water, acid solutions, alkaline solutions, molten metal, and even solid metal. Thermal environments that affect metal properties and fracture include exposure to low (cryogenic, for example) and high temperatures.

Selection and Preservation of Fracture Surfaces

The proper selection, preservation, and cleaning of fracture surfaces is vital to prevent important evidence from being destroyed or obscured. Surfaces of fractures may suffer either mechanical or chemical damage. Mechanical damage may arise from several sources, including the striking of the surface of the fracture by other objects. This can occur during actual fracture in service or when removing or transporting a fractured part for analysis.

Usually, the surface of a fracture can be protected during shipment by a cloth or cotton covering, but this may remove some loosely adhering material that might contain the primary clue to the cause of the fracture. Touching or rubbing the surface of a fracture with the fingers should definitely be avoided. Also, no attempt should be made to fit together the sections of a fractured part by placing them in contact. This generally accomplishes nothing and almost always causes damage to the fracture surface. The use of corrosion-inhibiting paper to package samples should be considered.

Chemical (corrosion) damage to a fracture specimen can be prevented in several ways. For instance, because the identification of foreign material present on a fracture surface may be important in the overall determination of the cause of the fracture, many laboratories prefer not to use corrosion-preventive coatings on a fracture specimen. When possible, it is best to dry the fracture specimen, preferably using a jet of dry, compressed air (which will also blow extraneous foreign material from the surface), and then to place it in a dessicator or pack it with a suitable dessicant. However, clean, fresh fracture surfaces should be coated when they cannot be protected from the elements. After failure of large structures several days may be required to remove critical specimens, and so coating the fracture surfaces would be the prudent decision.

Cleaning of fractured surfaces should be avoided in general, but must be done for SEM examination and often to reveal macroscopic fractographic features. Cleaning should proceed in stages using the least aggressive procedure first, then proceeding to more aggressive procedures if needed. Washing the fracture surface with water should especially be avoided. However, specimens contaminated with seawater or with fire-extinguishing fluids require thorough washing, usually with water, followed by a rinse with acetone or alcohol before storage in a dessicator or coating with a dessicant. Sometimes cleaning may also be required for removal of obliterating debris and dirt, or to prepare the fracture surface for SEM examination. Other acceptable cleaning procedures include use of a dry-air blast or of a soft-hair artist's brush; treating with inorganic solvents, either by immersion or by jet; treating with mild acid or alkaline solutions (depending on the metal) that will attack deposits but to which the base metal is essentially inert; ultrasonic cleaning; and application and stripping of plastic replicas.

Cleaning with cellulose acetate tape is one of the most widely used methods, particularly when the surface of a fracture has been affected by corrosion. A strip of acetate about 0.1 mm (0.005 in.) thick and of suitable size is softened by immersion in acetone and placed on the fracture surface. The initial strip is backed by a piece of unsoftened acetate, and then the replica is pressed hard onto the surface of the fracture using a finger. The drying time will depend on the extent to which the replicating material was softened, and this in turn will be governed by the texture of the surface of the fracture. Drying times of not less than 15 to 30 min are recommended. The dry replica is lifted from the fracture, using a scalpel or tweezers. The replicating procedure must be repeated several times if the fracture is badly contaminated. When a clean and uncontaminated replica is obtained, the process is complete. An advantage of this method is that the debris removed from the fracture is preserved for any subsequent examination that may be necessary for identification by x-ray or electron diffraction techniques. To be complete, the analyst should filter solvents used for cleaning to recapture insoluble particulates.

Sectioning. Because examination tools, including hardness testers and optical and electron microscopes, are limited as to the size of specimen they can accept, it is often necessary to remove from a failed component a fracture-containing portion or section that is of a size convenient to handle and examine. This is a destructive process and may spoil evidence in potential litigation cases.

Before cutting or sectioning, the fracture area should be carefully protected. All cutting should be done so that surfaces of fractures and areas adjacent to them are not damaged or altered; this includes keeping the fracture surface dry, whenever possible. For large parts, the common method of removing specimens is by flame cutting. Cutting must be done at a sufficient distance from the fracture site so that the microstructure of the metal underlying the surface of the fracture is not altered by the heat of the flame, and so that none of the molten metal from flame cutting is deposited on the surface of the fracture.

Heat from any source can affect metal properties and microstructures during cutting. Therefore, dry abrasive cutoff wheels should never be used near critical surfaces that will be examined microscopically. Therefore, sectioning should be performed with jewelers' saws, precision diamond-edged, thin cutoff wheels, hacksaws, band saws, or soft abrasive cutoff wheels flooded with water-based soluble oil solution to keep metal surfaces cool and corrosion-free. Dry cutting with an air-driven abrasive disk may also be used with care to remove small specimens from large parts if kept cool, along with coating the fracture surface for protection.

Secondary Cracks. When the primary fracture has been damaged or corroded to such a degree that most of the information relevant to the cause of the failure is obliterated, it is desirable to open any secondary cracks to expose their fracture surfaces for examination and study. These cracks may provide more information than the primary fracture.

In opening cracks for examination, care must be exercised to prevent damage, primarily mechanical, to the surface of the fracture. This can usually be accomplished if opening is done in such a way that the two surfaces of the fracture are moved in opposite directions, normal to the fracture plane. Generally, a saw cut can be made from the back of the fractured part to a point near the tip of the crack, using extreme care to avoid actually reaching the tip of the crack. This saw cut will reduce the amount of solid metal that must be broken. The final breaking of the specimen can be done in several ways:

- By clamping the two sides of the fractured part in a tensile-testing machine, if the shape permits, and pulling
- By placing the specimen in a vise and bending one half away from the other by striking it with a hammer in a manner that will avoid damage to the surfaces of the crack
- By gripping the halves of the fracture in pliers or vise grips and bending or pulling them apart

Cooling the part with liquid nitrogen often reduces the force and plastic deformation necessary to fracture the part. Fortunately, there is little confusion during subsequent examination as to which part of the fracture surface was obtained in opening the crack.

It is recommended that both the crack separation and the visible crack length be measured prior to opening. The analyst may have to use dye penetrant or other nondestructive evaluation technique to actually see the length of a tightly closed crack. Often, the amount of strain that occurred in the specimen can be determined from a measurement of the separation between the adjacent halves of a fracture. This should be done before preparation for opening a secondary crack has begun. The lengths of cracks may also be important for analyses of fatigue fractures or for consideration for the application of fracture mechanics.

Macroscopic Examination of Fracture Surfaces

One very important part of any failure analysis is the macroscopic examination of fracture surfaces. Performed at magnifications from 1 to 50 or 100×, it may be conducted by the unaided eye, a hand lens or magnifier, a low-power stereoscopic microscope, or a SEM. Macroscopic photography of up to 20× magnification requires a high-quality camera and special lenses; alternatively, a large magnifying glass may be used to enlarge a specific area in the photo, such as a crack or other small detail. A metallographic microscope with macroobjectives and lights may be used for somewhat higher magnifications. However, depth of field becomes extremely limited with light optics. For much greater depth of field, a SEM may be used for low-magnification photography as well as higher-magnification work. Stereo or three-dimensional photographs may also be made to reveal the topographic features of a fracture or other surface.

Frequently, a specimen may be too large or too heavy for the stage of the metallograph or the chamber of a SEM, and cutting or sectioning the specimen may be difficult or not allowed because of legal limitations at the time. In these instances, excellent results can be achieved by examining and—where appropriate—photographing replicas made by the method for cleaning fractures (see discussion under “Cleaning”). These replicas can be coated with a thin layer (about 20 nm, or 2×10^{-8} m, thick) of vacuum-deposited gold or aluminum to improve their reflectivity, or they may be shadowed at an angle to increase the contrast of fine detail. The replicas may be examined by incident-light or transmitted-light microscopy. Because they are electrically conductive, the coated replicas may also be examined with a SEM.

The amount of information that can be obtained from examination of a fracture surface at low-power magnification is extensive. A careful scan of the exterior surface of the part in the area adjacent to the fracture is required to determine whether specific stress raisers are present of a type that could have initiated the fracture. If any marks possess sharp reentrant angles, they constitute sites of stress concentration, a frequent cause of crack initiation. In this situation, the obvious remedy is more careful handling procedures and better inspection. Tool marks are another source of stress concentration. A fillet that has too small a radius, even though the surface of the fillet may be an excellent example of high-grade machining, is a recognized initiation site for fatigue cracks. Sharp-bottomed tool marks can initiate fatigue fractures even though the general contour of the area has a generous radius.

The shape, size, and cross section of a specimen or structural component can have a large effect on both the macroscopic and the microscopic appearance of the fracture surface, especially when pronounced stress raisers are present. Holes, corners, notches, machining marks, and, most of all, preexisting cracks actively influence fracture appearance. Pronounced stress raisers are more likely to be contained in a large part than in a small part, because large parts have greater volume and surface area.

The orientation of the fracture surfaces must be consistent with the proposed mode of failure and the known loads on the failed part. Failure in monotonic tension produces a flat (square) fracture normal (perpendicular) to the maximum tensile stress and frequently a slant (shear) fracture at about 45°. This 45° slant fracture is often called a “shear lip.” Many fractures are flat at the center, but surrounded by a “picture frame” of slant fracture. An example of this behavior is to be found in the familiar cup-and-cone fracture of a round tensile test bar.

In thin sheets, tube walls, and small diameter rods, slant shear fracture may occur because through-the-thickness stresses are minimized; that is, even though there may be a plane-strain condition, there may be minimal triaxial (hydrostatic) tensile stresses.

Macroscopic examination can usually determine the direction of crack growth and hence the origin of failure. With brittle, flat fractures, determination depends largely on the fracture surface showing a radial fanlike pattern of the type shown in Fig. 3. Cracks propagate parallel to shear lips if they are present. Where fracture surfaces show both flat and slant surfaces, this can be the terminal end of a fast-moving brittle fracture where the crack speed has slowed significantly. Crack extension can relax the stress so that final fracture occurs by slant shear fracture. Conversely, if a fracture has begun at a free surface, the fracture-origin area is usually characterized by a total absence of slant fracture or shear lip.

Fig. 3 Surface of a fatigue fracture in a 4330V steel part. Chevron marks point to origin of fatigue in lower left corner. Arrows identify shear rupture along the periphery.

Low-power examinations of fracture surfaces often reveal regions having a texture different from the region of final fracture. Fatigue, stress-corrosion, and hydrogen embrittlement fractures may also show these differences because the final failure is due to overload after the cross section is reduced by one of the aforementioned crack-initiation modes.

Figure 4(a) shows the fracture surface of a steel tube and is an excellent example of the type of information that can be obtained by macroscopic examination. The V-shaped chevron marks and fanlike marks clearly indicate that the fracture origin is at the point marked by the arrow. This region, unlike the rest of the fracture, has no shear lip. The flat fracture surface suggests that the stress causing the failure was tension parallel to the length of the tube. The origin of the fracture as seen at higher magnification in Fig. 4(b) shows several small fracture origins having a texture different from that of the remainder of the fracture surface.

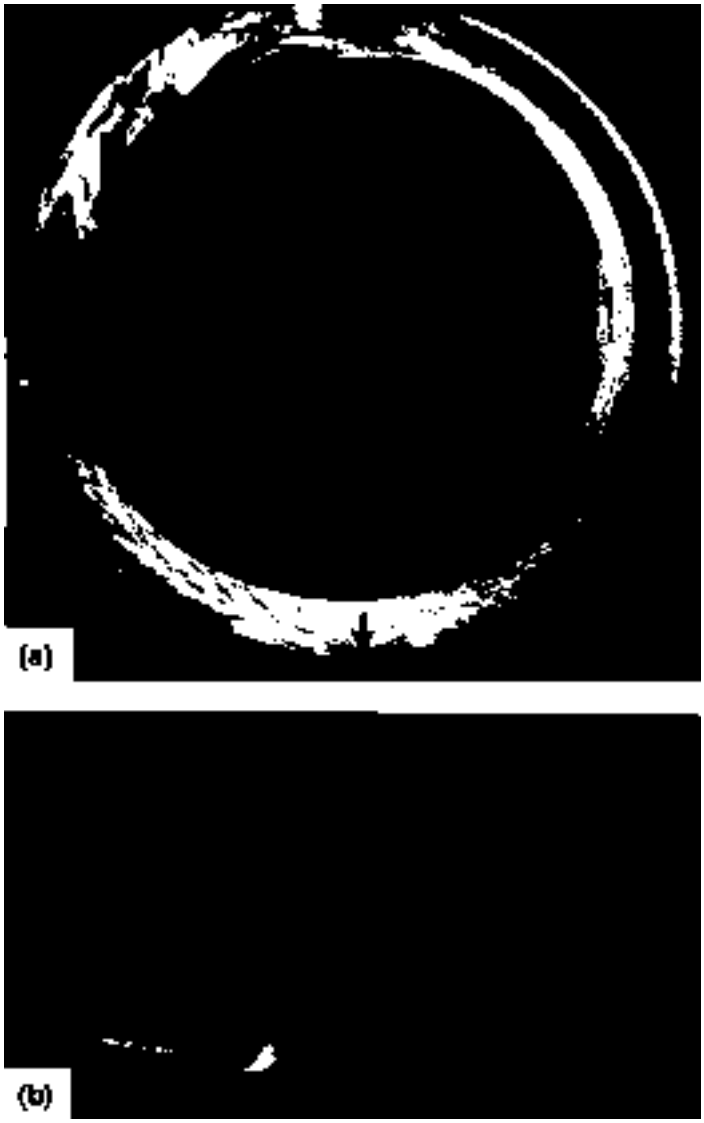


Fig. 4 Fracture of a steel tube. (a) Fracture surface at approximately actual size, showing point of crack initiation (at arrow), chevron and fanlike marks, and development of shear lips. (b) Fracture-origin area at 5×; note that fracture nuclei differ in texture from the main fracture surface.

Microscopic Examination of Fracture Surfaces

Microscopic examination of fracture surfaces is typically done with a SEM. A SEM has the advantage over light microscopy because of the large depth of field and very high magnifications attainable, typically 5000 to 10,000×. In addition, SEMs are often equipped with microanalytical capabilities, for example, energy-dispersive x-ray spectroscopes (EDS). Chemical analysis can be helpful in confirming the chemistry of microstructural features that may be confused with fracture features.

The primary limitation of SEM analysis is sample size. A SEM analysis must be conducted in a vacuum so the sample must be put into a chamber that typically holds a sample less than 20 cm (8 in.) in diameter. Although there are some fracture surface features that are commonly associated with particular failure modes, the novice failure analyst must be very careful in fractographic analyses. Some of the more classic examples of fracture surface topography that indicates a fracture mode are:

- Dimpled rupture typical of overstress failures of ductile metals and alloys
- Cleavage facets, typical of transgranular brittle fracture of body-centered cubic and hexagonal close-packed metals and alloys
- Brittle intergranular fracture typical of temper-embrittled steel, where fracture is due to segregation of an embrittling species to grain boundaries (such as oxygen in iron or nickel), due to intergranular stress-corrosion cracking, or due to hydrogen embrittlement

- Stage II striations, typical of *some* (not all) fatigue failures

Stress Analysis

It is sometimes quite apparent that an excessively high load or stress level was the direct cause or contributed significantly to the failure. Even so, an accurate stress analysis of the magnitude and type (axial, torsion, bending) of stress is required to substantiate the role of stress. In other failure analyses, the analyst may have strong evidence that the cause of a failure is related to excessively high static stresses (or cyclic stresses in the case of fatigue). In these cases, an analysis of the stress during normal operation (or abnormal operation if identified) must be conducted. Analytical, closed-form calculations based on engineering mechanics are often used by designers to predict stress levels in the early design stages. This method of using known "machine design and structural" formulas to predict the stress under a given load is also helpful to the failure analyst, especially in cases where this step may not have been used in the original design of the part. It is not uncommon to find products that have no record of any stress analysis in the development of the product. Even if such calculations were made, the failure analyst may not have access to them. The analyst must answer the questions "Was the component sized properly by the design stress analysis? Did the material have the properties assumed in the design? Did the part fail in a manner consistent with that assumed in design, or did it fail in a way not anticipated in the original design?"

In cases where unusual or abnormal loading is suspected, direct calculation of stresses will fall short and predict incorrect stress levels. In these cases, experimental stress analysis is used for determining machine loads and component stresses. This technique normally involves attachment of strain gages to similar parts in critical areas or typical areas where the failure has occurred. The strain gages are connected to a monitoring device either directly with wires or indirectly by radio signals for monitoring moving or rotating parts. In this way, the actual dynamic stresses can be determined.

For products with very complex shapes and high thermal gradients, a finite-element analysis (FEA) may be required to estimate the level of stress that most likely existed in the failed component. These analyses can stand alone or can be used to help select critical locations for strain-gage attachment. Finite-element analyses can be time consuming and expensive, but they are necessary for an accurate assessment of stress levels in areas of complex geometry of some components. This type of analysis is almost essential for determining stresses caused by thermal gradients such as those found in welding.

Overload failures are often a result of improper design or improper operation. A design analysis is essential in determining which of these is the root cause. Sometimes improper design is a result of incorrect information passed to the designer. In these cases, a failure is the only indication that the wrong input was used for the design. This is also true for fatigue failures. For proper design of rotating or moving parts, a detailed stress analysis is essential. It is much more difficult to predict dynamic stresses than static stresses.

Fracture Modes

Because the initial steps in failure analysis of a fracture involve visual and macroscopic observation, the first impressions should be based on obvious visual evidence. The simplest and most important observations relate to deformation: Was the metal obviously deformed? If it was deformed prior to fracture, yielding and fracture have occurred due to one or more gross overloads. It is predominantly a ductile fracture or a very high-stress, low-cycle fatigue fracture, as can be demonstrated by repeated manual bending of a paper clip or wire coat hanger. The deformation is directly related to the type of stress causing fracture: tension (stretched), bending (bent), torsion (twisted), or compression (shortened or buckled), or a combination of these stress types.

The absence of gross deformation of the failed part indicates that the fracture is predominantly brittle. A brittle fracture should not be confused with brittle material. The shape or geometry of the part made from a ductile metal can result in an overload failure with little overall shape change, or a failure mechanism can operate to start and grow a crack such as a fatigue crack or stress-corrosion crack. When such a crack grows to the point that the remaining cross-sectional area of the part is overloaded by the normal loads, the final overload failure has little macroscale deformation associated with it. Thus, on a macroscale, the failure of a ductile metal can appear brittle. Of course, overload failures of brittle material always appear brittle on a macroscale.

It is usually more difficult to analyze a brittle fracture because there are a large number of possible mechanisms that can cause fracture with little or no obvious deformation. For single overload fractures, these include such factors as stress concentrations, low temperatures, high rates of loading, high metal strength and hardness, SSC, hydrogen embrittlement, temper embrittlement, large section size, and others. For fatigue fractures, causative factors can include stress concentrations, tensile residual stresses, large stress amplitudes, large numbers of load applications, corrosive environments, high temperatures, low metal strength and hardness, wear, and others.

From this discussion, it should become clear that proper failure analysis is not simple but can become exceedingly complex, requiring considerable thought, examination, questioning, and reference to other sources of information in the

literature. However, identifying the failure mode is the key step in a failure analysis, and it is the essential part of determining the root cause.

Ductile Fracture. Overload fractures of many metals and alloys occur by ductile fracture. Overloading in tension is perhaps the least complex of the overload fractures, although essentially the same processes operate in bending and torsion as well as under the complex states of stress that may have produced a given service failure.

The classic example of ductile failure is a tensile test. In this fracture process, considerable elongation, that is, deformation, takes place before the geometric instability, necking, begins. Even after the deformation is localized at the neck, significant deformation occurs at the neck before the fracture process begins. After the neck forms, the curvature of the neck creates a region of tensile hydrostatic (or "triaxial") stress. This leads to initiation of an internal crack near the center of the necked region. In commercial grade alloys, discontinuities such as inclusions or second-phase particles are sources of early void formation by separation of the matrix and the particle. Some of these voids coalesce to develop a crack, which is perpendicular to the tensile axis. The crack spreads until the state of stress, ductility of the metal, and flow condition reach a condition that favors a shear displacement. The crack path then shifts to a maximum shear plane, which is at an angle to the tensile axis (close to 45° in cylindrical specimens). Sometimes this shear lip forms only on one side of the initial flat crack. When this occurs, the resulting fracture surface has a macroscopic appearance known as cup-and-cone. For brittle materials, the majority of the fracture surface is perpendicular to the tensile axis with little or no fracture surface lying on a plane of shear.

Ductile failures in biaxially loaded sheet and plate structures often consist entirely of a shear lip. Pipe and pressure vessels are examples of biaxially stressed components. Often, failures in these components may first appear brittle with limited ductility; however, close inspection usually reveals some general thickness reduction but no necking at the fracture surface.

High-magnification examination of ductile fracture surfaces usually reveals dimples, which tend to be equiaxed when fractures occur under tensile load. Slant fractures or ductile fracture on planes of high shear stress generate elongated dimples. Ductile fractures (i.e., those with macroscopic deformation) are usually transgranular.

Brittle Fracture. There are two general types of brittle fracture caused by a single overload: transgranular cleavage and intergranular separation. Each has distinct features that make identification relatively simple.

Transgranular cleavage can occur in body-centered cubic metals and their alloys (for example, ferritic steels, iron, tungsten, molybdenum, and chromium) and some hexagonal close-packed metals (for example, zinc, magnesium, and beryllium). Face-centered cubic metals and alloys (such as aluminum and austenitic stainless steels) are usually regarded as immune from this fracture mechanism. (See the article "Mechanisms and Appearances of Ductile and Brittle Fracture").

Iron and low-carbon steels show a ductile-to-brittle transition with decreasing temperature that arises from a strong dependence of the yield stress on temperature. Brittle fracture of normally ductile metals depends on several physical factors, including specimen shape and size, temperature, and strain rate. Thus, a component or structure that has given satisfactory service may fracture unexpectedly; the catastrophic brittle fracture of ships in heavy seas and the failure of bridges on unusually cold days are examples. Metallurgical changes, especially strain aging, may cause the brittle fracture of such items as crane hooks and chain links after long periods of satisfactory operation.

Cleavage fracture is not difficult to diagnose because the fracture path is by definition crystallographic. In polycrystalline specimens, this often produces a pattern of brightly reflecting crystal facets, and such fractures are often described as *crystalline* (perhaps improperly as metals are by definition crystalline). The general plane of fracture is approximately normal (perpendicular) to the axis of maximum tensile stress, and a shear lip is often present as a "picture frame" around the fracture. The local absence of a shear lip or slant fracture suggests a possible location for fracture initiation, for shear lips form during the final stages of the fracture process.

The fractography of cleavage fracture in low-carbon steels, iron, and other single-phase, body-centered cubic metals and alloys is fairly well established. Polycrystalline specimens contain numerous fan-shaped cleavage plateaus. The most characteristic feature of these plateaus is the presence of a pattern of river marks, which consist of cleavage steps or tear ridges and indicate the local direction of crack growth. The rule is that, if the tributaries of the "river lines" are regarded as flowing into the main stream, then the direction of crack growth is downstream. This is in contrast to macroscopic chevron marks, where the direction of crack growth, using the river analogy, would be upstream (see the article "Fracture Appearance and Mechanisms of Deformation and Fracture").

Other fractographic features that may be observed include the presence of cleavage on conjugate planes, tear ridges, ductile tears joining cleavage planes at different levels, and tongues, which result from fracture in mechanical twins formed ahead of the advancing crack. Cleavage fracture in pearlitic and martensitic steels is less easily interpreted because microstructure tends to modify the fracture surface. In fact, cleavage fracture surfaces of pearlitic steel have characteristics similar to fatigue striations, so one must be careful not to confuse the fracture mode.

Intergranular fracture can usually be recognized, but determining the primary cause of the fracture may be difficult. Fractographic and microscopic examination can readily identify the presence of second-phase particles at grain boundaries. Unfortunately, the segregation of a layer a few atoms thick of some element or compound that produces intergranular fracture often cannot be detected by fractography. Auger analysis and sometimes EDS are useful for very

thin layers. Some causes of intergranular brittle fracture are given below, but the list is not exhaustive. It does, however, indicate some of the possibilities that need to be considered, and either eliminated or confirmed, as contributing to the fracture:

- The presence at a grain boundary of a large area of second-phase particles (such as carbides in Fe-Ni-Cr alloys or MnS particles in an overheated steel)
- Segregation of a specific element or compound to a grain boundary where a layer a few atoms thick is sufficient to cause embrittlement (embrittlement caused by the presence of oxygen in high-purity iron, oxygen in nickel, or antimony in copper and temper embrittlement of certain steels are examples of intergranular embrittlement where detection of a second phase at grain boundaries is difficult)

The conditions under which a progressively growing crack may follow an intergranular path before final fracture occurs include SCC, embrittlement by liquid metals, hydrogen embrittlement, and creep and stress-rupture failures. These failure modes are discussed in more detail elsewhere in this Volume. In addition, the article "Intergranular Fracture" in this Volume discusses various causes.

Fracture Mechanics Applied to Failure Analysis

The application of fracture mechanics is often pertinent to the investigation of failures, as well as to the formulation of preventive measures. In general, there are two types of conditions that may lead to structural failure:

- Net-section instability where the overall structural cross section can no longer support the applied load
- The critical flaw size (a_c) is exceeded by some preexisting discontinuity or when subcritical cracking mechanisms (for example, fatigue, SCC, creep) reach the critical crack size

Failures due to net-section instability typically occur when a damage process such as corrosion or wear reduces the thickness of a structural section. This type of failure can be evaluated by traditional stress analysis or FEA, which are effective methods in evaluating the effects of loading and geometric conditions on the distribution of stress and strain in a body or structural system.
However, stress analyses by traditional methods or FEA do not easily account for crack propagation from preexisting cracks or sharp discontinuities in the material. When a preexisting crack or discontinuity is present, the concentration of stresses at the crack tip becomes asymptotic (infinite) when using the conventional theory of elasticity. In this regard, fracture mechanics is a useful tool, because it is a method that quantifies stresses at a crack tip in terms of a stress-intensity parameter (K):

$$K = Ys\sqrt{pa}$$

where Y is a geometric factor (typically on the order of about 1), σ is the gross stress across the fracture plane, and a is the crack length. The stress-intensity parameter K quantifies the stresses at a crack tip, and a critical stress-intensity value (K_c) thus can be defined as:

$$K_c = Ys_f\sqrt{pa_c}$$

where σ_f is the fracture stress occurring with a critical crack size, a_c. The critical stress intensity, also known as fracture toughness (K_c), is the value of stress intensity (K) that results in rapid, unstable fracture. Fracture toughness (K_c) depends on both the thickness of the section and the ductility of the material. For a given material, the fracture toughness (or critical stress intensity, K_c) decreases as section thickness is increased. The value of K_c decreases with increasing section thickness until a minimum value is reached. The toughness at this minimum, which is an inherent material property, is the plane-strain fracture toughness (K_{Ic}). Plane-strain fracture is a mode of brittle fracture without any appreciable macroscopic plastic deformation and is thus referred to as linear-elastic fracture mechanics (LEFM). The general conditions for LEFM analysis are expressed as:

$$\text{thickness} \geq 2.5(K_{Ic}/\sigma_y)^2$$

where σ_y is the yield strength.
Linear-elastic fracture mechanics is a useful tool in failure analysis as many (and perhaps most) structural failures occur by the combined processes of crack initiation followed by subcritical crack growth mechanism (for example, fatigue, stress corrosion, creep) until a critical crack (a_c) is reached. In this regard, fracture mechanics is an effective tool for evaluating critical flaw size (a_c) that leads to rapid unstable fracture and can help answer questions during a failure analysis, such as:

- Where should one look for the transition from subcritical crack growth to unstable rapid fracture?

- What was the load on the component at the time of failure?
- Was the correct material used and manufacturing/processing sound?
- Was the part designed properly?
- Did the environment influence the failure?

Of course, many situations may involve thin sections and/or very ductile materials, where the conditional constraint for LEFM may not apply. In this case, the measure of toughness is plane-stress fracture toughness (K_c) and requires the use of elastic-plastic fracture mechanics (EPFM), as the process of unstable fracture involves some plasticity. Plane-stress fracture toughness (K_c) is higher than plane-strain fracture toughness (K_{Ic}), but when thinner sections and more ductile materials are involved net-section instability becomes a factor. See the appendix in the article "Fracture Appearances and Mechanisms of Deformation and Fracture" for more information on the use of fracture mechanics in failure analysis.

Metallographic Examination

Metallographic examination of polished and of polished-and-etched sections by optical microscopy and by electron-optical techniques is a vital part of failure investigation and should be carried out as a routine procedure when possible. Metallographic examination provides the investigator with a good indication of the class of material involved and its structure. If abnormalities are present, these may be associated with undesirable characteristics that predispose to early failure. It is sometimes possible to relate them to an unsuitable composition or to the effects of service, such as aging in low-carbon steel that has caused precipitation of iron nitride, or gassing in copper. Microstructural examination may also provide information as to the method of manufacture of the part under investigation. It can reveal the heat treatment and possible deficiencies in heat treatment such as decarburization at the surface. Microstructural inspection can also reveal possible overheating through coarsening of carbides of superalloys and solution and precipitation of manganese sulfide. Other service effects, such as corrosion, oxidation, and severe work hardening of surfaces, also are revealed, and their extent can be investigated. The topographical characteristics of any cracks, particularly their mode of propagation, can be determined, for example, transgranular or intergranular. This provides information that can be helpful in distinguishing between different modes of failure. For example, fatigue cracks always propagate perpendicular to the maximum cyclic tensile stress while stress-corrosion cracking may propagate along grain boundaries.

Only a few general directions can be given as to the best location from which to take specimens for microscopic examination, because almost every failure has individual features to be taken into account. In most examinations, however, it must be determined whether the structure of a specimen taken adjacent to a fracture surface or a region at which a service defect has developed is representative of the component as a whole. This can be done only by the examination of specimens taken from the failure region and specimens taken from other locations. For instance, in the case of ruptured or bulged boiler tubes in which failure is usually restricted to one portion only, it is desirable to examine specimens taken from both sides of the fracture, from a location opposite the affected zone, and also from an area as remote from the failure as the size of the sample permits in order to determine whether the failure has been due to a material defect or to overheating—and, if the latter, whether this was of a general or localized nature. In investigations involving general overheating, sometimes the original condition of the material can be ascertained only from a sample cut from a part of the tube many feet away from the affected zone.

Metallographic specimens should be taken perpendicular to the fracture surface, showing the fracture surface in edge view. In cases where metal cleanliness may be an issue, the specimen orientation must be selected properly to determine inclusion density and morphology. This type of examination must be performed on the unetched metallographic specimen.

In the investigation of fatigue cracks, it may be desirable to take a specimen from the region where the fracture originated to ascertain if the initial development was associated with an abnormality, such as a weld defect, a decarburized surface, a zone rich in inclusions, or, in castings, a zone containing severe porosity. Multiple fatigue-crack initiation is very typical of both fretting and corrosion fatigue and may form in areas where there is constant stress across a section. Similarly, with surface marks, where the origin cannot be identified from outward appearances, a microscopic examination will show whether they occurred in rolling or arose from ingot defects, such as scabs, laps, or seams. In brittle fractures, it is useful to examine a specimen cut from where the failure originated, if this can be located with certainty. Failures by brittle fracture may be associated with locally work-hardened surfaces, arc strikes, local untempered martensite, and so forth.

For good edge retention when looking at a fracture surface, it is usually best to plate the surface of a specimen with a metal, such as nickel, prior to mounting and sectioning, so that the fracture edge is supported during grinding and polishing and can be included in the examination. Alternative means include hard metal or nonmetal particles embedded in the mount adjacent to the edges.

Analysis of Metallographic Sections. As with hardness testing and macroscopic examination, the examination of metallographic sections with a microscope is standard practice in most failure analyses, because of the outstanding capability of the microscope to reveal material imperfections caused during processing and of detecting the results of a variety of in-service operating conditions and environments that may have contributed to failure. Inclusions,

microstructural segregation, decarburization, carbon pickup, improper heat treatment, untempered white martensite, second phases such as γ′ in nickel-base superalloys and intergranular corrosion are among the many metallurgical imperfections and undesirable conditions that can be detected and analyzed by microscopic examination of metallographic sections.

Even in the absence of a specific metallurgical imperfection, examination of metallographic sections is invaluable to the investigator in the measurement of microstructural parameters such as case depth, grain size, thickness of plated coatings, and heat-affected zone size (HAZ)—all of which may have a bearing on the cause of failure.

Mechanical Testing

Hardness testing is the simplest of the mechanical tests and is often the most versatile tool available to the failure analyst. Among its many applications, hardness testing can be used to assist in evaluating heat treatment (comparing the hardness of the failed component with that prescribed by specification), to provide an estimate of the tensile strength of steel, to detect work hardening, or to detect softening or hardening caused by overheating, decarburization, or by carbon or nitrogen pickup. Hardness testing is also essentially nondestructive except when preparation of a special hardness test specimen is required, as in microhardness testing. Portable hardness testers are useful for field examination, but the type of hardness test must be appropriate for the sample. For example, Brinell is preferred over Rockwell for a gray cast iron part. One must ensure the proper load is used for the test specimen thickness.

Other mechanical tests are useful in confirming that the failed component conforms to specification or in evaluating the effects of surface conditions on mechanical properties. Where appropriate, tensile and impact tests should be carried out, provided sufficient material for the fabrication of test specimens is available. After all photography, fractography, and nondestructive testing have been performed, the tensile properties of the failed component(s) may be tested. (It may be necessary to obtain permission for this destructive testing if litigation is involved.) This involves sectioning the component and machining test specimens. (In all phases, from rough sectioning to the application of load and measurement of the final dimensions, it is advisable to photodocument the process.) It is especially important to understand that the failed component may have been exposed to environmental conditions not experienced by exemplar components. For example, if the component was involved in a crash, there may have been a fire that exposed the component to temperatures that would have altered the mechanical properties. Likewise, the forces that acted to produce an overload failure in one area of the component may have also plastically deformed the component in other areas. To determine the mechanical properties of the component material, a location for testing must be chosen that has not been exposed to detrimental conditions. Here again, exemplar testing is a valuable tool for the failure analyst and should be employed when data from a failed component is skewed by environmental conditions.

Exemplar components are components that match the failed component (i.e., the same part number and hopefully the same lot or batch). The use of exemplars can range from simply being a visual reference in a demonstration to being used for testing specimens to provide mechanical property data about a particular part or batch of parts. The closer the exemplars are to the failed component (geometrically and chronologically), the more reliable the comparison of test data. If a raw material discrepancy is suspected, it may be that other components of the same batch or lot have the same properties. It may also be useful to have exemplars from other batches for comparison testing.

The failure analyst should exercise care in interpreting mechanical test results. If a material has a tensile strength 5 to 10% below the minimum specified, this does not mean that low hardness or strength is the cause of its failure in service. Also, it should be understood that laboratory tests on small specimens may not adequately represent the behavior of a much larger structure or component in service. For instance, it is possible for a brittle fracture of a large structure to occur at or near ambient temperature, while subsequent laboratory tests of Charpy or Izod specimens show a transition temperature well below -18 °C (0 °F). The effects of size in fatigue, stress-corrosion, and hydrogen-embrittlement testing are not well understood. However, on the basis of the limited evidence available, it appears that resistance to these failure processes decreases as specimen size increases. Several investigators have found correlation problems of transition-temperature-type impact tests with service performance.

Occasionally, the mechanical properties may be acceptable over most of the component, but may vary at a bend or other discontinuity. Castings can have significant variations in properties from one location to another depending on the solidification practice for the casting. Thus, the location of the test specimen within the component can also be significant. Mechanical property tests for cast components are frequently performed on coupons separately cast for this purpose. Therefore, results from samples from the casting itself may not be directly comparable.

Other factors may affect material properties results. Material may have been tested prior to forming or other deformation. Subsequent coatings or case hardening may have improved or degraded mechanical properties. Variations from one location to another due to local material processing variations may help explain differences in mechanical properties between the bulk material and the failure origin.

Tensile tests, in many failure analysis investigations, do not provide enough useful information because relatively few failures result from metal that is deficient in tensile strength. Furthermore, samples cut from components that have failed in a brittle manner generally show adequate ductility under the conditions imposed during a tensile test.

Sometimes, however, there is justification for tensile testing of failed components to eliminate poor-quality material as a possible cause of failure. Often, these tensile tests for determining material quality are carried out by manufacturers and suppliers when examining components that have been returned to them for analysis.

The role of directionality in tensile testing of wrought metals should also be considered. Specimens cut transversely to the longitudinal axis of a component (such as a shaft, plate, or sheet) usually give lower tensile and ductility values than those cut along the longitudinal axis. This is due to the marked directionality and the resulting anisotropy produced during rolling or forging. When sectioning tension-test specimens from the failed component, special attention should be paid to the orientation of the specimen. Some components may have quality-assurance notes on the drawings that indicate where tensile specimens are to be taken. This is especially true of critical aircraft components. Anisotropic materials have properties that vary with test specimen orientation. Tensile-strength and yield-strength specifications are usually given in the longitudinal and transverse directions. Typically, and unless otherwise specified, the tensile specimen should be taken with its major axis parallel to the direction of grain flow; however, test specimens are typically taken in two of the three directions: longitudinal and long transverse. (The short transverse direction is typically not tested since it is difficult to obtain specimens of sufficient length in that direction.) It may be necessary to lightly polish the component surface and use a macroetchant for the material being worked with to determine the grain flow. The specimens are removed from the parent component by first cutting a rough specimen shape, then final machining to the specified form.

Residual stresses in the component may result in warped specimens or pinched cutting tools during rough machining. Care should be taken to document the residual stress observations, so that comparison to exemplar parts sectioned in the same manner can be made. Hopefully, final machining of the testpiece will correct the warped shape while still allowing the specimen to run parallel to the grain flow. Straightening of a warped tensile specimen during the test will result in a nonlinear indication on the initial portion of the stress-strain curve. This can be corrected to a line by using curve-fitting software or by drawing a line by eye back to the abscissa as shown in Fig. 5. There may also be nonuniformity in the specimen microstructure that affects the properties of the material, for example, a change in grain size due to cold work or changes in temper due to the HAZ of a weld.

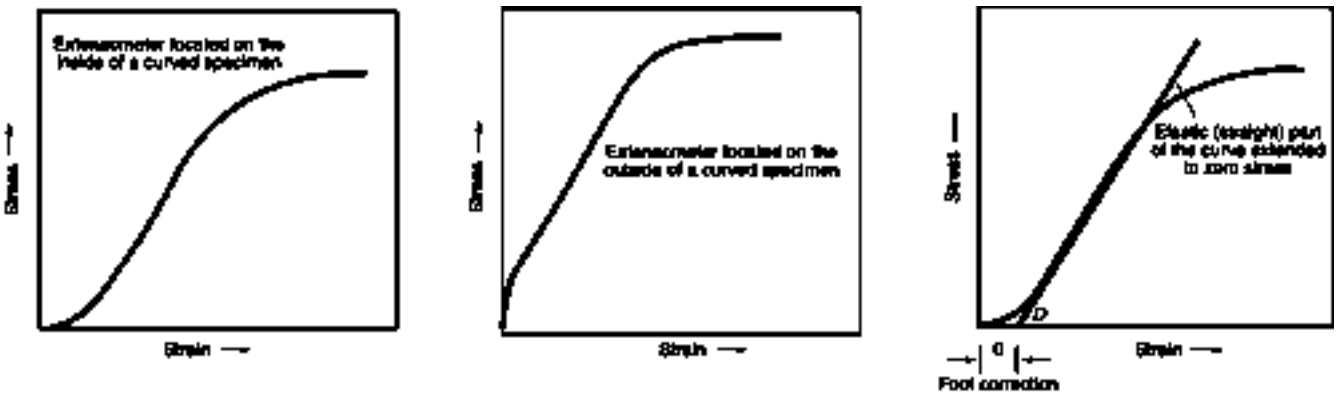


Fig. 5 Examples of stress-strain curves requiring foot correction. *D*, point where the extension of the straight (elastic) part diverges from the stress-strain curve. Source: Ref 1

It may be necessary to make some tests either at slightly elevated or at low temperatures to simulate service conditions. Also, it may be helpful to test specimens after they have been subjected to particular heat treatments simulating those of the failed component in service to determine how this treatment has modified mechanical properties. For example, treating a steel at a temperature in the embrittling range for about 1 h prior to impact testing will indicate any tendency to strain-age embrittlement. The determination of the ductile-to-brittle transition temperature may be useful in investigating brittle fracture of a low-carbon steel.

Component Proof Testing. Critical components are sometimes proof tested after manufacture. This is especially true of critical components that will be subjected to large loads. Proof testing involves loading a component past the recommended operating limits, and possibly slightly into the yield zone. This helps to ensure there are no manufacturing or materials defects that would cause premature failure. Proof loads are usually determined on a case-by-case basis and may be expressed in several forms, such as proof load to twice the operating load or proof load to 90% of the determined yield strength. Extreme care should be taken not to damage the component in the specification of a proof load. For example, stretching a component may alter a designed preload by relieving residual compressive stresses introduced during manufacture. There are usually documents with details of specific methods, tools, and loads used in proof testing and results of tests on critical components. If performing proof testing on exemplar parts from a given manufacturer, it will lend credibility to the results if the manufacturer's test procedures are followed. If the proof tests cite standard methods, such as ASTM, ISO, or other industry-accepted procedure, acquire a copy of the method and follow it closely. If results vary from those reported by the manufacturer, it may be necessary to prove the results are valid.

Conductivity of Aluminum Alloys. During a failure analysis investigation of a heat treated aluminum alloy, conductivity testing may be performed to evaluate proper heat treat condition, to assess areas for heat damage, or to estimate tensile strength. Conductivity is the reciprocal of electrical resistivity and is directly proportional to the mean free path of an electron in the crystal structure of the material. The mean free path is affected by the microstructure of the material, which is affected by heat treat condition.

Solution-annealed structures have soluble constituents that have precipitated out of solution, providing for a material with high conductivity. Aging causes fine precipitation of a second phase and thus decreases the conductivity of the material. Natural aging yields the lowest conductivity and intermediate strength, while artificial aging results in intermediate conductivity and highest hardness/strength. Overaging of the alloy causes the conductivity and hardness values to approach those of the solution-annealed condition.

The manufacturer does not typically specify conductivity values of an aluminum alloy component. However, with proper design and material selection, the alloy and temper of the component are specified, and corresponding values of conductivity may be found in reference data such as specification SAE-MIL-H-6088. Conductivity measurements must be paired with hardness measurements to determine if the failed component meets alloy and heat treatment requirements. An aluminum alloy and its temper can be determined by measuring hardness and conductivity and verifying against reference data.

Chemical Analysis

In a failure investigation, routine analysis of the material is usually recommended. Often it is done last because an analysis usually involves destroying a certain amount of material. There are instances where the wrong material was used, under which conditions the material might be the major cause of failure. In many cases, however, the difficulties are caused by factors other than material composition.

In most instances, slight deviations from specified compositions are not likely to be of major importance in failure analysis. However, small deviations in aluminum content can lead to strain aging in steel, and small quantities of impurities can lead to temper embrittlement. In specific investigations, particularly where corrosion and stress corrosion are involved, chemical analysis of any deposit, scale, or corrosion product, or a substance with which the affected material has been in contact, is required to assist in establishing the primary cause of failure.

Where analysis shows that the content of a particular element is slightly greater than that required in the specifications, it should not be inferred that such deviation is responsible for the failure. Often, it is doubtful whether such a deviation has played even a contributory part in the failure. For example, sulfur and phosphorus in structural steels are limited to 0.04% in many specifications, but rarely can a failure in service be attributed to sulfur content slightly in excess of 0.04%. Within limits, the distribution of the microstructural constituents in a material is of more importance than their exact proportions. An analysis (except a spectrographic analysis restricted to a limited region of the surface) is usually made on drillings representing a considerable volume of material and therefore provides no indication of possible local deviation due to segregation and similar effects.

Also, certain gaseous elements, or interstitials, normally not reported in a chemical analysis, have profound effects on the mechanical properties of metals. In steel, for example, the effects of oxygen, nitrogen, and hydrogen are of major importance. Oxygen and nitrogen may give rise to strain aging and quench aging. Hydrogen may induce brittleness, particularly when absorbed during welding, cathodic cleaning, electroplating, or pickling. Hydrogen is also responsible for the characteristic halos or fisheyes on the fracture surfaces of welds in steels, in which instance the presence of hydrogen often is due to the use of damp electrodes. These halos are indications of local rupture that has taken place under the bursting microstresses induced by the molecular hydrogen, which diffuses through the metal in the atomic state and collects under pressure in pores and other discontinuities. Various effects due to gas absorption are found in other metals and alloys. For example, excessive levels of nitrogen in superalloys can lead to brittle nitride phases that cause failures of highly stressed parts.

Various analytical techniques can be used to determine elemental concentrations and to identify compounds in alloys, bulky deposits, and samples of environmental fluids, lubricants, and suspensions. Semiquantitative emission spectrography, spectrophotometry, and atomic-absorption spectroscopy can be used to determine dissolved metals (as in analysis of an alloy) with wet chemical methods used where greater accuracy is needed to determine the concentration of metals. Combustion methods ordinarily are used for determining the concentration of carbon, sulfur, nitrogen, hydrogen, and oxygen.

Wet chemical analysis methods are employed for determining the presence and concentration of anions such as Cl^-, NO_3^-, and S^-. These methods are very sensitive.

X-ray diffraction identifies crystalline compounds either on the metal surface or as a mass of particles and can be used to analyze corrosion products and other surface deposits. Minor and trace elements capable of being dissolved can be determined by atomic-absorption spectroscopy of the solution.

X-ray fluorescence spectrography can be used to analyze both crystalline and amorphous solids, as well as liquids and gases.

Infrared and ultraviolet spectroscopy are used in analyzing organic materials. When the organic materials are present in a complex mixture (such as, for example, solvents, oils, greases, rubber, and plastics), the mixture is first separated into its components by gas chromatography.

Analysis of Surfaces and Deposits. Wavelength-dispersive x-ray spectrometers (WDSs) and EDSs are frequently used for providing information regarding the chemical composition of surface constituents. They are employed as accessories for SEMs and permit simultaneous viewing and chemical analysis of a surface. The Auger electron spectrometer is useful for detecting the elements in extremely thin surface layers. The Auger electron spectrometer can provide semiquantitative determinations of elements with atomic numbers down to 3 (lithium). The size of the area examined varies greatly with the test conditions; it may be from 1 to 50 μm in diameter.

For chemical analysis of surface areas as small as 1 μm in diameter, the electron-microprobe analyzer is widely used. This instrument can determine the concentration of all but the low atomic number elements, with a limit of detection below 0.1%. The area examined with the ion-microprobe analyzer is a few microns in diameter larger than that examined with the electron-microprobe analyzer. The ion-microprobe analyzer has the advantage of being able to detect nearly all elements (including those of low atomic weights) in concentrations as low as 100 ppm. It is sometimes used to volatilize materials, which are then passed through a mass spectrometer.

Electron microprobes and other modern analytical instruments are described in greater detail in *Materials Characterization,* Volume 10 of the *ASM Handbook.*

The instruments discussed previously are used for direct analysis of surfaces; other techniques can be used for analyzing material that has been removed from the surface. For example, if material is removed in a replica (perhaps chemically extracted), it can be analyzed structurally by x-ray diffraction or electron diffraction. Also, depending on the quantity of material extracted, many of the routine chemical analysis techniques may be applicable.

Spot testing uses chemical tests to identify the metal, the alloying elements present, deposits, corrosion products, and soil. Spot tests can be performed both in the laboratory and in the field; they do not require extensive training in analytical chemistry. The only requirement is that the substance be dissolvable; hydrochloric acid or even aqua regia may be used to dissolve the material.

Spot tests for metallic elements such as chromium, nickel, cobalt, iron, and molybdenum are usually done by dissolving a small amount of the alloy in acid and mixing a drop of the resulting solution with a drop of a specific reagent on absorbent paper or a porcelain plate. Spot colorings produced in this way indicate the presence or absence of the metallic radical under test. Samples may be removed from gross surfaces by spotting the specimen with a suitable acid, allowing time for solution and collecting the acid spot with an eyedropper.

Simulated-Service Testing

During the concluding stages of an investigation, it may be necessary to conduct tests that simulate the conditions under which failure is believed to have occurred. Often, simulated-service testing is not practical because elaborate equipment is required, and even where practical it is possible that not all of the service conditions are fully known or understood. Corrosion failures, for example, are difficult to reproduce in a laboratory, and some attempts to reproduce them have given misleading results. Serious errors can arise when attempts are made to reduce the time required for a test by artificially increasing the severity of one of the factors—such as the corrosive medium or the operating temperature. Similar problems are encountered in wear testing.

On the other hand, when its limitations are clearly understood, the simulated testing and statistical experimental design analysis of the effects of certain selected variables encountered in service may be helpful in planning corrective action or, at least, may extend service life. The evaluation of the efficacy of special additives to lubricants is an example of the successful application of simulated-service testing. The aircraft industry has made successful use of devices such as the wind tunnel to simulate some of the conditions encountered in flight, and naval architects have employed tank tests to evaluate hull modifications, power requirements, steerage, and other variables that might forestall component failure or promote safety at sea.

Taken singly, most of the metallurgical phenomena involved in failures can be satisfactorily reproduced on a laboratory scale, and the information derived from such experiments can be helpful to the investigator, provided the limitations of the tests are fully recognized. However, many company managers prefer to conduct trials to verify improvements before major conversions are approved. This is a more conservative approach, but it only takes one improper recommendation that results in an adverse result to justify trials of major changes.

Reference cited in this section

1. P.M. Mumford, Test Methodology and Data Analysis, *Tensile Testing,* P. Han, Ed., ASM International, 1992, p 55

Corrosion Failures

Corrosion is traditionally defined as the destructive chemical or electrochemical reaction of a metal with its environment. A broader, more modern definition of *corrosion* is the deterioration of a material or its properties due to reaction with its environment. The latter definition makes three important points. It does not limit corrosion to chemical or electrochemical processes, because some forms of corrosion do not involve these processes. It also characterizes corrosion damage as the deterioration of the material or its properties, because some forms of corrosion weaken the material without visible changes in appearance or measurable weight loss. Finally, the modern definition of corrosion acknowledges that nonmetallic materials may corrode. Examples of corrosion of nonmetallic materials are the aging of rubber due to the effects of heat and/or oils; swelling, oxidation, stress cracking, and ultraviolet deterioration of plastics; destruction of the binder in concrete by environmental agents; and the biological attack or rotting of wood. This section focuses on analysis of metal corrosion.

There are many forms of metallic corrosion. These include: uniform corrosion, localized attack (e.g., crevice corrosion and pitting), galvanic attack, cracking phenomena (e.g., stress-corrosion cracking, hydrogen embrittlement, liquid metal embrittlement, and corrosion fatigue), velocity phenomena (e.g., erosion, cavitation, and impingement), fretting, intergranular attack, and dealloying or selective leaching. These forms are discussed in more detail in the article “Forms of Corrosion” in this Volume.

The type of corrosion, the corrosion rate, and the severity and extent of corrosion are influenced by the nature of the environment, the metal surface in contact with the environment, and the mechanical stresses (magnitude and direction). These factors do not necessarily remain constant as corrosion progresses. They are affected by externally imposed changes and those resulting from the corrosion process itself. Other factors that greatly affect corrosion processes include temperature and temperature gradients at the metal/environment interface, crevices in the metal part or assembly, relative motion between the environment and the metal part, and the presence of dissimilar metals in an electrically conductive environment. Processing and fabrication operations such as surface grinding, heat treating, welding, cold working, forming, drilling, and shearing produce local or general changes on metal parts that, to varying degrees, affect their susceptibility to corrosion.

When a corrosion failure has occurred, several means of preventing or minimizing future failures are available. Very often more than one method is used at the same time. The more important corrective and preventive measures are:

- Change in alloy, heat treatment, or product form (e.g., solution annealing of austenitic stainless steels minimizes the risk of intergranular attack and stress-corrosion cracking)
- Use of resin coatings (e.g., acrylics, epoxies, phenolics, furanes and urethanes in the form of paints, potting compounds, adhesives, coatings, and linings)
- Use of inert lubricants (chemically inert resins such as silicones, esters, and fluorocarbons that sometimes can serve both as effective lubricants and as corrosion-resistant coatings and linings)
- Use of electrolytic and chemical coatings and surface treatments (e.g., anodizing of aluminum and aluminum alloys for protection in natural, “nonaggressive” environments)
- Use of metallic coatings (e.g., zinc-rich coatings)
- Use of galvanic protection (cathodic or anodic)
- Design changes for corrosion control
- Use of inhibitors
- Changes in pH and applied potential
- Continuous monitoring of variables.

Corrosion Failure Analysis

Complete investigations of corrosion failures can be very complex, although not all corrosion failures require a comprehensive, detailed failure analysis. Often the preliminary examination will determine the extent of investigation required. In general, the investigation should consider various possibilities without being unnecessarily costly or time consuming. Routine checks to determine that the specified material is used is important. Such checks have shown that “seamless” tubes that failed in service by developing longitudinal splits were actually welded tubes that corroded preferentially at the welds. Forgings that failed in service actually were castings in which failure was initiated by

corrosion at porous areas exposed at the surface of the metal. Also, Monel metal parts that corroded rapidly in an environment to which Monel is highly resistant actually were strongly magnetic and made of carbon steel. In another instance, failure of a braided copper wire was traced to the substitution of a carbon-black filler for the usual silica filler in a sheath covering the wire. Thus, galvanic action occurred between the carbon and the copper in the presence of moisture.

While knowledge of the part and its application is desirable in any failure analysis, accurate history is especially important in the investigation of corrosion failures.

Information about the type of environment to which the failed part was exposed is essential. For instance, corrosion behavior in plants along rivers is affected by both local and upstream chemical composition in the system. Other determining factors include the temperature, whether exposure to the environment is continuous or intermittent and whether these and other factors varied during the service life of the part.

If available, engineering drawings and material and manufacturing specifications for the part should be examined. Particular attention should be given to any part changes that may have been made. Missing information should be obtained from operating and inspection personnel, if possible. At the same time, the accuracy of any relevant documentary information, such as daily log sheets or inspection reports, should be verified. The investigator should determine if any tests or changes affected the physical evidence of the failure.

On-Site Examination and Sampling

On-site examination is generally the same for corrosion failures as for other types of failures. The region of failure itself should be examined visually using hand magnifiers and any other suitable viewing equipment that is available.

The areas immediately adjacent to and near the failure, as well as related components of the system, should be examined for possible effects on the failure. Remote, related equipment should be examined, especially in complex systems and where liquid or gases flow. Also, the possibility of introducing chemicals or other contaminants from upwind or upstream areas should be checked. In one case, for example, ammonia released to the atmosphere by a neighboring plant lead to the SCC failure of a carbon steel boiler.

If possible, the failed component and related components of the system should be photographed before samples are removed. Color photographs are highly desirable, especially when colored products of corrosion are present. A good practice is to calibrate the color at the beginning of every session by taking a picture of the film box, for example, using the same lighting conditions.

Accessories useful in on-site examinations include plastic bags for holding samples, a stainless steel spatula for digging out soft corrosion products, and a file capable of cutting through hard scales. A magnet can distinguish austenitic from martensitic and ferritic stainless steels, as well as steels from nonferrous alloys.

Sampling. When on-site sampling is done, the investigator should be guided by the information already obtained about the history of the failure. The bulk environment to which the failed part was exposed should be sampled, and suitable techniques should be used to obtain samples and make observations (such as pH, for instance) on the local environment at the point of failure.

In addition to taking samples from the failed area, samples from adjacent areas or apparently noncorroded regions should be obtained for comparison purposes. New or unused parts can provide evidence of the initial or unexposed condition of the part.

Removal of specimens and samples of corrosion products from the failed part or area requires the utmost caution. Care must be taken to avoid destroying valuable evidence or damaging the part and its related components. For example, torch cutting often is used for removal of specimens because it is fast and convenient. With torch cutting, cuts should be made away from the failure site to prevent alteration of the microstructure, thermal degradation of residues, and introduction of contaminants. If an abrasive cutoff wheel or a saw is used, the same precautions to avoid overheating apply. Also, coolants or lubricants that can contaminate or alter the part or any deposits present should not be used.

After the samples have been extracted, they must be suitably protected during transportation to the laboratory, such as with glass vials and polyethylene bags. One way to retain deposited material is to tape a covering of inert plastic over the critical area.

Preliminary Laboratory Examination

The procedures followed in the preliminary laboratory examination will vary, depending on whether an on-site examination has already been performed by the failure analyst and its completeness. An on-site examination by a well-equipped investigator will include much of the work that otherwise would have to be done in the preliminary laboratory examination.

When no on-site examination has been done, a failure analyst will be aided by records on the part and environment and the remainder of the failed part (or at the very minimum a good photographic record of it). Also helpful are undamaged or unused parts, related components, and samples of the environment.

Preservation of Evidence. Whether or not an on-site examination has been done, the samples should be handled so that maximum information can be gained before any sample is damaged, destroyed, or contaminated, preventing further tests. Also, a complete written and photographic record should be kept through all investigation stages.

Visual Examination and Cleaning. First, the sample is examined visually, most often with the aid of a low-power hand magnifier. At this stage, important features include the extent of damage, general appearance of the damage zone, and the color, texture, and quantity of surface residues. If substantial amounts of foreign matter are visible, cleaning is necessary before further examination. The residues can be removed in some areas, leaving portions of the failure region in the as-received condition to preserve evidence. When only small amounts of foreign matter are present, cleaning can be deferred so that the surface can be examined with a stereomicroscope before and after cleaning. Cleaning also can be deferred until necessary for surface examination at higher magnifications or for the preparation of metallographic specimens. Small amounts of residues can be removed using transparent tape or acetate replicas and retained for later analysis.

Washing with water or solvent, with or without the aid of an ultrasonic bath, usually adequately removes soft residues that obscure the view. Inhibited pickling solutions will remove adherent rust or scale. Usually, the cleaning solutions should be saved for later analysis and identification of the substance removed.

Nondestructive Tests. For parts in which internal damage may have resulted from corrosion or from the combined effects of corrosion, stress, and imperfections, nondestructive testing is desirable. Radiography and ultrasonics can be used to locate internal discontinuities, and magnetic-particle and liquid-penetrant techniques are used to locate surface imperfections. Methods such as eddy current and holography are used less frequently, mainly because these methods require the use of standards to accurately interpret the data.

Microscopic Examination. Examination by both light microscopy and electron microscopy can be used to observe minute features on corroded surfaces, to evaluate microstructure of the metallic parts, and to observe the manner and extent to which the metal was attacked by the corrodent. Viewing the cleaned surface with a stereomicroscope clearly shows gross topographic features such as pitting, cracking, or surface patterns that can provide information about the failure mechanism. This information includes whether corrosion was the sole phenomenon involved, the type of corrosion, and whether other mechanisms, such as wear and fracture, also were operative.

If the features cannot be observed clearly using a stereomicroscope, instruments such as deep-field photographic microscopes or a SEM may be used. These instruments produce images with a greater depth of field and, therefore, can resolve the topographical features of very rough surfaces. Transmission electron microscopy, using replicas, can resolve extremely fine features.

Microscopic examination of polished or polished-and-etched specimens can reveal microstructural features as well as damage such as cracking. If the corrosion products possess sufficient coherence and hardness to be polished, they should be retained. One way to keep the surface material in place is to impregnate the sample with a casting-type resin and allow it to harden before cutting samples. Usually, it is helpful to vacuum impregnate the sample during the casting process to be sure that any surface-connected voids are filled with casting plastic. To secure maximum quality of retention, polishing on napless cloths with diamond abrasives is used.

Chemical Analysis of the Corrosion Products

The job of the failure analyst is to establish what role, if any, the corrosion products played in the failure and to identify and analyze the metal or metals of which the failed part was made, the environment to which the failed part was exposed, inhomogeneities in the part surface, and foreign matter and metal surfaces.

Both conventional techniques (such as wet chemical analysis, emission spectroscopy, x-ray diffraction, infrared spectrophotometry, gas chromatography, and x-ray fluorescence spectrography) and special techniques (such as energy dispersive x-ray spectrometry, electron-microprobe analysis, ion-microprobe analysis, Auger-electron spectrometry, Mössbauer spectrometry, and electron diffraction) may be needed to define the composition and structure of various substances completely.

Identification and analysis of the metals of which the failed part was made usually is routine. Ordinarily, the purpose is not to look for minor deviations from the chemical composition specified for the part, but rather to check for possible major deviations in composition and to make sure that the correct alloy was used. However, for some austenitic stainless steels, the corrosion resistance and other properties of welded joints require close control over the composition of the stainless steel and the weld metal.

The bulk composition of the failure environment as well as the local composition of the environment at the metal interface are important in determining whether, or how, corrosion contributed to failure. Composition of the environment usually can be obtained chemically or spectroscopically.

Serious corrosion damage can result from the presence of inhomogeneities in the surface of a metal part used in a corrosive environment. A classic example is severe local attack on a stainless steel because of embedded particles of “tramp” iron in the surface of the stainless steel. Establishing the nature of a layer of material on the metal surface may be more difficult than analyzing the material, especially when a thorough on-site investigation cannot be made. The surface

layer can be merely a trace of innocuous soil or a residue of corrodents or of corrosion products. It also can differ in composition from the bulk metal.

Corrosion Testing

Various types of corrosion-testing techniques are used to investigate corrosion failures and to evaluate the resistance to corrosion of metals and alloys for service in specific applications. They include accelerated tests, simulated-use tests, and electrochemical tests. Other techniques include monitoring performance in pilot-plant operations and in actual service.

Some accelerated-test methods have been accepted as standard by both the military and industry. To shorten testing time, corrosion is accelerated in relation to naturally occurring corrosion, usually by increasing temperature, using a more aggressive environment, or increasing stress (e.g., SCC).

Because various factors influence natural corrosion processes and differ widely in time dependence, the results of laboratory tests must be interpreted carefully. They can be related to expected actual service behavior only when close correlation with long-term service results has been established.

Simulated-use tests are used frequently to analyze corrosion behavior of metals and alloys in specific applications. In these tests, either actual parts or test specimens are exposed to a synthetic or natural service environment.

Electrochemical tests provide data that can establish criteria for passivity or anodic protection against corrosion and determine critical breakdown or pitting potentials. The two general methods of electrochemical corrosion testing are controlled current and controlled potential. For either test method, ASTM G 3 provides useful guidance and standardization of the manner of recording and reporting electrochemical measurements.

In the controlled-current method, the current (a measure of the corrosion rate) is controlled and the resulting corrosion potential is measured. Several instruments are available for such tests, in which either logarithmic or linear polarization curves are developed. Both galvanostatic and galvanodynamic polarization measurements are employed to plot anodic and cathodic polarization curves.

In the controlled-potential method of electrochemical testing, the corrosion potential (oxidizing power) is controlled and the resulting corrosion current is measured. Equipment is available for both constant-potential (potentiostatic) and varying-potential (potentiodynamic) testing to determine overall corrosion-rate profiles for metal-electrolyte systems over a range of potentials.

For the most part, electrochemical tests are more valuable for evaluating the corrosion resistance of materials per se rather than for direct use as a failure analysis tool.

Tests for Corrective Action

Once the failure cause and mechanism have been established, the conclusions and the effectiveness of recommended corrective actions can be confirmed through testing. These tests should simulate as closely as possible the environmental and mechanical conditions to which the failed part was subjected in service. Ideally, the part should be tested in service; however, such testing frequently is not feasible because of factors such as length of time to failure or extraneous damage caused by failure of the part. Thus, simulated-service tests may have to be performed in the laboratory. The following parameters should be carefully controlled:

Environmental factors:

- Temperature, which may be steady or fluctuating, and also may affect stress
- Single-phase or two-phase environment, which may involve alternate wetting and drying
- Environmental composition, including major and minor constituents, concentration and changes thereof, dissolved gases, and pH
- Electrochemical conditions, which may involve galvanic coupling or applied cathodic protection

Mechanical factors:

- Loading, which may be static or cyclic—if cyclic, mean stress may be zero, tensile, or compressive; also, stress-wave shape and period must be defined
- Surface damage, which may occur by fretting, abrasion, cavitation, or liquid-impingement corrosion

Wear Failures

Wear, friction, and lubrication are complex, interwoven subjects that may all affect the tendency of a part to fail or to cease being able to perform its intended function. While all three are very important, only wear receives emphasis in this section, while friction and lubrication are covered less extensively. The science of wear is called *tribology,* from the Greek word *tribos,* meaning wear, and it forms the basis for a systems approach to wear failure analysis.
A systems approach to wear is necessary to understand the interactions among:

- The system material components—the bulk and surface properties of the wearing parts, the properties of the material causing the wear, and the interface medium
- The operating variables
- The interactions among the material components of the system
- The operating environment

It should already be clear that wear failures are not easy to analyze, describe, or prevent. There is no general agreement even as to the various types or forms of wear because they can be described in several different ways. One relatively simple way to classify the mechanical surface damage caused by wear is to describe eight major forms or mechanisms. This section discusses these eight form of wear, as follows:

- Abrasive wear
- Erosive wear
- Adhesive wear
- Fretting
- Cavitation
- Liquid-droplet impingement
- Rolling-contact fatigue
- Corrosive wear

About half of all wear failures are caused by abrasive wear, 15% by adhesive wear, and about 8% each by erosion and fretting. The other four mechanisms account for the remaining wear failures. The characteristics and mitigation of each are briefly described in the following section, along with lubricant failures.

Types of Wear

Abrasive Wear. The principal mechanism of abrasive wear is indicated by cutting. Abrasive wear is sometimes referred to as grinding wear. High-stress, low-speed particles or projections from hard materials cut and plow small grooves in softer materials. This type of wear is characteristic of damage to many ground contact tools, such as plows and cultivators, as well as other parts that dig into or rub against hard, abrasive materials. Again, it can lead to complete or partial failure and destruction of the part, particularly with respect to cutting tools, which can be badly dulled.
Abrasive wear can be mitigated by changing material to:

- High-carbon steel
- 20 to 30% Cr white cast iron
- WC or TiC composites

and by increasing surface hardness using:

- Hard chrome plate
- Hardfacing alloy weld overlay
- Electroless nickel-phosphorus alloy plate
- High-velocity oxyfuel (HVOF) ceramic coatings (WC, CrC, aluminum oxide)

Lubrication is of marginal use in combating abrasive wear, unless a lubricant layer thicker than the particle size can be maintained. Further examples and details are given in the article “Abrasive Wear Failures” in this Volume.
Erosive wear is similar to abrasive wear, except that the force is provided by the kinetic energy of the particles as they are carried in a fluid. Low-stress, high-speed particles impinge upon the surface at some angle, tending to cut (ductile metal) or fracture (brittle material) very small *wear chips,* or particles from the surface. Erosion is a particular problem with various types of impellers, propellers, fans, and other parts where particles in a fluid, such as air or water, strike the

surface. The very large number of very small particles tends to cause channeling, forming grooves and removing soft surface layers (e.g., paint). Erosive wear also tends to round sharp edges of various types of parts, reducing efficiency and eventually leading to complete destruction.

In general, erosive wear resistance is increased with the hardness of the surface. Frequently, a relatively thick hardfacing or a relatively thin, hard, wear-resistant coating is used, because it may not be desirable to increase the general hardness of the metal. The increased hardness may lead to brittle fracture, which could be a worse problem than the erosive wear. Some specific steps that can be taken are to change flow conditions by:

- Reducing fluid velocity, but particle dropout must be avoided
- Eliminating turbulence at misalignments, diameter changes, gaps at joints, and so forth
- Avoiding sharp bends less than about 3.5 pipe diameters

and change material to:

- Harder material, but must be harder than particles, for example: high-chromium white cast iron or nitride case-hardened steel
- Hard coatings, for example: cement-lined pipe; tungsten carbide, hard chrome, electroless nickel; or cobalt-base alloy flame, arc, or HVOF deposits
- Elastomer or rubber lining in slurry piping

Adhesive wear occurs by dynamic metal-to-metal contact between two surfaces sliding relative to one another, when there is poor or nonexistent lubrication. The principal mechanism of adhesive wear is described by the key word "microwelding," which is similar to friction welding. At low stress, bonding is on a microscale at high points, called *asperities,* at least initially. At later stages and higher stresses, bonding is more extensive, and severe damage occurs. Adhesive wear, then, involves microwelding between two metals that are mutually soluble; that is, they are inherently capable of being welded together. Since adhesive wear is often basically a lubrication problem, a lubricant is frequently involved and must be considered as part of the system. Discussion of lubricants and lubrication is contained in a later section of this article.

While the term *dynamic adhesive wear* is preferred, other terms are frequently used to describe varying degrees of damage. Some of these terms, in order of increasing severity, are:

- *Scuffing:* superficial scratches on the mating faces
- *Scoring:* grooves cut into the surface of one of the components
- *Galling:* severe tearing and deep grooving of one face and buildup on the mating surface
- *White layer:* in steels, the formation of a very hard (>800 HV) white etching phase by frictional heating
- *Seizure:* "friction welding" of the mating parts so they can no longer move

These terms are less accurate, and it is preferred to use the general term of adhesive wear. The "dynamic" aspect refers to the movement of one surface sliding past another surface, such as a shaft rotating in a sleeve bearing, making and breaking of threaded connections, or two gear teeth contacting under load. Adhesion is favored by chemically clean surfaces, nonoxidizing conditions, and by chemical and structural similarities between the sliding couple.

If a poor lubricant, or no lubricant, is present in the interface, the adhesion between the two surfaces rapidly escalates and very large wear scars may occur, accompanied by gross overheating. The heat comes from the friction between the two surfaces (see the information on friction in a later section of this article). In some cases, complete destruction of surfaces may occur. In order to have destruction of the surface, frictional heating must bring the local temperature at the interface into the 870 to 1090 °C (1600 to 2000 °F) range, or higher, for steel. At this temperature, great changes occur in the microstructure of hardened steels, which are temperature-sensitive. Rehardening will occur, with the formation of a very hard, brittle, untempered martensite, "white layer," at the surface, surrounded by softer, highly tempered martensite below the surface. The structure has a pattern that is quite similar to that of grinding burn.

To prevent or minimize adhesive wear, all facets of the situation must be considered; a system approach must be taken. The result may be that several courses of action will be taken. The considerations must include the nature of the wearing metals, the surfaces of these mating components, and the lubrication or other environment that may be present. Adhesive wear is the one wear mechanism that is most degraded by poor lubrication, or lack thereof, and the most benefited by good lubrication.

Lubrication to separate and cool the surfaces and remove wear debris can be used to minimize adhesive wear:

- Keep bulk lubricant cool to prevent overheating at the contact surfaces from frictional heat.
- Use lubricating oil with EP or other additives that form chemical films to prevent metal-to-metal contact.
- At higher temperatures, use greases or solid lubricants.

Materials with better inherent adhesive wear resistance can be selected:

- Martensitic and precipitation-hardened (PH) stainless steels have generally poor galling resistance.
- Austenitic stainless steels, with high work-hardening rate, are better.
- For carbon and alloy steels, higher hardness is beneficial.
- Stainless steel S21800 and nickel casting alloy CY5SnBiM are very resistant to adhesive wear.

Surface finish and coatings are important for minimizing adhesion:

- Relatively smooth surfaces do not have projections (asperities) that penetrate the lubricant film, which, in turn, should have as high of a viscosity as possible for the application.
- A surface that is too smooth (<10 μin.) will not carry the lubricant into the contact zone and will not tolerate debris.
- Hard coatings, plating, or overlays can be beneficial.
- Diffusion coatings, such as nitriding, tuftriding, and sulfidization, can be good.
- Use mutually insoluble metals with low shear strength to prevent adhesion. (Gold or silver plating is sometimes used on one of two mating high-speed gears, or other contacting parts.)

Fretting, sometimes referred to as static adhesive wear, is wear that occurs between two closely contacting surfaces having oscillatory relative motion of extremely small magnitude. The term *fretting* is derived from the French word "frottement," which means rubbing. Fretting wear has also been called fretting corrosion, friction oxidation, chafing fatigue, and wear oxidation. Fretting is the cause of discoloration and damage in close-fitting joints or between any two closely fitted parts that are then subjected to vibration during service. Some examples are:

- Press fits on shafts at coupling hubs, gears, wheels, and bearing inner races
- Bolts, rivets, and pins under the heads and in their holes
- Rolling-element bearings while stationary during shipping
- Wire rope between the strands
- Turbine and compressor blade roots
- Heat exchanger tubes in tube sheets and baffles

Instead of having a dynamic sliding motion as discussed previously, fretting involves adhesive wear in a joint that is essentially stationary except for the small vibratory elastic motion that inevitably occurs during machine operation. In other words, the joint is supposed to be stationary, but actually it is not, due to a minute metal-to-metal rubbing motion at the interface. Again, microwelding occurs, tearing tiny fragments from one or both surfaces. This wear debris remains between the surfaces and oxidizes to form a reddish-brown "cocoa" (mostly αFe_2O_3 iron oxide) on iron and steel, or a black powder (aluminum oxide) on aluminum. It is the wear debris that oxidizes when exposed to air that causes fretting to also be called *fretting corrosion*, although the "corrosion" seems to be incidental to the wear resulting from the adhesion.

One of the major problems with fretting is not the destruction of the joint itself, but the fact that it frequently leads to fatigue cracks and fracture originating in the interface. Fretting can cause the endurance limit of a steel to be reduced from, say, 515 to about 375 MPa (75 to about 55 ksi), often enough to cause a service failure. In many cases, the site where the fretting fatigue originates is different from the expected location of fracture, particularly if the expected failure location has been strengthened by heat treating or mechanical prestressing of the surface.

The major variables in fretting are the tightness of the interface and the magnitude of the vibration that is causing relative motion. The probability of the fretting leading to fatigue failure depends on the magnitude of the cyclic stresses in the fretted part and the reduction in the endurance or fatigue limit. Fretting is not easy to prevent or eliminate, although it can be minimized by one or more of the following procedures:

- Increase the interference of press-fit components, so as to increase the clamping force and prevent relative motion. Recommended values are given in design handbooks.
- Use plastic or rubber pads, gaskets, or bushings to absorb some or all of the vibration.
- Increase the hardness of one or both surfaces to increase the resistance to microfracture.
- Roughen the surfaces in an attempt to "lock" them together, although this probably will not work unless the pressure or interference fit is also increased between the two surfaces.
- Chromium plate one surface to reduce the coefficient of friction and try to prevent microwelding to the other surface.
- Heat treat with a salt bath or gas atmosphere diffusion treatment to form an εFeN compound in the surface that resists fretting of ferrous metals.

- Mechanically prestress the surface, such as by shot peening, to prevent the fretting from causing fatigue cracking and fracture.

Elimination of the vibration causing the fretting is often impossible, because it is a normal aspect of the operation of the equipment or because it is so small as to be beyond one's ability to correct (e.g., by better balancing of rotating components).

Cavitation is the removal of material from a surface by the formation and rapid collapse of bubbles of gas or vapor in an adjacent liquid. Sometimes called *cavitation corrosion* or *cavitation erosion*, this type of wear is somewhat unique in that only one metal surface is involved. The other material is a liquid, frequently water or water/antifreeze. The basic mechanism is true pitting fatigue, resulting in pits in the surfaces of parts such as diesel engine wet cylinder sleeves, hydraulic turbines, nozzles, impellers, and propellers, or any other part that moves rapidly through liquid. It can also occur in pressure let-down valves in rapidly flowing liquid piping systems. High-speed flow of liquid in these devices causes local hydrodynamic pressures to vary widely and rapidly. The resulting pits can be quite small or very large and deep, depending on the particular circumstances.

Cavitation pitting is caused by rapid, repetitive, relative movement between the metal and the liquid. When the local pressure is reduced, as when the metal and liquid momentarily move in opposite directions, small cavities are formed in the liquid at the interface with the metal in the low-pressure regions. These gas or vapor cavities form when the local pressure drops below the vapor pressure of the liquid or dissolved gas. Then, when the metal and the liquid move toward each other with reversal of movement, the cavities at the interface implode, or collapse, on the metal surface, causing very high local-contact stresses. If vibration or other relative movement is continued for a large number of load applications at these local points on the metal, the implosion forces cause fatigue fracture or pits to form in the metal surface. Continued service of this type can cause complete penetration of the metal and/or large pits in the metal. One of the characteristics of cavitation pitting is that the pitting occurs primarily in a pattern on the surface, with most of the pits concentrated in specific areas, depending on the local vibration or flow characteristics. Another characteristic of cavitation damage is that the pitting is very rough and jagged. For example, if cavitation is occurring in a centrifugal pump, the damage will generally be on the back side of each vane of the impeller at the same location near the inlet and will be jagged. Erosion damage, on the other hand, will be on the front or pressure side of the vanes and will generally be grooved.

Cavitation pitting of diesel engine wet cylinder sleeves can result as the sleeve wall moves in and out due to the internal pressure cycles. It can cause complete destruction of an engine if the coolant gets into the combustion chamber or can cause loss of efficiency of an impeller or propeller. Any corrosion is incidental to the pitting fatigue that causes the pits, although corrosion can easily occur on fresh, unprotected metal fracture surfaces.

Some of the variables involved in cavitation are:

- Liquid composition, density, surface tension, and viscosity
- Dissolved gas content
- Pressure relative to the vapor pressure of the liquid at the system temperature
- Temperature relative to the boiling point at the local pressure
- Vibration levels
- Flow velocity—there is a power-law relationship with an exponent between 3 and 7
- Material

Cavitation pitting is not easily prevented, because the vibration and liquid flow characteristics are not easily controlled. Some possible ways to avoid it are to streamline fluid flow by design changes to avoid sudden pressure drops, stiffen the part to change vibration characteristics, increase the vapor pressure of the liquid or the total pressure in closed systems, and reduce surface roughness. Pumps should be operated at or very near their design flow rate and head pressure; cavitation occurs in off-design operation. Material changes, such as using a metal with a higher fatigue strength, can slow but not eliminate damage. For example, a Stellite overlay can increase resistance by 10 times or more over bare type 304 stainless steel, but will not stop the cavitation. If none of these possible solutions work, it may be necessary to simply replace the parts during scheduled maintenance.

Liquid-droplet impingement is, in some ways, the inverse of cavitation, but it produces damage that is very similar to cavitation damage in appearance. Liquid-droplet impingement is metal loss from the repeated impact of liquid droplets that are being carried along in a gas or vapor stream. The damage is generated by the initial high-pressure pulse in the area of impact causing cracks and radially displaced metal, followed by breaking off of the asperities thus formed by subsequent rapid radial flow from subsequent droplet impacts. The droplets usually form by condensation from a vapor stream, as the temperature, the pressure, or both are reduced. The most common places for this damage to occur are the blades in the late stages of steam turbines and downstream of elbows in piping systems. Similar to cavitation, the area where metal has been removed by liquid-droplet impingement will be very rough and jagged.

The significant variables in liquid-droplet impingement damage include:

- Material resistance
- Flow velocity
- Pressure and temperature drops
- Droplet quantity and size
- Liquid specific gravity

The primary direction to take in combating this form of wear damage is to modify the flowing gas properties. Design or process changes can be used to eliminate, or at least reduce, the size and quantity of the droplets. Material changes to minimize damage could include:

- Increasing the local hardness in the damage zone, such as by flame or induction hardening
- Using weld overlay, laser clad, or brazed-in inserts of a material such as Stellite
- Applying very hard coatings, such as HVOF tungsten carbide

Rolling-contact fatigue is surface damage that results by fatigue from repeated rolling, or rolling and sliding contact between curved metal surfaces. The key words descriptive of this type of wear are “pitting fatigue” with origins on or near contacting surfaces. Although the cause of the fatigue and the locations of the origins are somewhat different in each subtype, these are true fatigue fractures that are caused by shear-stress origins due to contact stresses between two metal surfaces. Three mechanisms have been identified and are described: subsurface-origin pitting, surface-origin pitting, and subcase-origin spalling. This form of wear damage is most commonly seen in rolling-element (ball and roller) bearings, gears, and rolling-mill rolls.

The common thread linking the three types of rolling-contact fatigue is that each represents the limit of load-carrying capacity; improvements are made only with great difficulty, if at all, assuming that the materials are proper and the geometries are accurate. Where metal-to-metal contact is concerned, such as in gears or rolling-contact bearings, the materials are usually hardened steels, with hardness values near 60 HRC. More details on rolling-contact fatigue are given in the article “Fatigue Failures” in this Volume.

Corrosive wear really is not a separate type of wear, but is simply a combination of corrosion and wear occurring either alternatively or simultaneously. Corrosive wear can result in more rapid degradation of material than would be observed by either the corrosion or the wear acting individually. Two conditions can occur. In one case, corrosion deposits that form are more easily removed by one of the wear mechanisms, such as erosion. In the other case, the wear process can remove protective deposits or passive layers, exposing bare metal to the corrosive environment.

Corrosive wear is a major reason for gasoline engine wear, particularly if the engine is stopped before it has had a chance to heat up to operating temperature. The moisture that was formed as a product of combustion cannot evaporate from the cold engine, but remains inside and condenses on various internal parts. Corrosion may occur if the lubricant does not prevent it. Then, when the engine is again operated for a short time, the corrosion products will be worn off by operation, only to reform again because the moisture cannot evaporate from the cold engine.

Wear of parts operating in a corrosive environment is particularly difficult to prevent because of the difficulties in specifying and testing alternate metals. Frequently, various types of surface coatings are of great help in solving the problems of corrosive wear. In some cases, the use of lubricants containing corrosion inhibitors can be beneficial.

Lubricants

Almost any surface film acts as a lubricant, preventing cold welding of asperities on opposing surfaces or allowing opposing surfaces to slide across one another at a lower frictional force than would prevail if the film was not present. Lubricants may be either gas, liquid, or solid. The main function of a lubricant is to maintain separation of the surfaces by pressure within the film, which opposes the applied contact force. Another function of a lubricant is to carry away heat generated by two surfaces sliding under contact pressure. Liquid lubricants dissipate heat better than solid or semifluid lubricants, but the shear properties of all types of lubricant are critical to performance. A third function of liquid lubricants is to flush dirt and wear debris out of the contact zone, so they can be removed subsequently by filtration.

Lubricating oils are relatively free-flowing organic substances that lower the coefficient of friction in mechanical devices. They are available in a broad range of viscosities, and many are blended or contain additives to make them suitable for specific uses. In general, lubricating substances that are fluid at 20 °C (68 °F) are termed *oils;* lubricating substances that are solid or semifluid at 20 °C (68 °F) are termed *greases.*

Oils are derived from petroleum (mineral oils), plants, or animals (fixed oils). Commercial mineral, oil-base products consist mainly of saturated hydrocarbons (even though naphthene-base crudes are predominantly unsaturated) in the form of chain or ring molecules that are chemically inactive and do not have polar heads. These commercial products may or may not contain waxes, volatile compounds, “fixed” oils, and special-purpose additives.

Silicone-based oils have been developed that can be used for somewhat higher temperatures, where mineral oils or fixed oils would break down and degrade.

Fixed oils are usually considered to have greater “oiliness” than mineral oils. Oiliness is a term that describes the relative ability of any lubricant to act as a boundary lubricant. Electron-diffraction experiments have shown that molecules of an effective lubricating agent—a long-chain fatty acid of high molecular weight, such as stearic acid or oleic acid—are attached to a metallic surface by polar bonding. They stand up much like individual strands in a pile carpet. The result is a surface layer with high adhesion and high resistance to contact stress and low resistance to lateral shear along the surface.

Lubricating grease is a semifluid product consisting of a dispersion of a thickening agent in a liquid lubricant. In more practical terms, most greases are stabilized mixtures of mineral oil and metallic soap. The soap is usually a calcium, sodium, or lithium compound and is present in the form of fibers whose size and configuration are characteristic of the metallic radical in the soap compound.

Some greases may contain additives to make them more suitable for specific applications, such as metalworking or threads on fasteners. Some of these materials include graphite, molybdenum disulfide, polytetrafluoroethylene (PTFE or Teflon), copper, nickel, and zinc powders.

Solid lubricants provide certain advantages in high-vacuum, aerospace, or cryogenic applications, where liquids would evaporate or congeal, and at somewhat higher temperatures, where liquids could decompose. Numerous solid inorganic and organic compounds as well as certain composite materials may be classified as solid lubricants. Notable examples are graphite, molybdenum disulfide, and PTFE. Dozens of proprietary compounds may also be classified as solid lubricants.

Although solid lubricants may be applied to achieve design simplification or weight reduction, they usually are adopted because of their good stability, retaining their lubricity at elevated temperatures, in chemically active environments, and under exposure to nuclear radiation. However, there has been some evidence in the literature that molybdenum disulfide can break down in hot, moist environments to release hydrogen sulfide, which can cause hydrogen embrittlement of some steels.

Lubrication Modes. Several basic lubrication modes should be understood in order to analyze and correct wear failures. In all modes, contact surfaces are separated by a lubricating medium, which may be a solid, a semisolid, or a pressurized liquid or gaseous film. In *hydrodynamic lubrication,* the shape and relative motion of the sliding surfaces cause the formation of a fluid film having sufficient pressure to separate the surfaces. In *hydrostatic lubrication,* the lubricant is supplied under sufficient external pressure to separate the opposing surfaces by a fluid film. *Elastohydrodynamic lubrication* is a system in which the friction and film thickness between the two bodies in relative motion are determined by the elastic properties of the bodies, in combination with the viscous properties of the lubricant at the prevailing pressure, temperature, and shear rate. In *dry-film (solid-film) lubrication,* a coating of solid lubricant separates the opposing surfaces and the lubricant itself wears away. *Boundary lubrication* and *thin-film lubrication* are two modes in which friction and wear are affected by properties of the lubricant. In boundary lubrication, each surface is covered by a chemically bonded fluid or semisolid film, which may or may not serve to separate opposing surfaces, and viscosity of the lubricant is not a factor affecting friction and wear. In thin-film lubrication, the lubricant usually is not bonded to the surfaces, it separates opposing surfaces, and lubricant viscosity affects friction and wear.

Lubricant Failures. Devices depending on lubricants to minimize friction and wear can fail when the lubricant fails. Most lubricant failures occur by chemical decomposition, contamination, changes in properties from exposure to elevated temperature, or loss of film. Lubricating oils and greases can fail by any one of the aforementioned processes alone; however, in most situations, chemical decomposition, contamination, and temperature are all involved and are interrelated. In addition, pressure-lubricating systems are susceptible to certain types of failure that result in the inability of the system to provide the required flow rate where it is most needed. Thus, failure by lubricant starvation often results.

Decomposition. When oil is heated in the presence of air, oxidation occurs. Oxidation increases the viscosity and organic-acid concentration of mineral oils, causing varnish and lacquer deposits to form on hot metal surfaces. Under severe conditions, the deposits may be converted to hard, carbonaceous substances. Fixed oils absorb oxygen more readily than do mineral oils, and some may dry, thicken, and form elastic solids. Certain fixed oils (castor, olive, sperm whale, and lard oils) oxidize more slowly than others. These fixed oils are more widely used in blended oils because of their nondrying characteristics. Temperature also affects oxidation rates. In mineral and blended oils, for example, the rate doubles with each 10 °C (18 °F) rise in temperature. Oxidation rates also are higher when the oil is agitated or when catalysts such as copper or acids are present.

In general, solid-film lubricants fail by mechanical removal of microscopically thin layers. Wear debris, which consists primarily of lubricant particles, is generated by the sliding action of a sharp edge against the bonded film on a contact surface. The sharp edge shears a layer of the film (and sometimes the entire film) from the substrate. Contact between a rolling element and a sharp ridge can chip the film, often initiating more extensive failure. This process eventually causes a lack of dynamic stability (as would result from excessive clearance in a bearing) or adhesive wear of metallic contact surfaces.

Many bonded, solid lubricants derive their adhesion from binders, which are incorporated into the film in quantities up to about 20% by volume. Metals, oxides, silicates, or other ceramics are the most common binders. When wear debris contains binder particles, it abrades the remaining film more rapidly than when it consists solely of particles of the lubricating substance.

Contamination of the lubricant with water or chemicals can lead to lubricant decomposition, corrosion of contact surfaces, or both. Contamination with abrasive substances or debris can cause abrasive wear, especially when the size of the contaminant particles is as thick or thicker than the lubricating film.

In internal-combustion engines, water, halide acids, sulfur acids, and products of partial combustion of fuel hydrocarbons are picked up by the lubricating oil. These contaminants can cause various undesirable chemical reactions with oil and metallic surfaces, resulting in the formation of varnish deposits, sludges, viscous emulsions in the oil, or in corrosive wear of engine components.

Because of their chemical nature, fixed oils are particularly susceptible to chemical alteration by alkalis. Alkalis cause *saponification* (formation of soap) by direct chemical reaction with the fatty acids in fixed oils. This reaction alters the nature of the lubricant and, consequently, its lubricating properties.

The viscosity of virtually all oils and greases is affected by temperature and pressure. Almost any change in temperature significantly affects viscosity: the higher the temperature, the lower the viscosity. Increase of pressure increases viscosity, but the effect is seldom significant except at very high pressures.

A decrease in temperature of mineral oil from 100 to 0 °C (212 to 32 °F) can increase viscosity by 100 times or more. The viscosity of fixed oils is also affected by temperature, but to a lesser extent than for mineral oils. For the above decrease in temperature range, lard oil will increase its viscosity by approximately 30 times.

Greases that are solid or semisolid at room temperature gradually soften with increasing temperature and can become fluid and free-flowing at operating temperatures from 95 to 205 °C (200 to 400 °F), depending on the type of grease.

Oils containing substantial quantities of volatile compounds may lose these components by evaporation when operating temperatures are too high. This process not only alters the viscosity, but also upsets the chemical nature of the oil, thus changing other properties. In some instances, volatilization can take place within the lubricating film, as when frictional heat is ineffectively removed by the circulating fluid. Bubble formation within the film reduces the load-carrying capacity of the film, leading to adhesive or abrasive wear or, in severe cases, cavitation pitting of opposing surfaces.

Loss of Film. When boundary lubrication is provided by soft metallic soaps (iron stearate, for example), a rise in surface temperature can result in a marked increase in the coefficient of friction and the wear rate suddenly changes from mild to severe. The temperature at which this change occurs (*transition temperature*) is the point at which the soap desorbs from the metal surface and no longer provides a bonded, continuous surface film. Transition temperatures generally are within the range of 120 to 205 °C (250 to 400 °F), depending on the lubricant and the chemical composition of the metal substrate. Extreme-pressure lubricants function by reaction with the metal surface rather than by adsorption of components in the lubricant. They often are used as substitutes for soap-type boundary lubricants when operating temperatures exceed the transition temperature.

Prevention of Lubricant Failures. Often, lubricant failure can be traced to the selection of an inappropriate lubricant. Petroleum lubricating oils are available in various formulations, and they have various special properties. When these properties cannot be obtained by conventional refining techniques or are obtainable only at a very high cost by refining, they are imparted to the lubricant by additives. Often, certain oil additives impart entirely new performance characteristics to the base oil.

In many instances, lubricant failures (and subsequent component failures) can be prevented or corrected by changes in design of the device or in design of the lubrication system. Starvation of a bearing caused by inadequate lubricant flow or by clogging of oil passages sometimes can be corrected by increasing the size of the passages. Often, an increase or decrease in clearance between sliding surfaces will enable the lubricant to function more effectively. Shields, covers, and seals sometimes can prevent lubricant contamination from external sources. In other instances, filtration or absorption devices can be incorporated into the system to remove unwanted contaminants.

Analysis of Wear Failures

Analysis of wear failures depends greatly on knowledge of the service conditions under which the wear occurred. However, many determining factors are involved, requiring careful examination, both macroscopic and microscopic.

Wear failures generally result from relatively long-time exposure, yet certain information obtained at the time the failure is discovered is useful in establishing the cause. For example, analyzing samples of the environment (especially the lubricant), sludge from the lubricating system, or a used oil filter can reveal the nature and amount of wear debris or abrasive in the system. Therefore, analyses of wear failures involve some techniques in other types of failure. As previously mentioned, it is necessary to consider the complete system involved in the failure. Examination of the worn parts alone can lead to erroneous conclusions and inadequate recommendations.

Background Information

- Determine the nature of the application and the specified manufacturing methods of the parts, including surface treatments or coatings.
- Establish the design and the actual operating conditions: loads, speeds, temperatures, and so forth.

- Find out what lubricant, if any, was specified and how was it applied.
- Establish the ambient environment and service conditions.
- Determine whether the observed wear is normal and typical for the particular application, or if it is abnormal and unique.

Examination

- Define the surface configuration of both the worn surface and the original surface.
- Define the relative motions in the system, including direction and velocity.
- Define the force or pressure between the mating surfaces, or between the worn surface and the wear environment on both the macroscopic and microscopic scales.
- Identify the actual materials in the worn part, the environment, the abrasive, the wear debris, and the lubricant.
- Define the type and effectiveness of any lubricant: oil, grease, surface film, naturally occurring oxide layer, adsorbed film, other, or none.
- Determine the wear rate and coefficient of friction.
- Reach a conclusion as to the wear mechanism, or the combination of mechanisms: abrasive, adhesive, rolling-contact fatigue, and so forth.
- Devise a solution and formulate recommendations.

The following sections discuss some of the details and special methods that are involved in the analysis of wear failures.
Service Conditions. The first step in wear failure analysis is the initial identification of the type of wear or, if more than one type can be recognized, evaluation of the relative importance of each type as quantitatively as possible. This identification of the type or types of wear requires a detailed description of the service conditions based on close observation and adequate experience. A casual and superficial description of service conditions will be of little value.
Descriptions of service conditions often are incomplete, thus imposing a serious handicap on the failure analyst, especially if working in a laboratory remote from the service site. For instance, assume that an analyst must study the problem of a badly seized engine cylinder. Obviously, this is an instance of adhesive metal-to-metal wear or lubricated wear, because use of a suitable engine oil is implied. Furthermore, assume that during an oil change the system had been flushed with a solvent such as kerosene to rinse out the old oil and had been left filled with solvent instead of new oil inadvertently. Also assume that a slow leak, resulting in loss of the solvent, was not detected during the operating period immediately preceding seizure. The analyst probably would receive the damaged parts (cylinder block and pistons) after they had been removed from the engine, cleaned, and packed. If evidence of the substitute “lubricant” could not be established, clear determination of the cause of failure would be extremely difficult or perhaps impossible.
Another example could involve a rolling-element bearing that was running hot because of the lack of an adequate oil supply. The first thing maintenance people will often do when they find a hot bearing is add oil. So, when the bearing fails and is sent to the laboratory, it arrives showing clear signs of overheating, but is covered with oil.
Similarly, incomplete descriptions of service conditions can be misleading in analysis of abrasive wear. For example, in describing the source of abrasion that produces wear of mining and ore-handling equipment, generalized references to the ore, such as copper ore, are common. Such descriptions are too vague to be meaningful; the mineral being extracted usually has little effect on the abrasiveness of the mixture, whereas the bulk rock, or *gangue*, is the principal source of abrasive particles. Unless the gangue minerals are studied both qualitatively and quantitatively, a valid assessment of wear, whether normal or abnormal, is impossible.
In analysis of conventional lubricated wear, detailed description of the lubricant is essential and often must be supplemented by data regarding pressures applied to mating surfaces, operating temperatures, and surface conditions. When corrosion is a factor in lubricated wear, determination of the temperature, degree of aeration, hydrogen-ion concentration, velocity of the lubricant, and the composition and concentration of the corrodent in the lubricant may be difficult. Other analysis complications include the presence of substances that inhibit or accelerate corrosion. In some pumps and compressors, the fluid being processed may leak into the lubrication system, contaminating the oil and causing it to become corrosive. So, it is important in such cases to know the composition of the process stream and to submit a sample of the lubricant for analysis.
Initial Examination and Measurements. Examination of a worn part generally begins with visual observation and measurement of dimensions, usually requiring micrometers, calipers, and standard or special gages. Observations of the amount and character of surface damage often must be made on a microscopic scale. An optical comparator, toolmaker's microscope, recording profilometer, or other fine-scale measuring equipment may be required to assess adequately the amount of damage that has occurred.
Weighing a worn component or assembly and comparing its weight with that of an unused part can help define the amount of material lost. The difference in weight is a measure of the severity of the wear. This material is lost; it is transferred to an opposing surface by adhesive wear or eliminated by abrasive wear, erosion, cavitation, or combinations

of these mechanisms. Weight-loss estimates also help define relative wear rates for two opposing surfaces that may be made of different materials or have been worn by different mechanisms.

Screening of abrasives or wear debris to determine the particle sizes and weight percentage of particles of each size is often helpful. The combination of determination of particle size with chemical analysis of the various screenings can be useful when one component in an abrasive mixture primarily causes wear, or when wear debris and an abrasive coexist in the wear environment. The combination of screening with microscopy often can reveal such details as progressive alteration of the size and shape of abrasive particles with time, as might occur in a ball mill.

Physical measurements can define the amount and location of wear damage, but they seldom provide enough information to establish either the mechanism or the cause of the damage.

Surface damage can range from polished or burnished conditions to removal of a relatively large volume of material. Examination of the worn surface can provide much information, including:

- The amount of material removed
- The type of damage (scratching, gouging, plowing, adhesion, pitting, corrosion, spalling, or simple penetration)
- The existence and character of surface films
- Whether certain constituents are being attacked preferentially
- The direction of relative motion between a worn surface and abrading particles
- Whether abrading particles have become embedded in the surface

Because wear is a surface phenomenon, the original surface configuration of the components in contact influences wear by influencing resistance to relative motion.

Microscopy may be useful to study features of the worn surface including the configuration, distribution, direction of scratches or gouges, and indications of the preferential removal of specific constituents of the microstructure. Abrasive particles or wear debris should be studied under the microscope to observe their shape and the configuration of their edges (sharp or rounded) and to determine if they have fractured during the wear process. These particles can also be chemically analyzed in an SEM using EDS to determine their nature and indicate their source.

Examination of the worn surface by light microscopy at magnifications up to about 100 diameters usually is required to detect uneven or abnormal wear patterns and to reveal the direction of relative movement between the worn surface and the opposing surface or abrasive. Sometimes, higher magnifications are required, and a SEM may be necessary to study areas of slight wear. These techniques are described further in *Friction, Lubrication, and Wear Technology,* Volume 18 of the *ASM Handbook.*

Direct observation at magnifications greater than about 50 diameters can be difficult if the part does not fit in the stage of a metallurgical microscope. Sectioning to remove a portion of the worn surface for direct observation precludes repair and reuse of the part. Some stereobinocular microscopes are capable of such magnifications. Replication is another technique that can be used for light-microscope or SEM observations of worn surfaces of large parts. Replication using plastic films or harder cast materials offers the additional advantage that a reproduction of the surface can be obtained at a remote site and carried back to the laboratory for detailed study.

Changes in surface configuration that occur during the wear process affect subsequent stages of wear. As mentioned earlier, each manufacturing method produces surfaces that are characteristic of that process, whether it is grinding, machining, burnishing, or something else. The process of “wearing in,” which involves progressive reduction in surface roughness by adhesive or abrasive wear of opposing surfaces, generally is followed by a period of relatively little wear. The initial smoothing out of asperities, particularly in lubricated systems that operate under boundary lubrication, reduces the microscopic hills and valleys on the surface to a height about the same as the thickness of the lubricant film. The surfaces then ride on each other with no interference between peaks on the opposing surfaces, and wear essentially ceases. In normal service, then, there will be a characteristic pattern of this break-in wear that will define the expected contact between mating surfaces.

In other instances, particularly if the initial surfaces are somewhat rougher or if boundary lubrication is ineffective, adhesive wear may result in progressive surface roughening and eventual failure. If this process releases wear debris into the joint and the particle size of this debris exceeds the thickness of the lubricating film, combined adhesive and abrasive wear between the opposing surfaces and the wear debris can result in rapid deterioration.

Direction of Relative Motion. When only unidirectional sliding is involved, scratches or gouges produced on the worn surface are aligned with the direction of relative motion. In a sleeve bearing, for example, the scratches should run circumferentially on the inner surface of the bearing and on the mating shaft. Scratches resulting from wear that are oriented in other directions indicate misalignment, vibration, or looseness. These factors can contribute to the severity of the wear.

In devices that undergo combined rolling and sliding, knowledge of the relative velocities and directions of rolling and sliding is necessary for definition of the wear mechanism. The direction of *rolling* is defined as the direction in which the point of contact moves. The direction of rolling is always opposite to the direction of rotation of a rolling element. On a given surface, a condition of positive *sliding* exists if the direction of sliding is the same as the direction of rolling.

Negative sliding occurs on the mating surface, where the directions of rolling and sliding are opposite. Most surface-origin, rolling-contact fatigue failures originate in regions of negative sliding, because the shear stresses are usually more severe there than in regions of positive sliding. Negative sliding occurs on the dedenda of gear teeth, on the cam follower riding on a cam, and in other devices on the part that have the lower surface velocity in a rolling-sliding system.

Metallography. Once the worn surface has been characterized, it is appropriate to prepare metallographic sections through the wear damage and through comparable unworn areas. Because wear is a surface phenomenon, it is almost always necessary to use special techniques to avoid rounding of the edge of the metallographic specimen. One technique is nickel plating on the specimen before it is sectioned and mounted. Additions of powdered glass, hard steel shot, or pelletized alumina, to name a few examples, may be used in the mounting material to improve edge support. The mounted specimen must be polished with care using polishing cloths with little nap, such as silk or nylon, so that the edge is not rounded. For valid analyses of very thin surface layers, techniques such as taper sectioning are needed to allow metallographic observations and microhardness measurements. Taper sectioning involves sectioning the specimen at a shallow angle, and it produces a geometric magnification in the direction of sectioning of surface layers and coatings.

Etchants, in addition to preparing a specimen for the examination of microstructure, also reveal characteristics of the worn surface. Two features that can be revealed by etching a worn surface are phase transformations and metal transfer caused by localized adhesion to an opposing surface and the results of overheating caused by excessive friction. An example of a structure produced by overheating is the “white layer” (untempered martensite) that sometimes develops on steel or cast iron under conditions of heavy sliding contact. Etching a worn surface also can help in detecting the selective removal of specific constituents of the microstructure. It is very important that the worn surface be examined under the microscope and photographed before it is etched, because some topographic features may be easier to observe on the unetched surface.

Metallography will determine whether or not the initial microstructure of the worn part met specification, both in the base material and in any hardened case or surface coatings that may have been applied. It also will reveal existence of localized phase transformation, sheared or cold-worked surface layers, or, as described in the case study on chromium-plated cylinders, the presence of embedded abrasive particles.

Macroscopic and microscopic hardness testing indicate the resistance of a material to several wear mechanisms. Because harder materials are likely to cut or scratch softer materials, comparative hardness of two sliding surfaces may be important. Microhardness measurements on martensitic steels may indicate that frictional heat has overtempered the steel and, when used in conjunction with a tempering curve (a plot of hardness versus tempering temperature), can allow a rough estimate of surface temperature. In some failures, such as in case crushing, a microhardness traverse to verify the depth of the hard case will be needed. Bulk hardness measurements also can indicate whether or not a worn part was heat treated correctly.

Effect of Microstructure and Hardness on Wear. The microstructural heterogeneity of a wear surface influences the wear process, because constituents such as carbides, inclusions, intermetallic compounds, and dispersed phases have properties different from those of the matrix. Hard microconstituents such as carbides make a metal extremely resistant to abrasive wear if they are closely spaced in a relatively hard matrix.

Matrix hardness is important to wear resistance. If hard microconstituents are dispersed widely in a matrix that is not hard enough to have good wear resistance of its own, the matrix may wear away rapidly, leaving the hard particles projecting from the surface, where they can cut into a mating surface. For this reason, under dry-sliding conditions, fine pearlite exhibits considerably better wear resistance than does coarse pearlite or a mixture of ferrite and pearlite.

X-ray and electron diffraction analyses disclose the structure of a crystalline solid. These techniques are particularly valuable for analyzing abrasives, wear debris, or surface films because they can identify compounds, not merely elements. Modern diffraction units, with computerized data collection and analysis using search-match techniques, can help by identifying metals, corrosion products, abrasive minerals, and wear debris. Microstructural features such as retained austenite cannot always be seen by microscopic examination of an etched specimen; quantitative diffraction analysis can reveal the relative amounts of such unresolved constituents in the microstructure.

Chemical Analysis. One or more of the various techniques of chemical analysis—wet chemical analysis, spectroscopy, colorimetry, x-ray fluorescence, atomic absorption, EDS, or electron-beam microprobe analysis—usually is needed for properly analyzing wear failures. The actual compositions of the worn material, the wear debris, the abrasive, and the surface film must be known in order to devise solutions to most wear problems.

Although not so familiar to most failure analysts, a similar number of analytical methods are available for evaluating lubricants. If an ample quantity of lubricant is available, some of the bulk properties, such as viscosity, can be measured. Identification of the oil, its additives, and contaminants can be done by Fourier transform infrared (FTIR) analysis and inductively coupled photospectroscopy (ICP). An analysis of the lubricant can determine if the proper base stock and additives were present, and if any contamination or degradation has occurred in service.

Chemical analysis may be needed to establish or confirm the wear mechanism. The following example describes an instance in which adhesive wear caused by solid-phase welding between mating surfaces was verified by electron-beam microprobe analysis.

Geological studies of soils or similar abrasive mixtures of minerals are mandatory in analyzing wear of tillage tools, earthmoving equipment, and ore-handling devices. The abrasive characteristics of sandy soils are different from those of coarse, rocky soils, clay, or fine silt. Particle hardness and shape (jagged or rounded) are important. The degree of compaction determines the amount of pressure that forces abrasive soil particles against a surface that is being dragged through the mixture. In addition, moisture content is important because it determines, in part, the temperature, lubricity, and cohesiveness of the mixture. Many ores, slags, and similar bulk materials are extremely abrasive to equipment for handling and moving materials; geological analysis can define the abrasive character of these aggregates.

Formulating Conclusions and Report Writing

At a certain stage in every investigation, the evidence revealed by the examinations and tests is analyzed and premises are formulated. Obviously, many investigations will not become a series of clear-cut stages. If the probable cause of failure is apparent early in the examination, the pattern and extent of subsequent investigation will be directed toward confirmation of the probable cause and the elimination of other possibilities. Other investigations will follow a logical series of stages, as outlined in this article, and the findings at each stage will determine the manner in which the investigation proceeds. As new facts modify first impressions, different hypotheses of failure will develop and will be retained or abandoned as dictated by the findings. Where extensive laboratory facilities are available to the investigator, maximum effort will be devoted to amassing the results of mechanical tests, chemical analysis, fractography, and microscopy before the formulation of preliminary conclusions is attempted. Finally, in those investigations in which the cause of failure is particularly elusive, a search through reports of similar instances may be required to suggest possible clues.

Some of the work performed during the course of an investigation may be thought to be unnecessary. It is important, however, to distinguish between work that is unnecessary and that which does not bear fruitful results. During an examination, it is to be expected that some of the work done will not assist directly in determining the cause of failure; nevertheless, negative evidence may be helpful in excluding some potential causes of failure from consideration. Sometimes the evidence that excludes possible failure causes may be stronger than the available evidence supporting the probable cause for a failure.

On the other hand, any tendency to curtail work essential to an investigation should be guarded against. In some instances, it is possible to form an opinion regarding the cause of failure from a single aspect of the investigation, such as visual examination of a fracture surface or examination of a single metallographic specimen. However, before final conclusions are reached, supplementary data confirming the original opinion, if available, should be sought. Total dependence on the conclusions that can be drawn from a single specimen, such as a metallographic section, may be readily challenged unless a history of similar failures can be drawn upon. Yet, circumstances sometimes dictate that a conclusion be formulated on limited data. This is especially true in forensic investigations because of the usually limited extent of the problem and the need to preserve as much evidence as possible. Also, of course, economic limitations are always present. It is rarely justified for the failure analyst to conduct every type of test and investigation possible.

The following checklist, which is in the form of a series of questions, has been proposed as an aid in analyzing the evidence derived from examinations and tests and in formulating conclusions. The questions are also helpful in calling attention to details of the overall investigation that may have been overlooked:

- Has the failure sequence been established?
- If the failure involved cracking or fracture, have the initiation sites been determined?
- Did cracks initiate at the surface or below the surface?
- Was cracking associated with a stress concentration?
- How long was the crack present?
- What was the level or magnitude of the loads?
- What was the type of loading: static, cyclic, or intermittent?
- Is the fracture surface consistent with the type of loading assumed in the design?
- What was the failure mechanism?

- What was the approximate service temperature at the time of failure?
- Did high or low temperature contribute to failure? Were there temperature excursions?
- Did wear contribute to failure?
- Did corrosion contribute to failure? What was the type of corrosion?
- Was the proper material used? Is a better material required?
- Was the cross section adequate for the type of service? Were the stresses too high?
- Was the quality of the material acceptable in accordance with specifications?
- Were the mechanical properties of the material acceptable in accordance with specifications?
- Was the component that failed properly heat treated?
- Was the component that failed properly fabricated?
- Was the component properly assembled or installed?
- Was the component repaired during service and, if so, was the repair correctly performed?
- Was the component properly “run in”?
- Was the component properly maintained and properly lubricated?
- Was failure related to abuse in service?
- Can the design of the component be improved to prevent similar failures?
- Was the failure primary, or was it damaged by failure of another part?
- Are failures likely to occur in similar components now in service, and what can be done to prevent their failure?

In general, the answers to these questions will be derived from a combination of records and the examinations and tests previously outlined in this article. However, the cause or causes of failure cannot always be determined with certainty. In this instance, the investigation should determine the most probable cause or causes of failure, distinguishing findings based on demonstrated facts from conclusions based on conjecture.
In most cases, the conclusion should be on the first page, closely followed by recommendations.
Failure Analysis Report Preparation and Writing. A failure analysis report is the summation of all of the individual tests and analyses performed during the course of an investigation. The report coalesces the various results and presents them in a concise and logical format. It should be written to describe the particulars of the subject being studied, convey pertinent information gathered from testing and analysis, and logically and accurately interpret and explain data and results. A report allows the reader to follow the investigator's thought processes during the analysis and documents and explains the course the investigation followed. The author may be the sole investigator, a coinvestigator, a contributor, or other person designated as the primary point of contact. Basically, the body of the report tells the story of the investigation, explaining why certain tasks were performed, in what order, what the results were and their interpretations, and answers the questions originally posed by the customer at the onset of the study. Brevity is important, but not at the expense of completeness and clarity. Reports should be written with the target audience in mind. For example, are the intended readers technically oriented? A report intended for engineers and scientists would likely differ from a report submitted to marketing or finance personnel. Frequently, readers will scan only one or two sections of the report, usually either the abstract, summary, or conclusions section, prior to deciding whether to read the entire report. Consequently, these sections must adequately convey the scope and findings of the study.
Adequate preparation is required prior to actually writing a report. An author must compile all of the information generated during an investigation and review the data. Many investigators find it is helpful to assemble all pertinent data, such as photographs, spectra, and mechanical and physical test results, and lay the information out in front of them sequentially, corresponding to the order the tests were actually performed. Glean the useful data from the redundant or superfluous. Next, review some basic questions to decide exactly what information the report should convey. What did the customer request? Frequently customers ask for something, but it is the responsibility of the investigator to discern what they really want. Which tests and analyses were performed? Constraints may be placed upon the investigation—generally, time or cost. If an investigation is requested, but limited by deadlines or funding, the analyst must decide which tests and procedures will furnish the most useful information in the most timely fashion without exceeding any preestablished budget. If certain tests or analyses were excluded, whether for time, budgetary, or other reasons, it is often beneficial to explain why they were not performed. Certain failures require minimal amounts of testing and analysis; others involve multiple disciplines and extensive investigations. A series of questions,

posed prior to writing the final report, help to ensure that the analysis adequately addresses the most important considerations, regardless of whether the investigation is large scale or routine (Ref 2):

- Has all relevant background information been collected?
- What was the most likely sequence of failure?
- For failures involving cracking or fracture, was there a single initiation site, or multiple sites?
- Was crack initiation surface or subsurface?
- Was a stress concentrator a factor in cracking?
- Was the cracking fresh? If not, how long was the crack extant.
- What was the load intensity?
- Did cyclic, intermittent, or static loading contribute?
- What was the stress orientation?
- Was the failure temperature related? If so, what temperatures did the component encounter?
- Was the failure wear related?
- Did corrosion contribute to the failure? If so, what type of corrosion?
- Was an appropriate material used? Should a more suitable material be used? Was it within specification? It is worth remembering that a specified material may still be inadequate.
- Were the mechanical and physical properties of the material adequate? Within specification?
- Was the component properly heat treated? fabricated? assembled and/or installed? run/broken in? lubricated? maintained?
- Was the component ever repaired? If so, was the repair adequate?
- Did in-service abuse contribute to the failure?
- Would improving component design preclude similar failures?
- How can future failures of comparable components be avoided?

It is preferable for the report to list work chronologically, as the tasks were performed. The flow of the report should be from nondestructive to destructive. The results must satisfy the customer's original requirements. Were standard procedures followed and documented? Individual companies or organizations usually have guidelines concerning the format and content of technical reports, required signatures, distribution, and so forth. Likewise, legal considerations, such as what constitutes a business record and the length of time records and supporting data are retained, should adhere to company policies.

Technical reports are not prose, but the author can still imbue singularity into the work, while maintaining professionalism and cogency. Although there is no single universally accepted report format, most share common features and are tailored to satisfy the needs of the customer and employing organization. For instance, technical reports have historically been written in the past tense from a third-person perspective. However, more authors are now using first person to describe their analyses. They believe first person instills the reader with a better understanding of the flow of the investigation and helps them understand the thought processes behind the analyses performed. Many experienced analysts prefer to put their conclusions at the beginning of a report, so that the reader understands the pertinence of the information reported. Photographs, figures, charts, and so forth should augment and enhance the textual narrative of the report, not replace it; many experts feel that imbedding such graphics in the body of the report, rather than at the end, tends to help the flow of the report. Company procedures and practices may address such issues as voice and format. If no formal company policy exists, there are many organizations that provide standard formats and guidelines in a variety of media. The American National Standards Institute (ANSI) can provide hardcopy references and standards for guidance (Ref 3). Likewise, the Internet World Wide Web can provide excellent on-line reference sites (Ref 4, 5).

Elements of the Report. Traditionally, failure analysis reports contain, as a minimum:

- An introduction that includes background information on the subject of the report and restates the work requested and may serve as a written record for verbal work requests
- A section detailing investigative procedures and results
- Conclusions

Lengthier reports can comprise:

- A heading
- An abstract
- A table of contents
- An introduction
- Sections devoted to investigative procedures, results, discussions, conclusions, and recommendations
- Signature blocks
- Acknowledgments
- References
- Appendices

Most reports begin with a *heading.* The heading may include the company name, department, author, date, report number, and title of the report. The title should be in title case and be as descriptive as possible without being too verbose.

An *abstract* is useful to succinctly describe the scope and objectives of the investigation and the findings of the analysis. All information, descriptions, and conclusions stated in an abstract should be expounded upon in the body of the paper. An abstract should be brief, generally between 150 and 500 words. The abstract should not be a complete summary of the paper; instead, it should convey only the salient points of the report (Ref 3).

A *table of contents* should be included if the report contains many pages, typically ten or more. A table of contents sequentially lists the sections of the report and their respective page numbers.

The *background,* or *foreword,* section of the report supplies the reader with pertinent information and helps serve as an introduction. This section of the report should contain as much useful information as possible, including:

- Nomenclature of the parts involved
- Description of the work requested
- Descriptions of the events surrounding the subject failure
- Related documentation such as specifications, standards, drawings
- Manufacturing and service histories
- Applicable part numbers, materials, loads, and so forth

An *investigative procedures* section details the tests and methods utilized during an analysis. The progression of tests should follow the investigator's test plan. Typically, testing is performed on components sequentially, from nondestructive to destructive. Destructive testing, by its very nature, destroys potential evidence. If litigation is involved, or potentially involved, it is necessary to obtain agreement between all parties as to the scope of the work, and also then to obtain a court order to allow the destructive testing to be performed. This should be addressed and documented. A detailed as-received description is beneficial, complemented with photodocumentation. Photos, figures, charts, and similar visuals should be captioned with an explanatory statement, including applicable scales, units, and so forth. Any cleaning and preparation should be noted. The macroscopic features can be described, followed by the microscopic features. Microscopy generally progresses from optical to electronic. Metallographic and chemical analysis methods used during the investigation should be detailed. Equipment utilized to perform the various tests can be stated in this section. Physical and mechanical testing performed should be documented. Mechanical testing can include hardness and tensile testing. Physical testing can include density and electrical conductivity.

The *results* section of the paper reports the pertinent data and information obtained from the various tests performed during the investigation. The data are subsequently used in the discussion and conclusions sections; therefore, extemporaneous data should not be included. The results should be presented using descriptions, tables, figures, charts, or other graphics. Mathematical and computational procedures and formulaic manipulations and equations should be described textually. If multiple equations are used, each should be numbered sequentially for reference. Testing and data accuracy should be addressed. The appropriate appendix should be referenced concerning error analysis and computational and analytical details.

The *discussion* section of the report describes the interpretation and analysis of the findings obtained from the various tests. The most important results should be recapped. Relationships between various properties, facts, and findings should be explained. The characteristics of the subject specimens can be compared to those of laboratory-manufactured exemplars; this helps gage the accuracy of the tests and provides valuable reference

information. Plausible explanations relating to failure modes and causes should be addressed. The discussion should proceed from specific to general. Any variations in data should be explained. Possibility versus probability of various scenarios should be considered and discussed.

The *conclusions* section of a technical report should summarize the most significant results and findings of the investigation and present the most cogent and plausible explanations. It may be necessary to reiterate information discussed in prior sections to bolster any conclusions or theories presented. Conclusions are generally based on facts, but often informed speculation and opinion are required to satisfy customer requirements. State if a conclusion is based on supposition, and describe supporting factual data. As stated previously, frequently the abstract and conclusions sections are the only portions of a report that are read, so they need to be especially concise, accurate, and well written.

The three main outcomes of a failure analysis are: determining how the subject component failed (e.g., overload), establishing why the component failed (e.g., overtorque), and making *recommendations* to prevent future failures, when applicable and appropriate.

Recommendations should address, and rectify, the deleterious factors that contributed to the failure. It is worth noting that sometimes recommendations may be costly to implement or unpopular with others in an organization. However, it is essential that the investigator preclude failures, thereby increasing safety, reducing the likelihood of injury, and increasing service life of the given component.

A *signature block* is normally required for technical reports. Local company policies generally identify necessary signatories. Signatories' titles and contact information should be included, immediately below the appropriate signature. As a minimum, the primary investigator/author should sign the report. Any concurrence or editorial signatures should follow the primary signature.

An *acknowledgments,* or *contributors,* section is frequently included in technical papers. Coinvestigators, colleagues, technicians, analysts, and so forth who made contributions during an investigation can be acknowledged for their efforts.

Any work from other authors cited in a paper should be referenced. There are two major formats for *citations,* or *referencing work:* the number-identification system and the author-date system. The former ascribes specific numbers to the references, corresponding to entries in a successively numbered list of references. The latter parenthetically cites an author's name and publication dates in the body of the text and corresponds to an entry in an alphabetically arranged reference list. Generally, the order of information is: the author's name, the article title, and publication information. Specific formats and examples are discussed at length in various sources, such as the *Modern Language Association (MLA) Handbook*, 5th edition (1999) and *Chicago Manual of Style,* 14th edition (1993). Since the advent of the Internet, on-line sources are becoming more common. Although no standards have as yet been adopted, several present similar formats (Ref 6). Most writing style guides and standards organizations generally agree that an Internet reference should include the authors, a descriptive document title, date, and uniform resource locators (URLs).

Appendices should contain computational, factual, and analytical details referenced in the body of the report. Error calculations should be placed in the appendix. Multiple appendices should be designated by sequential letters (Appendix A, Appendix B, etc.). The appendices should contain sufficient information so that the reader can understand and evaluate the results without needing to refer to previous publications.

A *glossary of terms* used in the report can follow the appendix. Terms and their respective definitions should be in alphabetical order, each on a separate line.

A final review and edit of the report is always beneficial. Useful considerations when reviewing a report include:

- Alternative viewpoints and ideas can be addressed if additional people review a report. This must be balanced against the time involved during the editing cycle.
- Always perform spelling and grammar checks.
- Is all the information contained in the report pertinent to the subject investigation?
- Are the results and conclusions logical and accurate?
- Have the failure mode and cause been ascertained?
- Does the report have a logical flow and continuity of results?
- Are the recommendations plausible?
- Were all the original questions answered?
- Were there any contradictory results? If so, retest or explain.

- Were proper procedures followed and documented?
- Was the test plan followed as designed, or was divergence necessary?
- Were all graphics properly captioned? Were scales included with all photodocumentation?
- Were units used consistently?
- Was the equipment used in the investigation documented and in calibration?
- Is all the information in the report suitable for distribution? Were concerns regarding classified, restricted, or proprietary information addressed?

The Audience. Obviously, not every report will require coverage under each of these sections. Lengthy reports should begin with an abstract or executive summary. Because readers of failure analysis reports are often people in purchasing, operations, accounting, management, and even legal staff, the avoidance of technical jargon wherever possible is highly desirable. A glossary of terms may also be helpful. The use of appendices, containing detailed calculations, equations, and tables of chemical and metallurgical data can serve to keep the body of the report clear and uncluttered. The use of literature or book references is recommended along with references to pertinent standards such as ASTM, and so forth.

Customer needs vary. Some may request a formal report, others an abbreviated report or executive summary. Not all reports require each section previously mentioned. In certain instances, sections of the report can be combined for better flow and continuity. For example, the results and discussion sections, or the procedures and results portions, can be combined into one section. An internal company report may not require the detail that a potentially litigious investigation might merit. It is worth remembering, however, that even the most innocuous report may potentially be used in litigation. Therefore, semantics and nomenclature should be chosen carefully. Potentially inflammatory or insensitive wording should be avoided. As an example, the use of the word “flaw” versus “defect.” Not every flaw is a defect, but all defects are flaws (Ref 7).

Consult Other Disciplines. When available, take advantage of the resources of other knowledge and experience available to you. This can be someone in your company or even a vendor. An interdisciplinary approach to complex failure analyses of large structures or machines is often warranted. Activities such as nondestructive testing and FEA are usually subject areas in which the failure analyst is not proficient. Working with engineers and specialists in other disciplines can be required to reach the root cause. Most specialty support such as nondestructive testing or FEA analysis can be purchased as a service. However, simple collection of results without a meaningful dialogue of the engineering and metallurgical variables at play in a particular failure may lead to poor results and improper interpretation. Developing a team to solve complex problems can ensure the best possible outcome when the failure requires a broad analysis.

Follow-Up on Recommendations. The purpose of most industrial failure analysis is to determine the basic or root cause of failure of metal parts. Failures are, at best, a source of irritation and, at worst, a safety hazard, but they are always an economic loss. Therefore, study of the failure should result in carefully formulated recommendations in the report that are aimed at reducing or eliminating similar failures in the future. Such recommendations may involve adding warning labels, changes in design, metallurgy, manufacture, quality control, maintenance, repair practices, and anticipated usage of the product.

If recommendations are made in the report, they should be reviewed with appropriate personnel after the report has been issued so that the recommendations are not overlooked. There may be valid reasons for not implementing the recommendations, but at least those performing the failure analysis and writing the report will have fulfilled their responsibility.

References cited in this section

2. G.F. Vander Voort, Conducting the Failure Examination, *Met. Eng. Quart.*, May 1975

3. Scientific and Technical Reports: Elements, Organization, and Design,” ANSI/NISO Z39.18-1995, American National Standards Institute/National Information Standards Organization, 1995

4. S. Butts, “Technical Report Writing,” National Aeronautics and Space Administration (NASA), Glenn Research Center, 1 Oct 2001, http://ltid.grc.nasa.gov

5. B. St. Martins, “Using MLA Style to Cite and Document Sources,” 16 Oct 2001, http://www.bedfordstmartins.com/online/cite5.html#1

6. “Information and Documentation—Bibliographic References—Part 2: Electronic Documents or Parts Thereof,” ISO 690-2, International Organization for Standardization, 1995

7. R.K. McLeod, T. Heaslip, and M. Vermij, Defect or Flaw-Legal Implications, *Failure Analyses: Techniques and Applications*, J.J. Dickson, Ed., ASM International, 1992, p 253–261

Key Guidelines

The following list summarizes key guidelines for conducting a failure analysis investigation:

- There are 14 possible steps to follow in conducting a failure analysis—from collecting and examining samples to testing, determining the failure mode, writing the report, and developing follow-up recommendations.
- In working to produce economical, reliable products, it is vital for manufacturers to recognize the importance of analyzing failures that occur during product development testing or in service.
- Of primary importance in conducting a thorough failure analysis is the compilation of information regarding the manufacturing, processing, and service history of the failed part. This includes obtaining original specifications and drawings, if available.
- It is important for failure analysts to keep written and visual records of their investigations.
- An essential first step in any failure analysis is a thorough visual examination including consideration of fluids, soil, debris, paint, or corrosion found on the surface of the failed part.
- Nondestructive testing methods (magnetic-particle, liquid-penetrant, ultrasonic, radiography, and eddy-current testing) as well as mechanical tests (hardness, tensile, and impact, can be helpful in determining the cause of a failure.
- The influence of the chemical and thermal environment must be kept in mind during any failure analysis. Corrosion and property changes may be critical in an analysis.
- Macroscopic examination of the fracture surfaces can be considered the most important part of a failure analysis. It is typically performed at magnifications of 1 to 100× using the unaided eye, a hand lens, a low-power stereoscopic microscope, or SEM.
- Performing both macro- and microscale examination of unetched and etched specimens provides important evidence about the type and structure of the material involved in the failure.
- Brittle fractures are typically more difficult to analyze because of the large number of possible fracture mechanisms and the lack of obvious deformation.
- In any failure analysis, all avenues must be explored to ensure that the investigator's analysis is complete and that the final conclusions are well documented and accurate.

Acknowledgments

This article is based on contents from the following lessons in the course *Principles of Failure Analysis* (Materials Engineering Institute, ASM International):

- D. McGarry, “General Procedures for Failure Analysis”
- R. Parrington, “Corrosion Failures”

- G. Powell, “Wear Failures”

Contents on reporting writing were contributed by Steve McDanels, National Aeronautics and Space Administration. Contents on mechanical testing were provided by Daryll McKinley, McSwain Engineering.

References

1. P.M. Mumford, Test Methodology and Data Analysis, *Tensile Testing*, P. Han, Ed., ASM International, 1992, p 55

2. G.F. Vander Voort, Conducting the Failure Examination, *Met. Eng. Quart.*, May 1975

3. Scientific and Technical Reports: Elements, Organization, and Design,” ANSI/NISO Z39.18-1995, American National Standards Institute/National Information Standards Organization, 1995

4. S. Butts, “Technical Report Writing,” National Aeronautics and Space Administration (NASA), Glenn Research Center, 1 Oct 2001, • http://ltid.grc.nasa.gov•

5. B. St. Martins, “Using MLA Style to Cite and Document Sources,” 16 Oct 2001, http://www.bedfordstmartins.com/online/cite5.html#1

6. “Information and Documentation—Bibliographic References—Part 2: Electronic Documents or Parts Thereof,” ISO 690-2, International Organization for Standardization, 1995

7. R.K. McLeod, T. Heaslip, and M. Vermij, Defect or Flaw-Legal Implications, *Failure Analyses: Techniques and Applications*, J.J. Dickson, Ed., ASM International, 1992, p 253–261

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Introduction

FAILURE ANALYSIS is an investigative process in which visual observations are made and conclusions are drawn based on physical evidence and features present on a failed component. Recording the features observed in a failure analysis is critical both to document the component condition and to communicate the results of the failure analysis to others.

The primary means of documenting the conditions and features observed during a failure analysis investigation is photography. Failure analysis photography is a combination of science and art. Proper imaging techniques are required to produce an acceptable fracture surface photograph. Experience is required to produce a visually acceptable fracture surface photograph. The use of proper equipment, proper technique, and the study of the resulting photographs can lead to exceptional results. This article reviews photographic principles as applied to failure analysis and specific techniques employed in both the field and in the laboratory.

Visual Examination

Visual Observations. Failure analysis is a process centered in visual examination and analysis. The details observed on a fracture surface are based on differences in fracture surface topography, reflectivity, and color characteristics. Observation of these factors requires the failure analyst to use proper illumination techniques and suitable magnification to resolve and distinguish the fracture features present. During the failure analysis process, features on a microscopic scale must be imaged and interpreted; this interpretation needs to be considered in conjunction with macroscopic features, the gross condition of a component, and the environment in which the component was operated to determine the mode and cause of failure.

Fracture Feature Notation. Prior to beginning photography of a failed component, the component should be examined thoroughly to determine its condition; to evaluate which features are important and which features are extraneous, that is, postfracture mechanical damage; and to determine if special treatment or cleaning of the failed component is required. This process typically involves examination with the unaided eye and proceeds to higher magnification, usually by stereomicroscopic examination. A handheld magnifier may also be of assistance at this stage of the investigation. Illumination at varying degrees of obliqueness and orientation can be helpful in evaluating fracture conditions present.

The failed component and any fractured and separated pieces should then be photographed. At this stage, it is important to document how the fracture is related to the remainder of the component as well as to document the relative size of the component, using suitable reference scales in the photograph.

Photographic Documentation Strategy. The failure analysis process, dependent on visual examination, is also dependent on the recording of features observed at each step of the investigation. The failure analysis process may change the appearance of the fractured component by cleaning or sectioning for scanning electron microscopy examination and/or metallography. It is important to photograph the component at each stage of the process to retain the visual information available and the component condition at each step.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Field Photographic Documentation

Composition. Photographic documentation of failed components in the field requires photography of the failed component, the assembly, and the entire structure. Documentation should include the orientation and connection of various components and should include multiple views of each component. Field photographic documentation is used in the failure analysis laboratory for component and fracture orientation, feature reference, and condition assessment. Compositions should avoid extraneous objects in the image that would excessively complicate the photograph. Where possible, the use of multiple photographs taken at closing distances is beneficial in referencing component placement and orientation within an assembly during later parts of the investigation. Photographic lighting in the field should avoid harsh shadows that result in excessive contrast. Harsh shadows can be minimized by use of fill-flash.

Component Identification. It is important during the field photo documentation process to make sketches and record the name and description of components to assist in the laboratory failure analysis process as well as in report preparation and other communications. It is helpful to have component manuals present during the field documentation process, so that proper terminology can be used for the components being examined.

Techniques of instant photography may be beneficial in documenting components in the field and in recording component nomenclature on the photograph.

Field Photographic Documentation Specifications. The American Society for Testing and Materials (ASTM) provides guidelines for photographic documentation of components involved in a failure analysis that may become part of litigation. ASTM E 1188, "Standard Practice for Collection and Preservation of Information and Physical Items by a Technical Investigator," states that photographic documentation should begin as soon as possible after the incident. If items involved in the incident are disassembled or subjected to destructive testing, each step of the disassembly or testing shall be documented by contemporaneous photographs or videotaping. The failure analyst should constantly keep in mind the fact that the purpose of photography is to record the condition of a failed component at any given period of time for transfer of information in the future. The specific photographic documentation may be the only record of the condition of the component available subsequent to the field examination.

Other ASTM standards providing guidance for failure analysis work and failure analysis documentation include:

- ASTM E 678, "Standard Practice for Evaluation of Technical Data"
- ASTM E 860, "Standard Practice for Examining and Testing Items That Are or May Become Involved in Litigation"
- ASTM E 1020, "Standard Practice for Reporting Incidents"

Numerous photographic texts also are available to assist the failure analysis photographer. Some of these are cited in the Selected References listed in this article.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Laboratory Photographic Documentation

As-Received Condition. It is important to document the condition of the failed component on receipt in the failure analysis laboratory. This documentation is consistent with the guidelines given in ASTM E 860, "Examining and Testing Items That Are or May Become Involved in Litigation," as well as proper failure analysis technique. The as-received photographic documentation should accurately depict the condition of the failed component, including the fracture surface condition as well as the overall condition of the component. Accurate color in laboratory photographic documentation is critical, and periodic use of photographic color standards assists in this task.

The as-received photographic process should begin with direct lighting and a scale or ruler in the photograph for reference. After direct lighting, varying degrees of oblique lighting can be used to highlight the topographic features present on the fracture surfaces of failed components. To the extent possible, confusing backgrounds should be eliminated through the use of background paper or material. Stereophotography is extremely useful at this stage of the failure analysis process to record component configuration and spatial characteristics.

The overall component photodocumentation is followed by macrophotography of the component and fracture details, stereomicroscopic photography of the component and fracture details, and scanning electron microscopy of the component and fracture details.

Further analysis may require component sectioning and metallography. Optical metallographic photography is used to document the microstructure and internal structural condition of a failed component.

Component failure simulation testing, if performed, should be photographically documented in a manner similar to the documentation of the failed component condition, using still photography and videodocumentation. In some cases, high-speed videodocumentation may be used in the failure simulation testing.

Laboratory Component Lighting. Most failure analysis laboratory photographic documentation is performed using flash photography. Flash photography ensures that the component is adequately lighted and that the

colors of the component, including any degradation, are accurately recorded. Color-balanced flood-lights can also be used for laboratory photodocumentation. Color photography reveals heating effects, corrosion effects, mechanical damage, coating condition, and other types of component degradation not readily distinguishable using black-and-white photography. Floodlighting may also be used in laboratory photographic documentation to ensure that fracture features are properly highlighted and that the fracture condition is accurately recorded. Side lighting of fracture surface features is especially useful in revealing fracture surface texture that contains information related to the failure mode. In those cases, side lighting of the fracture enhances the topographical features present on the fracture surface and provides a means of communicating the fracture features to others.

For relatively small components, the photographic copystand provides an excellent means for documenting component and fracture surface condition by varying the angle and direction of the incident lighting. The copystand should be rigid to avoid vibration when long exposures are used for close-up photography. The copystand allows lighting from single or dual sides of the component. Orientation of the component with respect to the incident lighting highlights different characteristic features on the fracture surface. Highly reflective components can create glare or reflected images of the camera or lights. Special techniques can be employed to deal with highly reflective conditions and are described later in this article.

Magnification Determination. The relative size of a failed component and the magnification of the photograph are important to incorporate in the laboratory failure analysis process. This is most easily accomplished by placing a scale adjacent to the failed component and imaging the scale with the failed component, while keeping the lens-to-film plane constant. The photographic magnification then becomes the ratio of the actual scale size to the size of the scale in the printed image. Printing of the failure analysis photograph to any enlargement size allows the magnification of the enlargement to be determined by measuring the ratio of the actual scale size to the enlarged scale size.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Photographic Equipment

Camera Bodies. The most common camera body used in failure analysis photography is the single-lens reflex 35 mm camera. The 35 mm failure analysis camera may have either manual focus and manual exposure or automatic focus and automatic exposure control. The automatic camera body that allows both aperture and shutter priority control is preferable for field and laboratory failure analysis work. The camera body that allows interchangeable lenses is most suitable for failure analysis work, because of the need to perform close-up photography as well as overall, wide-angle, and telephoto-range photography.

Camera Lenses. The most versatile lens used for failure analysis photography is the zoom lens, with a focal-length range of approximately 28 to 120 mm. A typical zoom lens useful for failure analysis photography is shown in Fig. 1. Such a range allows the formatting of the image, so that the primary object in the field of view is the failed component or assembly. Extraneous features in the photograph can be detracting and add difficulty in communicating failure analysis information. The other lens that finds particular use in failure analysis photography is the macro- or close-up lens. A typical macrolens is shown in Fig. 2. Close-up photography is often required in failure analysis to document the detailed features present on a fracture surface. In order to accomplish this, either a macrolens in the 50 or 100 mm focal-length range or a zoom lens with a macrocapability is useful. Adapter rings and extension tubes can also be used to adjust the focal length and provide an increase in magnification. The higher focal lengths have the disadvantage of reduction of depth of field. Either type of macrolens may use manual or automatic focus. In many conditions, the automatic lens is switched to the manual mode in order to achieve focus at very close focal distances.

Fig. 1 Typical zoom lens used for overall failure analysis photography

Fig. 2 Typical macrolens used for close-up failure analysis photography

Camera Flashes. The primary flash used for failure analysis photography must be capable of being mounted on the camera as a “hot shoe” flash or on an extension cord for side lighting. A typical flash arrangement with an off-camera extension cord is shown in Fig. 3. It is preferable that flash photography be performed in a through-the-lens metering mode to achieve optimal exposure. For most failure analyses, use of the off-camera flash in a side-lighting condition provides beneficial shadowing that brings out fracture detail and makes interpretation of fracture features easier. Another flash useful for failure analysis photography is the ring flash, which provides for optimal flash lighting and exposure in macrophotography. A typical ring flash arrangement is shown in Fig. 4. The ring flash also allows selection of flash from a single side of the lens or from above or below the lens. Ring flash lighting is especially useful in recording fracture surface topographic features and surface markings, such as part numbers.

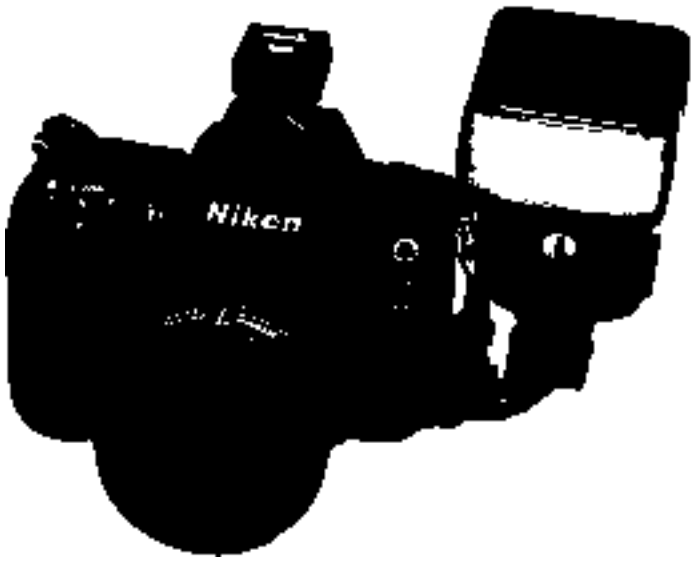


Fig. 3 Typical off-camera flash arrangement for failure analysis photography

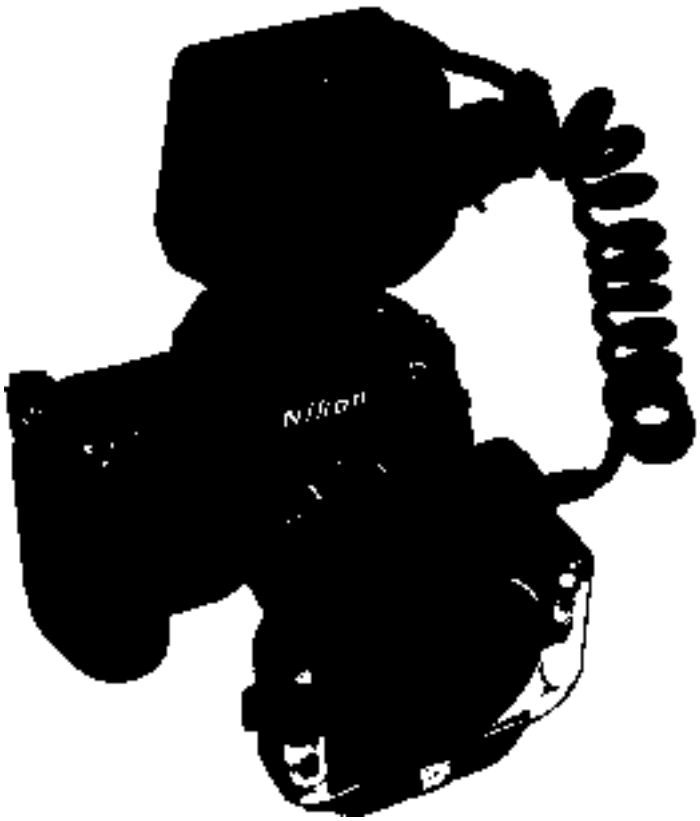


Fig. 4 Typical ring flash arrangement for failure analysis photography

Camera Accessories. The primary accessory important to the failure analyst is a suitable carrying case for camera equipment. A typical case for field use is shown in Fig. 5. Such a case should provide protection for the camera equipment being used in both field and laboratory conditions. Many times, the failure analyst travels to other laboratories or the field, and the camera case must adequately protect the camera equipment in those conditions. Camera cases may be soft-sided, plastic, or metal. The case should be equipped with adequate padding to separate camera components and to ensure cameras are free of damage during transit.

Fig. 5 Typical carrying case for field and laboratory photographic equipment

Film Photography

Film Imaging. The primary photographic recording media used in failure analysis is 35 mm film. The 35 mm format, coupled with high-quality film, produces excellent enlargements to poster-sized prints. Film is available as color negatives, color transparencies, and as black-and-white negatives. The most common film used in failure analysis is the color negative, which results in a color print. Fine-grain film is preferable in failure analysis because of the ability to make enlargements. Color prints are able to precisely record the conditions observed by the failure analyst and to provide a means of communicating those observations to others.

Film speed is important to the failure analyst because of the nature of the film itself. High-speed films typically have coarser grain and are less capable of producing high-resolution enlargements. Color print films with speeds in the range of 100 to 200 American Standards Association (ASA) are optimal for the failure analysis application. They offer both reasonable shutter speeds and have sufficiently fine grain to produce enlargements up to 75 cm (30 in.) per side.

Film Emulsion. Certain films also incorporate an extra layer of emulsion for recording light-colored detail, such as in portrait photography. These films are especially useful in failure analysis because of their ability to capture detail in light-colored, highly reflective areas often encountered in fracture surface photography.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Digital Photography

Digital Imaging. Scientific digital photography is similar in many ways to color film photography. In film photography, the image is recorded on silver halide crystals in the film emulsion. In digital photography, the image is recorded on a microelectronic sensor made up of an array of light-sensitive elements similar to the silver halide crystals in a film emulsion. The captured image is stored in a removable media that is subsequently used to transfer the image file to a computer or printer for printing.

Digital photography provides immediate availability of images viewable on the camera imaging screen or on a computer screen. Inadequacies in illumination or exposure can be quickly detected and corrected. It is especially useful in documenting stages of a failure analysis where cleaning or sectioning is involved. Each step of the process can be documented with digital photography, and the image captured at each stage can be verified.

Image Size/Resolution. The resolution in a digital image is determined by the number of pixels in the imaging sensor; the higher the number of pixels, the greater the image resolution. High-resolution photographs are attainable with multiple megapixel sensor arrays. Image file sizes in the multiple megabyte range produce excellent photographs with enlargement capability.

File Formats. The two primary digital camera file formats used in failure analysis photography are the TIFF and JPEG file structures. The TIFF and bitmap file formats are uncompressed and provide maximum resolution. The TIFF format is the most commonly used uncompressed format for scientific photography. The JPEG is a compressed file storage format with compression that typically results in a reduced file size with no discernable difference in image characteristics or quality. Care should be taken to avoid multiple JPEG compression steps or high-compression formats that can result in loss of image quality.

Digital Cameras and Devices. Digital cameras have features ideal for failure analysis work. Those include a liquid crystal diode display screen for display of the image recorded, removable storage media, zoom lenses, and macro-focus capability. Such a camera offers all of the standard features found on a professional 35 mm single-lens reflex film camera. It has the same flexibility of interchangeable lenses and off-camera flash. The digital camera has the added advantage of the immediate viewing of the captured image and the rapid production of high-quality prints.

Lenses for Digital Cameras. The professional digital camera body accepts the same lenses as used on professional single-lens reflex film cameras. Typically, the zoom lens is beneficial for recording the condition of the component, and a macro-lens is useful for recording up-close detail.

Digital Storage Media. The digital storage media used in professional digital cameras has several formats, depending on the media selected by the camera manufacturer. Digital media has sufficient storage capability to record the equivalent of multiple rolls of photographic film images. However, care must be taken to not bring the digital media into close proximity to any magnetic device, and it must be protected from physical damage, especially to the electrical contacts.

Digital scanners are useful in failure analysis for introduction of photographs into reports and for electronic transmission of photographs obtained during failure analysis. Professional scanners are preferred because of their increased resolution and image quality. Image resolution should be carefully selected to avoid excessive file sizes that are unreasonable for storage, reproduction, and electronic transmission. An image resolution of 200 to 400 dots/in. is suitable for failure analysis photograph scanning, and both film and digital photographs can be scanned with excellent results.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Basics of Photography

Aperture selection in failure analysis photography is one of the most critical aspects of the photographic process, because the aperture setting directly controls the depth of field as well as the exposure time. A large aperture opening (small aperture number) results in a shallow depth of field. In those conditions, it is possible that the entirety of the fracture surface may not be in focus. A small aperture opening (high aperture number) results in a large depth of field, which is beneficial in documenting the overall failed component condition. However, the small aperture has limited light transmission and may result in very long exposure times. Long exposure times can lead to loss of sharpness if the camera or support has any vibration present. The typical aperture setting for most failure analysis photography is F8 to F11. Higher aperture settings may be used in conditions where the camera can be mounted either on a copystand or tripod and floodlighting can be used. In

those situations, the aperture setting can go to F32 and higher without loss of image sharpness. For automatic cameras, the aperture priority setting is preferred in failure analysis, because of the need to ensure that the entirety of the fracture surface is in focus in the photograph.

Shutter Speed Selection. The shutter speed is critical in creating sharp photographs. The highest shutter speed possible is preferred, because it minimizes the possibility of image blur from camera movement during exposure. A minimum shutter speed of $\frac{1}{100}$ s produces adequately sharp photographs. If a suitable shutter speed cannot be attained by hand-holding the camera, a flash should be used. Most flash units synchronize to the camera shutter speed of $\frac{1}{60}$ s, but because of the short flash duration, flash photography does not suffer from loss of sharpness due to camera movement. Floodlight photography can suffer from shutter-speed problems if an excessively large aperture number is selected, which results in very long exposure times. With a camera mounted on a tripod or on a copystand, exposure times as long as 2 s can still yield sharp photographic images.

Feature Exposure Selection. For most failure analysis photography, the correct exposure is one that accurately records the condition of the component. Under certain conditions, multiple exposures are taken to highlight certain features of the component for specific purposes. One of those conditions would be the lighting of the inside surface of a hollow component. In that situation, the outside of the component may be underexposed or overexposed, but the internal surface of the hollow component would be properly illuminated. The general guideline is that exposure should document the overall condition of the component, followed by adjustment, in order to accurately record the visual characteristics of the part of the failed component under investigation.

Filter Selection. The main filter used in failure analysis photography is the ultraviolet haze or UV filter. The UV filter protects the lens from scratching and from accumulation of deposits and dirt. The UV filter also aides in sharpening the photograph under hazy conditions. Polarizing filters may be used in highly reflective conditions. If tungsten lighting is used, then a blue filter, such as an 80B, is commonly used to correct the color temperature shift caused by the tungsten lighting.

Richard H. McSwain, McSwain Engineering, Inc.

Photographic Lighting

Natural lighting is typically used in photographic field documentation of failed components and structures. It should be remembered that the color temperature of the lighting varies through the day, and color temperature change affects the appearance of the photograph. Early-morning photographs tend to have a blue tint, and late-afternoon photographs have a golden tint. Use of color standard cards can assist in ensuring that the color recorded is the “true” color of the component. Natural lighting also suffers from variations in image contrast due to sunlight conditions, changing from bright conditions to overcast conditions. Bright sunlight conditions create harsh shadows that may prevent the proper printing of the photograph to show all features present. Severely overcast conditions may result in long exposure times for the required aperture setting and may lead to a loss of sharpness in the image. Fill-flash is often useful outdoors for both the bright sunlight conditions, to fill in heavy shadows, as well as in overcast conditions, to produce adequate overall lighting.

Backlighting is often used for emphasizing a component contour or shape. Such lighting can be accomplished using a lighted panel viewer or other technique to project light from behind a component. It should be kept in mind that the type of lighting used must be compensated for color temperature in order to achieve a proper color photograph. Use of a clear plastic sheet, supported to provide separation from the table top and lit from below, is also useful in producing photography without shadows.

Flash lighting is the most commonly used lighting technique in failure analysis photography. It is well suited to failure analysis because it is directional, and the shadows produced emphasize the fracture features present on a failed component, where the flash is used off-camera. It is critical in using an off-camera flash to properly direct the flash to the component being photographed. This can be accomplished by use of camera flash

brackets or by carefully holding the flash unit during photography. Care must be taken, as well, to not overexpose components with the flash. Flash overexposure with digital photography is especially critical because of the limited exposure latitude of a digital image compared to a film image.

Copystand Lighting. The two side lights on a copystand provide uniform lighting, with excellent contrast for recording the features of a failed component. Copystand lighting also provides the option of lighting from only one side of the failed component, to better reveal fracture surface details.

Fiber Optic Lighting. Fiber-optic techniques are useful in directing light to a certain location, including cavities and hollow areas, as well as providing a means to emphasize fracture surface topography. The fiber-optic lighting color temperature should be compensated for by using the appropriate filter on the camera lens. Fiber-optic lighting using a borescope can be especially useful in highlighting cavities that are difficult to light by normal means.

Fracture Surface Photography

Fracture Surface Side Lighting. The topography of a fracture surface is one of the most important features present in a failed component. It is the means by which the failure analyst determines the mode of failure. Surface topography is best revealed through proper lighting techniques. For general fracture surface documentation, dual incandescent side lighting with the copystand provides an excellent lighting condition to document the overall condition of a component. High-intensity moveable lighting and other light sources may be required for specific component conditions.

A typical arrangement of a dual-side incandescent lighting copystand is shown in Fig. 6, where lights “A” and “B” both illuminate the specimen. The resulting photograph with dual-side illumination is shown in Fig. 7(a). Dual-side lighting is limited in revealing surface topography, because topographic feature shadows are filled by the opposing light source.

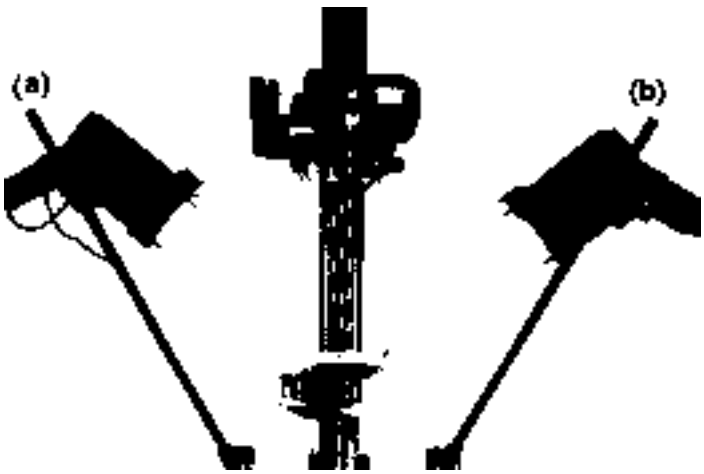


Fig. 6 Copystand for dual-side incandescent lighting

Fig. 7 Fracture-surface side lighting. (a) Photograph of a fatigue fracture resulting from dual-side incandescent lighting with lights A and B (Fig. 6). (b) Photograph of a fatigue fracture resulting from single-side incandescent lighting from light A (Fig. 6)

The arrangement for a single-side incandescent lighting copystand is to illuminate the specimen with just one of the lights. With illumination from just light “A” in Fig. 6, the resulting photograph clearly reveals the fractographic features present (Fig. 7b). When using single-side lighting, the specimen should be rotated under the lens, and the effects of the lighting should be evaluated to reveal the most detail in the fracture surface.

In general, the light should flow across the grain of the fracture and not parallel to it. Figure 8(a) shows the effect of sample side lighting of the fatigue fracture with the light parallel to the fracture features, and Fig. 8(b) shows the same fracture with the side lighting perpendicular to the fatigue fracture feature. The side lighting perpendicular to the fracture features reveals much more of the fracture detail typical of fatigue. A white card can be used in the position of the opposing light to reflect light into harsh shadowed areas, as shown in Fig. 9(a) and (b).

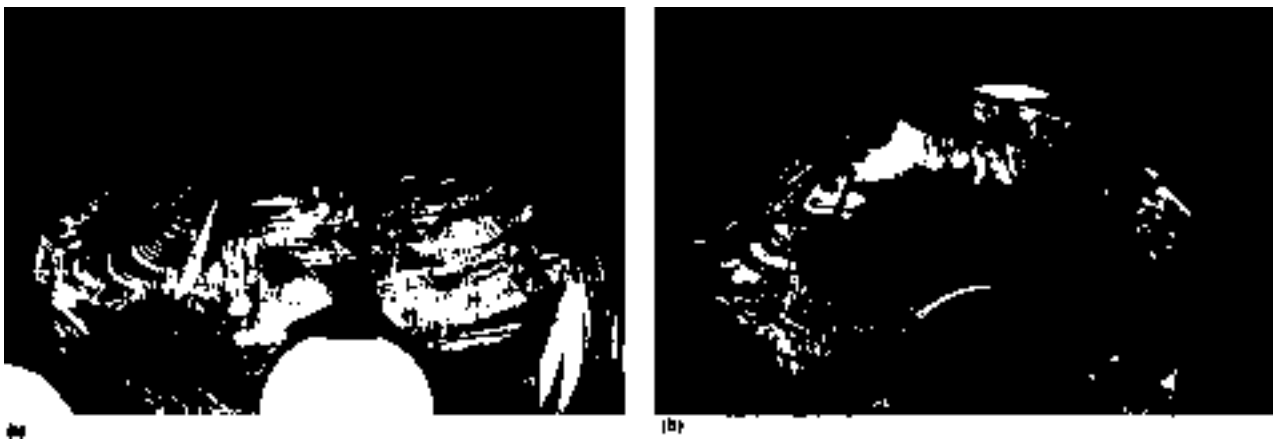


Fig. 8 Photographs from single-side incandescent light of a fatigue fracture with (a) parallel lighting and (b) perpendicular lighting

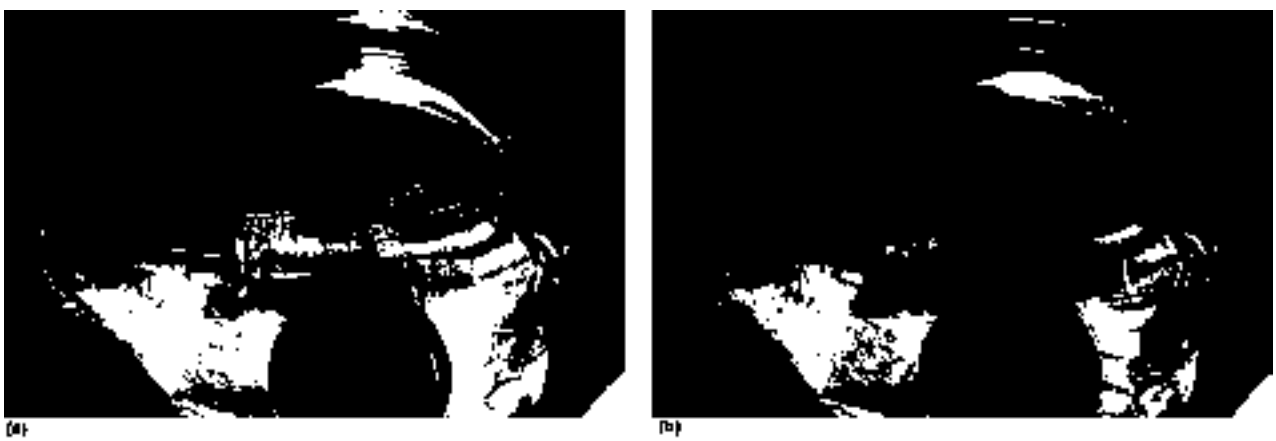


Fig. 9 Photographs from single-side perpendicular lighting of a fatigue fracture. (a) Single-side incandescent perpendicular lighting. (b) Single-side incandescent perpendicular lighting with an opposing white reflecting card

Fracture Surface Ring Lighting. Ring lighting produces excellent uniform light coverage of a fracture surface. A typical ring light attached to a camera lens and the resulting fracture photograph are shown in Fig. 10(a) and (b). The ring light is less able to highlight fracture features but is beneficial in photo-documentation at closer focal lengths.

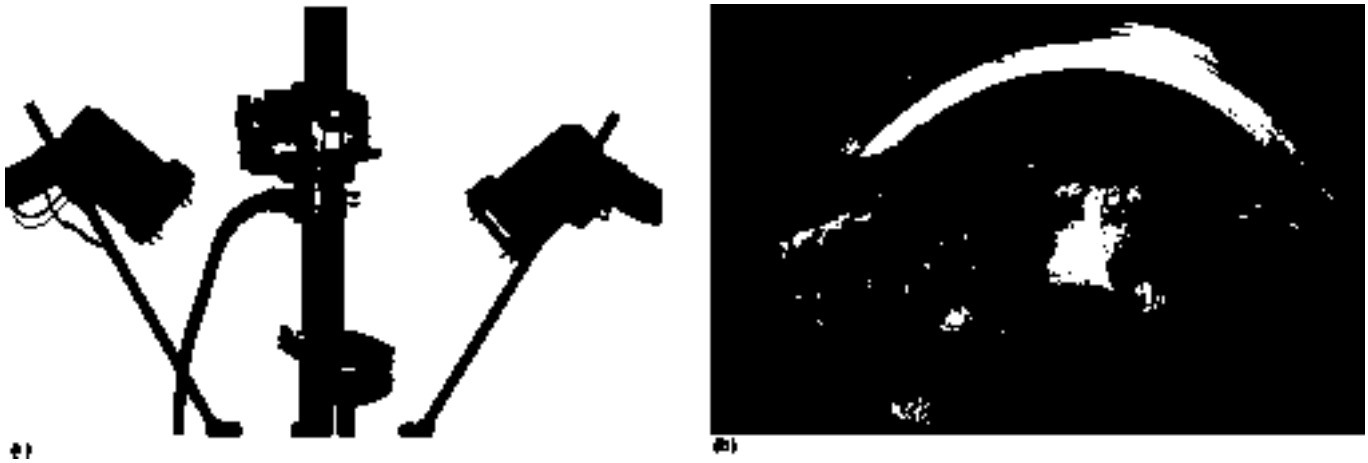


Fig. 10 Fracture surface fiber optic ring lighting. (a) Fiber optic ring light copystand setup. (b) Resulting fiber optic ring light photograph of a fatigue fracture

Fracture Surface Fluorescent Lighting. Fluorescent lighting produces a more diffuse lighting when photographing highly reflective fracture surfaces and provides excellent overall documentation of the component conditions. A typical arrangement of a single-side fluorescent lighting copystand is shown in Fig. 11(a). The resulting fractograph is shown in Fig. 11(b). A typical fluorescent ring light attached to a camera lens and the resulting fracture photograph are shown in Fig. 12(a) and (b). The fluorescent light is limited in showing detail from shadowing of fracture surface features but is especially useful when photographing highly reflective fracture surfaces, such as fatigue in aluminum.

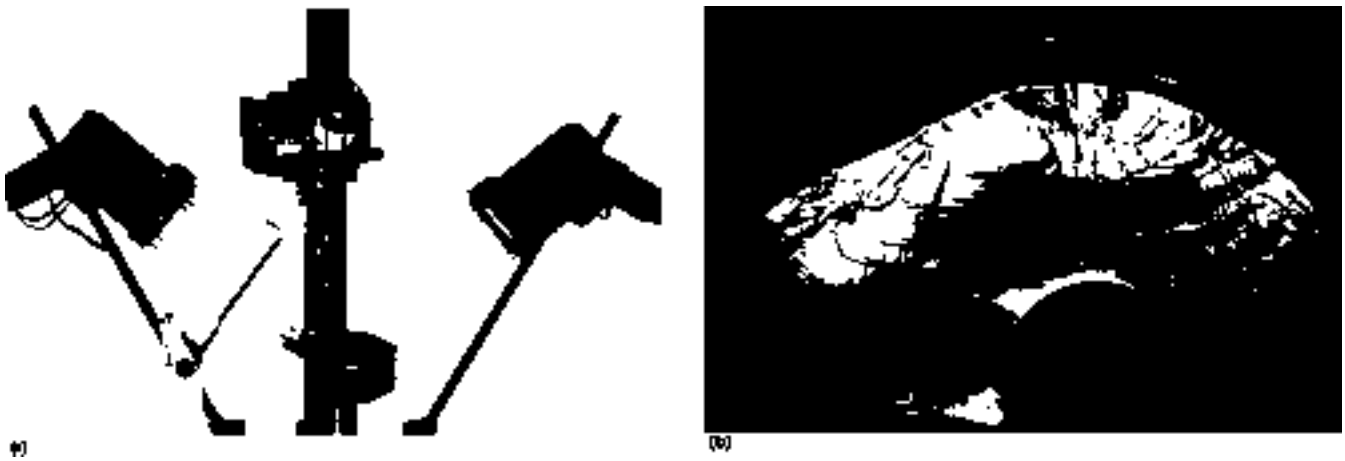


Fig. 11 Fracture surface fluorescent lighting. (a) Single-side fluorescent copystand lighting setup. (b) Resulting single-side fluorescent lighting photograph of a fatigue fracture

Fig. 12 Fracture surface fluorescent ring lighting. (a) Fluorescent ring light copystand setup. (b) Resulting fluorescent ring lighting photograph of a fatigue fracture

Fracture Surface Fiber Optic Lighting. In some cases, the side-lighting effect may need emphasis, as in the case of a fatigue fracture origin. An incandescent fiber-optic light source is useful in such conditions. A typical fiber-optic side-lighting arrangement and resulting fracture photograph are shown in Fig. 13(a) and (b). The resulting fractograph is high in contrast, and the fracture surface topography features are maximized.

Fig. 13 Fracture surface fiber optic side lighting. (a) Single-side fiber optic copystand lighting setup. (b) Resulting fiber optic lighting photograph of a fatigue fracture

Fracture Surface Diffuse Lighting. Highly reflective components, such as bearings and polished metallographic specimens, are difficult to photograph because of unwanted reflections. The use of diffuse incandescent illumination is beneficial when photographing highly reflective surfaces. The light diffusion is accomplished using a white paper “tent.” A typical tent lighting arrangement and the resulting photograph are shown in Fig. 14(a) and (b). The outcome is a reduction in unwanted reflections and a resulting acceptable photograph.

Fig. 14 Fracture surface diffuse lighting. (a) Dual-side incandescent diffuse copystand lighting setup. (b) Resulting diffuse lighting photograph of a fatigue fracture

Fracture Surface UV Lighting. There are certain requirements where UV lighting may be necessary to document the condition of a component. One such requirement is instrument panel photo documentation, where fluorescent materials are used in the manufacturing process. Another requirement is in the documentation of fluorescent penetrant nondestructive inspection indications. A typical UV lighting arrangement and resulting photograph are shown in Fig. 15(a) and (b). Film photography often has difficulty in capturing the UV fluorescent indications, and long exposure times are often required. Digital photography is more useful in capturing fluorescent indications under UV light, especially with cameras in which the media speed setting can be maximized.

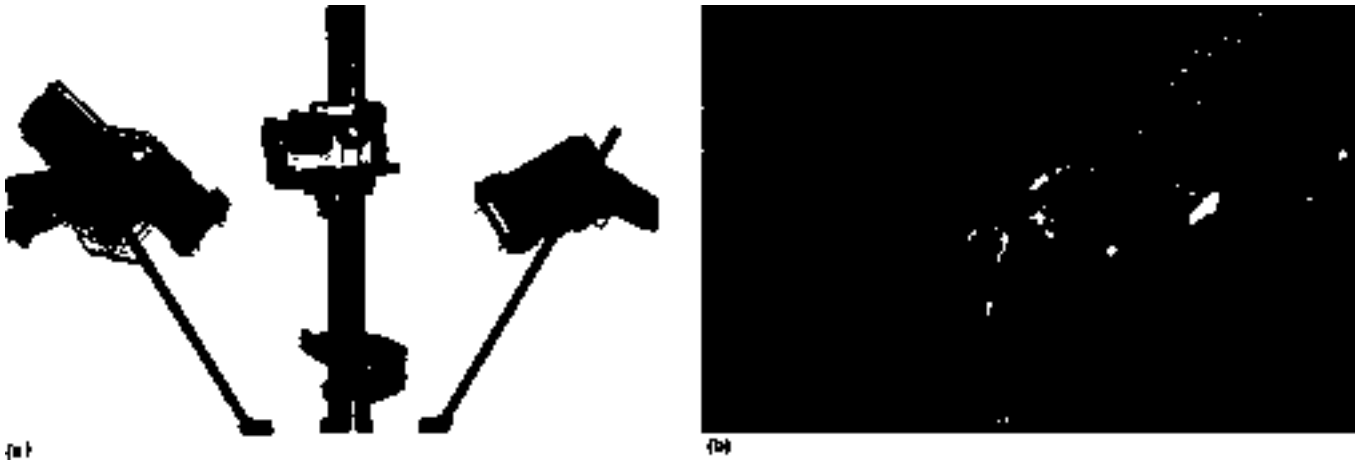


Fig. 15 Ultraviolet lighting. (a) Single-side ultraviolet copystand lighting setup. (b) Resulting ultraviolet lighting photograph of a fluorescent penetrant crack indication

Macrophotography

Macrolenses. Macrophotography in failure analysis work is performed using a zoom lens with macrocapability or a macrolens at either the 50 or 100 mm range. Typically, the maximum magnification achievable with the macrophotography lens is 1 to 1 magnification. A typical macrophotography arrangement and resulting fracture photograph are shown in Fig. 16(a) and (b).

Fig. 16 Macrophotography. (a) Typical macrocamera lens arrangement used in failure analysis photography. (b) Macrophotograph of a fatigue fracture

Macrophotography lighting is critical, because the camera is in close proximity to the component being photographed, and normal lighting techniques may not adequately reach the fracture surface or the component. Because of those requirements, either a macroflash is used or some type of side lighting that allows the proper lighting of the fracture surface without shadowing from the lens.

Depth of field in macrophotography is critical. At very close focal distances, the depth of field is critical. Typically, higher f-stop numbers are selected to ensure that the entire fracture surface is in focus.

Reference Locator Photographs. In macrophotography, it is useful to have available a reference locator photograph that shows the overall appearance of a component undergoing a failure analysis. As close macrophotographs are taken, the reference overall photograph can be marked to record the location and condition of each of the macrophotographs taken.

Microscopic Photography

Optical stereomicroscopy is routinely used in failure analysis photography to examine and characterize fracture surface features. Images observed in the stereomicroscope must be recorded for documentation of the features present and to transfer that information to others. A typical stereomicroscope photographic arrangement is shown in Fig. 17, and the resulting photograph is shown in Fig. 18. The primary techniques for stereomicroscopy photography is film photography, instant film photography, digital photography, and, in some applications, videography. All of the techniques except film photography provide immediate feedback as to image quality. Stereomicroscopic photography of a fracture surface requires adjustment of the microscope to give adequate depth of field. This is accomplished by adjusting the field aperture. All of the techniques of fracture surface lighting equally apply to stereomicroscopic photography, and similar results are obtainable.

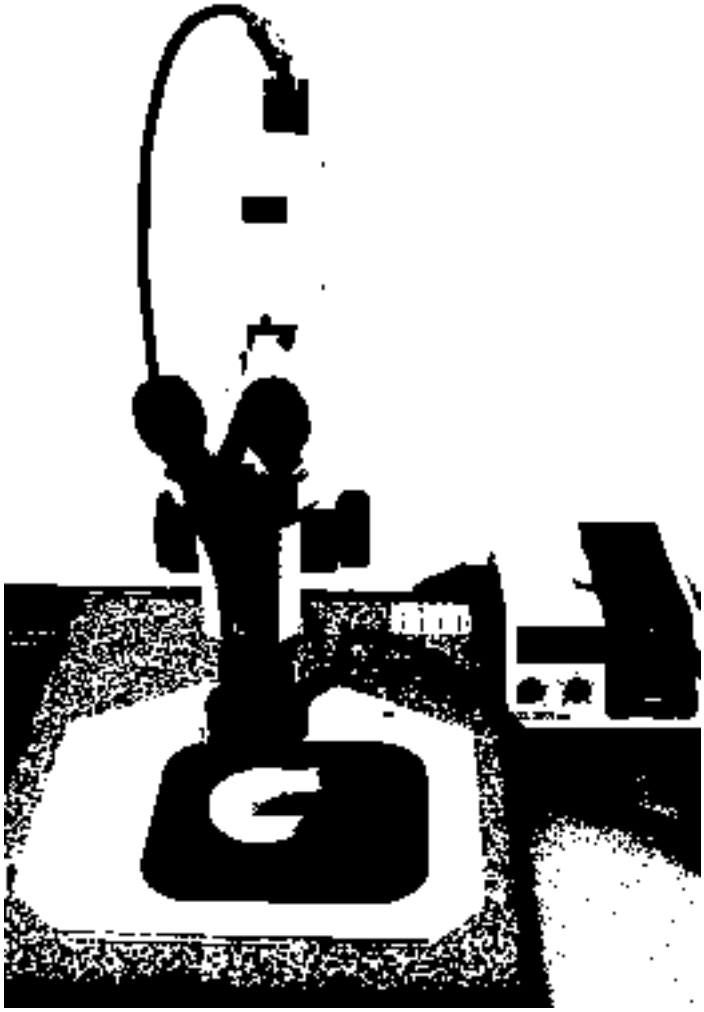


Fig. 17 Typical stereomicroscope used in failure analysis photography

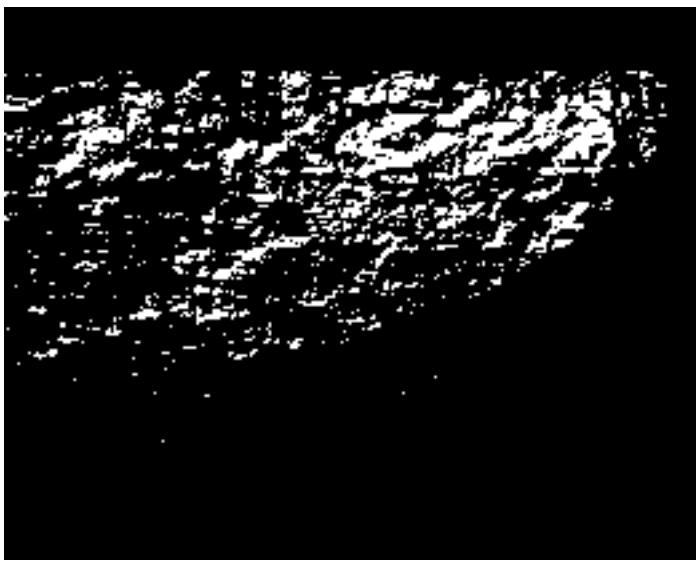


Fig. 18 Stereomicroscope photograph of a fatigue fracture

Optical metallography is used in failure analysis photography to characterize the macro- and microstructure of a failed component. As in optical stereomicroscopy, the features of the microstructure observed with optical metallography must be recorded and then printed. A typical optical metallograph photographic arrangement and resulting photograph are shown in Fig. 19(a) and (b). The photographic techniques employed in optical metallography include digital photography, film photography, and instant film photography. Optical metallography has historically used monochromatic lighting and black-and-white photography. However, color photography is advantageous for showing heating effects, corrosion

product, and microstructural features with color. Additional and expanded information on optical metallography is in *ASM Handbook*, Volume 9.

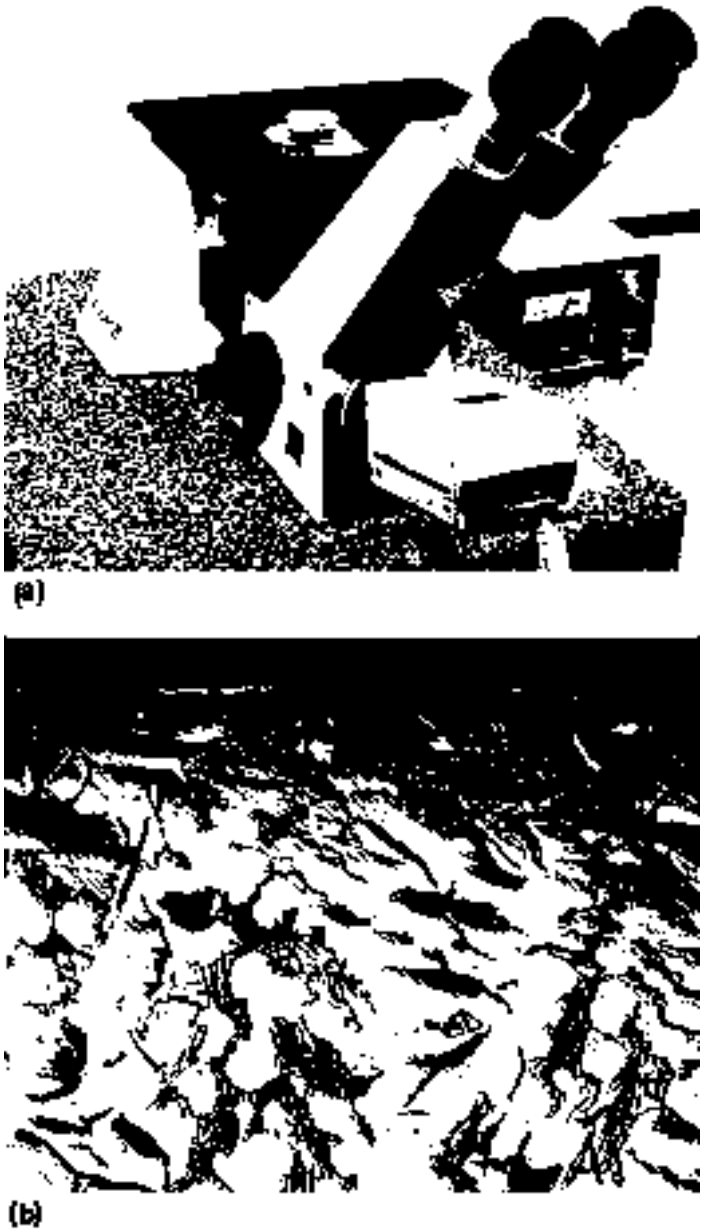


Fig. 19 Optical metallograph. (a) Typical optical metallograph used in failure analysis photography. (b) Optical metallograph photograph of a microstructure

Scanning electron microscopy is used in failure analysis photography for the examination of surfaces, including fracture surfaces, and the recording of fracture features. Scanning electron microscopy is used in failure analysis photography for the examination of fracture surfaces and other surface features and the recording of features present, as well as high-magnification recording of the microstructure. A typical scanning electron microscope and resulting scanning electron microscope photograph of a fracture surface are shown in Fig. 20(a) and (b). The scanning electron microscope is a higher-magnification extension of the optical stereomicroscope as well as the optical metallograph. Scanning electron microscope images are monochromatic and, as such, do not benefit by color photography. The primary techniques of photography associated with scanning electron microscopy are digital photography and instant film photography. X-ray imaging spectroscopy associated with scanning electron microscopy typically uses digital photography for the recording of x-ray images and maps. Videographic cameras may also be used to record entire scanning electron microscopy examination sessions.

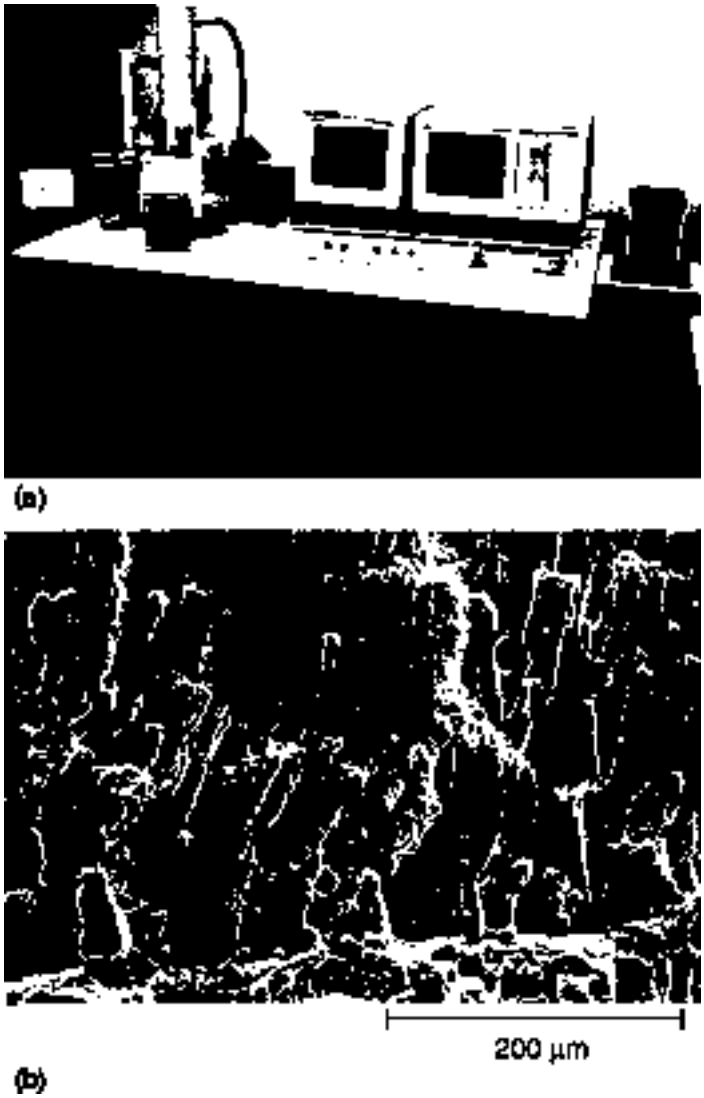


Fig. 20 Scanning electron microscopy. (a) Typical scanning electron microscope used in failure analysis photography. (b) Scanning electron microscope photograph of a fatigue fracture

Richard H. McSwain, McSwain Engineering, Inc.

Special Methods

Stereophotography can be used in failure analysis work for documenting component condition and recording information in three dimensions. Stereophotography consists of the capturing of two images, either simultaneously with a stereocamera or sequentially by the repositioning of a single camera. The resulting image pairs are viewed with a stereoviewer, enabling the observation of three-dimensional features and characteristics. Similar techniques are used with the scanning electron microscope by the capture of sequential images, with the second photograph taken after tilting the sample. A professional stereocamera is shown in Fig. 21. Stereophotographs may also be accomplished by using a single camera and a slide bar, as shown in the resulting stereopair image in Fig. 22.

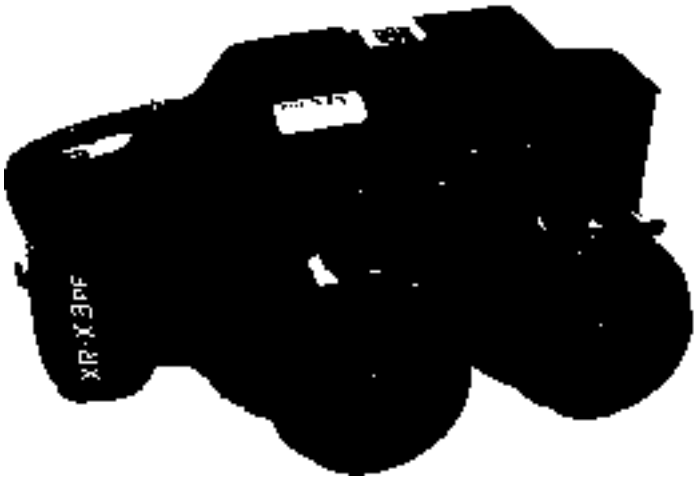


Fig. 21 Typical 35 mm stereocamera used in failure analysis photography

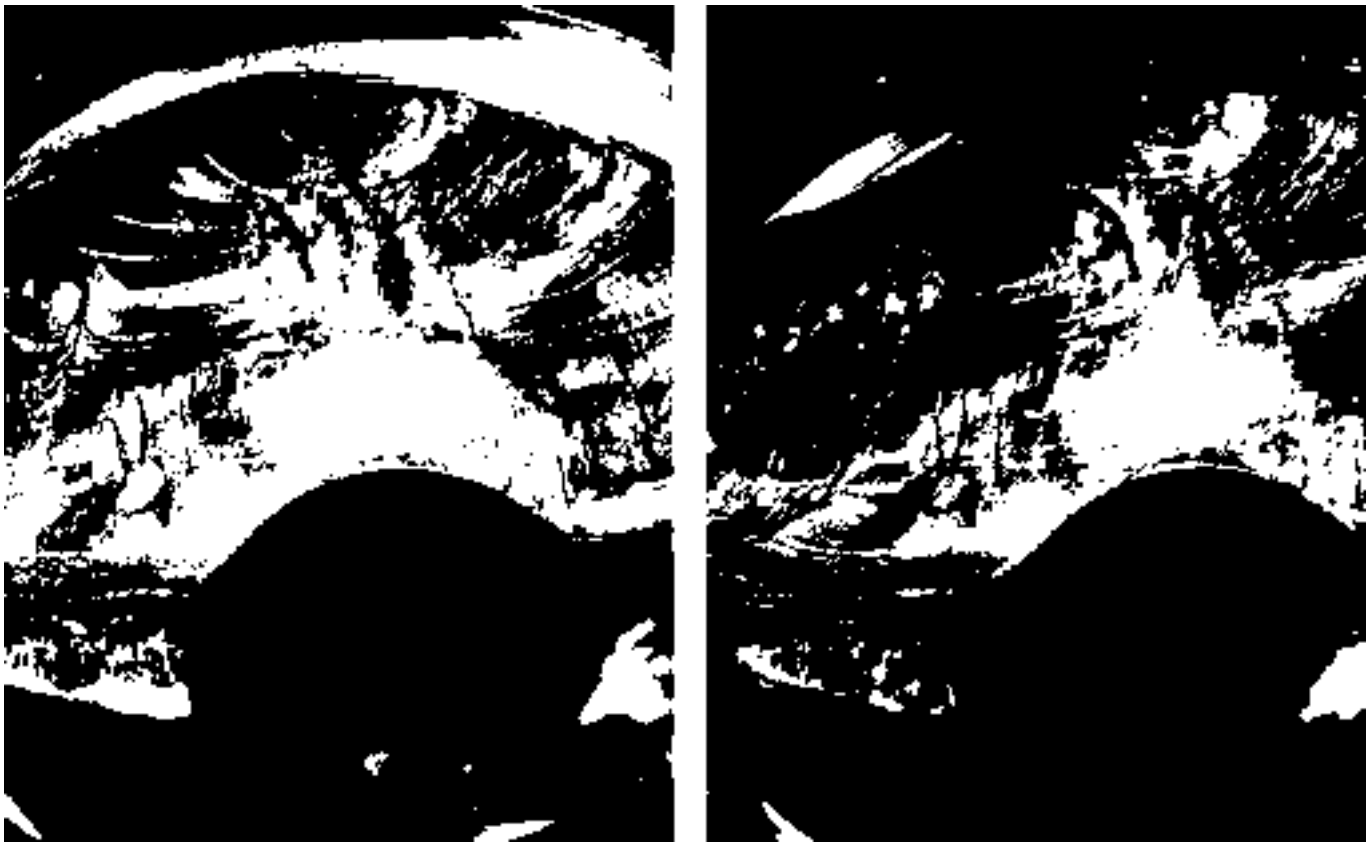
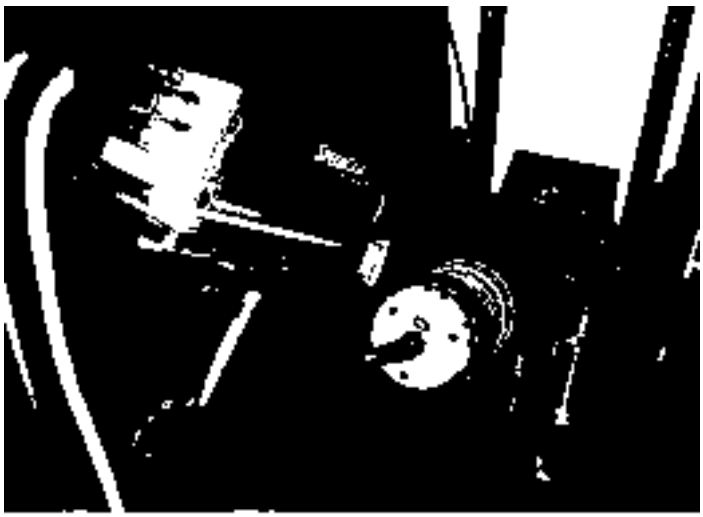


Fig. 22 Stereophotograph of a fatigue fracture

Videography is used in failure analysis photography to capture dynamic events as well as to document rapidly changing component conditions. Current videography cameras are available with either analog or digital technology. Digital videocameras have the advantage of the input of video-images into a computer for frame-by-frame analysis. Videography is also useful in failure analysis photography for low-light-level documentation, including UV lighting conditions as used in nondestructive inspection and infrared imaging in very low light levels. Videography is especially useful in failure analysis photography to document the disassembly of components, sectioning of components, and cleaning of components and fractures. Figure 23 is a digital image resulting from a typical digital videocamera.

Fig. 23 Videocamera photograph of a fatigue fracture

High-speed videography has use in failure analysis photography for the recording of high-speed events during exemplar component testing, impact testing, and fracture testing. High-speed cameras employ either high-speed film recording or high-speed digital technology. High-speed digital cameras have recording speeds of 1000 to 10,000 frames/s and higher recording times of 1 to 4 s. The recording technique is similar to digital video, but, instead of recording image data onto tape, the images are stored in random access memory. The images are typically transferred to a computer by either cable or digital storage media. The lighting requirements for high-speed video are much greater than for normal photography or videography because of the short duration exposure times. A typical digital high-speed video-camera being used in a dynamic testing application is shown in Fig. 24(a), and a resulting high-speed video-photograph is shown in Fig. 24(b).



(a)

(b)

Fig. 24 High-speed videography. (a) Typical high-speed digital videocamera used in failure analysis photography. (b) High-speed videophotograph of an impact test

Selected References

- P. Andrews, *The Digital Photography Manual,* Carlton Books Ltd., 2000
- S. Bavister, *Digital Photography,* Collins & Brown, London, 2000
- J. Bidner, *Digital Photography, A Basic Guide to New Technology, The Kodak Workshop Series,* Silver Pixel Press, 2000
- M. Busselle, *Better Picture Guide to Travel Photography,* Rotovision SA, Watson-Guptill Publishing, 1997
- J. Calder and J. Garrett, *The New 35 mm Photographer's Handbook,* 3rd revised ed., Three River Press, 1999
- S. Dalton, *Secret Worlds, Photographs by Stephen Dalton,* Firefly Books Ltd., 1999
- K. Eismann and D. McClelland, *Real World Digital Photography,* Peachpit Press, 1999
- J.G. Ferwerda, *The World of 3-D: A Practical Guide to Stereo Photography,* 3-D Book Productions, The Netherlands, 1990
- L. Frost, *The Question & Answer Guide to Photo Techniques,* David & Charles, 1995 and 1997
- J. Hedgecoe, *New Introductory Photography Course,* Focal Press, Butterworth-Heinemann, 1996
- J. Hedgecoe, *The Art of Color Photography,* Focal Press, 1998
- R. Hicks and F. Schultz, *Pro Lighting, Product Shots, A Guide to Professional Lighting Techniques,* Quarto Publishing, Rotovision, Inc., 1994
- R. Hicks and F. Schultz, *Pro Lighting, Special Effects, A Guide to Professional Lighting Techniques,* Quarto Publishing, Watson-Guptill Publications, 1995
- R. Hicks and F. Schultz, *Pro Lighting Still Life, A Guide to Professional Lighting Techniques,* Quarto Publishing, 1996
- J. Hilton, *Pro Photo Close Up Photography,* Rotovision SA, May 1998
- R. Hirsch, *Exploring Color Photography,* 3rd ed., McGraw-Hill, Times Mirror Higher Education Group, 1997
- M. Langford, *Basic Photography,* 6th ed., Focal Press, Butterworth-Heinemann, 1997
- B. London and J. Upton, *Photography,* 6th ed., Addison Wesley Longman, Inc., 1998
- A. May, *Digital Photography,* Dorling-Kindersley, London, 2000
- A. McWhinnie, *Reader's Digest Complete Photography Manual,* R.D. Associates, Inc., 2000
- *Photography Through the Microscope,* 9th ed., Eastman Kodak Co., 1998
- S.F. Ray, Ed., *High Speed Photography and Photonics,* Focal Press, Butterworth-Heinemann, 1997
- D.R. Redsicker, *The Practical Methodology of Forensic Photography,* CRC Press, Inc., 1994
- G. Schaub, *The Digital Darkroom, Black and White Techniques Using Photoshop,* Silver Pixel Press, 1999
- S. Sholik and R. Eggers, *Macro and Close Up Photography Handbook,* Amhurst Media Inc., 2000
- W. White, Jr., *Close-Up Photography, The Kodak Workshop Series,* Silver Pixel Press, 1984

Chemical Analysis of Metals in Failure Analysis

Debbie Aliya, Aliya Analytical

Introduction

CHEMICAL ANALYSIS is often a useful tool for failure analysis. There are two main categories of chemical analysis that are often used by failure analysts:

- *Bulk composition evaluation:* often performed in order to determine whether the correct alloy was used in the subject component
- *Microchemical analysis:* to find evidence of contamination, to evaluate the composition of microphases revealed on a metallographic specimen, or to evaluate corrosion products

Extreme care must be used in interpretation of chemical analysis work performed as part of a failure investigation. Minor deviations from specified composition must not be interpreted as the sole cause of a failure, without much additional supporting evidence. All chemical analysis methods have their limitations and proper uses. This article describes some of the more common elemental composition analysis methods and explains the concept of referee and economy test methods. Clues for communicating with the metals chemistry service provider are also given. There are many potential pitfalls in analyzing materials that have been in service, and it is important that the chemist or technician understands the history of the component, so that it may be prepared for analysis properly.
This article also describes some special analytical techniques, including field composition checkers and some nondestructive methods that can be used for in-process sorting that becomes necessary when a failure investigation detects mixed stock. This is not so important to the investigation itself but can be helpful to those who experienced the failure to prevent recurrences. Some information on detection of hydrogen in steels for hydrogen embrittlement evaluation is also provided.

Chemical Analysis of Metals in Failure Analysis

Debbie Aliya, Aliya Analytical

Bulk Composition Verification

Before the details of the analytical method are discussed, it should be noted that there are few analytical methods that can be properly performed without knowledge of the intended composition. Most of the analytical methods in use today are instrumental methods whose accuracy depends on calibrations based on alloys of similar composition. The response of the instrument to the alloys present is rarely linear with respect to composition. Thus, a general idea of the expected amount of any element is required in order to determine the composition in an accurate and precise manner. Of course, if enough material is provided, the chemist can repeat the test several times, approaching the proper calibration setup each time. However, this is costly in time, and there is still the risk of missing an element that was specified. If the analyst does not trust the chemist to actually perform the test and provide the results obtained, then another chemist should be found. Do not expect a competent analyst to work “blindly” out of a misguided belief that this will produce the “most objective and correct” results.
If the analyst does not know what the exact specified material is, knowing that it is a carbon steel, for example, versus an alloy steel or a stainless steel will be helpful. Obviously, the experienced metals chemist will be able to tell that something is a piece of aluminum or a copper alloy or a piece of steel. More unusual alloys may cause problems. However, for many methods, where individual elements must be run separately, it could be very expensive to test for every element that might be present in any specific alloy within the family of alloys. If the levels of trace elements that are not generally specified or reported on the certification documentation of the producing mill are desired, the analyst must specify what amount is being sought. For example, embrittlement from amounts of tin less than 0.001% can cause a problem, if the tin is in the “wrong places” in the microstructure. Many of the standard or economy methods simply will not detect these small amounts.
Likewise, very high concentrations of elements can cause false readings with many of the common methods. If the chemist has not calibrated with a similar specimen with a higher content than the specimen in question, an analytical error could easily go undetected. Many composition analysis methods must be calibrated with a material not only having a similar amount of the alloying element in question, but also having the same matrix element and a similar concentration

of alloying element. For example, it may not be a good idea to use a nickel-base superalloy with 15% chromium as a calibration check for an iron-base stainless steel with 15% chromium.

Another important characteristic that the chemist should be made aware of is whether the material has been wrought or is in an as-cast condition. As-cast products usually have more segregation than wrought products, so the sampling needs to be done from more layers of the material than may be acceptable for wrought product.

If the material has been subject to a surface treatment such as carburization of steel, the chemist should be told, so that an erroneous carbon content based on analysis of layers including the enriched surface layers is not reported.

When a component or assembly is subject to damage that prevents its continued functioning, those involved often want to know whether the correct material was used to manufacture the subject items. There is a long history of use of verification of composition as “acquittal” of a component that has stopped functioning. This practice is in general based on ignorance of the principles of engineering, which require that a specific loss of functionality be linked to a specific deficiency in the component. On the other hand, lack of conformance to a particular requirement in no way automatically implies a cause of failure. Also, conformance to all specified requirements does not automatically imply that the product was designed and manufactured properly. Despite the fact that determining the bulk composition is not often a key piece of information in a failure analysis, it is often a useful piece of information, in particular, when the analyst is not very familiar with the alloy system in question.

For example, the potential consequences of exceeding impurity limits for zinc and aluminum casting alloys are discussed in various articles in *Properties and Selection: Nonferrous Alloys and Special-Purpose Materials*, Volume 2 of *ASM Handbook*. If some of the consequences appear to have happened, it might be useful to perform a composition analysis. However, the analyst needs to be aware that the composition can be altered by service conditions. Particularly where corrosion occurs, some elements are generally removed in greater amounts than others. An example is “dezincification” of brass. Thus, the fact that the composition does not meet the specification today does not prove that it did not meet the specification at the time of manufacture. Great care is required in interpreting results of bulk-composition test results performed on actual products, especially when they have been in service.

Even when no corrosion is involved, the analyst needs to be aware that most listed steel specifications, for example, are based on ladle analysis. The ladle analysis is more representative of the entire heat than the relatively tiny specimen usually used in product composition testing. Most steel is sold based on this ladle analysis. However, it is entirely possible, especially for steel produced by the older rimming or semikilled ingot processes, for the composition to change prior to complete solidification. Even in steel that is produced by a “killed” or “deoxidized” process, chemical reactions can continue in the ladle as the metal is being cast. Although the mills often test multiple specimens from a single ladle, they rarely report more than one of the analysis results on the certification. As long as all the elements remain “in spec,” the entire heat will be sold with the same “ladle analysis.” For carbon and alloy steels and zinc casting alloys in product form, a broader range of allowed concentrations of intentional and residual elements is published by the Society of Automotive Engineers.

Also notable is that many product form metal specifications have other requirements than the composition. Whether a component met intended mechanical property requirements may be much more significant to the failure investigation than the composition, unless the failure was a corrosion failure of a supposedly corrosion-resistant alloy. The complete applicable product specification must be carefully read and interpreted before an out-of-specification composition is “blamed” for the failure.

In addition to corrosion reactions, another source of composition change that happens after the certifying test is performed occurs when intentional carburization is carried out to improve wear resistence. Many small parts, such as self-tapping screws, are virtually impossible to analyze for original base-metal carbon content, since the carbon added during carburizing often penetrates through much of the material. This example illustrates the importance of knowing what the specified processing sequence was for any component subjected to a bulk composition analysis. A microstructure analysis is often more informative.

There are many difficulties in analyzing the base metal composition of manufactured components. Often the component has been painted or plated. The plating or paint must be removed without changing the composition of the underlying metal. While this may be done relatively easily with paint on a heavy steel component by burning off the paint, this burning may alter the composition of a very light gage material. A solvent paint removal process may be required in that situation. Please note that only an experienced metals chemist should recommend proper specimen preparation techniques, as some might be quite dangerous if not performed with adequate knowledge of proper safety precautions. Another example of a difficult plating removal process would be a copper/nickel/chromium plating on a zinc die casting. Since the zinc is more soluble in acid than any of the plating layers, another approach is obviously required, such as sanding or drilling. Obviously, the odd shape of many products makes sampling difficult at best. Small specimens that have had part of their composition altered by carbon or nitrogen diffusion or that have been plated or painted tend to be more problematic than larger pieces of structural alloys.

Some analysts remelt and chill cast their metals specimens in order to get a disk large enough to analyze by the faster and thus more economical test methods, such as optical emission spectroscopy (OES) or glow discharge spectroscopy. As long as the chips or bits that are remelted are truly representative of the base metal (i.e., plating or other surface coating

has been removed, no carburized or decarburized material is unintentionally included, no chunks that have been altered by corrosion processes have been included, etc.) and the process is performed properly in a protective atmosphere, this can be a legitimate technique. However, many things can go wrong in this process, especially if the chemist is not familiar with the details of metal processing methods and service conditions that may have altered the composition.

In fact, any tool that is used to machine chips that are analyzed can actually contaminate the specimen. Extreme care by an experienced metals chemist is the minimum requirement for important analysis of metals composition in failure analysis work. If the milling to obtain chips for analysis is done in a machine shop where other machining is done, it may be necessary to set up a protective "tent" (for example, out of aluminum foil) around the chips being collected, to prevent contamination by airborne chips from other machines.

Economy Methods. The economy methods of metals composition analysis are usually methods that can be preformed on a "small chunk" of the metal in question. The economic benefits of these types of methods include the fact that no time is spent milling, drilling, weighing, and dissolving the specimen. As long as a clean flat area is available on the surface of the part, analysis may proceed after a quick grinding operation on a special sanding belt. The need for the special sanding belt is because the abrasive grit must not contain any of the elements that are going to be analyzed. For example, a zirconium oxide abrasive is often used to prepare steels for this type of analysis, because zirconium is rarely specified in steels. Thus, if a small piece of zirconium oxide does get embedded in the metal, it will not affect the analysis.

The actual size of the "small chunk" required varies with the particular test method. A common requirement for an optical emission spectrophotometer (OES) is a 6 mm ($\frac{1}{4}$ in.) diameter, flat-ground spot on a specimen at least 1.6 mm ($\frac{1}{16}$ in.) thick. Thinner material may burn through during the test, and a smaller flat area will not properly seal the vacuum chamber where the emitted light waves are analyzed. A common requirement for a glow discharge spectrophotometer (GDS) is a 13 mm ($\frac{1}{2}$ in.) diameter, flat-ground spot on a specimen at least 1.6 mm ($\frac{1}{16}$ in.) thick.

These analytical methods are quick to perform, especially once the instrument has been calibrated for the type of alloy in question. The main drawback to these methods is that most metals do not have a completely uniform composition from surface to core. Most metals are subject to some degree of segregation during solidification. Thus, sulfur is more likely to be concentrated in the core of a piece of steel, while many high-temperature-forming processes deplete carbon at the surface of steels. While the differences may not be great in modern wrought product, cast products can be very heavily segregated. Extreme care is required when interpreting such results.

Referee Methods. Most referee methods are either "wet chemical" or "combustion" methods. In these methods, the referee designation is usually due to the fact that the sampling process allows a more representative specimen to be analyzed. The accuracy of the "economy methods" described previously is very good on a properly maintained and calibrated piece of equipment. However, the methods performed on solid specimens can only measure the layers available. A specimen properly sampled for wet chemical or combustion analysis will contain all the layers of the material, from surface to core, except when the surface layers are specifically intended to be discarded. This might be the case for a heavy casting or forging.

The failure analyst should remember that the composition at the surface may be of more interest than that of the bulk. For example, if a leaf spring has fractured in fatigue, it may be valuable to analyze the outer 0.1 mm (0.004 in.) for carbon separately by combustion, even if the bulk composition is analyzed with an economy method. Again, even a properly performed standard-certification verification can be very misleading to an inexperienced failure analyst.

Common wet chemical analysis methods for metals include atomic absorption, where a beam of light is directed through a solution of known concentration prepared by dissolving metal chips in the proper acid mixture. The light is absorbed preferentially at certain wavelengths by certain elements. Thus, the more of a certain element present, the less light will get through to the detector. Again, proper calibration is extremely important. Another method of wet chemical analysis is the inductively coupled plasma spectrometer. Again, a solution of dissolved metal is fed into the instrument. The solution is dispersed and heated to plasma temperatures. The metal atoms, excited by the heat, emit light waves that have known correspondences to particular elements.

A drawback of all wet methods is that if there are microconstituents that are not easily soluble in the commonly available acids, erroneous results can easily occur. It is not the instrument that is at fault, but the fact that the method used to dissolve the metal was inadequate. These types of analysis errors may be very difficult to see.

Many tool steels and other highly alloyed corrosion resistant materials are difficult to analyze for this reason. Again, if correct results are important, the analyst should choose a chemist who has proven competency in analyzing the particular alloy in question. One way of demonstrating competency is participation in the "round robin" programs administered by the National Institute of Standards and Technology or the Collaborative Testing Service.

Please note that the competent chemist can take advantage of this usually problematic differential solubility to gain additional information about some materials. For example, aluminum is often added to steel to create conditions conducive to maintaining a fine grain size, even in the heat-affected zones of welds. This effect is created because aluminum nitrides keep the grain growth to a low level, even during exposure to the relatively high temperatures encountered during welding. However, aluminum is very highly oxidizable at molten steel temperatures, so it often

happens that some or much of the aluminum added to the steel ends up as an oxide rather than a nitride. The aluminum oxides are useless as grain-size controls. Aluminum oxides are not easily soluble in acids, while the aluminum nitrides are easily soluble. Thus, running the test for aluminum twice, once using an "economy" method that measures all of the aluminum (if the aluminum is not macrosegregated) and once using a wet method on a solution that has been strained to remove the insoluble aluminum oxides, can provide information on why an apparently "conforming" aluminum-killed steel has a coarsened grain structure.

Another referee composition method, used particularly often with ferrous alloys, is one of the various combustion methods. Here, a known quantity of solid chips of the metal is heated in a crucible while an oxygen-containing gas is passed over the crucible. The oxygen combines with carbon and sulfur present in the metal, forming carbon dioxide, carbon monoxide, sulfur dioxide, and so on. Some sort of detection device is used to, in effect, "count" the molecules of compound gases. These numbers are (generally, with modern instruments) automatically converted to a percent of carbon and/or sulfur in the original metal specimen.

Alternatives to Economy and Referee Methods. All of the methods described so far are in some manner destructive. When a less destructive method is required or when a failure investigation of an in-process incident reveals that the material in house has been "mixed" with an unspecified material, spot-check test kits may be useful. Most of these tests have limited ability to distinguish certain grades from other somewhat closely related grades. For example, a drop of potassium thiocyanate/stannous chloride solution will turn pink if placed on 316 stainless steel, which contains molybdenum, while it will not have any appearance change if placed on 304 stainless, which has no molybdenum. Other examples are also shown in Table 1 and in Ref 1 and 2. Kits containing premixed solutions are available from specialty companies.

Table 1 Qualitative chemical tests used in materials identification

Element tested for	Reagent	Positive result	Applications
Chromium	Diphenylcarbazide	Violet color	Identify chromium-free steel
Cobalt	Ammonium thiocyanate/acetone	Blue color	Identify cobalt-base alloys
Copper	Dithizone	Purple color	Sort copper-bearing stainless steels
Iron	Potassium ferricyanide	Blue precipitate	Sort low-iron high-temperature alloys
Lead	Sulfuric acid	White precipitate	Sort leaded bronze
Molybdenum	Potassium thiocyanate/stannous chloride	Pink color	Sort molybdenum-bearing stainless steels
Nickel	Dimethylglyoxime	Red precipitate	Sort 300 series from 400 series stainless steel

Hydrogen Embrittlement. On the subject of bulk chemical analysis of metals, hydrogen analysis in steels can be important in parts suspected to have broken or cracked because of hydrogen embrittlement. While reliable methods are becoming more available, the analysis of hydrogen in high-strength steel in the extremely low concentrations often associated with hydrogen embrittlement is still very costly. In addition, the amount of hydrogen in the steel is usually not of any particular significance. It is generally the atomic ("nascent") hydrogen that is present in the grain boundaries that causes the biggest problems. The hydrogen analysis methods available cannot distinguish between hydrogen atoms and hydrogen molecules and cannot distinguish between hydrogen within the grains and hydrogen at the grain boundaries. In addition to this, the hydrogen that has actually weakened the grain boundaries of a broken part has probably dissipated into the atmosphere by the time it gets to the chemistry lab. Thus, for routine investigations, hydrogen analysis is unlikely to provide much useful information. Hydrogen embrittlement diagnosis is better confirmed by other methods, including microstructure analysis to preclude other more observable causes of brittle intergranular cracking and confirmation that the cracking was delayed with respect to the application of stress or was caused by a sustained stress. Auger analysis may be used in bigger investigations of more costly failures, to preclude embrittlement of grain boundaries by other low melting point elements, such as sulfur, antimony, tin, and phosphorus. Any of these tests would probably shed more light on the causes of fracture than hydrogen analysis of a broken steel object.

Procuring a Chemical Analysis. If a piece of metal requires analysis and it is important to determine whether the potential chemist has the expertise needed, it may be necessary to ask some questions before selecting the chemist. The key items that should be discussed are the specimen size required, turnaround time, the cost, the accuracy and precision of the analysis and whether the chemist has a proven "digestion" method for any wet analysis being contemplated. The OES and GDS methods generally require a 6 to 13 mm ($\frac{1}{4}$ to $\frac{1}{2}$ in.) diameter flat surface. The common wet analysis methods generally require about 0.5 to 2.5 g (0.02 to 0.09 oz) of material. If the instrument to be used for the analysis is not calibrated with a specimen of the same matrix as the sample to be analyzed, great errors can occur. This is of particular

concern for wet chemical analysis, where the specimens are put into solution before being analyzed. For example, a 15% solution of chromium in water may give a different peak height at the emission wavelength than a 15% solution of chromium with 84% iron would. This has obvious applications for analysis of stainless steels. This type of difficulty, as well as difficulties in getting certain types of second phase particles into solution, are probably two of the biggest concerns that the failure analyst should have when procuring bulk chemical analysis, especially of the less common alloys.

If services are procured from an outside company, the analyst also needs to determine how long the test will take and how much it will cost. For example, consider a hypothetical failure of a heavy duty industrial fan that is supposed to have been fabricated from 316 stainless steel. The analyst calls a laboratory and asks if the lab can verify the composition. The first question may be how large a piece to send. What are the dimensions of the specimens? The second question is why is this analysis requested. Does the analyst just want to know if it meets the specification for 316 stainless? Or does the analyst need to know an actual, very precise average value for some of the elements? In this regard, several possible alternatives might be considered. For example, one alternative is that the purpose is just to know if it is a 316 stainless steel. In this case, the lab representative may suggest analysis by GDS with results available in two days, once the specimen is received.

However, if the analyst is suspicious about the strain-hardening behavior and wants to know the representative analysis on the nickel content, then additional testing may be desirable. In this case, the lab may suggest GDS testing and also inductively coupled plasma (ICP) testing to double-check the nickel content. This additional analysis, of course, will cost more and may take more time. Additional cost and time may also be required if ICP testing is performed for additional elements in addition to nickel. For example, if alloying content requires closer scrutiny, then the lab may suggest an ICP test with carbon and sulfur tested with a combustion method. This may take five days to get results, once the specimen is received.

The preceding example is not intended to show how much a particular cost or turn around time should be for a particular test. The intent is to indicate the trend of increasing cost and time for the more definitive methods and to show what subjects should be covered if both parties are to have a good chance of being satisfied by the transaction. Laboratories vary widely in how much they charge and how long they take to analyze a component. Some labs may run some alloys only once a week, for example, because of the time it takes to calibrate the instrument. They prefer to calibrate for fifteen similar alloy specimens than to calibrate five times for three similar alloy specimens.

Format of a Bulk Metal Composition Analysis Report. The following table is an example of a suggested format of a bulk metal composition analysis report. Note that the chemist has flagged (with an asterisk) the element (in this case, Mn) that does not meet the given specification:

Element (symbol)	**Actual % (by weight)**	Specified range
Carbon (C)	0.43	**0.37–0.43**
Manganese (Mn)	0.57*	**0.35–0.55**
Phosphorus (P)	0.007	**0.010 max**
Sulfur (S)	0.014	**0.025 max**
Silicon (Si)	0.17	**0.15–0.30**
Aluminum (A1)	0.0015	**Not specified**
Chromium (Cr)	**1.32**	**1.15–1.40**

Many labs will double-check the concentration of an element that falls outside the specified range before reporting that the material does not meet the specification. Some laboratory accreditation boards require that the element be checked with a different test method the second time, if one is available. Note also that the elements are listed in the order that they would be listed in the applicable specification. This makes it easy to follow. Many chemical analysis laboratories follow the less-easy-to-interpret practice of listing the elements in the order in which they were tested or in an apparently random order.

When evaluating composition, it should be noted that standard specifications permit a composition range, or in some instances a maximum value for a given element. Slightly exceeding nominal limits for a given element may not be as critical a potential root cause as to exceeding the limit of a maximum value element, for example, phosphorus.

References cited in this section

1. R. Lessard, Chemical Spot Testing, *Metals Handbook Desk Edition*, American Society for Metals, 1985, p 33•49–33•51

2. Nondestructive Rapid Identification of Metals and Alloys by Spot Test, STP 550, ASTM, 1974

Debbie Aliya, Aliya Analytical

Microchemical Analysis in Failure Analysis

Often failure analysis work requires microchemical analysis to find evidence of contamination, to evaluate the composition of microphases revealed by a metallographic specimen or at a grain boundary exposed by a fracture surface, or to evaluate corrosion products. Important information on specimen handling is provided at the end of this article. Failure to properly protect the specimens at all phases of the failure analysis can make correct interpretation of the results of this type of test impossible.

Backscattered Electron Imaging. Most scanning electron microscopes have a useful feature called backscatter electron mode. While the secondary electrons, the normal imaging mode, provide topological information about the specimen, the backscatter mode gives information about topology as well as composition. But the user should not switch over entirely to backscattered images. The images are less sharp than those that can be obtained by the secondary imaging. It is useful to use the two imaging modes to show differences between unknown materials. Figure 1 shows some flaky metallic contaminant that was of an unknown composition. Particle brightness appears similar in the image from the secondary electron imaging (SEI) mode of Fig. 1(a), while there are clearly two categories of particles (brighter and darker) in the image from the backscatter mode in Fig. 1(b). The electrons impinging on the specimen interact with the specimen in several ways. Higher-atomic-number elements “reflect” or backscatter more electrons than lower-atomic-number elements. This makes higher-atomic-number–rich areas of a specimen appear brighter.

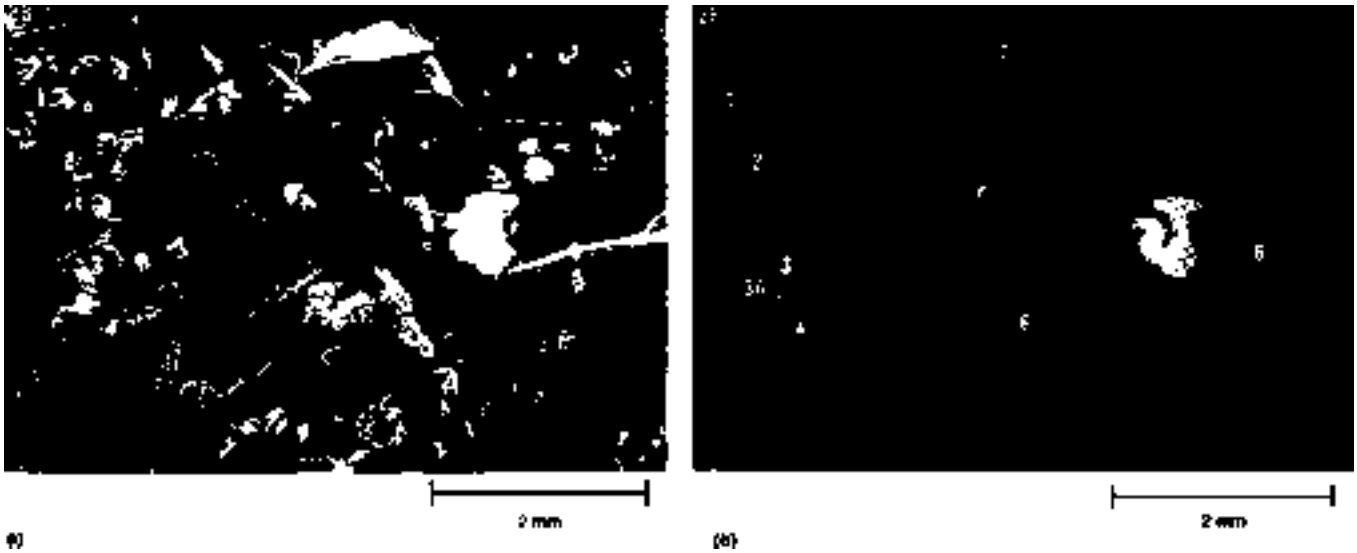


Fig. 1 SEM image of flaky metallic contaminants determined to be a mix of steel and aluminum particles by backscatter mode image. (a) Secondary electron imaging mode where particles are not distinguished by composition. (b) Image from backscatter mode, which shows the lower atomic number particles (in this case aluminum 1, 2, 3, 4, 5, 6, and 8) as darker than the iron particles (3A and 7). Note also that the fibers seen in (a) are invisible in this backscatter imaging mode.

The SEM backscatter mode is often useful for rough composition analysis. The backscattered image in Fig. 1(b) provides a fast and economical method of getting a general idea of how many particle types are present. The actual elemental composition was determined with the EDS microchemical analysis feature for several particles independently. The subsequent EDS analysis showed that the fibers (which are invisible in the backscatter image) were mainly composed of carbon and oxygen, with significantly lower atomic numbers than the aluminum (dark flakes) or the steel (bright flakes). The particles and fibers in both photos are held in place by a special double stick tape whose only EDS detectible element is carbon.

Energy-Dispersive Spectrometry Microchemical Analysis. The EDS feature of a scanning electron microscope is an even more powerful tool for microchemical analysis work. The EDS method works when electrons from the SEM analyzing

beam knock electrons out of their orbits around the atom. Electrons that are in higher-energy orbitals have a tendency to fill in the empty orbitals. The excess energy that cannot be carried into the closer orbital is released from the atom by means of a photon. Many different elements emit photons of energy levels detectable by the EDS detector.

Some elements emit photons of several particular energy levels. Figure 2 is a typical scan of a piece of some organic material. The vertical scale represents the number of photons detected during the analysis time. This scan was obtained using an EDS unit with a “low element” detector. The scan has also been corrected to remove the background. Many older EDS systems leave a differently shaped curve, even if no detectable elements are present. The use of the low element detector allows the carbon and oxygen to be detected. Zinc and sodium are indistinguishable in trace amounts with this test method, so the analyst has labeled the peak in question with both possibilities. The palladium (Pd) was used to sputter coat the specimen prior to putting it in the SEM to make it conductive and suitable for analysis. Other people use graphite or gold. Graphite has the obvious limitation that if you are looking for carbon, it confuses the issue. Gold also has many peaks that are at the same positions as other elements that are often of interest to the analyst. The unlabeled peak at the far left is the internal calibration peak for the EDS unit.

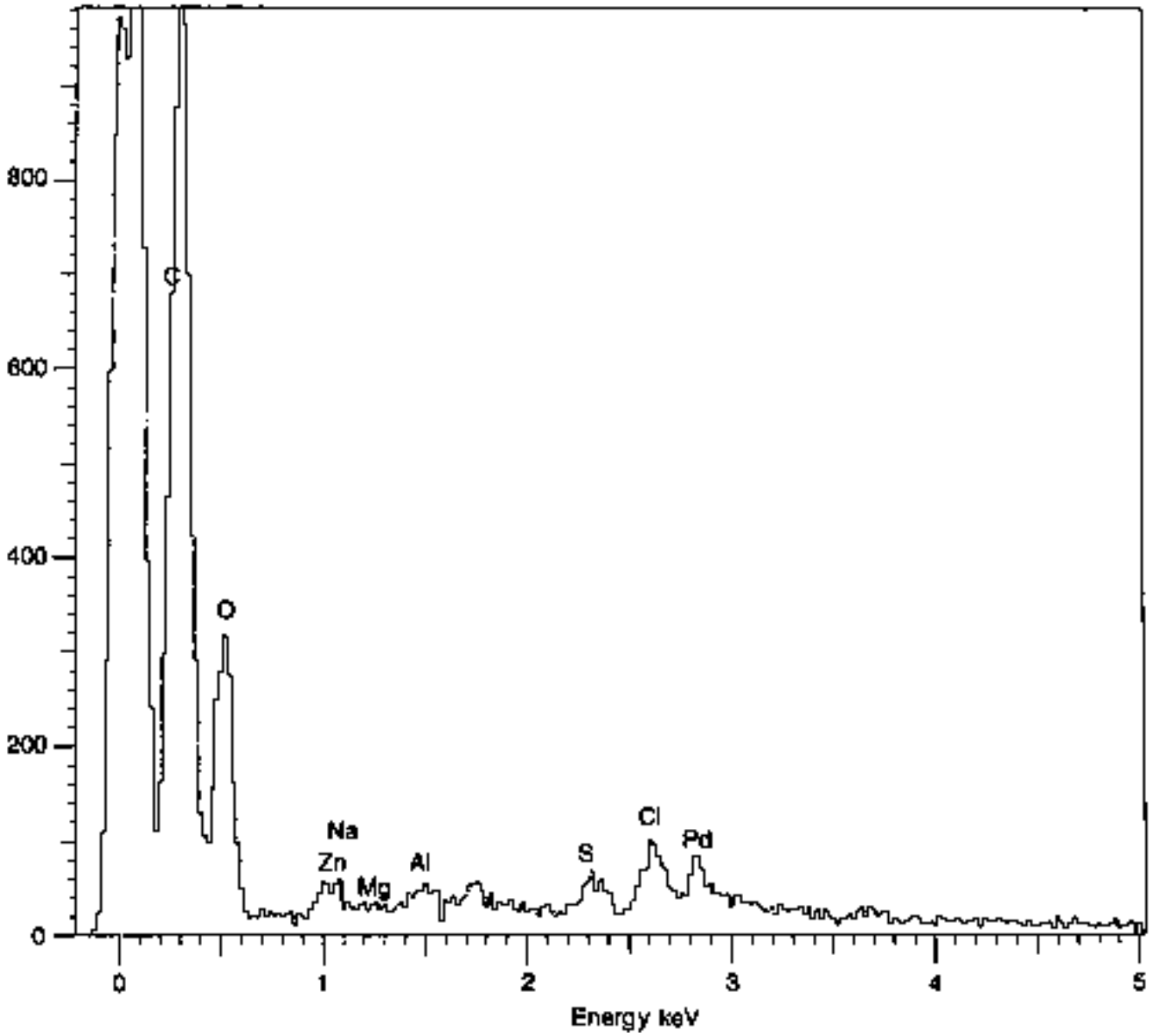


Fig. 2 Typical EDS scan of a polymeric hydrocarbon material

Most EDS units now available have computers to help the analyst determine the elements present, based on the energy levels of the emitted photons. Examination of the scan illustrates a significant point. Energy-dispersive spectrometry analysis only tells what elements are present. It gives no definitive information about molecule types. Hydrogen is not detectable at all. This means that polyethylene, polypropylene, and polystyrene will all look the same as paraffin and many liquid oils, because only the carbon will show up.

If photons of more than one energy level are emitted by a given type of atom, then the ratio of the number of photons at each energy level is constant, as long as the test conditions (in particular, accelerating voltage) are constant. Figures 3(a) and (b) show the scans for the bright and dark flakes shown in Fig. 1(b). The scans match to iron and aluminum, respectively. In both plots, the vertical scale is exaggerated to show the trace elements more clearly, and thus the tops of the iron peak (Fig. 3a) and aluminum peak (Fig. 3b) are cut off. The darker flakes have peak emission rate at an energy level that may indicate the presence of bromine or aluminum. Since bromine is not usually solid, aluminum is the likely contributor. In addition, bromine has a second peak at another energy level, and because no second peak was found, even when analyzed at high accelerating voltages, the flakes appear to be aluminum.

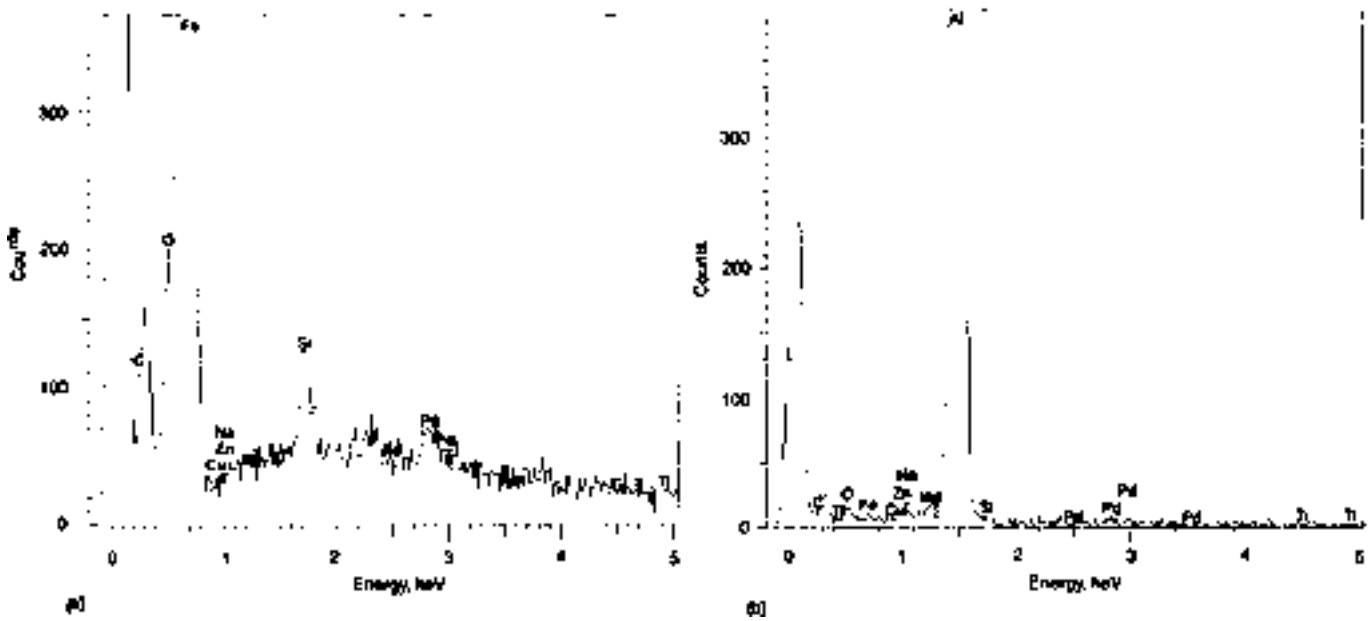


Fig. 3 EDS scans from particles detected in Fig. 1 showing composition differences. (a) Portion of EDS scan obtained from one of the (bright) iron based particles in Fig. 1(b). Some traces of carbon are also seen here, probably an oily residue on the flake. The silicon may be due to a dirt contaminant. (b) Portion of the scan obtained from one of the (dark) aluminum based particles. The unlabeled peak at the far left of both plots is the internal calibration peak for the EDS unit.

If small peaks are being sought while additional elements are present, especially if corrosion products are involved, it may be difficult to decide whether the peak is due to aluminum or a trace of bromine that may have contaminated the specimen. This is because the smaller peaks at the higher energy levels may not protrude above the baseline electronic noise. This indicates another important point about EDS analysis. It is not foolproof. The computer will give you an answer to every question that you ask it. However, you must have some understanding of chemistry to be able to interpret the results. “Murphy's Rules of Microchemical Analysis,” a humorous attempt to prevent the beginner from making some of the more tempting errors of inexperience, states, “The probability of detecting argon decreases with the experience of the analyst.” Since the traditional SEM and EDS units operate in a relatively high vacuum, it is indeed unlikely that argon will be present. Yet the computers list its emission energy at 2.957 keV.

Figure 4(a) shows a typical scan obtained from a piece of steel. Note that the main element present is iron with three peaks. There is also a trace of carbon detected, probably from oil on the surface of the part. There is no nitrogen present at a detectable level, and the location where it would have been found was labeled to emphasize this in Fig. 4(a). A small trace of manganese is sometimes detectable with the EDS but was not detected in this case. While most steels have specified levels of carbon, sulfur, phosphorus, silicon, and other elements, the EDS cannot usually detect these elements, as they are too dilute.

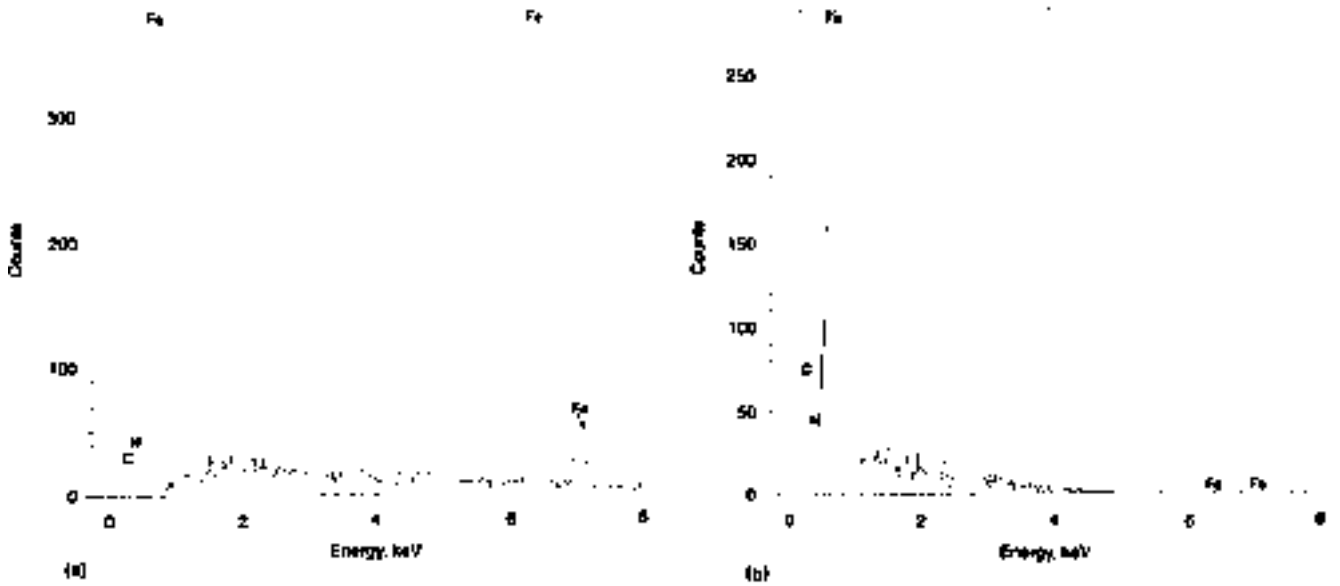


Fig. 4 Plot from EDS scan of low carbon steel sheet at (a) 15 keV and (b) 5 keV. The high energy iron peaks (above 5 keV) are missing in the spectrum in (b) produced from 5 keV electrons. The carbon peak is also higher in Fig. 4(b), suggesting a trace of carbon, probably from oil, on the surface of the part. The unlabeled peak at the far left is the internal calibration peak for the EDS unit.

Figure 4(b) shows another scan obtained from the same piece of steel but with a lower accelerating voltage for the electrons in the SEM. Note that the height ratios of the iron peaks have changed. This illustrates another important feature of EDS analysis. It is necessary to know what the analysis conditions are. In this case, the difference between the two scans is accelerating voltage of the impinging electrons. A voltage of 15 keV is a common test mode for high quality imaging on electrically conductive materials. The scan in Fig. 4(b) was produced with an accelerating voltage of only 5 keV. Note that the two larger energy peaks at approximately 6.3 keV are absent. A 5 keV electron can never allow a 6.3 keV photon to be emitted. Only the lower energy photons at 0.70 keV can escape from the specimen.

The lower accelerating voltage also means that the electrons will not penetrate as deeply. So if the layers nearest the surface are of interest, lower accelerating voltages may be helpful. For example, very thin oxides may be undetectable at 15 keV, because most of the signal will come from the subsurface layers. If the oxide layer is only 0.1 μm thick, it will be much easier to detect when the beam penetrates to approximately 0.4 μm than when it penetrates to 1 μm.

It is also important to point out that EDS analysis is considered to be a semiquantitative method. Part of the reason for this is the fact that the light elements are either not detectable at all or not detectable unless they are present in very concentrated amounts. Most of the atom types under atomic number 22 (titanium) need to have at least a few percent in the area being penetrated by the analyzing beam in order to produce a detectable signal. Heavier atoms may be detectable in lower concentrations, depending on what other elements are present. So if the specimen is oil with some fine bits of steel along with some suspended silica particles, carbon, iron, silicon, and oxygen may be detected if a modern “low-element” detector is used. If a less modern detector is used, only iron and silicon will be detected.

In any case, even the semiquantitative analysis is only valid for the heavy elements. In this case, the analysis might give iron 85% and silicon 15%. The detector will force the total to add to 100%. Even this ratio can be useful in some cases. However, the bigger problem with interpreting semiquantitative-analysis results from an EDS is that the specimen size is unknown. It is not a trivial matter to know exactly how deeply the electrons that are causing photon emission are penetrating. The shape of the volume from which the photons are being emitted is not easy to determine either. And, of course, unobserved subsurface second-phase particles may also contribute to the EDS signal.

Microscopically flat and homogeneous specimens can give quite good quantitative results with EDS. Such specimens are rare in failure analysis work. However, analysts working with a limited budget and who are doing SEM work for other reasons may find it informative to run a semiquantitative analysis on an available smooth and clean area of the component adjacent to the fracture. This information could provide a clue, for example, to distinguish 304 stainless from 316 stainless steel, without the added expense of a proper bulk chemical analysis. Especially in routine investigations where the requesting party has documentation showing that the proper material was used, this semiquantitative “microchemical” method can substitute for an economy bulk chemical analysis.

For this reason, analysis of suspected contaminants contained within a crumbly corrosion product on a fracture surface should generally be confined to the strictly qualitative mode. Interpreting the quantitative results from such a test would be very difficult at best.

Wavelength-Dispersive Spectrometry Microchemical Analysis. In modern x-ray microanalysis, the analyst actually has a second (less common in commercial laboratories) choice for the x-ray spectrometer: the wavelength-dispersive spectrometer, or WDS. The main difference between the EDS and WDS microprobe techniques is the method of detecting the x-rays. The original design by Castaing utilized a wavelength-dispersive spectrometer and was the model for the first commercial systems (see the article "Electron Probe X-Ray Microanalysis" in *Materials Characterization* Volume 10 of *ASM Handbook*. These systems consisted of an x-ray detector and single-crystal diffractometers that moved synchronously around the sample to detect the various wavelength x-rays (Fig. 5). This configuration is excellent for the quantitative analysis of flat samples but is awkward and time-consuming to set up and is difficult to use on irregular topography, such as a fracture surface. With the advent of semiconductor devices, a new type of detector was developed that, with no moving parts, could simultaneously count and determine the energy of the x-rays (Fig. 6). These detectors are the ones used for EDS modules on SEM devices.

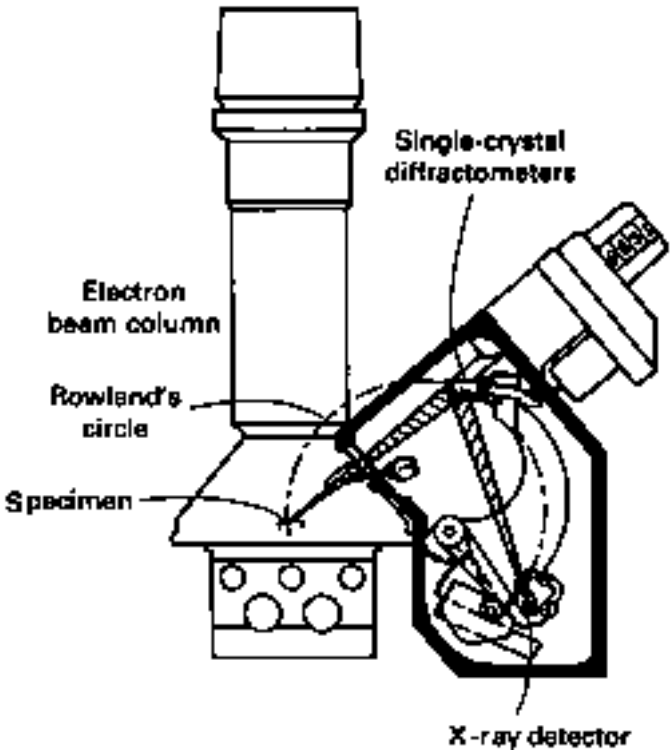


Fig. 5 Schematic diagram of the components of a wavelength-dispersive x-ray spectrometer

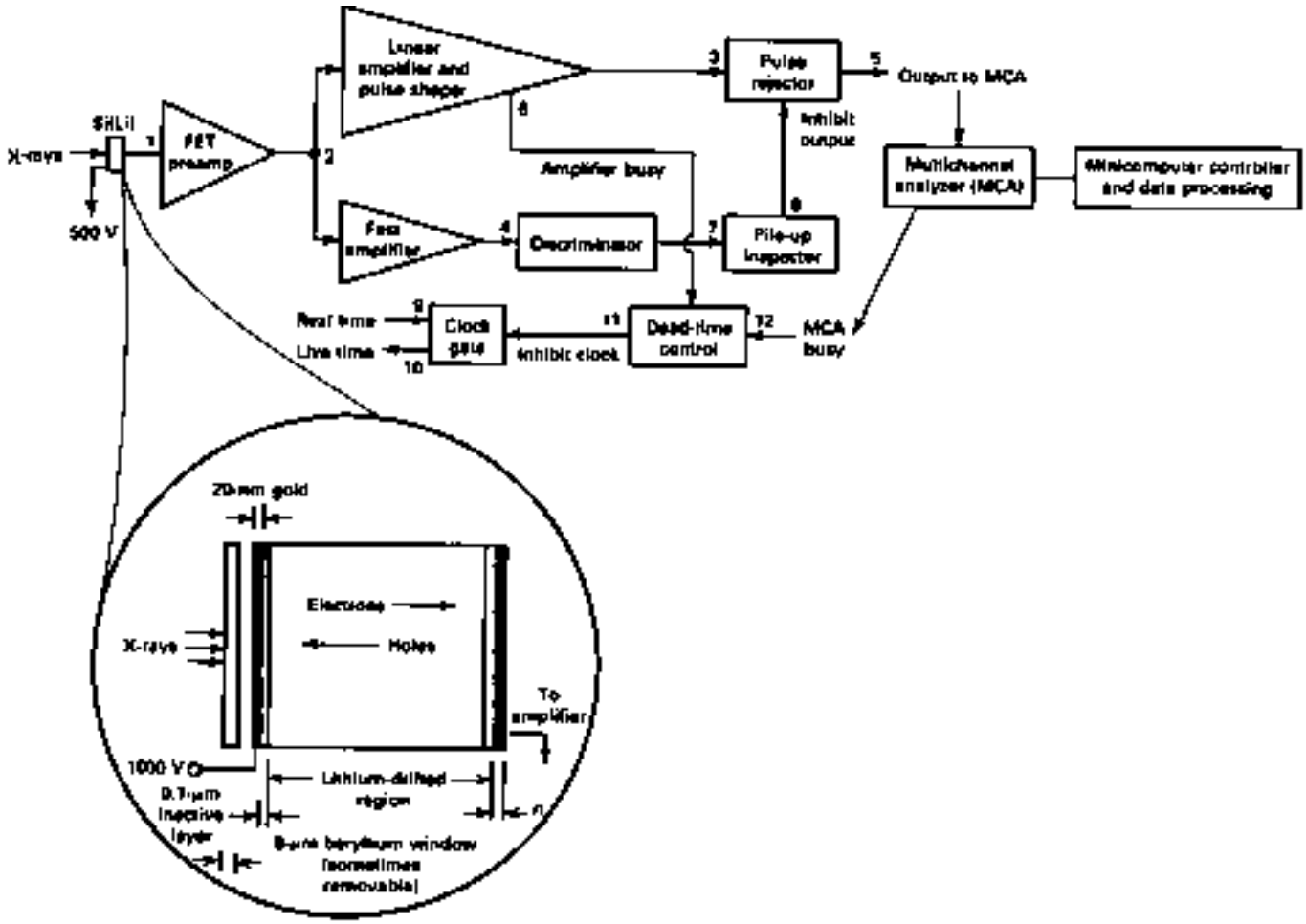


Fig. 6 Schematic of a complete energy-dispersive x-ray spectrometer used in electron-probe x-ray microanalysis. Various pulse processing functions and the multichannel analyzer are shown. FET, field effect transistor

Even though the EDS method is more common, the wavelength dispersive spectroscopy (WDS) method has utility in certain applications because of its ability to distinguish certain elements that are difficult in some cases to distinguish with the EDS method. In particular, sulfur and molybdenum are difficult to distinguish with EDS, as the positions of their peaks are so close to each other. With WDS, the peaks are quite discernible. Differences between EDS and WDS capabilities are described in more detail in the article "Electron Probe X-Ray Microanalysis" in *Materials Characterization* Volume 10 of *ASM Handbook.*

Specimen Handling Information*. The methods discussed in this section on microchemical analysis are often used as surface chemical analysis methods. This is because the same instrumental techniques often deal with microscopic features, trace chemical contamination, and/or thin surface films. Therefore, the analyst must always be concerned with sample handling to be sure that fragile features are not altered before critical information about the surface is obtained.

Specimen collection and storage methods must be planned to prevent the possibility of physical damage or chemical contamination and to ensure that sample integrity is preserved. If specimens for analysis are removed from a larger object, the method for removal should not allow excessive heating or contamination of the specimen area. Cutting without coolants at a distance that does not heat the sample is critical for surface analysis samples, as well as for those for metallurgical analysis. If this is not possible or convenient, acetate tape softened with acetone may be used to remove partially adherent contaminant prior to sectioning in a standard wet metallurgical (abrasive) cutoff saw.

Specimens should be stored in clean containers. This seems an obvious caveat, but even new containers can contain residual chemical contaminant that could be a confusing factor in the failure analysis. For example, mold release agents in new plastic bags can be transferred to a sample and detected later by chemical surface analysis. In addition, specimens should be stored in a cool dry environment to prevent superficial corrosion or other surface contamination prior to analysis.

Even the handling practices for routine fracture analysis work must be considered sources of possible damage to microscopic fracture features or surface chemistry. Poking or probing of fracture surfaces with fingers or tools should be avoided to prevent damage to microscopic features. Furthermore, although it seems to be human nature to reassemble the

pieces created by a fractured component, analysts (and others) must resist the urge to do so. Critical microscopic fracture features can be completely obliterated by this simple act, even on hardened alloy steel components.

Poor sample handling practices are often a problem in surface chemical analysis. Surface sensitive chemical analysis techniques can analyze thin layers only a few atoms thick. Chemical treatments commonly applied by manufacturers to the surface of metal components, such as those used for "corrosion protection," will prevent surface analysis methods from being an effective tool. Oils and perspiration from touching the surface with bare hands (Fig. 7) or spittle (Fig. 8) contaminate surfaces such that accurate analysis for failure evaluation cannot be performed. Residue from adhesive tapes also adds substantial artifact contamination to a surface, which can confound the failure analysis when the chemical nature of the surface is important.

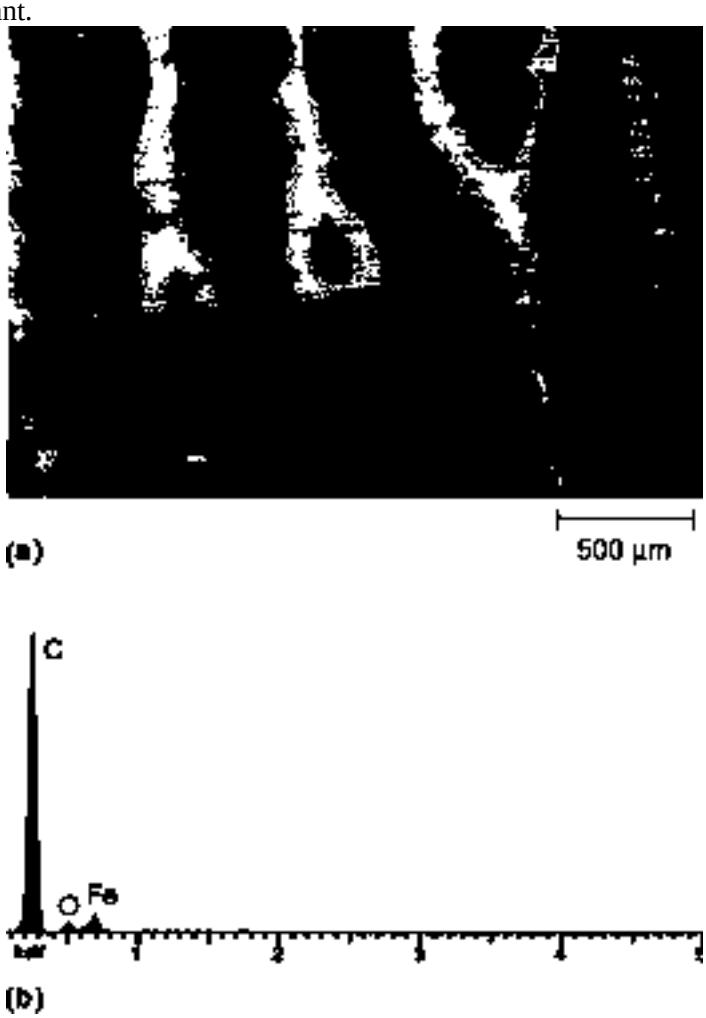


Fig. 7 (a) SEM image and (b) EDS spectrum (10 keV) for fingerprint residue on a clean steel surface. Sometimes other elements, from perspiration, such as chlorine, or potassium and sodium may also appear in fingerprints.

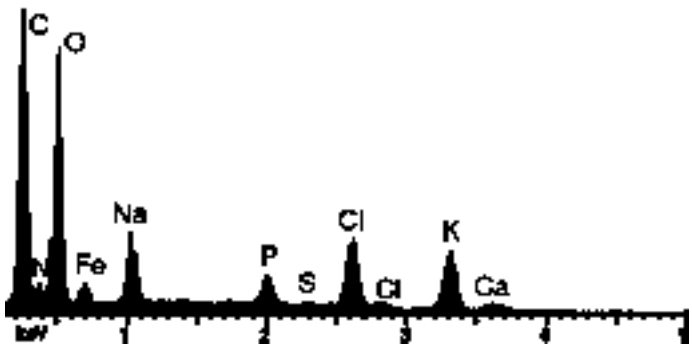


Fig. 8 EDS spectrum (10 keV) for analysis of residue from dried human spittle

It is essential to plan for surface evaluation early in the failure analysis and consider the effect of prior specimen preparation or analyses on possible later tests. Some surface evaluation techniques can cause damage to a surface that will affect the results of subsequent analyses. For example, heat from high intensity lights used for visual examination, photography, or light microscopy can drive off volatile chemical residues from the surface or distort a polymer component. The concentrated energy from an electron beam can distort or melt polymeric materials during SEM examination.

A less obvious example is that examination of a sample by SEM can result in a thin film of hydrocarbon contamination (Fig. 9), as well as a shallow layer of radiation damage that will break chemical bonds in the material struck by the electron beam. These two factors will prevent accurate identification of chemical compounds on the surface by subsequent surface chemical analyses. Therefore, consideration of potential subsequent analyses of a surface should be considered before routinely jumping directly to SEM examination.

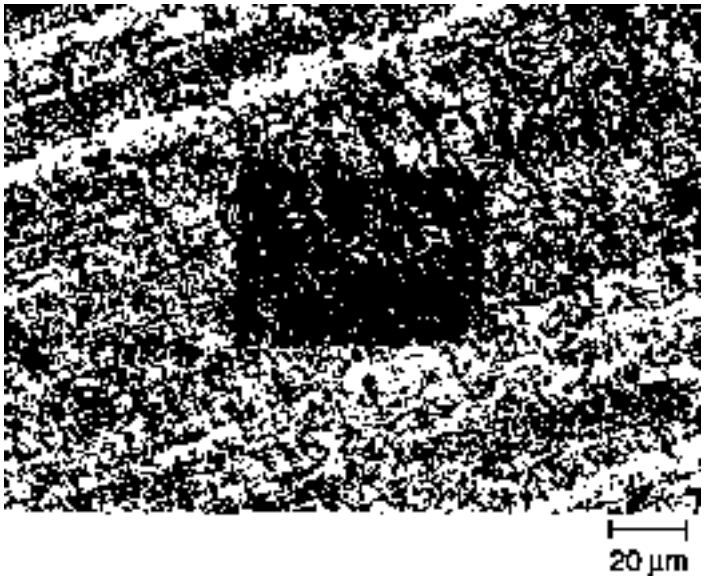


Fig. 9 SEM image showing carbon residue buildup from the electron beam scan area in the SEM

Footnote

* *Larry Hanke, Materials Evaluation and Engineering, Inc., provided this section on specimen handling.

Chemical Analysis of Metals in Failure Analysis

Debbie Aliya, Aliya Analytical

References

1. R. Lessard, Chemical Spot Testing, *Metals Handbook Desk Edition,* American Society for Metals, 1985, p 33•49–33•51

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Introduction

THE ULTIMATE OBJECTIVE of a failure analysis is to ascertain the mode and the cause of the failure, regardless of the material from which the part was fabricated. The investigation is performed in generally the same manner, whether the failed component was produced from metal or plastic or a combination of these materials. Thus, the general steps required to conduct a comprehensive failure investigation are the same, and these are outlined in Fig. 1. In general, the failure analysis process is analogous to putting together a jigsaw puzzle. A failure analysis requires assembling bits of information into a coherent and accurate portrayal of how and why the part failed. Reaching the objectives of the plastic failure analysis, namely, the determination of the mode and cause of the failure, or expressed alternatively, evaluating how the part failed and why it failed, requires a scientific approach and a broad knowledge of polymeric materials.

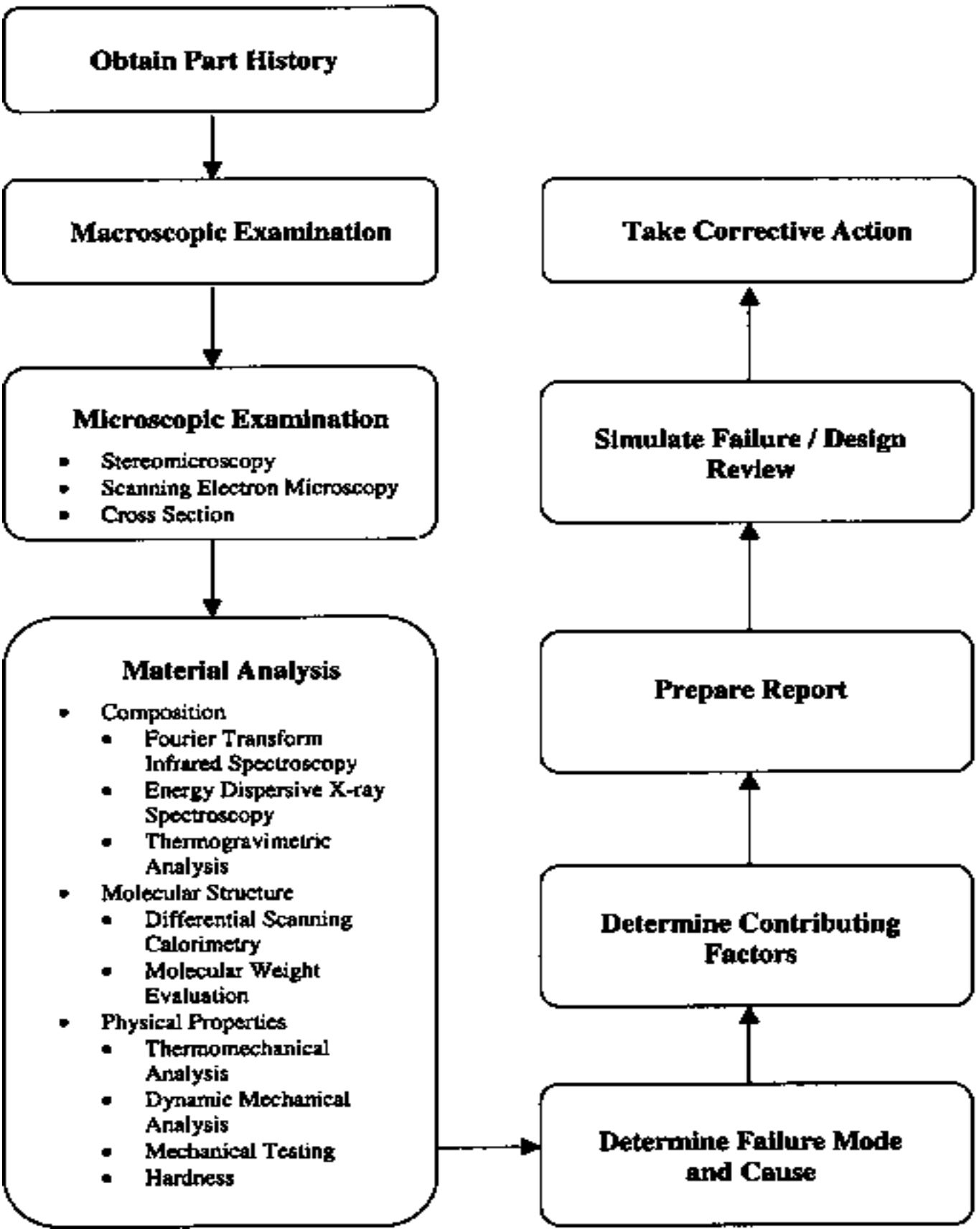


Fig. 1 Steps for performing failure analysis. The steps are the same regardless of the material

Plastic components can fail via many different modes, including catastrophic mechanisms, such as brittle fracture, ductile overload, creep rupture, environmental stress cracking, molecular degradation, and fatigue. In the case of failure involving fracture, the determination of the failure mode involves identifying how the crack initiated and how it subsequently extended. This is usually ascertained using a number of visual-based techniques, such as stereomicroscopy, scanning electron microscopy (SEM), and the preparation of mounted cross sections. Noncatastrophic failure modes are also relevant, and these include discoloration, distortion, and contamination. Assessing the mode of the failure is often not as difficult as establishing why the part failed. Evaluating why the part failed usually requires analytical testing beyond the visual-based techniques. In many cases, a single cause cannot be identified, because multiple integrated factors may have contributed to the failure. All of the factors that affect the performance of a plastic component can be classified into one of four categories: material, design, processing, and service conditions (Ref 1). These factors do not act independently on the component but instead act in concert to determine the performance properties of a plastic component. This is represented graphically in Fig. 2 (Ref 1).

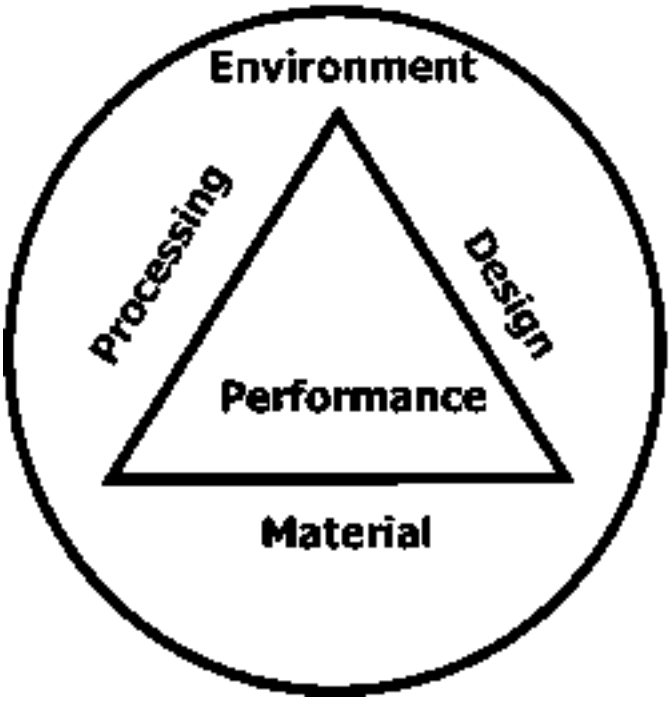


Fig. 2 Graphical representation of the four factors influencing plastic part performance

The principal differences between how failure analyses are performed on metal and plastic materials center on the techniques used to evaluate the composition and structure of the material. Unlike metals, polymers have a molecular structure that includes characteristics such as molecular weight, crystallinity, and orientation, and this has a significant impact on the properties of the molded article. Additionally, plastic resins usually contain additives, such as reinforcing fillers, plasticizers, colorants, antidegradants, and process aids. It is this combination of molecular structure and complex formulation that requires specialized testing (Ref 2). While the chemical composition of a failed metal component can often be evaluated using a single spectroscopic technique, a similar determination requires multiple analytical approaches for a part produced from a plastic resin.

This article reviews those analytical techniques most commonly used in plastic component failure analysis. The description of the techniques is not designed to be a comprehensive review and tutorial but instead is intended to make the reader familiar with the general principles and benefits of the methodologies. The descriptions of the analytical techniques are supplemented by a series of case studies. The technique descriptions refer to the case studies, and the two are written in a complimentary manner to illustrate the significance of the method. The case studies also include pertinent visual examination results and the corresponding images that aided in the characterization of the failures.

References cited in this section

1. J.A. Jansen, Conducting a Plastic Component Failure Investigation: Examples from the Appliance Industry, *International Appliance Technology Conference,* March 2002, p 2

2. J.A. Jansen, Plastic Component Failure Analysis, *Adv. Mater. Process.*, May 2001, p 56, 58, 59

Jeffrey A. Jansen, Stork Technimet, Inc.

Fourier Transform Infrared Spectroscopy

Fourier transform infrared spectroscopy (FTIR) is a nondestructive microanalytical spectroscopic technique that involves the study of molecular vibrations (Ref 2). The analysis results provide principally qualitative, but also limited quantitative, information regarding the composition and state of the material evaluated. Fourier transform infrared spectroscopy uses infrared energy to produce vibrations within the molecular bonds that constitute the material evaluated. Vibrational states of varying energy levels exist in molecules. Transition from one vibrational state to another is related to absorption or emission of electromagnetic radiation (Ref 3). These vibrations occur at characteristic frequencies, revealing the structure of the sample. Fourier transform infrared spectroscopy produces a unique spectrum, which is comparable to the fingerprint of the material. It is the principle analytical technique used to qualitatively identify polymeric materials.

Method

Several different sampling techniques, all involving either transmission or reflection of the infrared energy, can be used to analyze the sample material. This allows the evaluation of materials in all forms, including hard solids, powders, liquids, and gases. Depending on the spectrometer and the corresponding accessories, most samples can be analyzed without significant preparation or alteration. In the analysis of polymeric materials, transmittance, reflectance, and attenuated total reflectance (ATR) are the most common sampling techniques. Additionally, a microscope can be interfaced with the spectrometer to focus the infrared beam and allow the analysis of samples down to 10 μm. Regardless of the sampling technique, the beam of infrared energy is passed through or reflected off of the sample and directed to a detector. The obtained spectrum shows those frequencies that the material has absorbed and those that have been transmitted, as illustrated in Fig. 3. The spectrum can be interpreted manually, or more commonly, compared with voluminous library references with the aid of a computer.

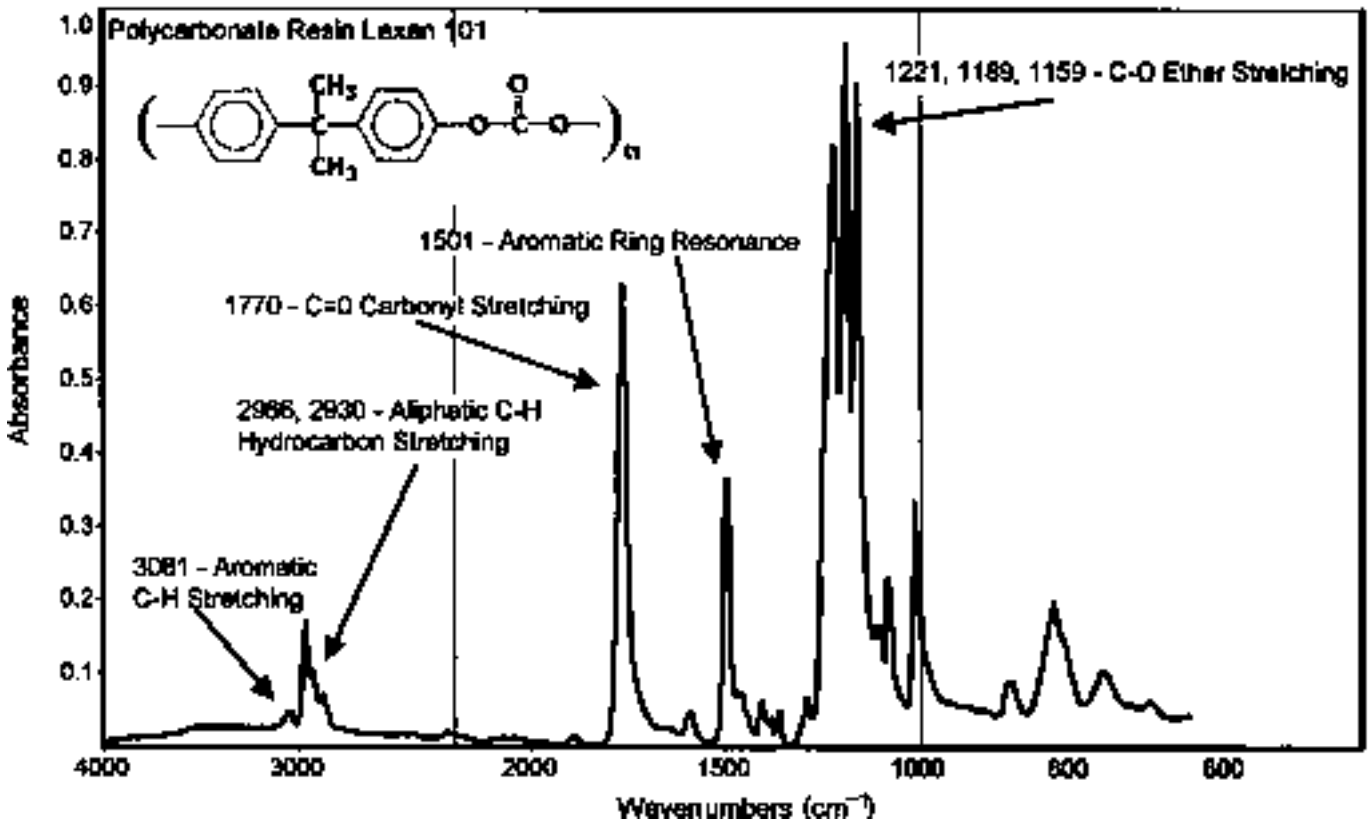


Fig. 3 A typical FTIR spectrum illustrating the correlation between structure and absorption bands

Results

The results generated through FTIR analysis are referred to as an infrared spectrum. The spectrum graphically illustrates the relative intensity of the energy absorbed on the *y*-axis versus the frequency of the energy on the *x*-axis. The frequency of the energy can be represented directly in microns (μm) or, more popularly, as reciprocal centimeters (cm^{-1}) referred to as wavenumbers. The discrete spectral features present with a FTIR spectrum are known as absorption bands.

Uses of FTIR in Failure Analysis

Material Identification. Possibly the most important use of FTIR in failure analysis is the identification of the base polymer used to produce the sample. The determination of the composition of the failed component is an essential part of the investigation. Because different polymers have a wide variation in their physical, mechanical, chemical resistance, and aging properties, the use of the wrong resin can yield detrimental results in many applications. Fourier transform infrared spectroscopy is well suited for the identification of polymers having different molecular structures, and this is illustrated in Fig. 4. Confirming that the failed article was produced from the specified material is the primary consideration of the failure analyst in assessing the cause of the failure. Thus, FTIR is often the first analytical test performed during a plastic failure analysis. The use of FTIR in characterizing the composition of the plastic-resin base polymer is illustrated in examples 1, 4, 7, and 9 in this article. One area where FTIR is inadequate is in differentiating between polymers having similar molecular structures, such as the members of the nylon family, and poly(ethylene terephthalate) and poly(butylene terephthalate). In these cases, other techniques, such as differential scanning calorimetry, must be used to augment the FTIR results.

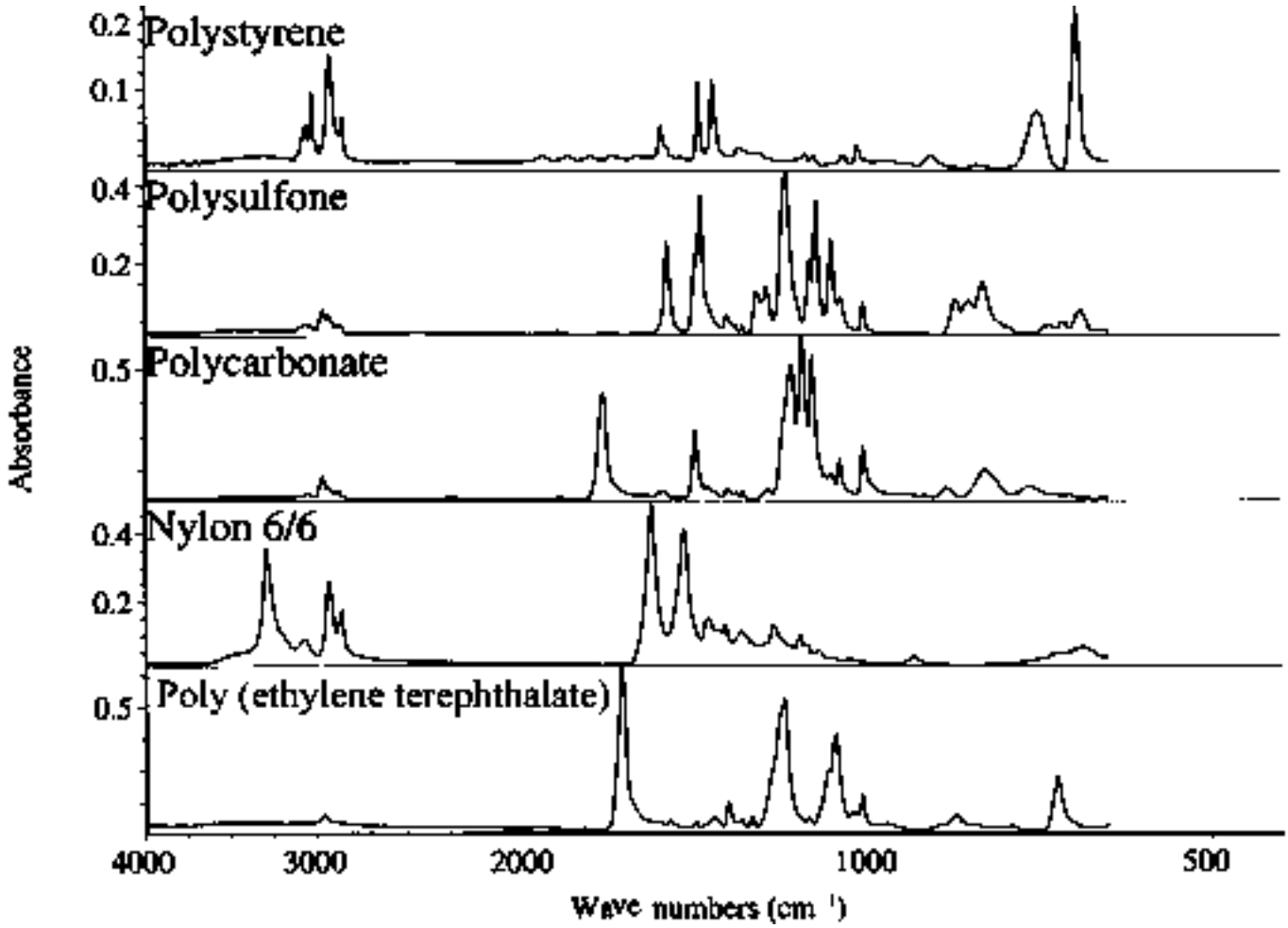


Fig. 4 Fourier transform infrared spectral comparison showing distinct differences between the results obtained on various plastic materials

Aside from the determination of the base polymer, FTIR is used to characterize other formulation constituents. Fourier transform infrared spectroscopic analysis can provide information regarding the presence of additives and filler materials. Due to the nonlinearity of infrared absorptivity of different molecular bonds, it is not possible to accurately state minimum concentration detection limits. However, it is generally considered that materials present within a compounded plastic resin at concentrations below 1% may be below the detection limits of the spectrometer. Given this restriction, it is likely that most major formulation additives, such as plasticizers, can be characterized, while low-level additives, including antioxidants, may go undetected. Given that FTIR is principally used for the analysis of organic materials, its use in the evaluation of inorganic filler materials is somewhat limited. However, some commonly used fillers, such as calcium carbonate, barium sulfate, and talc, produce unique, identifiable absorption spectra. Example 6 in this article shows the analysis of plastic-resin formulation constituents.

Contamination. Similar to its ability to identify the plastic formulation constituents, FTIR is extremely useful in the determination of contaminant materials within the failed part material. While contamination is never an intended part of a plastic compound, its presence certainly can have a tremendous impact on the properties of the molded component. Through the electronic manipulation of the obtained FTIR results, including spectral subtraction, extraneous absorption bands not attributed to the base resin can be used to characterize contaminant materials. Fourier transform infrared spectroscopy is useful in the identification of contaminant material, whether it is mixed homogeneously into the resin or present as a discrete inclusion. The role of FTIR in the identification of contaminants is discussed in examples 3 and 8 in this article.

Degradation. Fourier transform infrared spectroscopy is a valuable tool in assessing a failed component material for degradation, such as oxidation and hydrolysis. Molecular degradation, often involving molecular weight reduction, has a significant detrimental impact on the mechanical and physical properties of a plastic material. This degradation can result from several stages in the product life, including resin compounding, molding, and service. As a polymeric material is degraded on a molecular level, the bonds comprising the material are altered. Fourier transform infrared spectroscopy detects these changes in the molecular structure. While FTIR cannot readily quantify the level of degradation, it is useful in assessing whether the material has been degraded and determining the mechanism of the degradation. Specifically, several spectral bands and the corresponding molecular structure can be ascertained, including carbonyl band formation

representing oxidation, vinylidene group formation as an indication of thermal oxidation, vinyl, vinylene functionality for photooxidation, and hydroxyl group formation indicating hydrolysis (Ref 4). Case studies showing the effectiveness of FTIR in assessing molecular degradation are presented in examples 1, 13, and 15 in this article.

Chemical Contact. Parallel to the application of FTIR in addressing polymeric degradation, the technique is also useful in evaluating the failed sample material for chemical contact. Plastic materials can be affected in several ways through contact with chemical agents. Depending on the polymer/chemical combination, solvation, plasticization, chemical attack, or environmental stress cracking can occur. In the case of property alteration through solvation or plasticization, FTIR can be helpful in identifying the absorbed chemicals. Because these chemicals are present within the failed plastic material, the likelihood of distinguishing the agents is high. Based on the observed spectral changes, mechanisms and chemical agents producing chemical attack, including nitration, sulfonation, hydrolysis, and aminolysis, can be detected (Ref 4). Environmental stress cracking, the synergistic effect of tensile stress while in contact with a chemical agent, is one of the leading causes of plastic failure. The chemical agent responsible for the cracking may be identified using FTIR. However, given that such materials are often volatile organic solvents, the chemical may not be present within the sample at the time of the analysis. Examples 2, 9, and 14 in this article illustrate the identification of chemicals that had been in contact with a failed plastic component.

References cited in this section

2. J.A. Jansen, Plastic Component Failure Analysis, *Adv. Mater. Process.*, May 2001, p 56, 58, 59

3. A.T. Riga and E.A. Collins, Analysis of Structure, *Engineering Plastics*, Vol 2, *Engineered Materials Handbook*, ASM International, 1988, p 825, 826

4. J. Scheirs, *Compositional and Failure Analysis of Polymers*, John Wiley & Sons, 2000, p 109, 138, 153, 393, 415

Jeffrey A. Jansen, Stork Technimet, Inc.

Differential Scanning Calorimetry

Differential scanning calorimetry (DSC) is a thermoanalytical technique in which heat flow is measured as a function of temperature and/or time. The obtained measurements provide quantitative and qualitative information regarding physical and chemical changes involving exothermic and endothermic processes, or changes in the heat capacity in the sample material (Ref 2). Differential scanning calorimetry monitors the difference in heat flow between a sample and a reference as the material is heated or cooled (Ref 5). The technique is used to evaluate thermal transitions within a material. Such transitions include melting, evaporation, crystallization, solidification, cross linking, chemical reactions, and decomposition. A typical DSC result is presented in Fig. 5. Differential scanning calorimetry uses the temperature difference between a sample material and a reference as the raw data. In the application, the instrumentation converts the temperature difference into a measurement of the energy per unit mass associated with the transition that caused the temperature change. Any transition in a material that involves a change in the heat content of the material can be evaluated by DSC (Ref 5), the limitation being that commercially available equipment may not be able to detect transitions within materials that are present at concentrations below 5% (Ref 4).

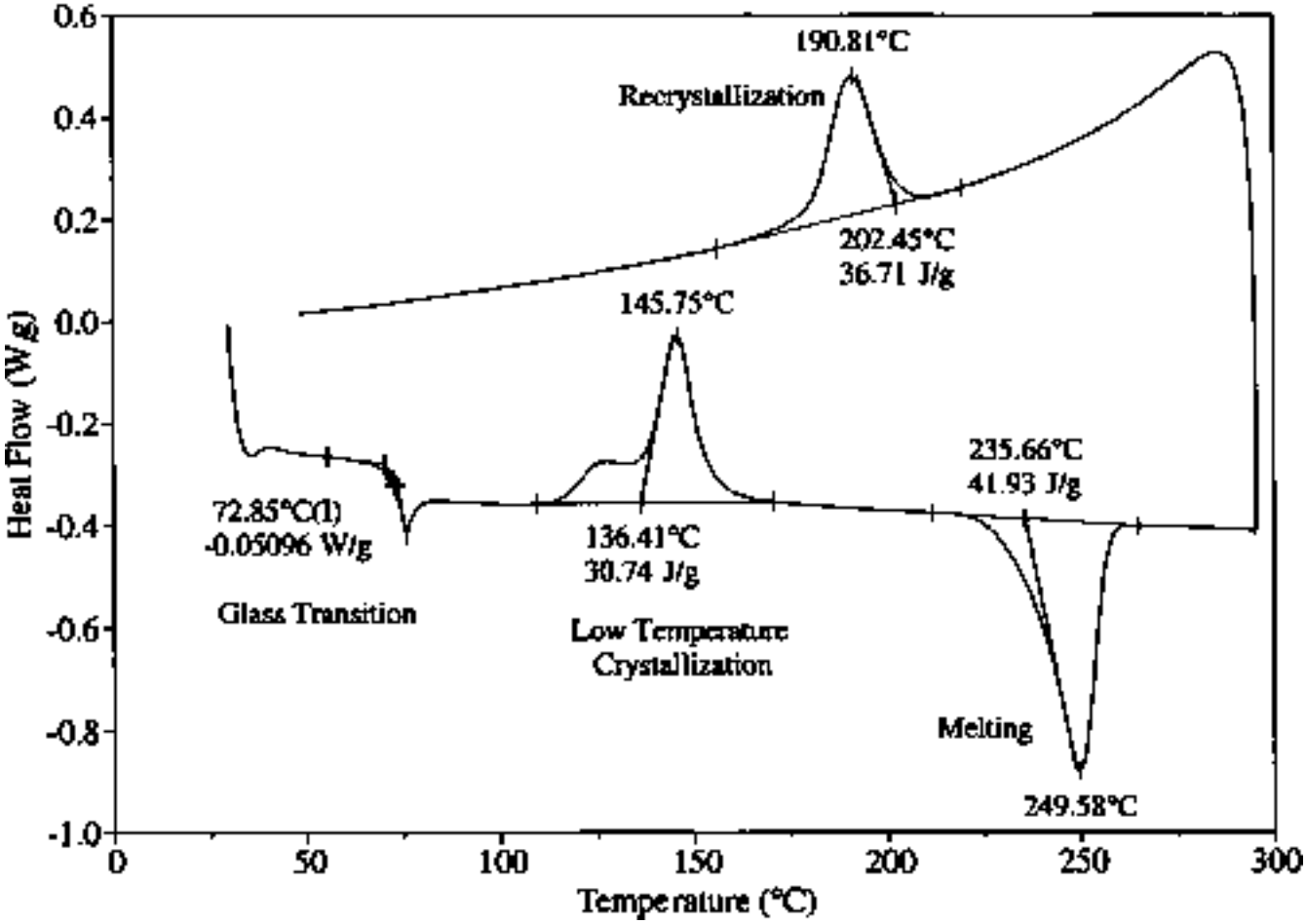


Fig. 5 Differential scanning calorimetry thermogram showing various transitions associated with polymeric materials. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

Method

Sample preparation for DSC analysis includes placing the specimen within a metal pan. The pan can be open, crimped, or sealed hermetically, depending on the experiment. A reference, either in the form of an empty pan of the same type or an inert material having the same weight as the sample, is used. The most commonly used metal pan material is aluminum; however, pans made of copper and gold are used for special applications. The sample and reference pans rest on thermoelectric disc platforms, with thermocouples used to measure the differential heat flow (Ref 5). Specimen size typically ranges between 1 and 10 mg, although this can vary depending on the nature of the sample and the experiment. The normal operating temperature range for DSC testing is -180 to 700 °C (-290 to 1290 °F), with a standard heating rate of 10 °C/min (18 °F/min). A dynamic purge gas is used to flush the sample chamber. Nitrogen is the most commonly used purge gas, but helium, argon, air, and oxygen can also be used for specific purposes. Often, two consecutive heating runs are performed to evaluate a sample. A controlled cooling run is performed after the initial analysis in order to eliminate the heat history of the sample. The first heating run assesses the sample in the as-molded condition, while the second run evaluates the inherent properties of the material.

Results

The plotted results obtained during a DSC analysis are referred to as a thermogram. The thermogram shows the heat flow in energy units or energy per mass units on the *y*-axis as a function of either temperature or time on the *x*-axis. The transitions that the sample material undergoes appear as exothermic and endothermic changes in the heat flow. Endothermic transitions require heat to proceed, while exothermic transitions give off heat.

Uses of DSC in Failure Analysis

Melting Point and Crystallinity. The primary use of DSC in polymer analysis is the detection and quantification of the crystalline melting process. Because the crystalline state of a polymeric material is greatly affected by properties including stereoregularity of the chain and the molecular weight distribution as well as by processing and subsequent environmental exposure, this property is of considerable importance (Ref 5). The melting point (T_m) of a semicrystalline polymer is measured as the peak of the melting endotherm. A composite thermogram showing the melting transitions of several common plastic materials is presented in Fig. 6. The T_m is used as a means of identification, particularly when other techniques, such as FTIR, cannot distinguish between materials having similar structures. This can be useful in identifying both the main resin and any contaminant materials. The material identification aspects of DSC are illustrated in examples 4, 5, 7, 8, 10, 11, 12, and 15 in this article.

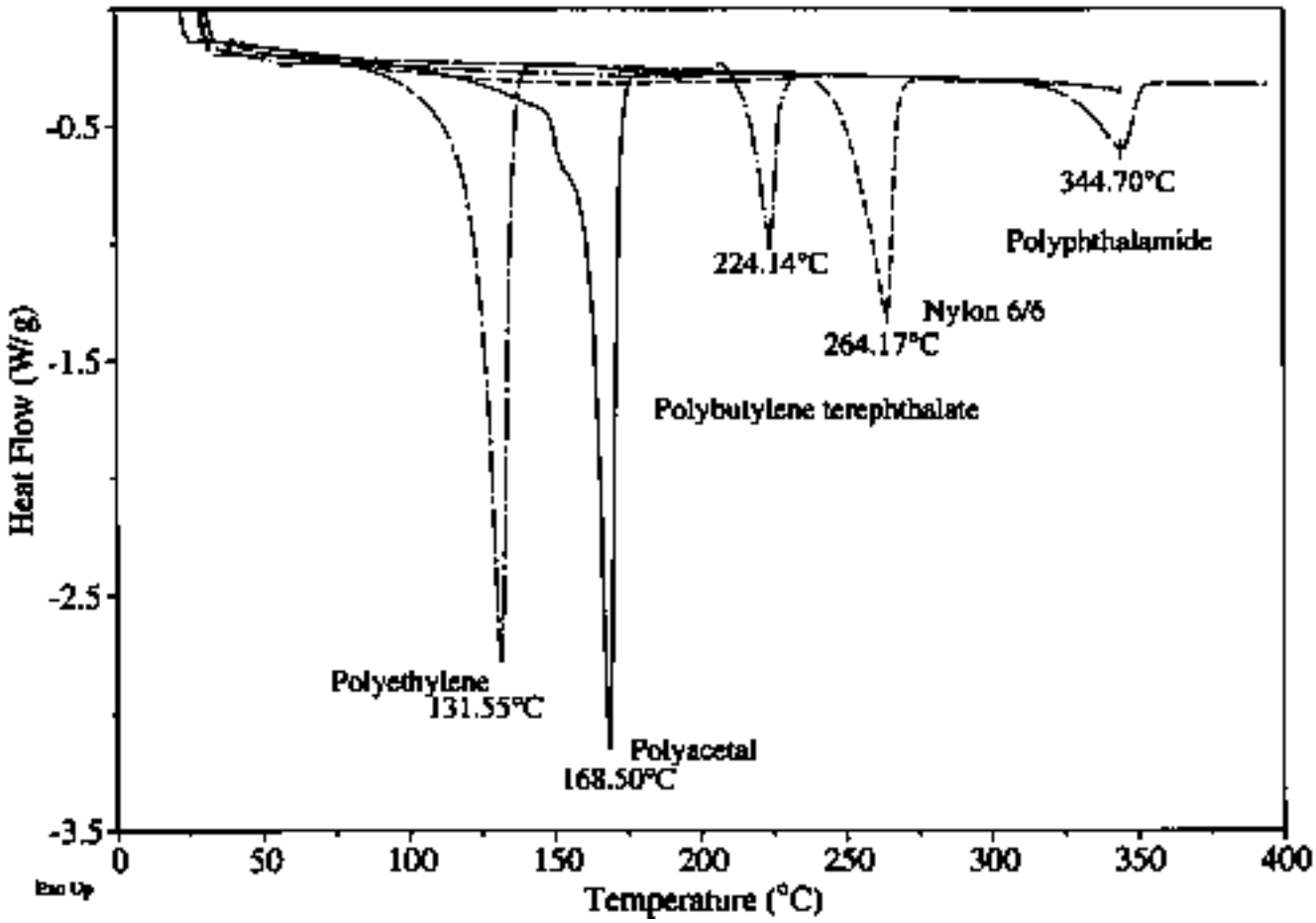


Fig. 6 Differential scanning calorimetry used to identify polymeric materials by determination of their melting point.

The heat of fusion represents the energy required to melt the material and is calculated as the area under the melting endotherm. The level of crystallinity is determined by comparing the actual as-molded heat of fusion with that of a 100% crystalline sample. The level of crystallinity that a material has reached during the molding process can be practically assessed by comparing the heat of fusion obtained during an initial analysis of the sample with the results generated during the second run, after slow cooling. The level of crystallinity is important, because it impacts the mechanical, physical, and chemical resistance properties of the molded article. In general, rapid or quench cooling results in a lower crystalline state. This is the result of the formation of frozen-in amorphous regions within the preferentially crystalline structure. Examples 11 and 12 in this article show applications involving DSC as a means of assessing crystallinity.

Recrystallization, or the solidification of the polymer, is represented by the corresponding exothermic transition as the sample cools. The recrystallization temperature (T_c) is taken as the peak of the exotherm, and the heat of recrystallization is the area under the exotherm. Some slow-crystallizing materials, such as poly(ethylene terephthalate) and polyphthalamide, undergo low-temperature crystallization, representing the spontaneous rearrangement of amorphous segments within the polymer structure into a more orderly crystalline structure. Such exothermic transitions indicate that the as-molded material had been cooled relatively rapidly. Example 9 in this article shows how low-temperature crystallization was detected via DSC.

Glass Transition in Amorphous Plastics. Polymers that do not crystallize and semicrystalline materials having a significant level of amorphous segments undergo a phase change referred to as a glass transition. The glass transition

represents the reversible change from/to a viscous or rubbery condition to/from a hard and relatively brittle one (Ref 6). The glass transition is observed as a change in the heat capacity of the material. The glass transition temperature (T_g) can be defined in several ways but is most often taken as the inflection point of the step transition. A composite thermogram showing the glass transitions of several common plastic materials is presented in Fig. 7. The T_g of an amorphous resin has an important impact on the mechanical properties of the molded article, because it represents softening of the material to the point that it loses load-bearing capabilities.

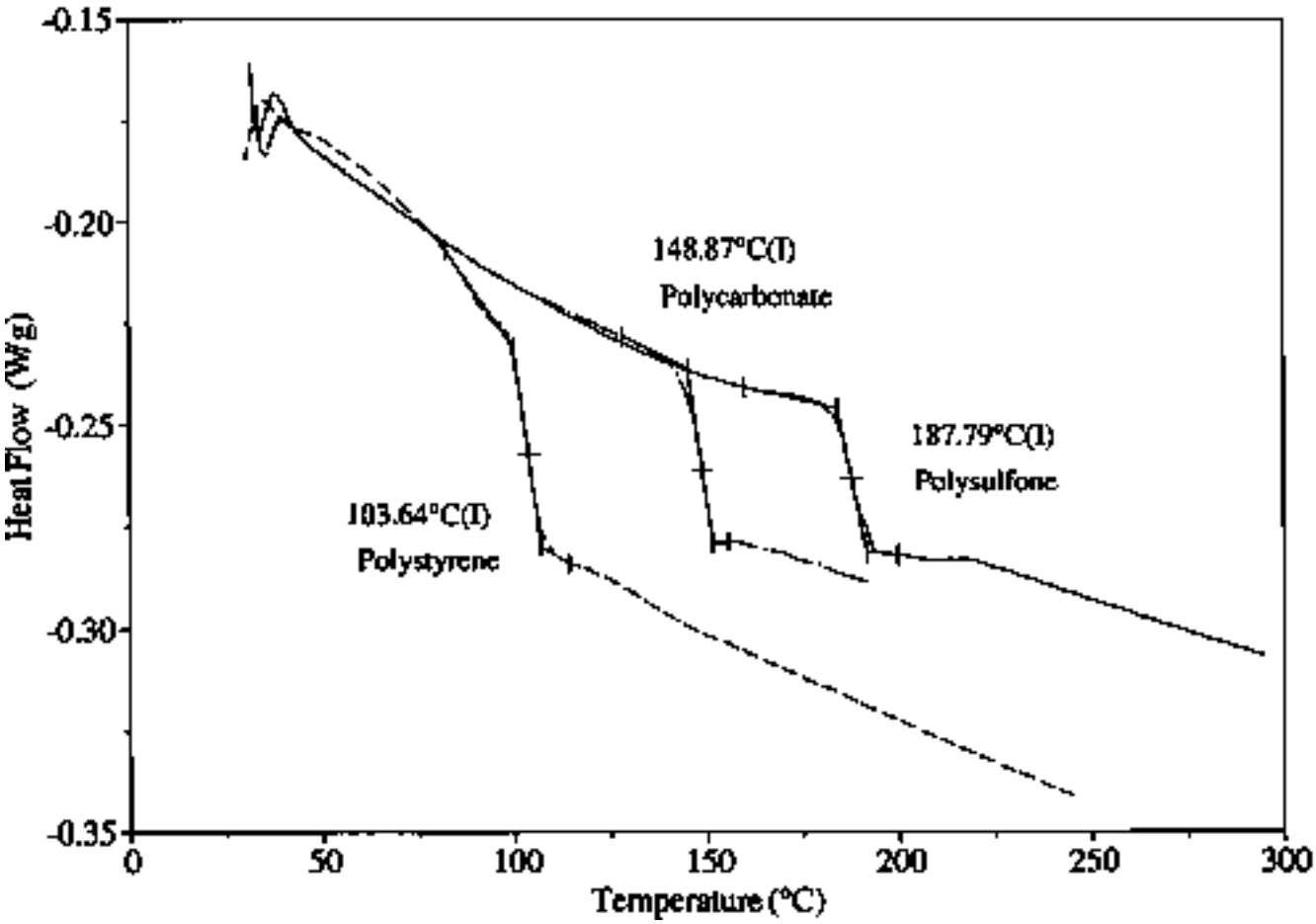


Fig. 7 Differential scanning calorimetry used to detect glass transitions within amorphous thermoplastic resins. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

Aging, Degradation, and Thermal History. As noted by Sepe (Ref 5), "DSC techniques can be useful in detecting the chemical and morphological changes that accompany aging and degradation." Semicrystalline polymers may exhibit solid-state crystallization associated with aging that takes place at elevated temperatures. In some polymers, this may be evident as a second T_m at a reduced temperature. This second T_m represents the approximate temperature of the aging exposure. Other semicrystalline materials may show an increase in the heat of fusion and an increase in the T_m. The thermal aging of both the resin and the failed molded part is illustrated in example 10 in this article.

Amorphous resins exhibit changes in the glass transition as a result of aging. In particular, physical aging, which occurs through the progression toward thermodynamic equilibrium below the T_g, produces an apparent endothermic transition on completion of the glass transition.

Degradation and other nonreversible changes to the molecular structure of semicrystalline polymers can be detected as reduced values for the T_m, T_c, or heat of fusion. Instances of degradation detected by DSC are presented in examples 7, 14, and 15 in this article. Similarly, degradation in amorphous resins can be observed as a reduction in the T_g or in the magnitude of the corresponding change in heat capacity.

Further, the resistance of a polymer to oxidation can be evaluated via DSC by standard methods or experiments involving high-pressure oxygen or air exposure. Such evaluations usually measure the oxidative induction time or the temperature at which oxidation initiates under the experimental conditions. This can be used to compare two similar materials or to determine whether a plastic resin has undergone partial oxidation.

References cited in this section

2. J.A. Jansen, Plastic Component Failure Analysis, *Adv. Mater. Process.*, May 2001, p 56, 58, 59

4. J. Scheirs, *Compositional and Failure Analysis of Polymers*, John Wiley & Sons, 2000, p 109, 138, 153, 393, 415

5. M.P. Sepe, *Thermal Analysis of Polymers*, RAPRA Technology, Shawbury, U.K., 1997, p 3, 4, 8, 17, 19, 22, 24, 33

6. L.C. Roy Oberholtzer, General Design Considerations, *Engineering Plastics*, Vol 2, *Engineered Materials Handbook*, ASM International, 1988, p 21

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Thermogravimetric Analysis

Thermogravimetric analysis (TGA) is a thermal analysis technique that measures the amount and rate of change in the weight of a material as a function of temperature or time in a controlled atmosphere. The weight of the evaluated material can decrease due to volatilization or decomposition or increase because of gas absorption or chemical reaction. Thermogravimetric analysis can provide valuable information regarding the composition and thermal stability of polymeric materials. The obtained data can include the volatiles content, inorganic filler content, carbon black content, the onset of thermal decomposition, and the volatility of additives such as antioxidants (Ref 4).

Method

Thermogravimetric analysis instruments consist of two primary components: a microbalance and a furnace. The sample is suspended from the balance while heated in conjunction with a thermal program. A ceramic or, more often, a platinum sample boat is used for the evaluation. As part of the TGA evaluation, the sample is usually heated from ambient room temperature to 1000 °C (1830 °F) in a dynamic gas purge of nitrogen, air, or a consecutive switch program. The composition of the purge gas can have a significant effect on the TGA results and, as such, must be properly controlled. The size of the sample evaluated usually ranges between 5 and 100 mg, with samples as large as 1000 mg possible. Minimal sample preparation is required for TGA experiments.

Results

The results obtained as part of TGA evaluation are known as a thermogram. The TGA thermogram illustrates the sample weight, usually in percent of original weight, on the *y*-axis as a function of time or, more commonly, temperature on the *x*-axis. The weight-change transitions are often highlighted by plotting the corresponding derivative on an alternate *y*-axis.

Uses of TGA in Failure Analysis

Composition. Thermogravimetric analysis is a key analytical technique used in the assessment of the composition of polymeric-based materials. The quantitative results obtained during a TGA evaluation directly complement the qualitative information produced by FTIR analysis. The relative loadings of various constituents within a plastic material, including polymers, plasticizers, additives, carbon black, mineral fillers, and glass reinforcement, can be assessed. The assessment of a plastic resin composition is illustrated in Fig. 8. This data is important as part of a failure analysis in order to determine if the component was produced from the correct material. The weight-loss profile of the material is evaluated, and, ideally, the TGA results obtained on the material exhibit distinct, separate weight-loss steps. These steps are measured and associated with transitions within the evaluated material. A thorough knowledge of the decomposition and chemical reactions is required to properly interpret the obtained results. In most situations, however, distinct weight-loss steps are not obtained, and in these cases, the results are complemented by the corresponding derivative curve. Noncombustible material remaining at the conclusion of the TGA evaluation is often associated with inorganic fillers. Such residue is often further analyzed using energy-dispersive x-ray spectroscopy (EDS) in order to evaluate its

composition. The use of TGA in characterizing plastic composition is presented in examples 8, 10, 12, and 15 in this article. Additionally, example 11 illustrates the quantification of an absorbed chemical within a failed plastic component.

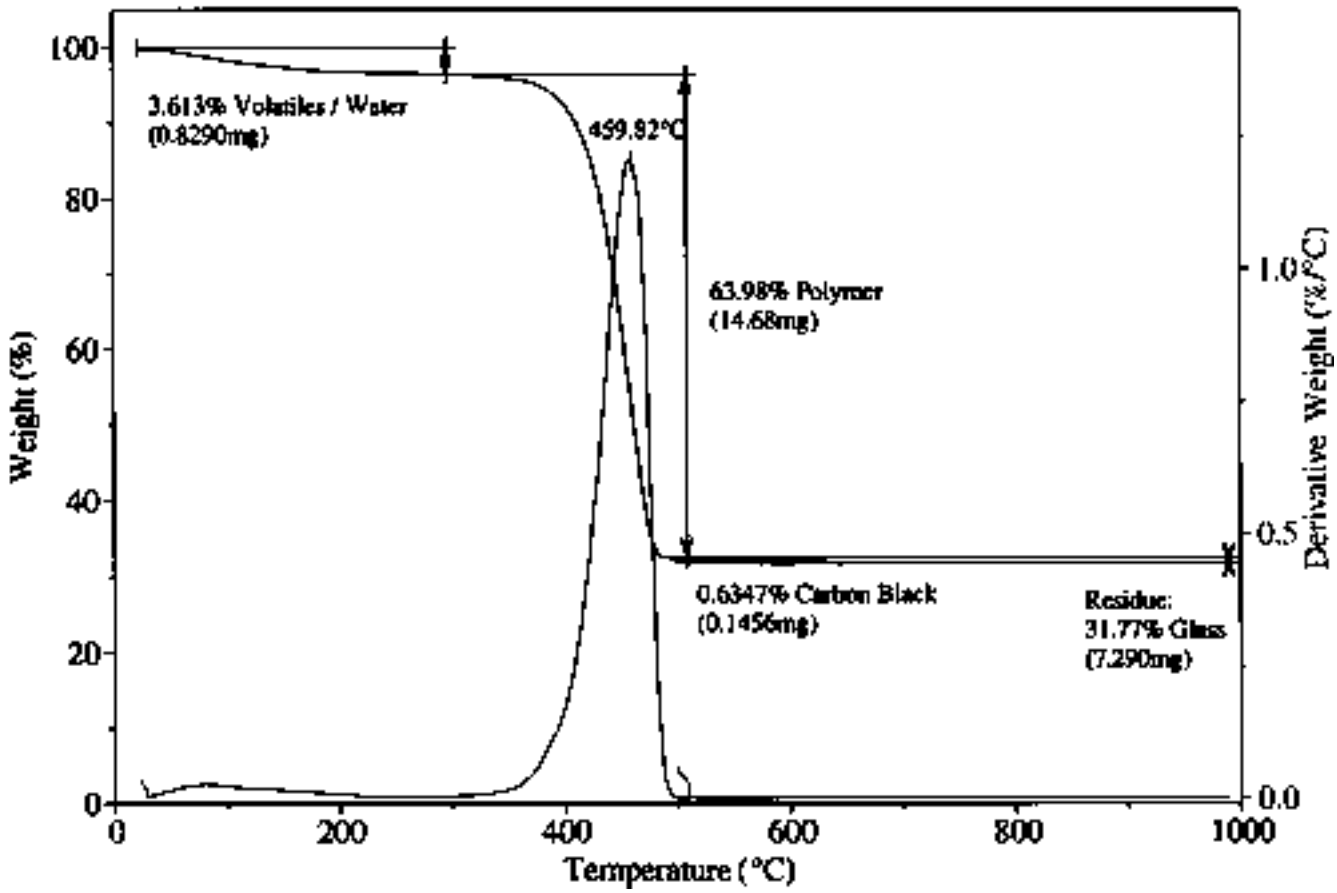


Fig. 8 Thermogravimetric analysis thermogram showing the weight-loss profile for a typical plastic resin

Thermal Stability. Thermogravimetric analysis data can also be used to compare the thermal and oxidative stability of polymeric materials. The relative stability of polymeric materials can be evaluated by assessing the onset temperature of decomposition of the polymer. Quantitatively, these onset temperatures are not useful for comparing the long-term stability of fabricated products, because the materials are generally molten at the beginning of decomposition (Ref 5). However, a comparison of the obtained TGA thermograms can provide insight into possible degradation of the failed component material. Example 6 in this article illustrates a comparison of the thermal stability of two polymeric materials, while example 13 shows the effects of molecular degradation. Degradation experiments involving polymeric materials can also provide information regarding the kinetics of decomposition. Such studies provide information regarding the projected lifetime of the material. Such measurements, however, provide little information pertinent to a failure analysis.

Evolved Gas Analysis. Thermogravimetric analysis evaluations can also be performed whereby the evolved gaseous constituents are further analyzed using a hyphenated technique, such as FTIR or mass spectroscopy (MS). Such TGA-FTIR or TGA-MS experiments are referred to as evolved gas analysis.

References cited in this section

4. J. Scheirs, *Compositional and Failure Analysis of Polymers*, John Wiley & Sons, 2000, p 109, 138, 153, 393, 415

5. M.P. Sepe, *Thermal Analysis of Polymers*, RAPRA Technology, Shawbury, U.K., 1997, p 3, 4, 8, 17, 19, 22, 24, 33

Jeffrey A. Jansen, Stork Technimet, Inc.

Thermomechanical Analysis

Thermomechanical analysis (TMA) is a thermal analysis method in which linear or volumetric dimensional changes are measured as a function of temperature, time, or force (Ref 2). Thermomechanical analysis is used to study the structure of a polymeric material by evaluating the implications of the material dimensional changes.

Method

Standard solid samples evaluated via TMA should be of regular shape, having two flat, parallel sides. Additionally, fiber and film samples can also be tested with minimal preparation. Experiments conducted to evaluate expansion and contraction of solid materials are performed on a quartz stage. The sample is placed on the stage, with a quartz probe resting on the opposing end. Thermomechanical analysis data can be acquired in compression modes, including expansion, penetration, dilatometry, rheometry, and flexure or tension mode (Ref 2). The analysis of film and fiber samples requires special fixturing, similar in principle to a universal mechanical tester. For all analysis configurations, the stage assembly is surrounded by a furnace and a cooling device. The normal operating range for TMA experiments is -180 to 1000 °C (-290 to 1830 °F), with a 5 °C/min (9 °F/min) heating rate commonly used. A compressive force is normally applied to the probe configuration throughout the evaluation for purposes of preload and stability.

Results

Plotted results generated through a TMA analysis, similar to the printed data obtained from all of the thermal analysis techniques, are referred to as a thermogram. The thermogram presents the sample dimension, either as length or a percentage of original length, on the *y*-axis, and as a function of temperature, time, or force on the *x*-axis. Temperature is the standard independent variable. Changes in the sample are presented as expansion or contraction.

Uses of TMA in Failure Analysis

Coefficient of Thermal Expansion. The coefficient of thermal expansion (CTE) is the change in the length of a material as a response to a change in temperature. The derivative of the slope of the line showing the dimensional changes with respect to temperature represents the CTE. This is a significant property when plastic materials are used under highly constrained conditions. This is commonly the case when plastic parts are used in conjunction with components produced from other materials, such as metals and ceramics. In general, the CTEs of polymeric materials are substantially greater than those of metals and ceramics. Thus, comparative testing of mating materials can produce data used to illustrate and even calculate the potential interference stresses on the materials in a multimaterial design. The evaluation of the CTEs of mating plastic and metal components is illustrated in examples 10 and 14 in this article.

Material Transitions. According to Sepe (Ref 5). “The CTE is an important property in itself; however, it is of particular value in polymers, because sudden changes in CTE can signal important transitions in the material structure.” Within semicrystalline polymers, the T_g, signaling the conversion from a hard, brittle material to a rubbery condition, is accompanied by an increase in the CTE. A thermogram representing a typical semicrystalline resin is shown in Fig. 9. The physical properties of the material can be expected to be significantly different across this transition. Amorphous resins soften at the T_g, and because of this, samples undergo compression under the inherent load of the testing conditions. A thermogram showing the glass transition in an amorphous resin is shown in Fig. 10. The evaluation of the glass transition is presented in example 14 in this article. Thermomechanical analysis is generally accepted as a more accurate method for assessing the T_g of polymeric materials, relative to DSC. “By using the prescribed attachments and the appropriate force, TMA can be used to determine two commonly measured properties of plastic materials: the heat-deflection temperature and the Vicat softening temperature.” (Ref 5).

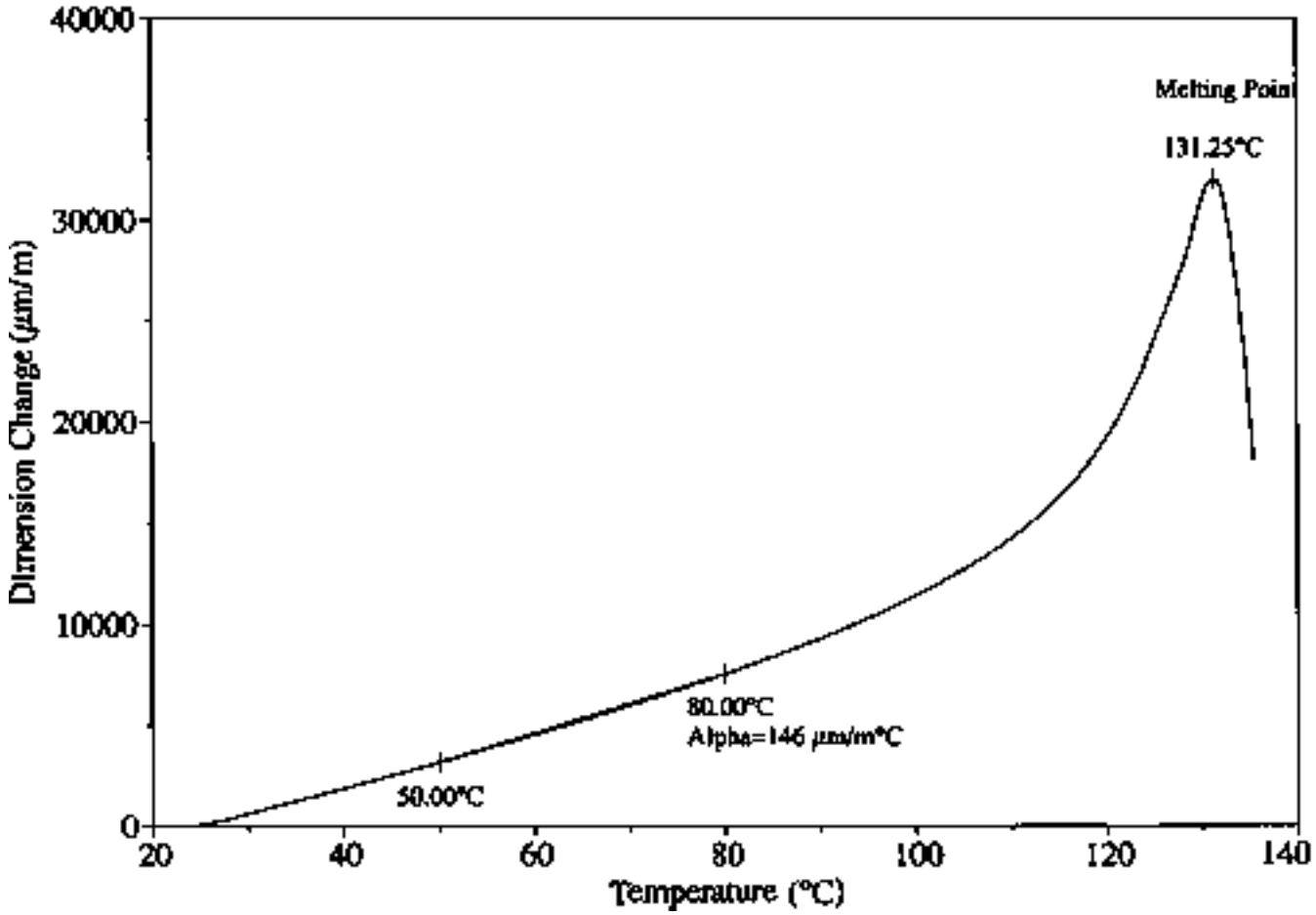


Fig. 9 Thermogravimetric analysis thermogram representing a typical semicrystalline plastic resin

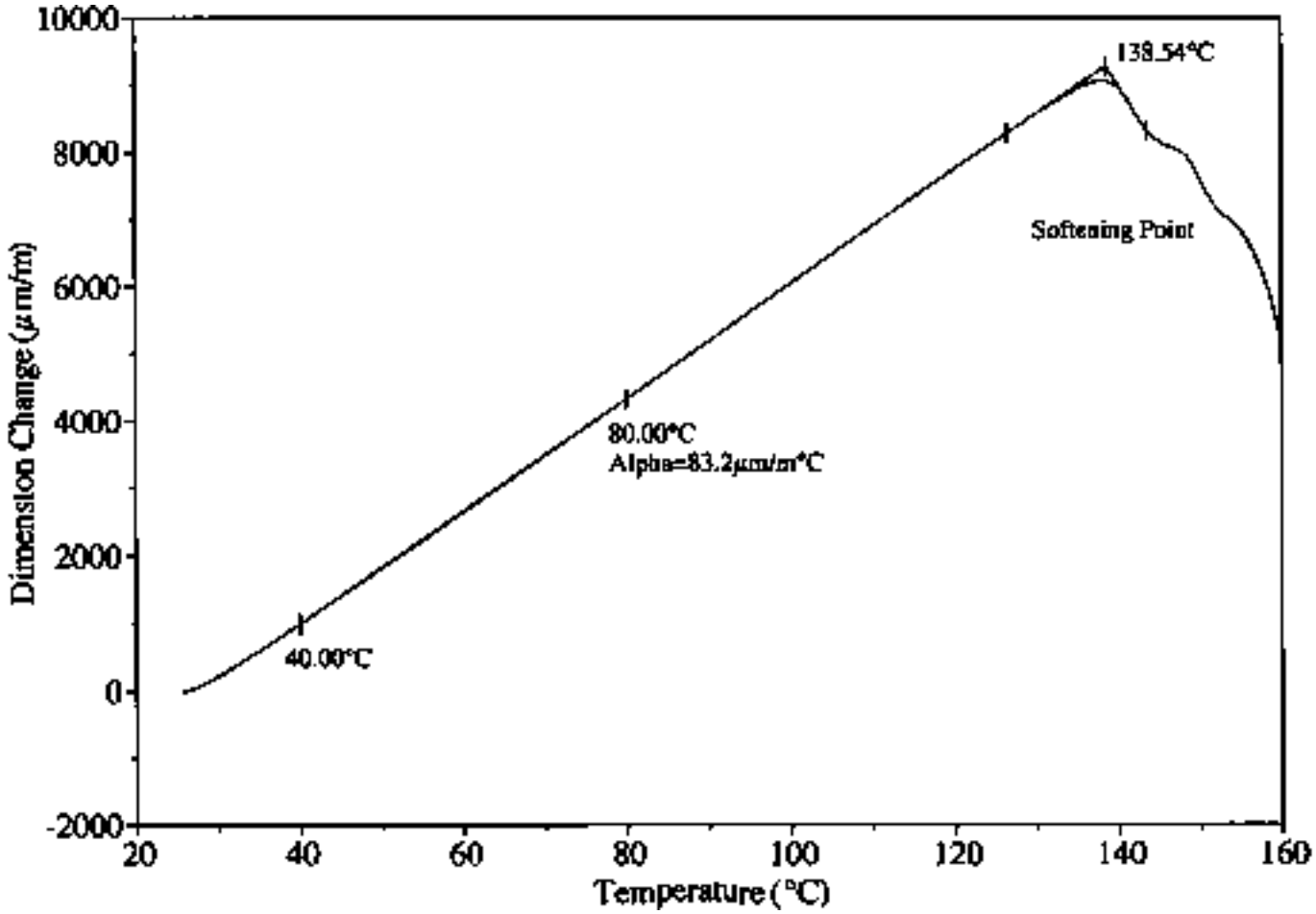


Fig. 10 Thermogravimetric analysis thermogram representing a typical amorphous plastic resin

Molded-In Stress. Internal molded-in stress is an important source of the total stress on a plastic component and is often sufficient to result in the failure of plastic materials. Such stresses are particularly important in amorphous resins, which are prone to environmental stress cracking. Molded-in stresses are commonly imparted through the forming process, especially injection molding and thermoforming. Molded-in stress is observed in amorphous resins as a marked expansion in the sample dimension at temperatures approaching the T_g, as illustrated in Fig. 11. This expansion is associated with rapid expansion as the internal stresses are relieved. This stress relief is due to molecular reorientation on attaining sufficient thermal freedom. In the absence of molded-in stress, the sample would compress due to the loss of load-bearing capabilities as the material undergoes glass transition.

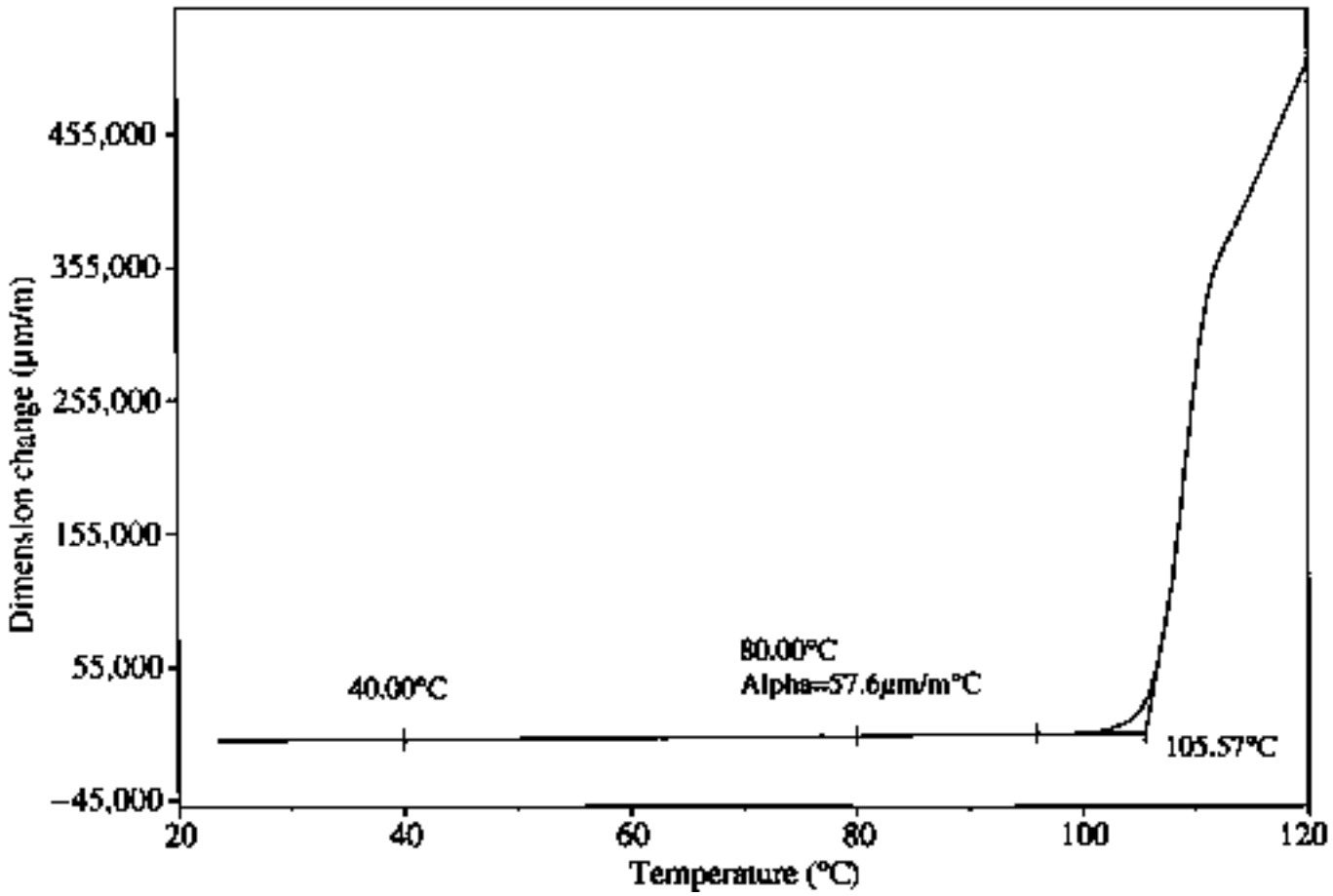


Fig. 11 Thermogravimetric analysis thermogram showing a high level of residual stress in an amorphous plastic resin

Chemical Compatibility. The chemical compatibility of a plastic material with a particular chemical agent can be assessed using TMA. In particular, the volume swell of a polymeric material by a chemical can be tested. The sample material is constrained in a quartz vessel, and the chemical agent is added. Dilatometry is used to measure the volume expansion of the material over time.

References cited in this section

2. J.A. Jansen, Plastic Component Failure Analysis, *Adv. Mater. Process.*, May 2001, p 56, 58, 59

5. M.P. Sepe, *Thermal Analysis of Polymers*, RAPRA Technology, Shawbury, U.K., 1997, p 3, 4, 8, 17, 19, 22, 24, 33

Jeffrey A. Jansen, Stork Technimet, Inc.

Dynamic Mechanical Analysis

Dynamic mechanical analysis (DMA) is a thermoanalytical technique that assesses the viscoelastic properties of materials. Dynamical mechanical analysis evaluates the stiffness, as measured by modulus, as a function of temperature or time. Polymeric materials display both elastic and viscous behavior simultaneously, and the balance between the elastic recovery and viscous flow changes with temperature and time (Ref 5). Measurements can be made in several modes, including tension, shear, compression, torsion, and flexure. The results obtained as part of a DMA experiment provide the storage modulus, loss modulus, and the tangent of the phase-angle delta (tan delta). Dynamic mechanical analysis is not routinely used as a failure analysis technique, but it can provide valuable material information.

Method

Dynamic mechanical analysis experiments can be performed using one of several configurations. The analysis can be conducted to apply stress in tension, flexure, compression, shear, or torsion. The mode of the analysis determines which type of modulus is evaluated. The measurement of modulus across a temperature range is referred to as temperature sweep. Dynamic mechanical analysis offers an advantage over traditional tensile or flexural testing in that the obtained modulus is continuous over the temperature range of interest. In addition, special DMAs can also be conducted to evaluate creep through the application of constant stress or stress relaxation by using a constant strain. Dynamic mechanical analysis studies can be performed from -150 to 600 °C (-240 to 1110 °F), usually employing a 2 °C/min (4 ° F/min) heating rate.

Results

The results obtained as part of a DMA evaluation are plotted to illustrate the elastic or storage modulus (E') and the viscous or loss modulus (E'') on the y-axes and as a function of temperature on the x-axis. Less frequently, time is used, depending on the type of experiment. Additionally, the tangent of the phase-angle delta (E''/E') is also calculated. A typical DMA thermogram is presented in Fig. 12.

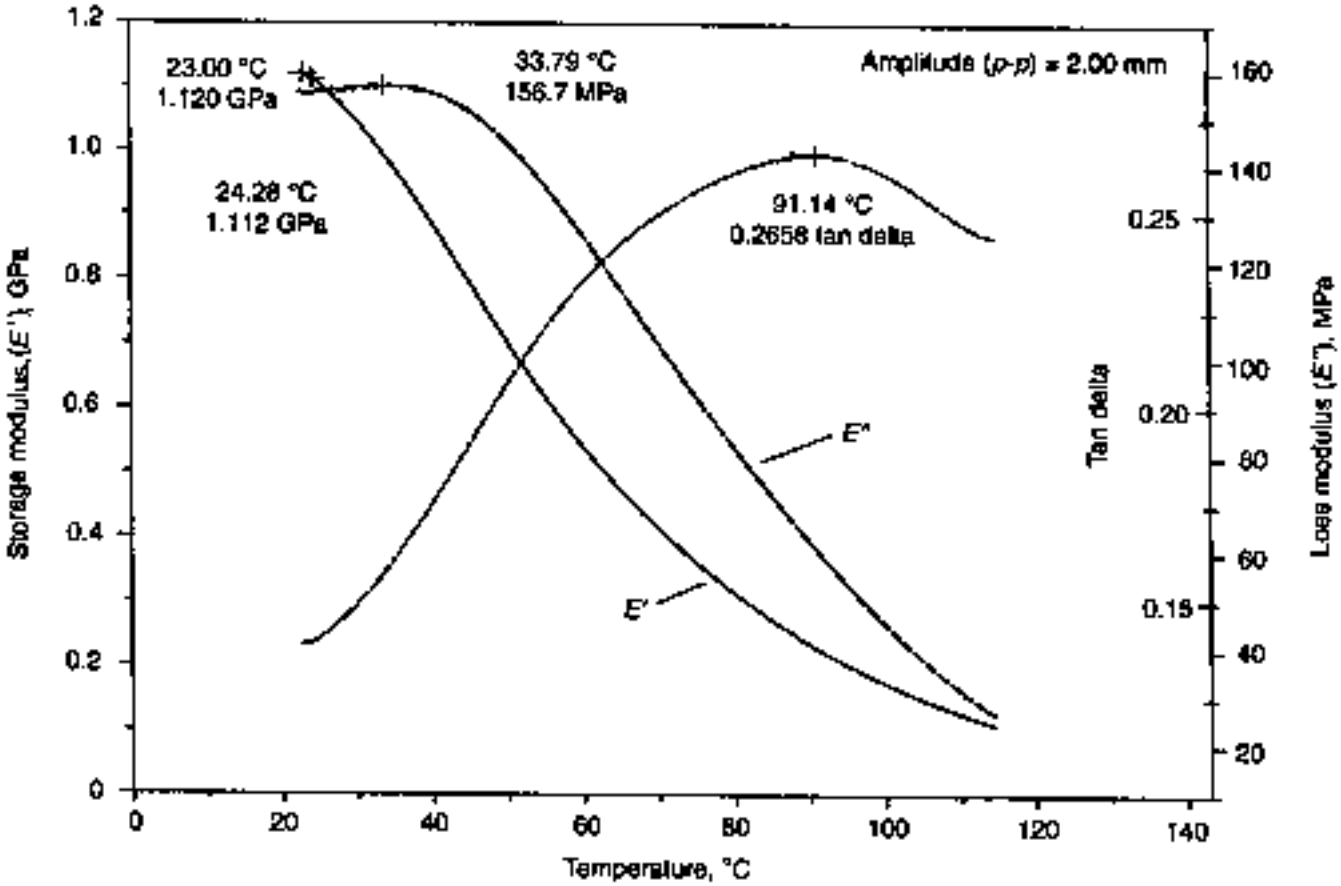


Fig. 12 Dynamic mechanical analysis thermogram showing the results obtained on a typical plastic resin. Tan delta is ratio of the loss modulus to the storage modulus.

Uses of DMA in Failure Analysis

Temperature-Dependent Behavior. The temperature-dependent behavior of polymeric materials is one of the most important applications of DMA. In a standard temperature-sweep evaluation, the results show the storage modulus, loss modulus, and the tan delta as a function of temperature. The storage modulus indicates the ability of the material to accommodate stress over a temperature range. The loss modulus and tan delta provide data on temperatures where molecular changes produce property changes, such as the glass transition and other secondary transitions not detectable by other thermal analysis techniques. The superiority of DMA over DSC and TMA for assessing the glass transition is well documented (Ref 5). Secondary transitions of lesser magnitude are also important, because they can relate to material properties such as impact resistance. The ability of a plastic molded component to retain its properties over the service temperature range is essential and is well predicted by DMA.

Aging and Degradation. Changes in the mechanical properties of plastic resins that arise from molecular degradation or aging can be evaluated via DMA. Such changes can significantly alter the ability of the plastic material to withstand service stresses. While the cause and type of degradation cannot be determined, DMA can assess the magnitude of the changes. This can provide insight into potential failure causes.

Solid and Liquid Interactions. Sepe (Ref 5) notes that “DMA is sensitive to structural changes that can arise when a solid polymer absorbs a liquid material.” This effect can arise from the absorption of water or organic-based solvents. Dynamic mechanical analysis experiments can assess changes in the physical properties of a plastic material that can result from such absorption, including loss of strength and stiffness. Example 11 in this article shows the changes in mechanical properties of a plastic resin associated with chemical absorption. The experiments can also evaluate the recovery after the removal or evaporation of the absorbed liquid.

Reference cited in this section

5. M.P. Sepe, *Thermal Analysis of Polymers*, RAPRA Technology, Shawbury, U.K., 1997, p 3, 4, 8, 17, 19, 22, 24, 33

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Methods for Molecular Weight Assessment

The aspect of molecular structure, and specifically molecular weight, makes polymeric materials unique among materials commonly used in engineering applications, including metals and ceramics. Molecular weight and molecular weight distribution are probably the most important properties for characterizing plastics (Ref 4). These parameters have a significant impact on the entirety of characteristics of a plastic resin, including mechanical, physical, and chemical resistance properties. Molecular weight assessment can be used to evaluate the characteristics of a base resin or to assess the effects of compounding, molding, or service on the material. Changes in molecular weight can occur throughout the material life cycle and can significantly impact the performance of the molded part. Changes can result in molecular weight decreases through such mechanisms as chain scission, oxidation, and hydrolysis, or as increases through destructive cross linking. Because of this, the characterization of molecular weight is an important aspect of a thorough failure analysis.

Gel permeation chromatography (GPC), which is also referred to as size exclusion chromatography, is an analytical method used to characterize the molecular weight distribution of a polymeric material. Similar to all chromatographic techniques, GPC uses a packed column to segregate various constituents. One or multiple columns used in conjunction are used to separate the polymeric and oligomeric materials within the plastic resin. The polymer is further separated by molecular weight, producing essentially a histogram representing the molecular weight distribution of the material. From these results, a numerical average molecular weight can be calculated. Detectors, based on refractive index or ultraviolet detection, are used to identify the changes in molecular weight.

Gel permeation chromatography offers the advantage, unlike melt viscosity and solution viscosity techniques, of producing results that directly represent the actual molecular weight and molecular weight distribution of the plastic resin. Another advantage is that GPC requires a relatively small sample size, 30 to 120 μg, for a complete evaluation (Ref 7). The technique, however, is often complicated to perform, using sophisticated instrumentation, and difficult to interpret. Example 10 in this article reviews the use of GPC in a failure investigation.

Melt Flow Index. The melt flow index or melt flow rate (MFR) describes the viscosity of a plastic material in the molten state. The sample material is heated through the melting or softening point and extruded through a die having a standard-sized orifice. Different materials use various test conditions, including temperature and load. The method for determining the MFR is described in ASTM D 1238. Melt flow rate is the simplest technique for assessing the molecular weight of a plastic material and is inversely proportional to the molecular weight of the polymer (Ref 4). Melt flow rate is widely used to describe the molecular weight of a plastic resin and is commonly cited by suppliers on a material data sheet. The units used to indicate MFR are grams per 10 min. Examples 7, 11, 12, and 14 in this article describe the use of melt flow in assessing molecular weight in a failure analysis.

While MFR is relatively easily determined and is commonly used to describe molecular weight, the technique has several negative aspects. Melt flow rate does not measure the molecular weight distribution of the analyzed material and represents only the average molecular weight of the material. Because of this, the blending of polymers having different molecular weight distributions and average molecular weights can result in equal determinations between very different materials having distinct properties.

Solution Viscosity. The traditional approach for determining only the molecular weight of a resin, but not the molecular weight distribution, involves dissolving the polymer in a suitable solvent. However, the more structurally complicated macromolecules require the use of hostile solvents, tedious sample preparations, and

costly time delays to obtain limited, single datapoint values. For example, the solution viscosity determination of polyvinyl chloride (PVC), according to ASTM D 1243, requires either a 1 or 4% concentration in cyclohexanone or dinitrobenzene, while polyamides, or nylons, require formic acid. Other engineering polymers might require tetrahydrofuran, dimethylformamide, dimethylsulfoxide, or other equally hostile solvents (Ref 8). The obtained solution viscosity values are only indications of molecular weight and do not reflect the absolute weight values (Ref 8). Example 9 in this article illustrates the use of solution viscosity in a failure investigation.

References cited in this section

4. J. Scheirs, *Compositional and Failure Analysis of Polymers*, John Wiley & Sons, 2000, p 109, 138, 153, 393, 415

7. "Polymer Characterization: Laboratory Techniques and Analysis," Noyes Publications, 1996, p 15

8. S.B. Driscoll, Physical, Chemical, and Thermal Analysis of Thermoplastic Resins, *Engineering Plastics*, Vol 2, *Engineered Materials Handbook*, ASM International, 1988, p 533

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Mechanical Testing

Because a wide range of mechanical tests are available to evaluate plastics and polymers, they initially do not seem to constitute a rational set. The totality of mechanical tests can be partitioned into logical groups in several distinct ways (Ref 9). One very useful way to classify the various mechanical test methods is to distinguish between tests that evaluate long-term properties, as opposed to those that evaluate short-term properties. Short-term tests include those that assess what are generally considered to be material properties. These include tensile tests, flexural tests, and the evaluation of impact resistance. Short-term tests, while generally easy to conduct and interpret, lack the ability to predict or assess the long-range performance properties of a material. As such, short-term tests are frequently listed on material data sheets.

Tests for Short-Term Properties. The most commonly performed mechanical test used to evaluate plastic material properties is the tensile test. This testing is performed on a dumbbell-shaped specimen and is outlined in ASTM D 638. Tensile testing provides data regarding the yield point in the form of yield strength and elongation at yield, the break properties as tensile strength at break and elongation at break, and the stiffness of the material as elastic modulus. Additionally, the tensile test generates information regarding the proportional limit.

A second short-term mechanical method that is used to evaluate plastic materials is flexural testing. Flexural testing simulates bending of the test sample. The test specimen is evaluated on a universal mechanical tester, and the tests can be performed using a three- or four-point bend configuration. Flexural testing provides two pieces of data: flexural modulus and break strength. This testing is performed in accordance with ASTM D 790.

Several different types of tests are used to evaluate the impact properties of a plastic material. These include pendulum-based tests, such as Izod and Charpy tests, and falling weight tests, such as the dart penetration configuration. Unlike tensile and flexural testing, the results obtained from impact testing do not provide fundamental material properties. Instead, impact testing results are more performance-based. Given these different methodologies of assessing the impact properties of a plastic material, the falling weight or dart impact tests are generally considered to be superior to the pendulum configurations. Falling weight tests

evaluate the sample material in two dimensions and not one, because the specimen is a plate rather than a beam. The data obtained during an instrumented falling weight impact test include the energy to maximum load, representing the energy required to initiate cracking, and the total energy to failure. The ratio of these two is an indication of the ductility of the material. Additionally, an examination of the test specimens is used to classify the failure mode from brittle to ductile. Falling weight impact testing is described as part of ASTM D 3029.

Tests for Long-Term Properties. Fatigue testing of plastic materials exposes the samples to cyclic stresses in an attempt to evaluate the samples in a manner that would produce fatigue failure while in service. Testing procedures are used to simulate flexural fatigue and tensile fatigue. The analyses are normally conducted in a way that does not excessively heat the specimen, thus altering the failure mode. The results of a fatigue test are shown in the form of a stress-number of cycles curve.

A second long-term test methodology assesses the creep resistance of the material. Creep testing exposes the sample to a constant stress over a prolonged period of time. This is done to simulate the effects of static stresses on the performance of a material in service. The extension or strain of the sample over time is measured. Traditional creep testing can take an extended period of time. Similar results can, however, be obtained through a DMA creep study, which can be performed in the course of a few days.

Mechanical Testing as Part of a Failure Analysis. The use of mechanical testing in a failure analysis is limited. The preparation of specimens from the failed component may not be possible. Further, published standard mechanical data, including yield strength, elastic modulus, and flexural modulus, are very dependent on the specimen configuration and testing conditions. Given that most published data are generated on specially molded test specimens, the testing of samples excised from molded articles may not provide an adequate comparison. In some cases, it is not apparent whether observed differences are the result of material deficiencies or variations in test specimen configuration. Instead, mechanical testing is most useful in comparing a known good or control sample with a failed part. Many times, this is best accomplished through some sort of proof load testing. Proof load testing involves measuring the strength and dimensional changes as a function of an applied load. In most cases, this testing involves producing a catastrophic failure within the test sample. The use of proof load testing as part of a failure analysis is illustrated in example 12 in this article.

Reference cited in this section

9. S. Turner, Mechanical Testing, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 545

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Considerations in the Selection and Use of Test Methods

Through the application of analytical testing and a systematic engineering approach, it is possible to successfully ascertain the nature and cause of a plastic component failure. The testing, however, must be performed in a sound manner, with the obtained data only being as good as the analysis method. Further, the data presented by the analytical methods are often complicated and, in many cases, require an experienced analyst to be properly interpreted.

The aforementioned analytical tests are not meant to be an all-encompassing list of the methods used to evaluate failed plastic components. Certainly, there are numerous testing methodologies that provide data pertinent to a plastic component failure analysis. Other analysis techniques, including EDS and SEM, are important tools in a plastic component failure analysis. Scanning electron microscopy and EDS methods are covered elsewhere in this Volume. More specialized chromatographic methods, including gas chromatography and gas chromatography-mass spectroscopy, are extremely useful in assessing low-concentration additives within a plastic resin. Nuclear magnetic resonance spectroscopy is useful in polymeric analysis, providing information related to composition beyond FTIR. Nuclear magnetic resonance can provide data regarding stereoregularity, carbon content, chemical composition, and copolymer structure (Ref 10). Additionally, surface analysis spectroscopic techniques, such as secondary ion mass spectroscopy, x-ray photoelectron spectroscopy, and electron spectroscopy for chemical analysis, are specifically used to characterize very shallow surface layers. These techniques can be used to analyze material composition but are particularly suited for the analysis of surface contaminants (Ref 10). While these analytical techniques can provide valuable data as part of a plastics failure analysis,

the tests described in this article are considered to be the most important in the majority of cases. Given the charge that this article be treated in a practical manner, test methods used less often were omitted. A summary showing both the treated and omitted analysis methods and the corresponding information gained is included in Table 1 (Ref 3).

Table 1 Practical information derived from polymer analysis methods

Test method	**Properties measured**	Use in failure analysis
Fourier transform infrared spectroscopy (FTIR)	Molecular bond structure	**Material identification, contamination, degradation, chemical contact**
Differential scanning calorimetry (DSC)	Heat of fusion, melting point, glass transition temperature, heat capacity	**Material identification, level of crystallinity, aging/degradation, thermal history**
Thermogravimetric analysis (TGA)	Weight loss over temperature or time	**Composition, thermal stability, evolved gas analysis**
Thermomechanical analysis (TMA)	Dimensional changes over temperature	**Coefficient of thermal expansion, material transitions, molded-in stress, chemical compatibility**
Dynamic mechanical analysis (DMA)	Elastic modulus, viscous modulus, tan delta	**Temperature-dependent behavior, aging/degradation, solid-liquid interactions**
Gel permeation chromatography (GPC)	Weight-average molecular weight, molecular weight distribution	**Degradation, suitability of material for use**
Melt flow rate (MFR)	Melt viscosity	**Degradation, compliance with material specification**
Solution viscosity	Intrinsic viscosity	**Degradation**
Mechanical testing	Strength and elongation properties, modulus	**Compliance with material specification, mechanical properties**
Scanning electron microscopy (SEM)	Surface and particle morphology	**Fracture mode**
Energy dispersive x-ray spectroscopy (EDS)	Elemental concentrations	**Material composition, fillers, additives**
Nuclear magnetic resonance (NMR)	Molecular bond structure	**Material identification**
Mass spectroscopy (MS)	Molecular structure	**Material identification, additives**
X-ray photoelectron spectroscopy (XPS)	Elemental concentrations	**Chemical composition of surfaces**
Auger electron spectroscopy (AES)	**Elemental concentrations**	**Chemical composition of surfaces**

Source: Ref 3

References cited in this section

3. A.T. Riga and E.A. Collins, Analysis of Structure, *Engineering Plastics*, Vol 2, *Engineered Materials Handbook*, ASM International, 1988, p 825, 826

10. M. Ezrin, Plastics Analysis: The Engineer's Resource for Troubleshooting Product and Process Problems and for Competitive Analysis, *Plast. Eng.*, Feb 2002, p 45, 46

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Case Studies

Example 1: Embrittlement of a Polycarbonate Bracket. A plastic bracket exhibited relatively brittle material properties, which ultimately led to catastrophic failure. The part had been injection molded from a medium-viscosity polycarbonate resin and had been in service for a short duration prior to the failure.

Tests and Results. A visual examination of the bracket revealed a series of surface anomalies, and it was suspected that the presence of the defects was related to the premature failure. The component base material was analyzed using micro-FTIR in the attenuated total reflectance (ATR) mode. The obtained spectrum exhibited absorption bands characteristic of polycarbonate, as shown in Fig. 13. No evidence of material contamination was found. A similar analysis was performed on the part surface in an area that showed the anomalous surface condition. The spectrum representing the surface was generally similar to the results obtained on the base material. However, the surface spectrum showed a relative increase in the intensity of a spectral band between 3600 and 3350 cm^{-1}, indicative of hydroxyl functionality. Additionally, the spectrum also showed changes in the relative intensities of several bands, as compared to the results representing the base material. A spectral subtraction was performed, and the results produced a good match with a library reference of diphenyl carbonate. This is illustrated in the spectral comparison presented in Fig. 14. Diphenyl carbonate is a common breakdown product produced during the decomposition of polycarbonate.

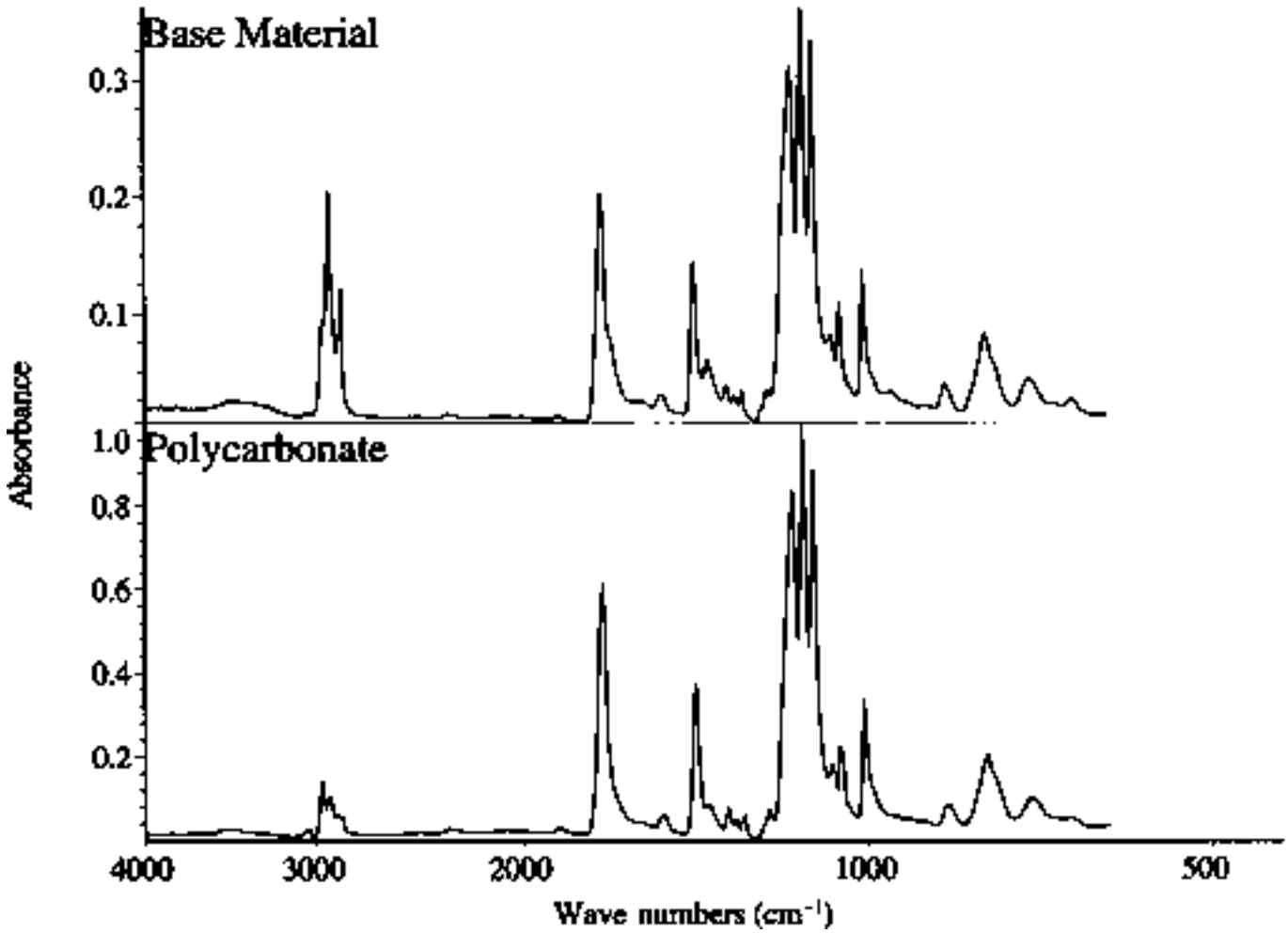


Fig. 13 The FTIR spectrum obtained on the bracket base material, exhibiting absorption bands characteristic of polycarbonate

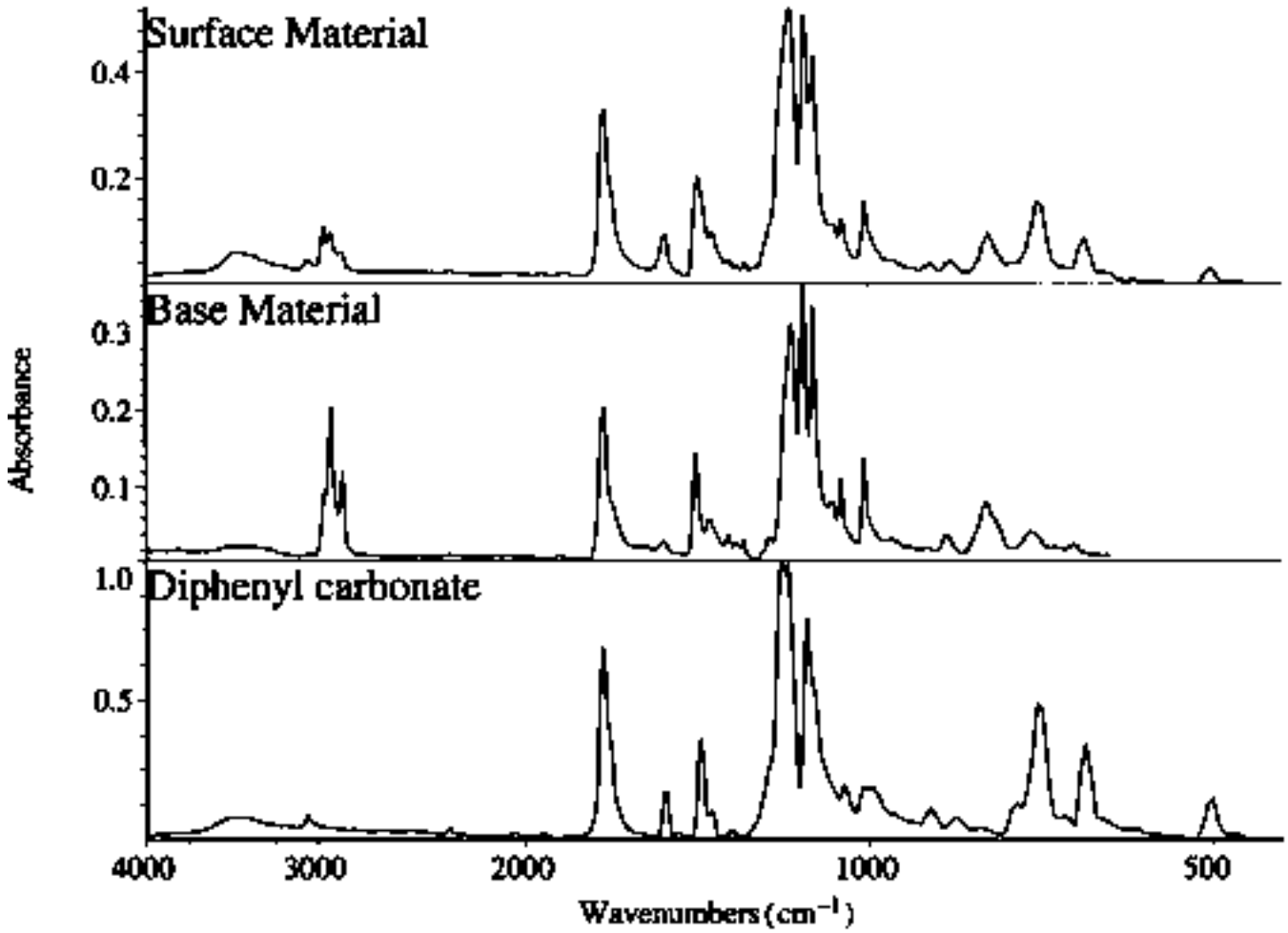


Fig. 14 Spectral comparison showing differences between the base material and surface spectra, attributed to diphenyl carbonate

Conclusions. Overall, the obtained results suggested that the anomalous surface condition observed on the bracket represented molecular degradation of the polycarbonate. This is consistent with the brittle properties exhibited by the component. The most likely cause of the molecular degradation was improper drying and/or exposure to excessive heat during the injection molding process.

Example 2: Chemical Attack of Acrylonitrile-Butadiene-Styrene Grips. A set of plastic grips from an electric consumer product failed while in service. The grips had been injection molded from a general-purpose grade of an acrylonitrile-butadiene-styrene (ABS) resin. The parts had cracked while in use after apparent embrittlement of the material.

Tests and Results. An examination of the grips confirmed a severe level of cracking, covering the majority of the grip surface. Handling of the parts revealed that the grip material exhibited very little integrity, unlike the usual ductility associated with ABS resins. A white discoloration was also observed on the otherwise red grips. The surface of the grips was evaluated using SEM, revealing isolated areas that showed significant degradation in the form of material loss, as shown in Fig. 15. The observed morphology suggested selective degradation of the polybutadiene domains within the ABS resin.

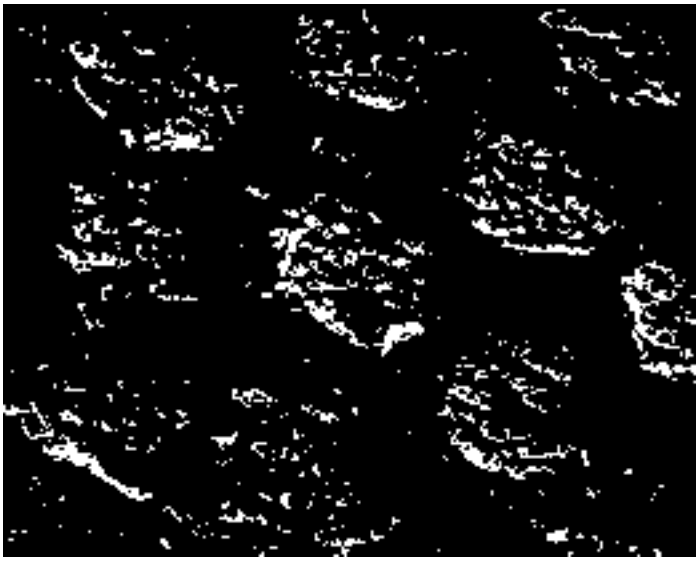


Fig. 15 Scanning electron image showing isolated degradation of the grip material. 30×

Micro-FTIR in the ATR mode was used to analyze the base material and the surfaces of the grips. The results obtained on the base material were characteristic of an ABS resin. Analysis of the surface of the part produced a somewhat different result. The spectrum representing the grip surface contained absorption bands associated with ABS; however, the results contained additional bands of significant intensity. A spectral subtraction was performed, thereby removing the bands associated with the ABS resin. The obtained subtraction spectrum produced a very good match with glyceride derivatives of fats and oils. This identification is illustrated in Fig. 16.

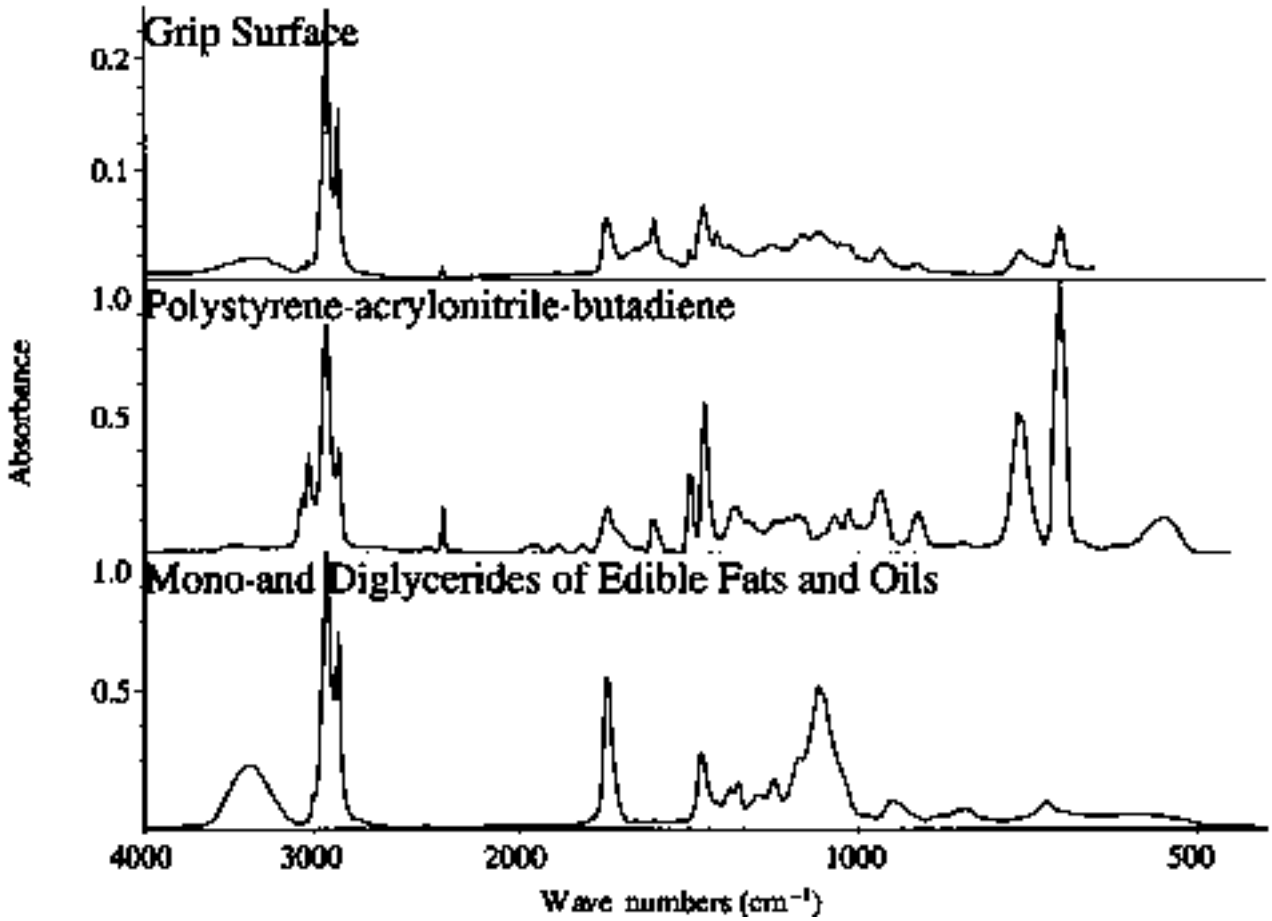


Fig. 16 The FTIR spectrum obtained on the grip surface. The spectrum contains absorption bands indicative of glyceride derivatives of fats and oils in addition to bands associated with the base ABS resin.

Conclusions. It was the conclusion of the analysis that the grips failed via brittle fracture associated with severe chemical attack of the ABS resin. A significant level of glyceride derivatives of fatty acids, known to degrade ABS resins, was found on the part surface. The glyceride derivatives selectively attacked the polybutadiene domains within the molded ABS part, leading to apparent embrittlement and subsequent failure.

Example 3: Inclusion within an ABS Handle. The handle from a consumer product exhibited an apparent surface defect. The handle had been injection molded from a medium-viscosity-grade ABS resin. The anomalous appearance was objectionable to the assembler of the final product and resulted in a production lot being placed on quality-control hold.

Tests and Results. The surface of the part was examined using an optical stereomicroscope. The defect appeared as a localized area of lightened color, and the zone immediately surrounding the anomaly was slightly recessed. A mounted and polished cross section was prepared through the part, revealing distinct included material within the base molding resin. The inclusion, as shown in Fig. 17, did not appear to contain a significant level of the blue pigment, as present in the base material.

Fig. 17 Micrograph showing the included material within the handle. 24×

The preparation of the cross section not only allowed a thorough inspection of the defect but also served to facilitate further analysis of the material. The sample was initially analyzed using EDS. The results obtained on the included material showed exclusively carbon and oxygen, precluding an inorganic contaminant. The base resin and the included material were further analyzed using FTIR in the reflectance mode. The spectrum representing the base material contained absorption bands indicative of an ABS resin. Analysis of the included material produced distinctly different results. The spectrum obtained on the included material was characteristic of polybutadiene, the rubber-modifying agent present in ABS. This identification is presented in Fig. 18.

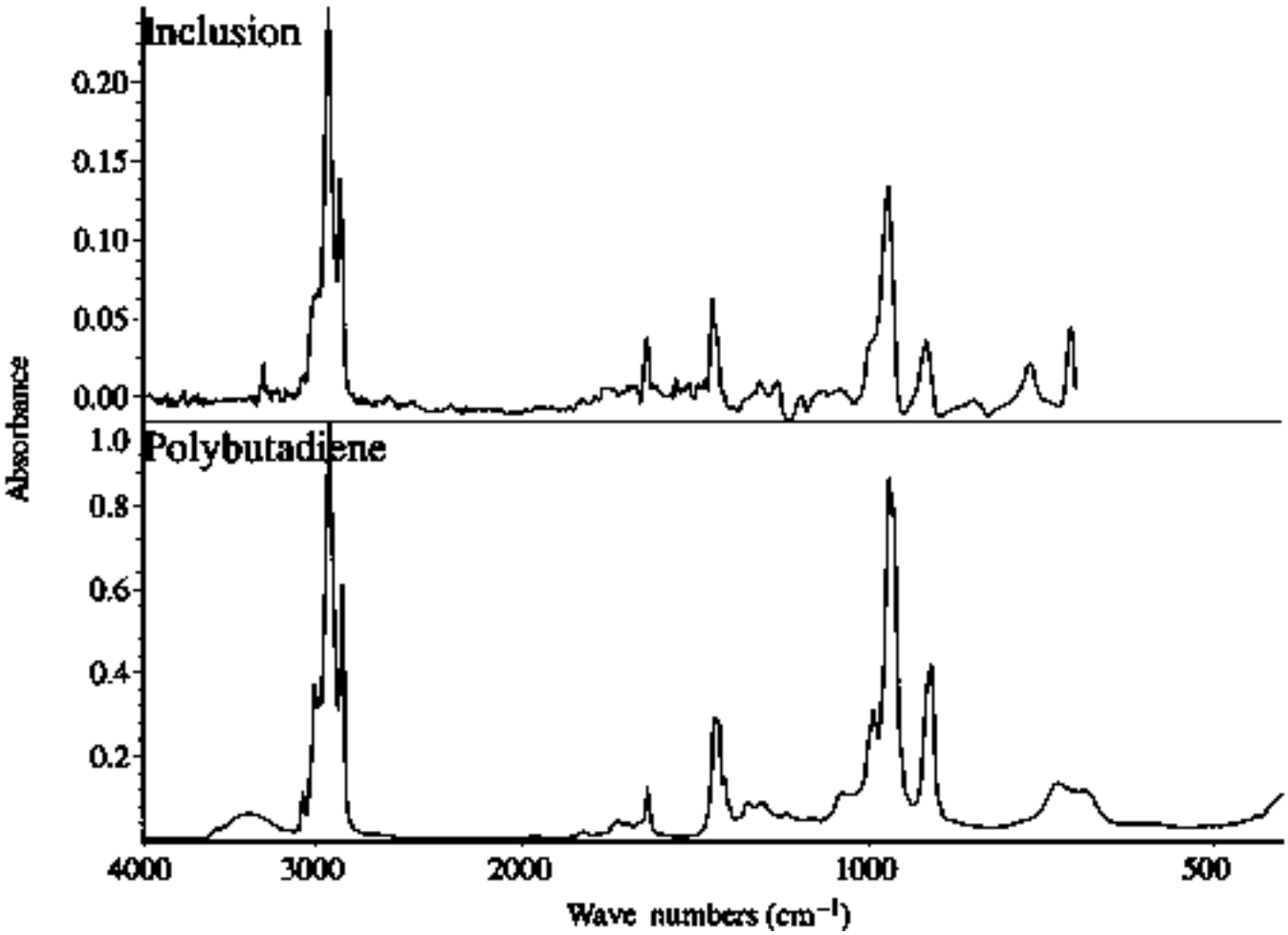


Fig. 18 The FTIR spectrum obtained on the included particle, characteristic of polybutadiene

Conclusions. It was the conclusion of the evaluation that the handle contained an inclusion, which produced the apparent surface anomaly. The included material was identified as polybutadiene. The most likely source of the included polybutadiene was an undispersed gel particle formed during the production of the molding resin.

Example 4: Relaxation of Nylon Wire Clips. A production lot of plastic wire clips was failing after limited service. The failures were characterized by excessive relaxation of the clips, such that the corresponding wires were no longer adequately secured in the parts. No catastrophic failures had been encountered. Parts representing an older lot, which exhibited satisfactory performance properties, were also available for reference purposes. The clips were specified to be injection molded from an impact-modified grade of nylon 6/6. However, the part drawing did not indicate a specific resin.

Tests and Results. A visual examination of the clips showed that the failed parts were off-white in color, while the control parts had a pure white appearance. An analysis of both sets of parts was performed using micro-FTIR in the ATR mode. A direct comparison of the results produced a good match, with both sets of spectra exhibiting absorption bands that were characteristic of a nylon resin. The comparison, however, revealed subtle differences between the two sets of clips. The spectrum representing the reference parts showed a relatively higher level of a hydrocarbon-based impact modifier, while the results obtained on the failed parts showed the presence of an acrylic-based modifier. The differences in the spectra suggested that the two sets of clips were produced from resins having different formulations, particularly regarding the impact modifier.

The clip materials were further analyzed using DSC. The thermogram representing the reference part material, as shown in Fig. 19, exhibited an endothermic transition at 264 °C (507 °F), characteristic of the melting point of a nylon 6/6 resin. Additionally, the results contained a second melting point, of lesser magnitude, at 95 °C (203 °F). This transition was indicative of a hydrocarbon-based impact modifier, as indicated by the FTIR results. The thermogram obtained on the failed clip material also showed a melting point characteristic of a nylon 6/6 resin. However, no evidence was found to indicate a transition corresponding to the hydrocarbon-based modifier found in the control clip material.

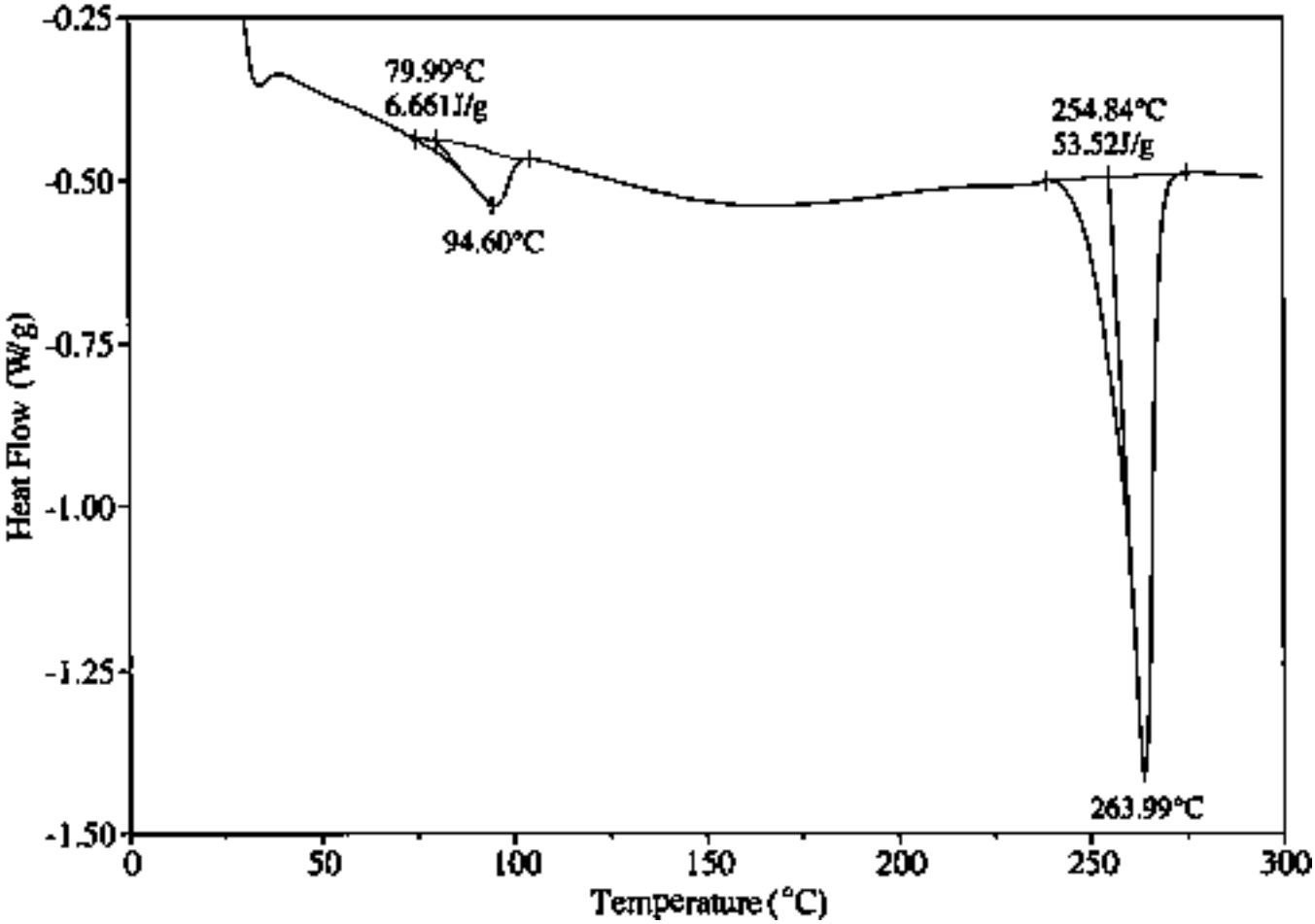


Fig. 19 The DSC thermogram representing the reference clip material, exhibiting an endothermic transition characteristic of the melting of a nylon 6/6 resin. The results also showed a second melting transition attributed to a hydrocarbon-based impact modifier.

Conclusions. It was the conclusion of the analysis that the control and failed clips had been produced from two distinctly different resins. While both materials satisfied the requirements of an impact-modified nylon 6/6 resin, differences in the impact modifiers resulted in the observed performance variation. From the results and the observed performance, it appeared that the material used to produce the failed clips had different viscoelastic properties, which produced a greater predisposition for stress relaxation.

Example 5: Embrittlement of Nylon Couplings. Molded plastic couplings used in an industrial application exhibited abnormally brittle properties, as compared to previously produced components. The couplings were specified to be molded from a custom-compounded glass-filled nylon 6/12 resin. An inspection of the molding resin used to produce the discrepant parts revealed differences in the material appearance, relative to a retained resin lot. Specifically, physical sorting resulted in two distinct sets of molding resin pellets from the lot that had generated the brittle parts. Both of these sets of pellets had a coloration that varied from that of the retained reference resin pellets. A sample of retained molding resin, which had produced parts exhibiting satisfactory performance, was available for comparative analysis.

Tests and Results. Micro-FTIR in the ATR mode was used to analyze the molding resin samples. The results obtained on the three molding resin samples were generally similar, and all of the spectra exhibited absorption bands characteristic of a nylon resin. Further analysis of the resin samples using DSC indicated that the control material results exhibited a single endothermic transition at 218 °C (424 °F), consistent with the melting point of a nylon 6/12 resin, as specified.

The DSC thermograms obtained on the two resin samples that produced brittle parts also exhibited melting point transitions associated with nylon 6/12. However, additional transitions were also apparent in the results, indicative of the presence of contaminant materials. The results obtained on one of the resin samples, as presented in Fig. 20, showed a secondary melting point at 165 °C (330 °F), indicative of polypropylene. The thermogram representing the second resin sample, as included in Fig. 21, displayed a second melting transition at 260 °C (500 °F), characteristic of a nylon 6/6 resin.

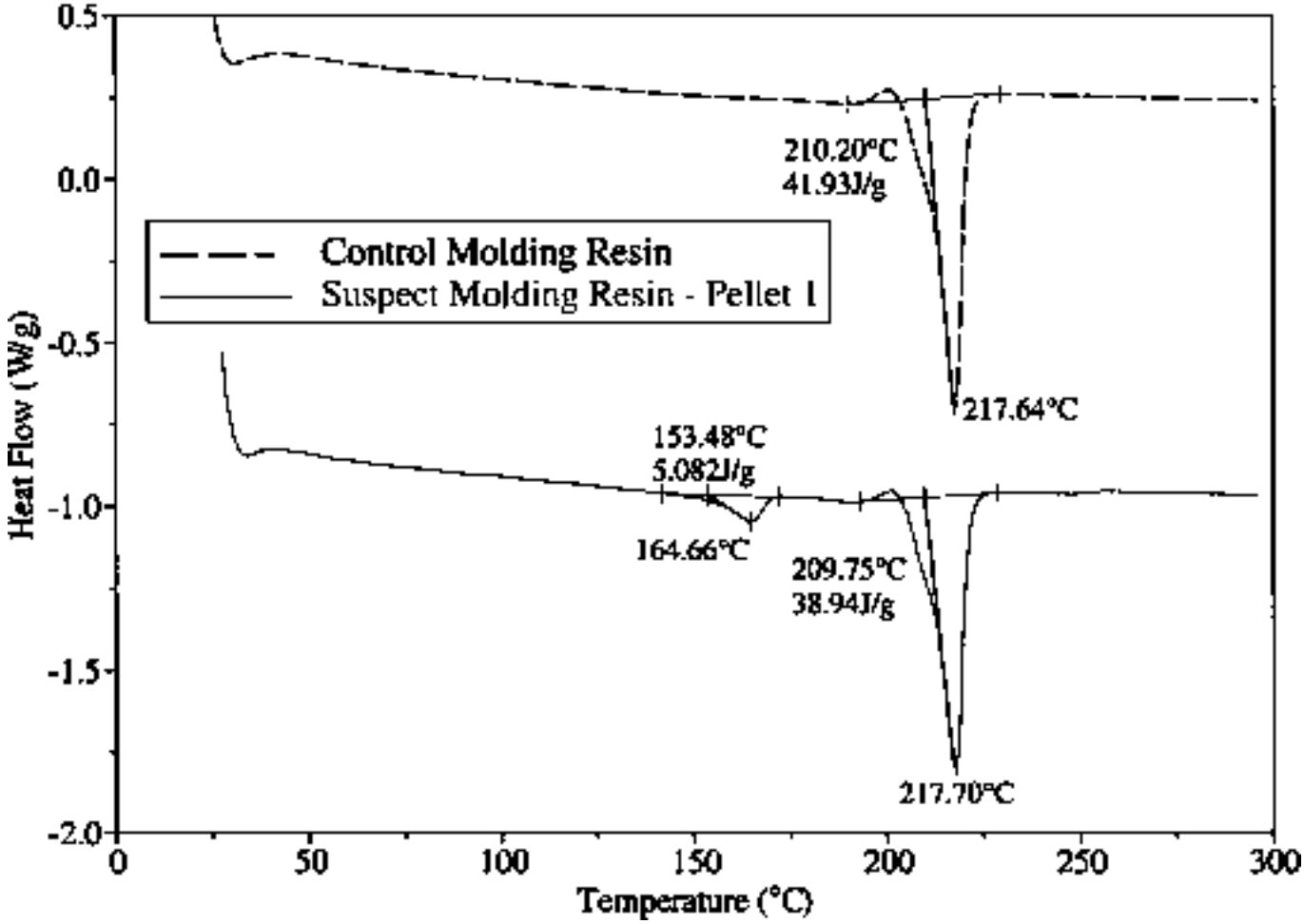


Fig. 20 The DSC thermogram representing a molding resin pellet that had produced brittle parts. The thermogram shows a major melting transition associated with nylon 6/12 and a weaker transition attributed to polypropylene.

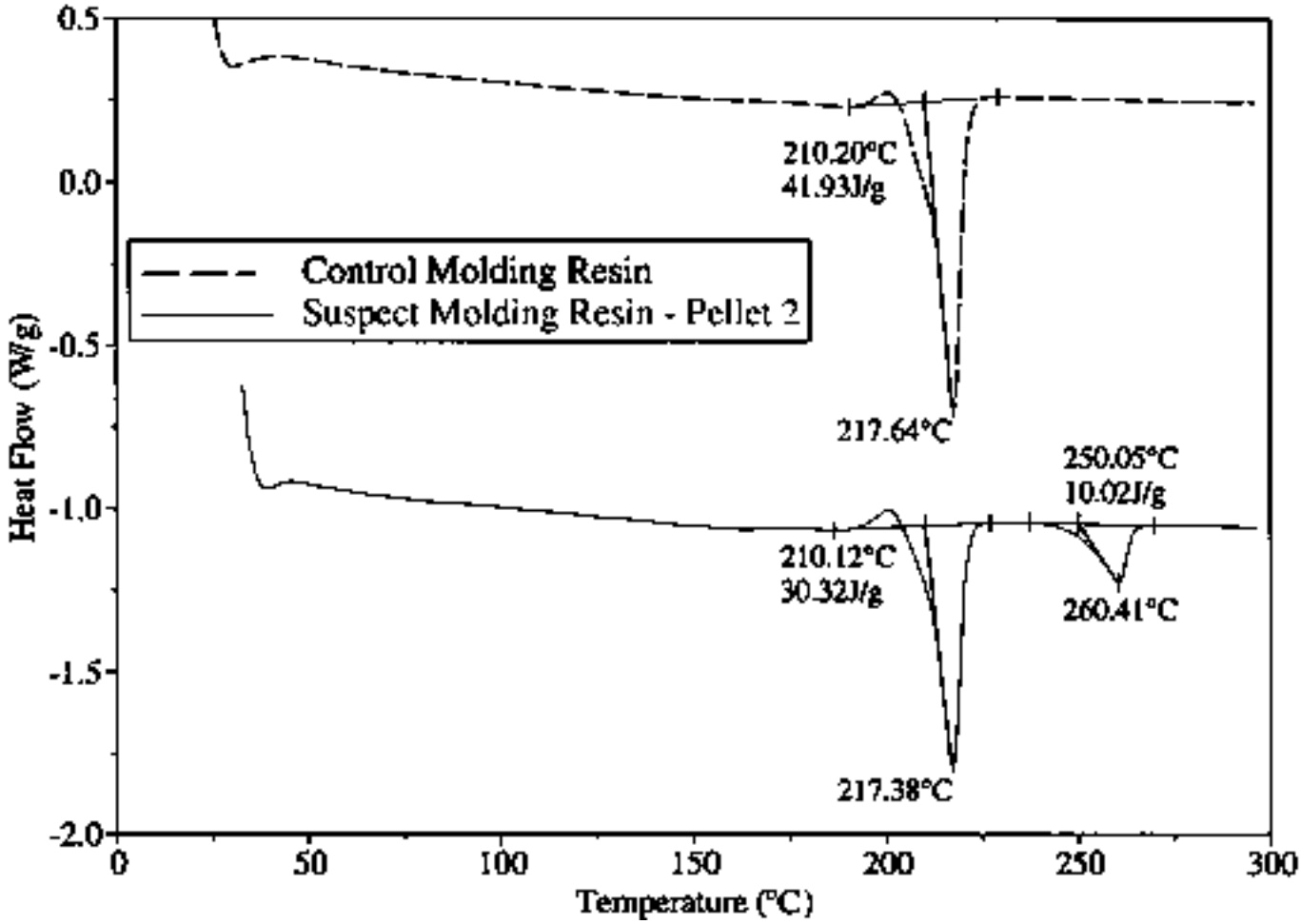


Fig. 21 The DSC thermogram representing a second molding resin pellet that had produced brittle parts. The thermogram shows a major melting transition associated with nylon 6/12 and a weaker transition attributed to nylon 6/6.

Conclusions. It was the conclusion of the analysis that the molding resin used to produce the brittle couplings contained a significant level of contamination, which compromised the mechanical properties of the molded components. Two distinct contaminants were found mixed into the molding pellets. The contaminant materials were identified as polypropylene and nylon 6/6. The source of the polypropylene was likely the purging compound used to clean the compounding extruder. The origin of the nylon 6/6 resin was unknown but may represent a previously compounded resin.

Example 6: Failure of Plasticized Poly(vinyl chloride) Tubing. A section of clear polymeric tubing failed while in service. The failed sample had been used in a chemical transport application. The tubing had also been exposed to periods of elevated temperature as part of the operation. The tubing was specified to be a poly(vinyl chloride) (PVC) resin plasticized with trioctyl trimellitate (TOTM). A reference sample of the tubing, which had performed well in service, was also available for testing.

Tests and Results. The failed and reference tubing samples were analyzed using micro-FTIR in the ATR mode, and the results representing the reference tubing material were consistent with the stated description: a PVC resin containing a trimellitate-based plasticizer. However, the spectrum representing the failed tubing material was noticeably different. While the obtained spectrum contained absorption bands characteristic of PVC, the results indicated that the material had been plasticized with an adipate-based material, such as dioctyl adipate. This identification is shown in Fig. 22.

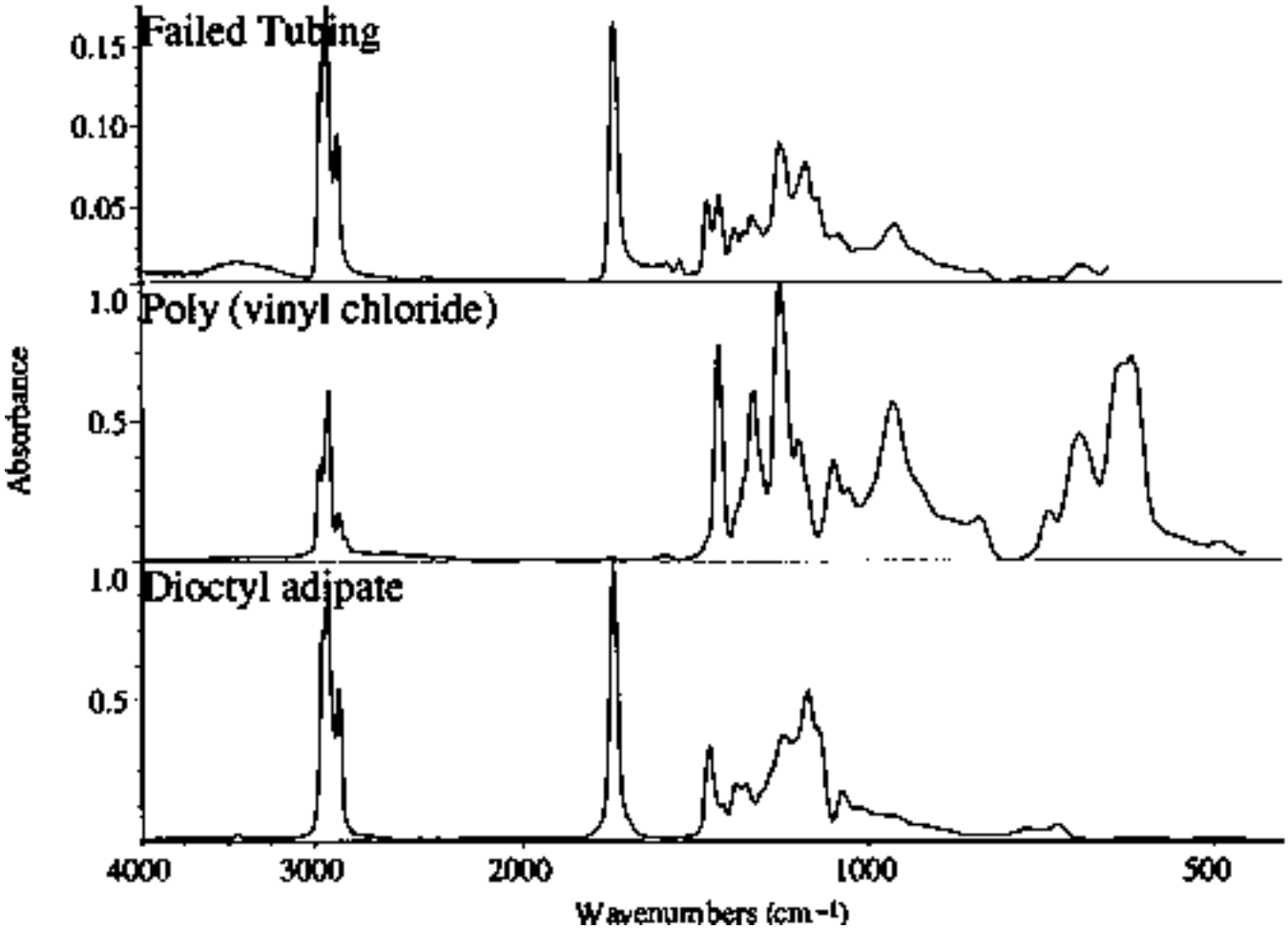


Fig. 22 The FTIR spectrum obtained on the failed tubing material. The spectrum exhibits absorption bands indicative of a PVC resin containing an adipate-based plasticizer.

In order to assess their relative thermal stability, the two tubing materials were analyzed via thermogravimetric analysis (TGA). Both sets of results were consistent with those expected for plasticized PVC resins. The thermograms representing the reference and failed sample materials showed comparable plasticizer contents of 28 and 25%, respectively. The results also showed that the reference material, containing the trimellitate-based plasticizer, exhibited superior thermal resistance relative to the failed material, containing the adipate-based material. This was indicated by the elevated temperature of weight-loss onset exhibited by the reference tubing material.

Conclusions. It was the conclusion of the evaluation that the failed tubing had been produced from a formulation that did not comply with the specified material. The failed tubing was identified as a PVC resin with an adipate-based plasticizer, not TOTM. The obtained TGA results confirmed that the failed tubing material was not as thermally stable as the reference material because of this formulation difference, and that this was responsible for the observed failure.

Example 7: Cracking of Poly(butylene terephthalate) Automotive Sleeves. A number of plastic sleeves used in an automotive application cracked after assembly but prior to installation into the mating components. The sleeves were specified to be injection molded from a 20% glass-fiber-reinforced poly(butylene terephthalate) (PBT) resin. After molding, electronic components are inserted into the sleeves, and the assembly is filled with a potting compound. A retained lot of parts, which had not cracked, were available for reference purposes.

Tests and Results. The reference and failed parts were analyzed using micro-FTIR in the ATR mode. The spectra obtained on both sets of parts contained absorption bands characteristic of a thermoplastic polyester, such as PBT or poly(ethylene terephthalate) (PET). Different types of polyester resins cannot be distinguished spectrally, because of the similar nature of their structures. However, subtle but distinct differences were apparent in the results, suggestive of degradation of the failed part material.

Differential scanning calorimetry was performed on the sleeve materials using a heat/cool/heat methodology. Testing of the reference material produced initial heating results indicative of a PBT resin, as illustrated by the melting point at 224 °C (435 °F). Analysis of the failed sleeve samples produced a melting transition at a significantly reduced temperature, 219 °C (426 °F). Additionally, the failed material transition was broader, and overall, the results suggested molecular degradation of the failed sleeve material. A comparison of the initial heating thermograms is presented in Fig. 23.

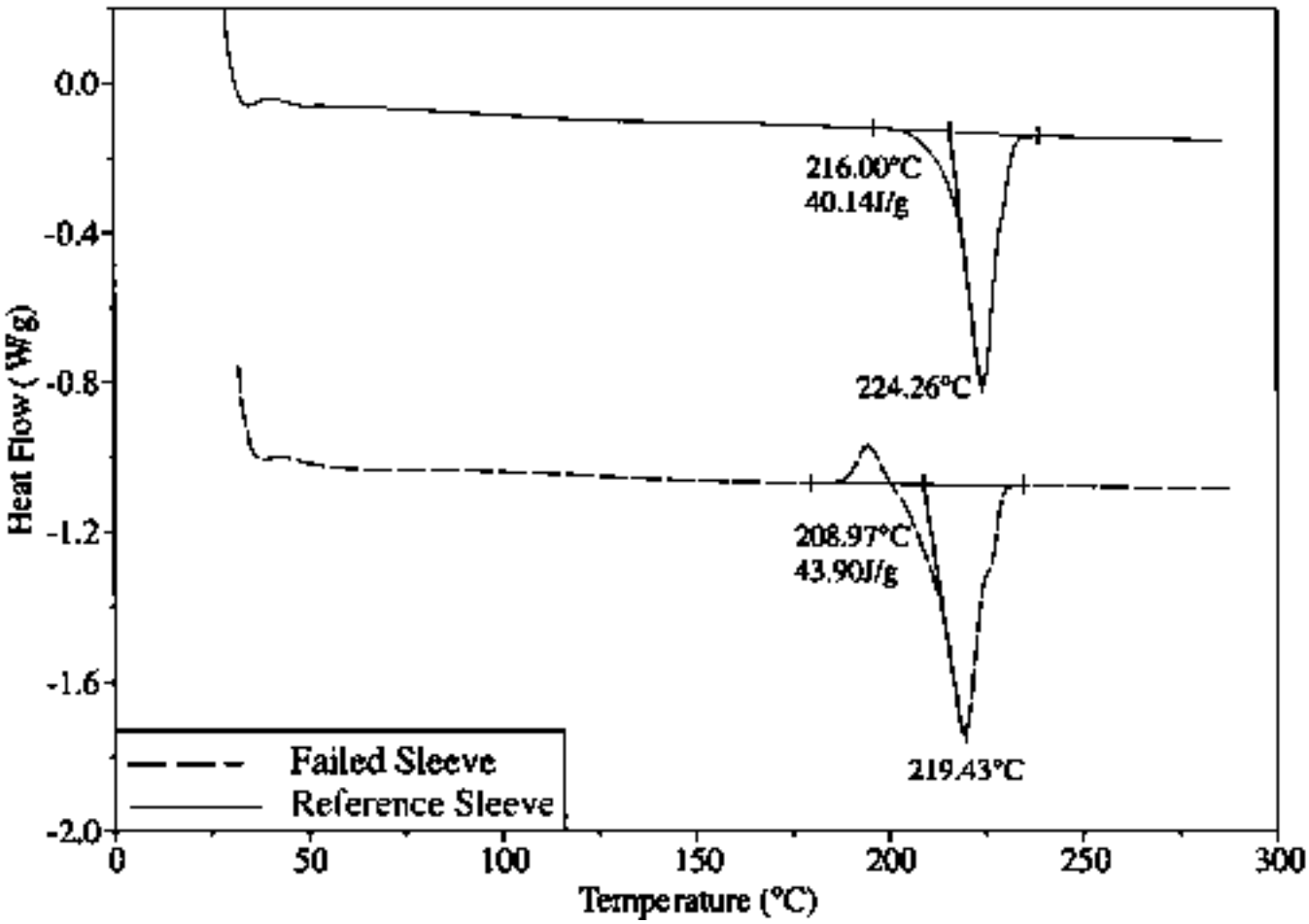


Fig. 23 A comparison of the initial heating run results, suggesting degradation of the failed sleeve material

The identification of degradation was supported by the second heating DSC results, obtained after slow cooling. The second heating thermogram representing the failed sleeve material showed additional differences relative to the results obtained on the reference material. The failed material did not produce the bimodal melting endothermic transition normally associated with PBT after slow cooling. This was thought to be the result of molecular degradation, which produced shorter polymer chain lengths, therefore reducing steric hindrance. A comparison of the second heating thermograms is included in Fig. 24.

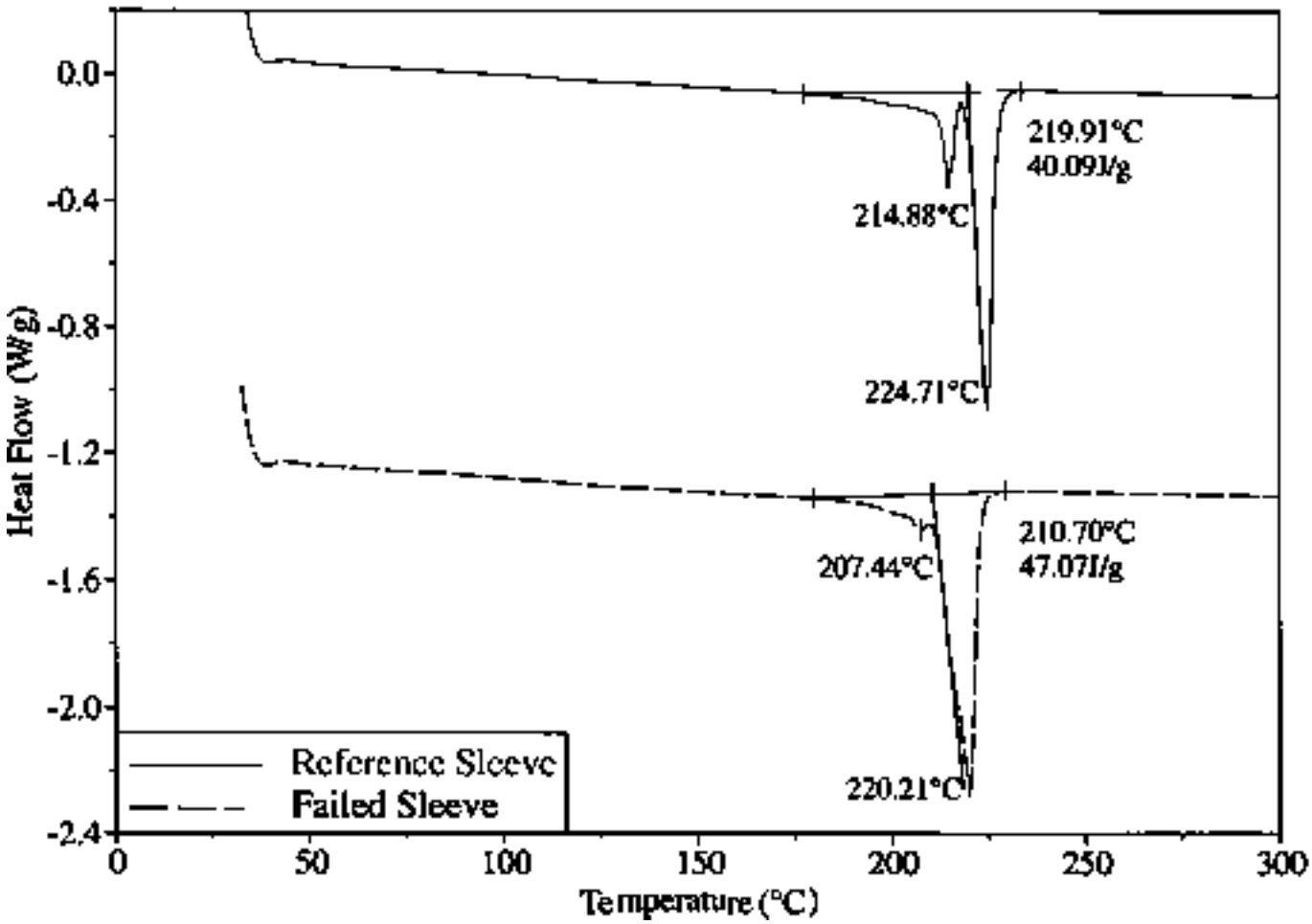


Fig. 24 A comparison of the second heating run results, further suggesting degradation of the failed sleeve material

The sleeve materials were further analyzed using TGA. The thermograms obtained on the reference and failed samples were generally consistent, including equivalent glass contents. Additionally, the results were in agreement with those expected for a PBT resin.

The melt flow rates (MFRs) of the reference and failed sleeve materials were determined. Because no molding resin was available for comparison purposes, the nominal range from the specification sheet, 14 to 34 g/10 min, was used. The testing showed that the failed sleeve material had been severely degraded, producing a MFR of 128 g/10 min. This was in agreement with the DSC data and indicated severe molecular degradation of the PBT resin. A review of the results generated by the reference parts also showed significant molecular degradation. While the extent of the degradation was less, the obtained melt flow rate, 50 g/10 min, still demonstrated a substantial reduction in the average molecular weight.

Conclusions. It was the conclusion of the evaluation that the failed sleeves had cracked due to embrittlement associated with severe degradation and the corresponding molecular weight reduction. The degradation was clearly illustrated by the reduced melting point and uncharacteristic nature of the associated endothermic melting transitions as well as the substantial increase in the MFR of the molded parts. The reduction in molecular weight significantly reduced the mechanical properties of the sleeves. The cause of the degradation was not evident, but the likely source appears to be the molding operation and exposure to elevated temperature for an extended period of time. It is significant to note that the reference parts also showed a moderate level of molecular degradation, rendering them susceptible to failure over a longer duration.

Example 8: Cracking of ABS Protective Covers. Numerous protective covers, used in conjunction with an electrical appliance, failed during assembly with the mating components. The failures were traced to a particular production lot of the covers and occurred during insertion of the screws into the corresponding bosses. The parts had been injection molded from an ABS resin to which regrind was routinely added. Retained parts, which exhibited normal behavior during assembly, were available for comparative analysis.

Tests and Results. A visual examination of the failed parts revealed relatively brittle fracture features, without significant ductility, as would be apparent as stress whitening or permanent deformation. Core material taken from the reference and failed parts was analyzed using micro-FTIR in the ATR mode. Both obtained spectra exhibited absorption bands associated with an ABS resin. However, the spectrum representing the failed part showed additional absorption bands. A

spectral subtraction was performed, thereby removing the absorbances attributed to the ABS resin from the spectrum obtained on the failed part. The spectral subtraction results were consistent with a thermoplastic polyester, such as PET or PBT. However, these two materials cannot be distinguished spectrally, because of similarities in their structures. As such, the melting point is usually used to differentiate between these materials. The FTIR results indicated the presence of contaminant material exclusively within the ABS resin used to mold the failed covers.

In order to further identify the contaminant material, a sample taken from the failed part was analyzed via DSC. The obtained DSC thermogram, as presented in Fig. 25, showed a glass transition at approximately 101 °C (214 °F), consistent with the expected results for an ABS resin. The results also showed an additional endothermic transition at 222 °C (432 °F), indicative of a PBT resin. The failed cover material was also analyzed using TGA in order to assess the level of the contamination. The TGA analysis was performed using high-resolution temperature programming, and the results revealed adequate separation of the ABS and PBT resins. Based on the results, the contamination was estimated to account for approximately 23% of the failed cover material.

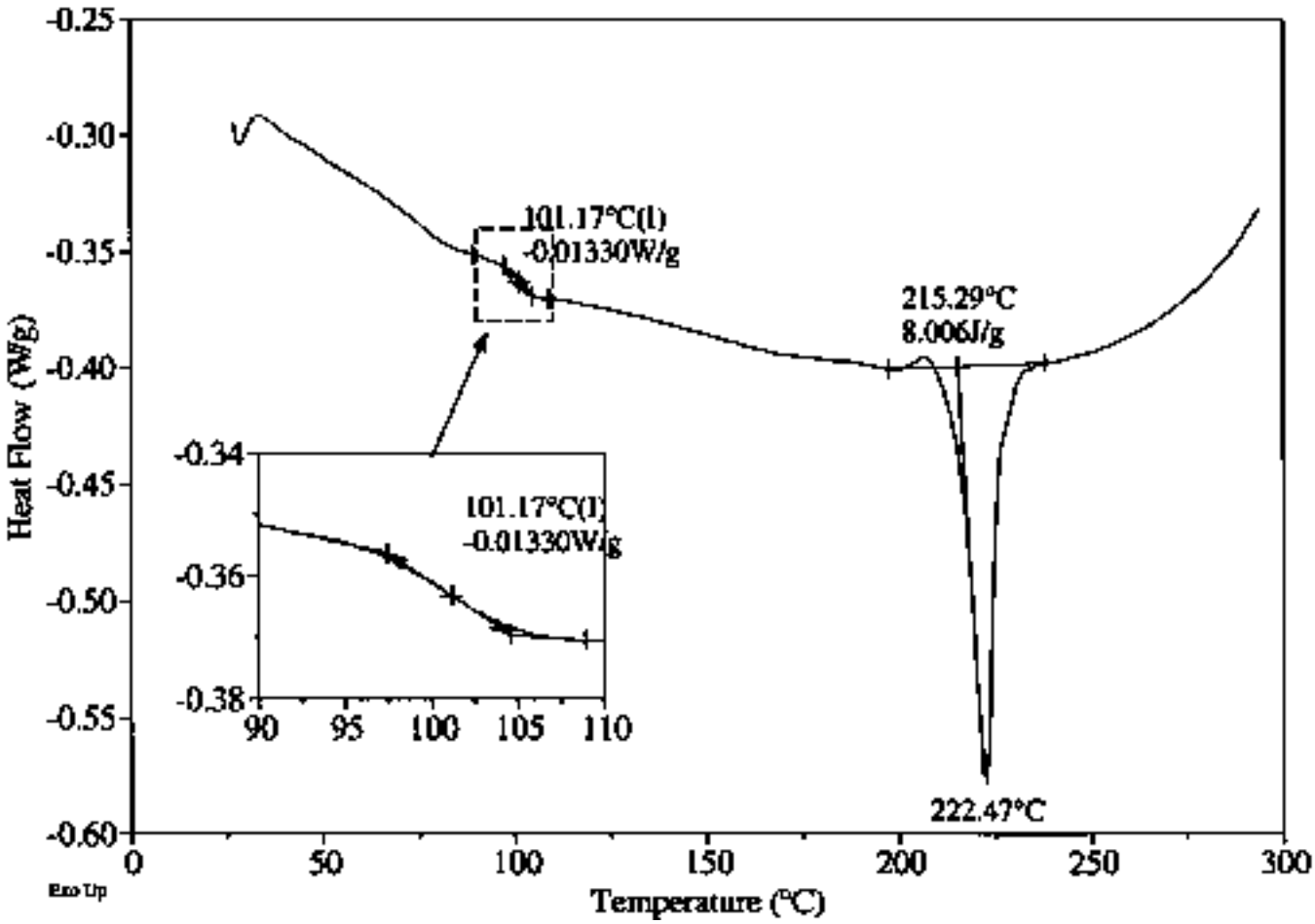


Fig. 25 The DSC thermogram obtained on the failed cover material. The thermogram shows an endothermic transition associated with PBT. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

Conclusions. It was the conclusion of the evaluation that the appliance covers failed via brittle fracture associated with stress overload. The failures, which occurred under normal assembly conditions, were attributed to embrittlement of the molded parts, due to contamination of the ABS resin with a high level of PBT. The source of the PBT resin was not positively identified, but a likely source appeared to be the use of improper regrind.

Example 9: Failure of Polycarbonate/PET Appliance Housings. Housings from an electrical appliance failed during an engineering evaluation. The housings had been injection molded from a commercial polycarbonate/PET (PC/PET) blend. The parts were being tested as part of a material conversion. Parts produced from the previous material, a nylon 6/6 resin, had consistently passed the testing regimen. The housing assembly included a spring clip, which applied a static force on a molded-in boss extending from the main body of the housing. Grease was applied liberally within the housing assembly during production.

Tests and Results. A visual inspection of the tested parts showed catastrophic failure within the molded-in boss. The failures were consistent across all of the parts and were located at an area where the spring clip contacted the housing

boss. While the final fracture zone exhibited limited features associated with ductility in the form of stress whitening, no such characteristics were apparent at the locations corresponding to the crack origins. The fracture surfaces were further examined via SEM. The SEM inspection showed the presence of multiple crack initiation sites along the side of the boss that had mated with the spring clip. No evidence of significant ductility was found with the crack initiation locations, as represented in Fig. 26. The overall features observed on the fracture surface were indicative of environmental stress cracking.

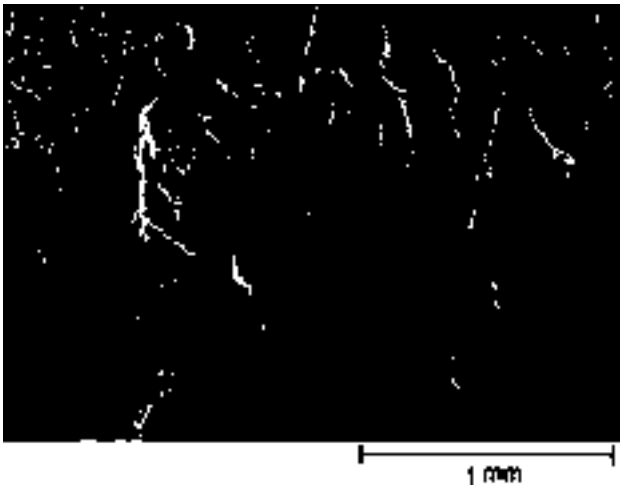


Fig. 26 Scanning electron image showing brittle fracture features at the crack initiation site, characteristic of environmental stress cracking. 24×

Micro-FTIR in the ATR mode was performed on the housing material, and the resulting spectrum was in agreement with the stated resin description, a blend of PC and polyester. No signs of material contamination were found. The housing material was further evaluated using DSC. The thermogram obtained during the initial heating run, as shown in Fig. 27, exhibited an endothermic transition at 253 °C (487 °F), characteristic of the melting point of a PET resin. The initial heating run results also showed a low-temperature exothermic transition associated with the crystallization of the PET resin. These results indicated that the material had not been fully crystallized during the molding process. The results generated during the second heating run, after slow cooling, did not show the low-temperature crystallization. The glass transition associated with the PC resin was observed in the second heating run.

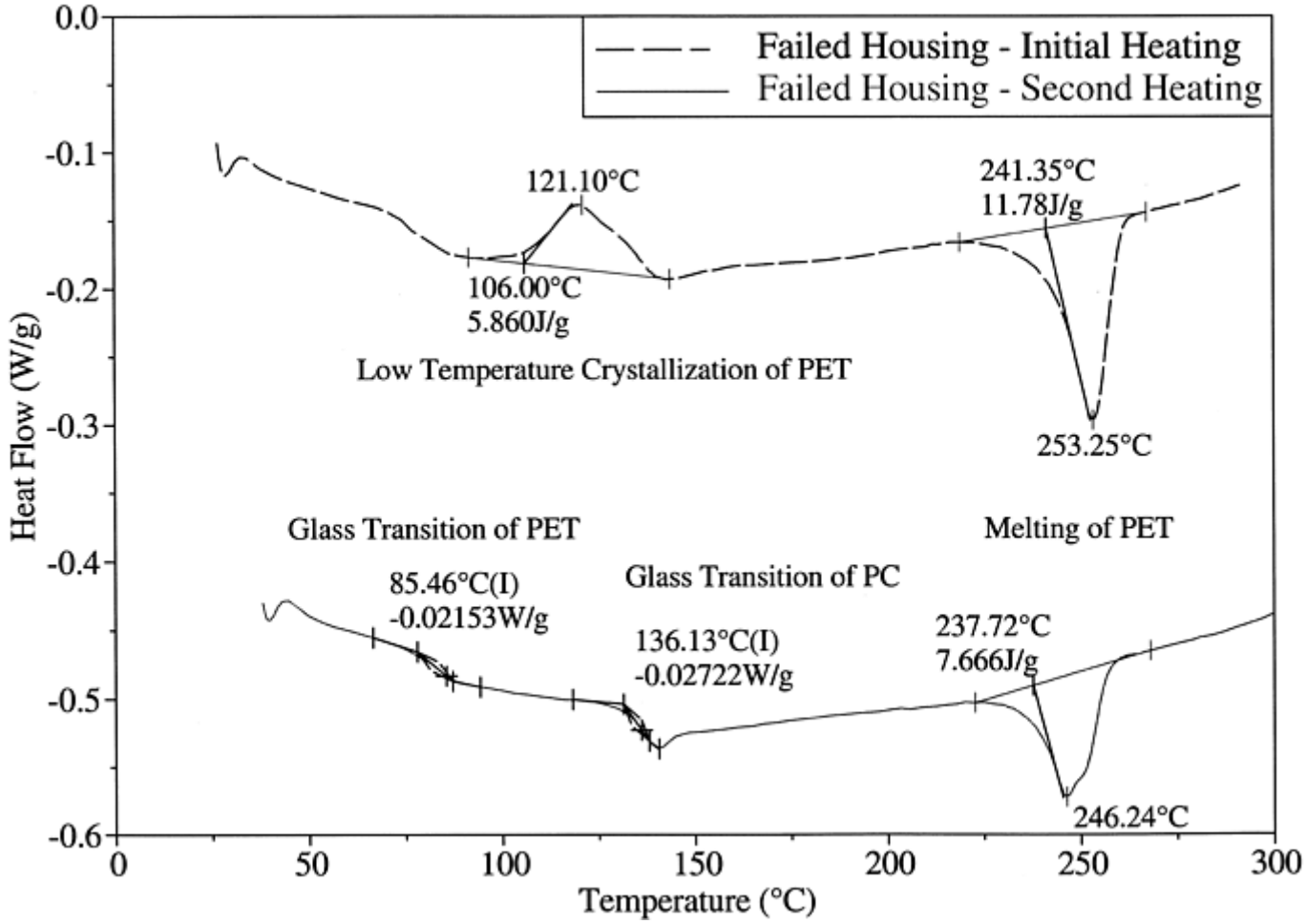


Fig. 27 The initial heating DSC thermogram, exhibiting a melting transition consistent with a PET resin. A low-temperature crystallization exothermic transition was also apparent. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

In order to assess the molecular weight of the housing material, the intrinsic viscosity of the resin was measured. A comparison of the results with historical data revealed a substantial reduction in the viscosity of the failed part material. This indicated that the housing material had undergone significant molecular degradation during the injection molding process.

The grease present within the housing assembly was analyzed using micro-FTIR. The FTIR test results indicated that the grease was composed of a relatively complex mixture. The lubricant contained a hydrocarbon-based oil, a phthalate-based oil, lithium stearate, and an amide-based additive. These results were significant, because phthalate esters are known to be incompatible with PC resins.

Conclusions. It was the conclusion of the analysis that the appliance housings failed through environmental stress cracking. The required chemical agent was identified as a phthalate-based oil present within the grease used to lubricate the assembly. Specifically, the phthalate oil was not compatible with the PC portion of the resin blend. The source of the stress responsible for the cracking appears to be the interference related to the spring clip. While the previous parts, produced from the nylon 6/6 resin, were also under similar stresses, this resin was not prone to stress cracking in conjunction with the lubricant. Thus, the resin conversion was the root cause of the failures. Additionally, the test results also showed that the injection molding process left the material susceptible to failure. Specifically, the molded parts had been undercrystallized, reducing the mechanical strength of the molded articles, and, more importantly, the resin had been degraded, producing a reduction in the molecular weight and reducing both the mechanical integrity and chemical-resistance properties of the parts.

Example 10: Failure of PET Assemblies. Several assemblies used in a transportation application failed during an engineering testing regimen. The testing involved cyclic thermal shock, immediately after which cracking was observed on the parts. The cracking occurred within the plastic jacket, which had been injection molded from an impact-modified, 15% glass-fiber-reinforced PET resin. The plastic jacket had been molded over an underlying metal coil component. Additionally, a metal sleeve was used to house the entire assembly. Prior to molding, the resin had reportedly been dried

at 135 °C (275 °F). The drying process usually lasted 6 h, but occasionally, the material was dried overnight. The thermal shock testing included exposing the parts to alternating temperatures of -40 and 180 °C (-40 and 360 °F). The failures were apparent after 100 cycles. Molding resin and non-failed parts were also available for analysis.

Tests and Results. The failed assemblies were visually and microscopically examined. The inspection showed several different areas within the overmolded jacket that exhibited cracking. The cracked areas were located immediately adjacent to both the underlying metal coil and the outer metal housing. The appearance of the cracks was consistent with brittle fracture, without significant signs of ductility. The examination also revealed design features, including relatively sharp corners and nonuniform wall thicknesses, that appeared to have likely induced molded-in stress within the plastic jacket.

The fracture surfaces were further inspected using SEM, and the examination revealed features generally associated with brittle fracture, as shown in Fig. 28. No evidence of microductility, such as stretched fibrils, was found. The fracture surface features indicated that the cracking had initiated along the outer jacket wall and subsequently extended through the wall and circumferentially around the wall. Throughout the examination, no indication of postmolding molecular degradation was found.

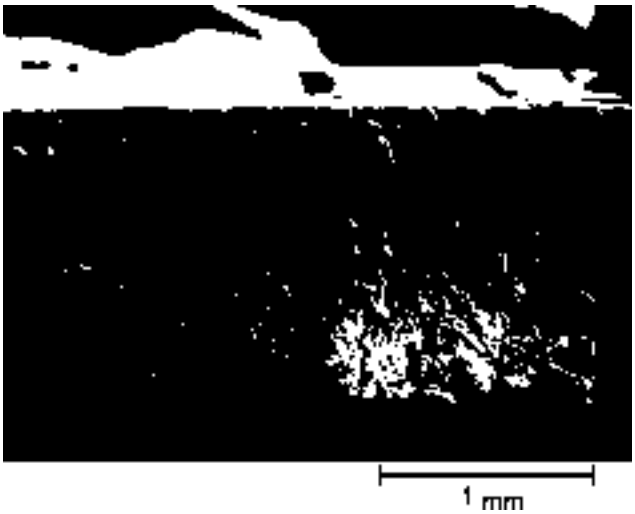


Fig. 28 Scanning electron image showing brittle fracture features on the failed jacket crack surface. 20×

Micro-FTIR was performed in the ATR mode on a core specimen of the jacket material. The resulting spectrum was consistent with a thermoplastic polyester resin. Such materials, including PET and PBT, cannot be distinguished spectrally, and a melting point determination is usually used to distinguish these materials. The failed jacket material and reference molding resin were analyzed using TGA, and the results obtained on the two samples were generally consistent. This included relatively comparable levels of volatiles, polymer, carbon black, and glass reinforcement. Further, the results were in excellent agreement with those expected for the stated PET material.

The failed jacket and reference materials were evaluated via DSC. Analysis of the failed jacket material produced results that indicated a melting transition at 251 °C (484 °F), consistent with a PET resin. However, a second endothermic transition was also present. This transition, at 215 °C (420 °F), suggested the melting of annealed crystals, indicating that the part had been exposed to a temperature approaching 215 °C (420 °F). The thermal shock testing appeared to be the only possible source of this thermal exposure.

Analysis of the molding resin also produced results consistent with a PET resin. The results also exhibited a second melting endotherm at 174 °C (345 °F). Again, this transition was associated with melting of annealed crystals for material exposed to this temperature. The apparent source of the exposure was the drying process. This was well in excess of the stated drying temperature.

Further analysis of the assembly materials using thermomechanical analysis (TMA) produced significantly different results for the PET jacket and the steel housing material. Determination of the coefficients of thermal expansion (CTEs) showed approximately an order of magnitude difference between the two mating materials.

An assessment of the molecular weight of the failed jacket samples as well as a non-failed part and the molding resin samples was performed using several techniques. A combination of MFR, intrinsic viscosity, and finally, gel permeation chromatography (GPC) was used, because of conflicting results. The MFR determinations showed that the drying process produced a considerable increase in the MFR of the resin, corresponding to molecular degradation in the form of chain scission. This was contrasted by the results generated by the intrinsic viscosity testing. These results showed an increase in the viscosity of the dried resin relative to the virgin resin. This increase was suggestive of an increase in molecular weight, possibly through partial cross linking. Testing of the resin samples and the molded parts via GPC produced results that reconciled the discrepancy. The GPC results showed that the drying process produced competing reactions of chain

scission and cross linking. The net result was severe degradation of the dried resin, which predisposed the molding material to produce jackets having poor mechanical properties. The GPC testing showed that the molded jackets were further degraded during the injection molding process.

Conclusions. It was the conclusion of the investigation that the assemblies failed via brittle fracture associated with the exertion of stresses that exceeded the strength of the resin as-molded. The stresses were induced by the thermal cycling and the dimensional interference caused by the disparity in the CTEs of the PET jacket and the mating steel sleeve. However, several factors were significant in the failures. It was determined that the resin drying process had exposed the resin to relatively high temperatures, which caused substantial molecular degradation. The drying temperature was found to be approximately 173 °C (344 °F), well in excess of the recommendation for the PET resin. Further degradation was attributed to the molding process itself, leaving the molded jacket in a severely degraded state. This degradation limited the ability of the part to withstand the applied stresses. Additionally, the testing itself exposed the parts to temperatures above the recognized limits for PET, and this may have significantly lowered the mechanical properties of the part.

Example 11: Cracking of a Polyethylene Chemical Storage Vessel. A chemical storage vessel failed while in service. The failure occurred as cracking through the vessel wall, resulting in leakage of the fluid. The tank had been molded from a high-density polyethylene (HDPE) resin. The material held within the vessel was an aromatic hydrocarbon-based solvent.

Tests and Results. A stereomicroscopic examination of the failed vessel revealed brittle fracture surface features. This was indicated by the lack of stress whitening and permanent deformation. Limited ductility, in the form of stretching, was found exclusively within the final fracture zones. On cutting the vessel, significant stress relief, in the form of distortion, was evident. This indicated a high level of molded-in stress within the part. The fracture surface was further inspected using SEM. The observed features included a relatively smooth morphology within the crack origin location, which was indicative of slow crack initiation. This area is shown in Fig. 29. Features associated with more rapid crack extension, including hackle marks and river markings, were found at the midfracture and final fracture areas, as represented in Fig. 30. The entirety of the fracture surface features indicated that the cracking had initiated along the exterior wall of the vessel. The cracking extended transversely through the wall initially, and subsequently, circumferentially around the wall. Throughout the examination, no signs of postmolding molecular degradation or chemical attack were found.

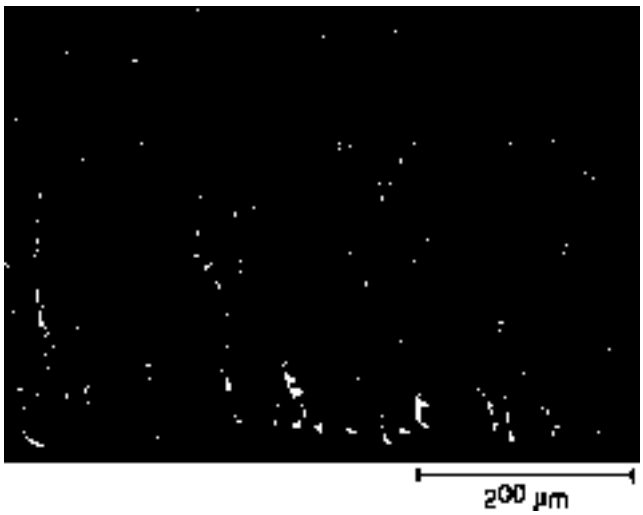


Fig. 29 Scanning electron image showing features associated with brittle fracture and slow crack growth within the crack initiation site. 100×

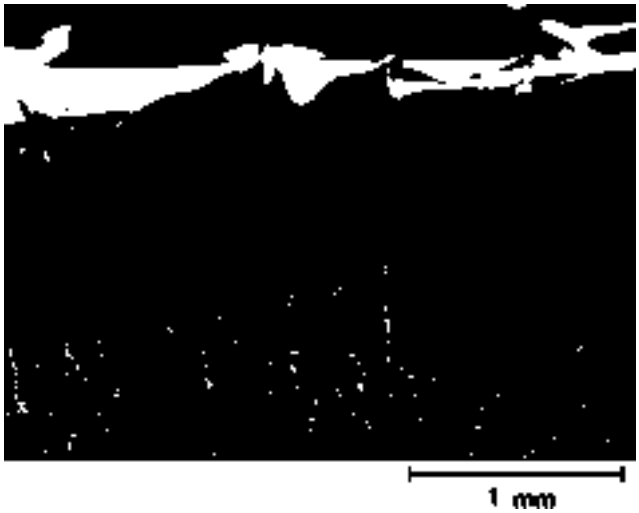


Fig. 30 Scanning electron image showing features indicative of rapid crack extension within the final fracture zone. 20×

The failed vessel material was analyzed using micro-FTIR in the ATR mode. The obtained spectrum exhibited absorption bands characteristic of a polyethylene resin. No evidence was found to indicate contamination or degradation of the material.

Material excised from the failed vessel was analyzed using DSC. The results showed a single endothermic transition associated with the melting point of the material at 133 °C (271 °F). The results were consistent with those expected for a HDPE resin. The results also showed that the HDPE resin had a relatively high level of crystallinity, as indicated by the elevated heat of fusion.

Thermogravimetric analysis was performed to further evaluate the failed vessel material. The TGA testing showed that the HDPE absorbed approximately 6.3% of its weight in the aromatic hydrocarbon-based solvent. Overall, the TGA results were consistent with those expected for a HDPE resin.

The MFR of the vessel material was evaluated, and the testing produced an average result of 3.8 g/10 min. This is excellent agreement with the nominal value indicated on the material data sheet, 4.0 g/10 min. As such, it was apparent that the vessel material had not undergone molecular degradation. The specific gravity of the resin was measured. The material produced a result of 0.965. This indicated that the material had a relatively high level of crystallinity, as suggested by the DSC results.

In order to assess the effects of the hydrocarbon-based solvent on the HDPE vessel, the material was evaluated using dynamic mechanical analysis (DMA). The vessel material was analyzed in two conditions. Material samples representing the vessel material in the as-molded condition as well as material from the failed vessel were evaluated. A comparison of the DMA results showed that in the saturated, equilibrium state, the HDPE resin lost over 60% of its elastic modulus at room temperature, because of the plasticizing effects of the solvent. A comparison of the DMA results, indicating the reduction in mechanical properties, is shown in Fig. 31.

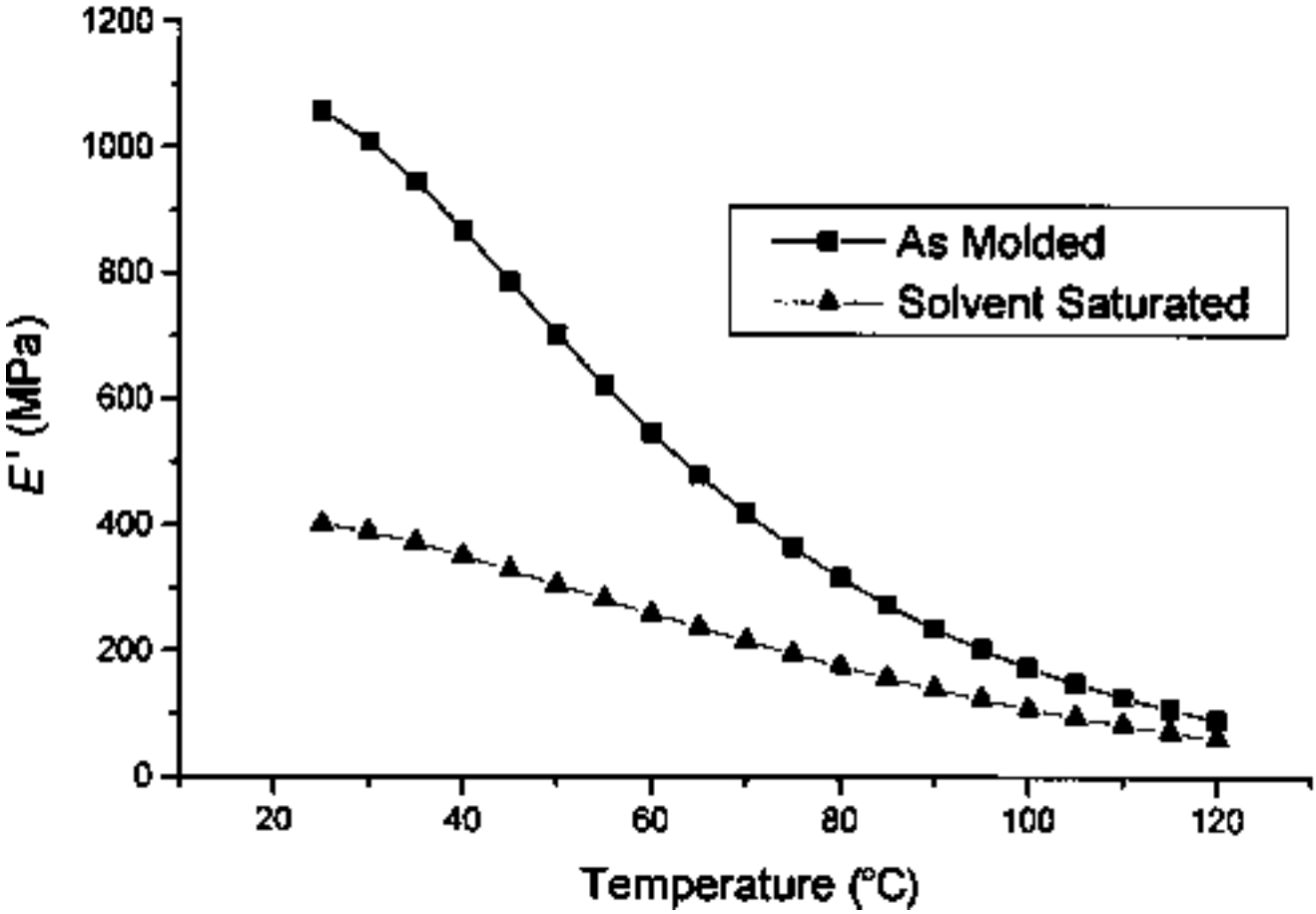


Fig. 31 A comparison of the DMA results, showing a loss of over 60% in the elastic modulus as a result of the effects of the solvent

Conclusions. It was the conclusion of the investigation that the chemical storage vessel failed via a creep mechanism associated with the exertion of relatively low stresses. Given the lack of apparent ductility, the stresses responsible for the failure appear to have been below the yield strength of the material. The source of the stress was thought to be molded-in residual stresses associated with uneven shrinkage. This was suggested by the obvious distortion evident on cutting the vessel. The relatively high specific gravity and the elevated heat of fusion are indicative that the material has a high level of crystallinity. In general, increased levels of crystallinity result in higher levels of molded-in stress and the corresponding warpage. The significant reduction in the modulus of the HDPE material, which accompanied the saturation of the resin with the aromatic hydrocarbon-based solvent, substantially decreased the creep resistance of the material and accelerated the failure. The dramatic effects of the solvent had not been anticipated prior to use.

Example 12: Failure of Polyacetal Latch Assemblies. Components of a latch assembly used in a consumer device exhibited a relatively high failure rate. The latches are used as a safety restraint, and failure in the field could result in severe injury. The failures were occurring after installation but prior to actual field use. Specifically, the failures occurred as cracking within retaining tabs used to secure a metal slide. The cracking was limited to an older design, with newer components showing no signs of failure. The latch assembly components were injection molded from an unfilled commercial grade of a polyacetal copolymer. As part of the evaluation, both failed parts representative of the older design and newer components were available for testing.

Tests and Results. A visual examination of the failed parts confirmed cracking within the retaining tab adjacent to the metal slide. The failures were present at consistent locations on all of the parts. The crack surfaces showed evidence of macroductility in the form of stress whitening within the final fracture zone exclusively. Throughout the visual examination, it was also apparent that the parts exhibited a very sharp corner formed by the retaining tab and the main body of the latch assembly body. Sharp corners are considered a poor design feature in plastic components, because they can result in severe stress concentration and can produce areas of localized poor fusion.

The fracture surface was further evaluated using SEM. The SEM examination showed a clear crack origin at the corner formed by the retaining tab. The crack origin areas exhibited brittle fracture features without signs of significant microductility. Secondary crazing was also apparent at the crack origin location. A typical crack initiation site is shown in Fig. 32. The overall features were suggestive of cracking caused by a relatively high strain rate event and/or very high stress concentration. The midfracture surface showed an increase in the apparent ductility, as evidenced by an overlapping morphological structure. The final fracture zone showed significant deformation and stretching, indicative of ductile

overload. A laboratory failure was created by overloading the tab from a nonfailed part in a manner consistent with the insertion of the corresponding metal slide. The laboratory fractures exhibited surface features that were in excellent agreement with those exhibited by the failed parts, as apparent in Fig. 32.

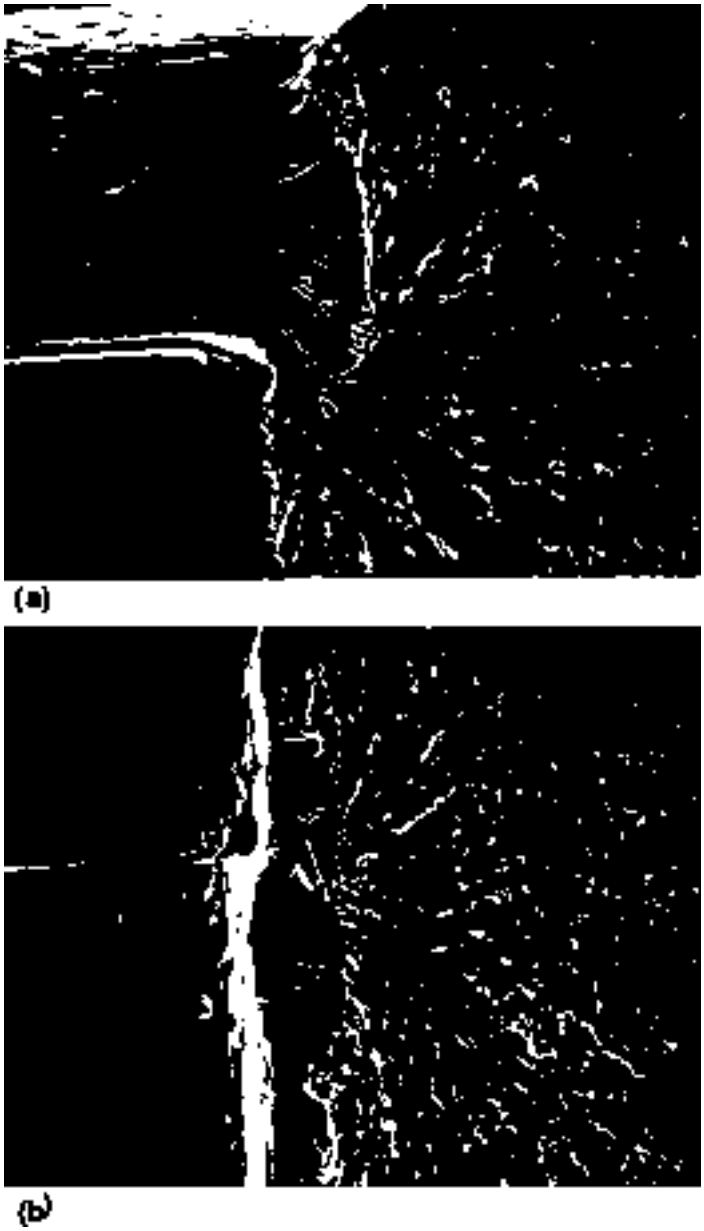


Fig. 32 Scanning electron images showing excellent agreement between the features present within the crack initiation sites of (a) the failed latch assembly and (b) the laboratory fracture. Both surfaces showed relatively brittle fracture features. 59×

The failed latch assembly material was analyzed using micro-FTIR in the ATR mode, and the obtained spectrum exhibited absorption bands characteristic of a polyacetal resin. It is significant to note that polyacetal copolymers and homopolymers cannot be differentiated spectrally, and a melting point determination is often used to distinguish between these materials.

Differential scanning calorimetry was used to analyze the latch material. The obtained results showed that the material underwent a single endothermic transition at approximately 165 °C (330 °F), characteristic of the melting point of a polyacetal copolymer. The results also showed that the part was somewhat undercrystallized. This was evident through a significant increase in the heat of fusion between the initial heating run and the second heating run, after slow cooling. Undercrystallization can reduce the mechanical strength of the molded article and is usually the result of molding in a relatively cold tool. The level of undercrystallization found in the failed parts, however, was moderate in nature and not thought to be a major factor in the failures. Thermogravimetric analysis was also performed on the latch material, and the obtained results were consistent with those expected for an unfilled polyacetal copolymer.

The latch material was also analyzed to determine its MFR. Parts representing the older, failed components and the newer, current design were evaluated. Both sets of molded parts produced results ranging from 10.7 to 11.0 g/10 min. This was in good agreement with the nominal MFR for the molding resin, 9.0 g/10 min. Throughout the analytical testing of the failed latch material, no evidence was found to indicate contamination or degradation of the molded parts.

Mechanical testing was performed in order to assess the effect of the recent design change. Because of the configuration of the parts, standard mechanical testing could not be performed. Instead, a proof load test was devised to directly assess the stress required to produce failure within the tab, at an area consistent with the failure latch assembly. A direct comparison was made between the two sets of parts. The parts representing the older design, with the sharp corner at the retaining tab, produced an average value of 78.7 N (17.7 lbf) at failure. More importantly, the parts within this group produced an average tab extension of 0.76 mm (0.03 in.) at failure. The evaluation of the part representing the new design generated significantly different results. Specifically, the failures occurred at a higher load, 92.1 N (20.7 lbf), and a greater tab extension, 2.5 mm (0.10 in.). A comparison of the mechanical test results is shown in Fig. 33. This mechanical evaluation clearly illustrated the advantage afforded by the design change, effectively increasing the tab radius.

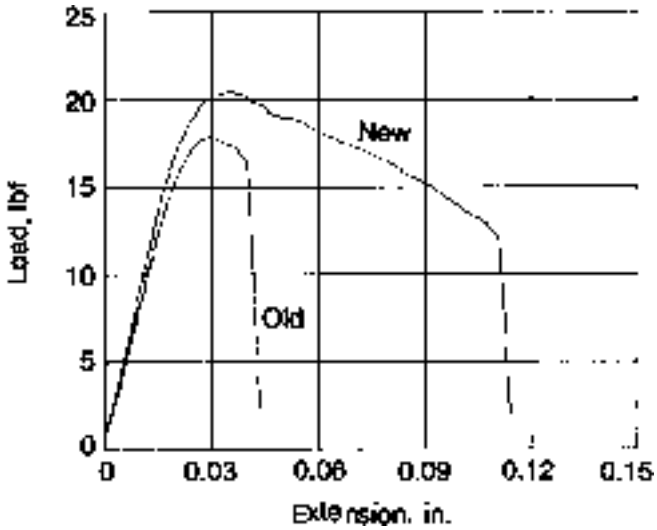


Fig. 33 A comparison of the mechanical test results, showing a significant improvement in the parts produced from the new design

Conclusions. It was the conclusion of the evaluation of the failed latch assemblies that the parts failed via brittle fracture associated with stress overload. The stress overload was accompanied by severe apparent embrittlement resulting from a relatively high strain rate event and/or significant stress concentration. The relatively sharp corner formed by the retaining tab was shown to be a primary cause of the failures, with the newer, redesigned parts producing superior mechanical test results.

Example 13: Failure of a Nylon Filtration Unit. A component of a water filtration unit failed while being used in service for approximately eight months. The filter system had been installed in a commercial laboratory, where it was stated to have been used exclusively in conjunction with deionized water. The failed part had been injection molded from a 30% glass-fiber- and mineral-reinforced nylon 12 resin.

Tests and Results. A visual examination of the filter component revealed significant cracking on the inner surface. The cracking ran along the longitudinal axis of the part and exhibited an irregular pattern. The surfaces of the part presented a flaky texture, without substantial integrity, and displayed significant discoloration. The irregular crack pattern, flaky texture, and discoloration were apparent on all surfaces of the part that had been in contact with the fluid passing through the component. Several of the crack surfaces were further examined using SEM. The fracture surface exhibited a coarse morphology, as illustrated in Fig. 34. The reinforcing glass fibers protruded unbounded from the surrounding polymeric matrix. The fracture surface also showed a network of secondary cracking. Overall, the observations made during the visual and SEM inspections were consistent with molecular degradation associated with chemical attack of the filter component material.

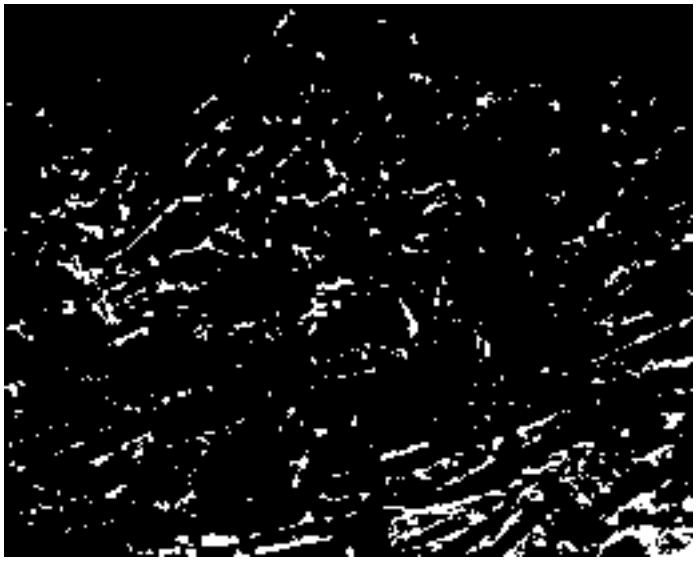


Fig. 34 Scanning electron image showing features characteristic of severe degradation of the filter material. 118×

To allow further assessment of the failure, a mounted cross section was prepared through one of the cracks. The cross section, as presented in Fig. 35, showed a clear zone of degradation along the surface of the part that had contacted the fluid passing through the filter. The degradation zone extended into the cracks, which indicated massive chemical attack. The prepared cross section was analyzed using energy-dispersive x-ray spectroscopy, and the results obtained on the base material showed relatively high concentrations of silicon, calcium, and aluminum, with lesser amounts of sulfur and sodium in addition to carbon and oxygen. The results were consistent with a mineral- and glass-filled nylon resin. Analysis of the surface material, which exhibited obvious degradation, showed a generally similar elemental profile. However, significant levels of silver and chlorine were also found. This was important, because aqueous solutions of metallic chlorides are known to cause cracking and degradation within nylon resins.

Fig. 35 Micrograph showing the cross section prepared through the filter component. 9×

The filter component material was further analyzed using micro-FTIR in the ATR mode. Analysis of the base material produced results characteristic of a glass- and mineral-filled nylon resin. However, analysis of the surface material showed additional absorption bands characteristic of substantial oxidation and hydrolysis of the nylon. A spectral comparison showing this is presented in Fig. 36. The presence of these bands is consistent with the high level of molecular degradation noted during the visual and SEM examinations.

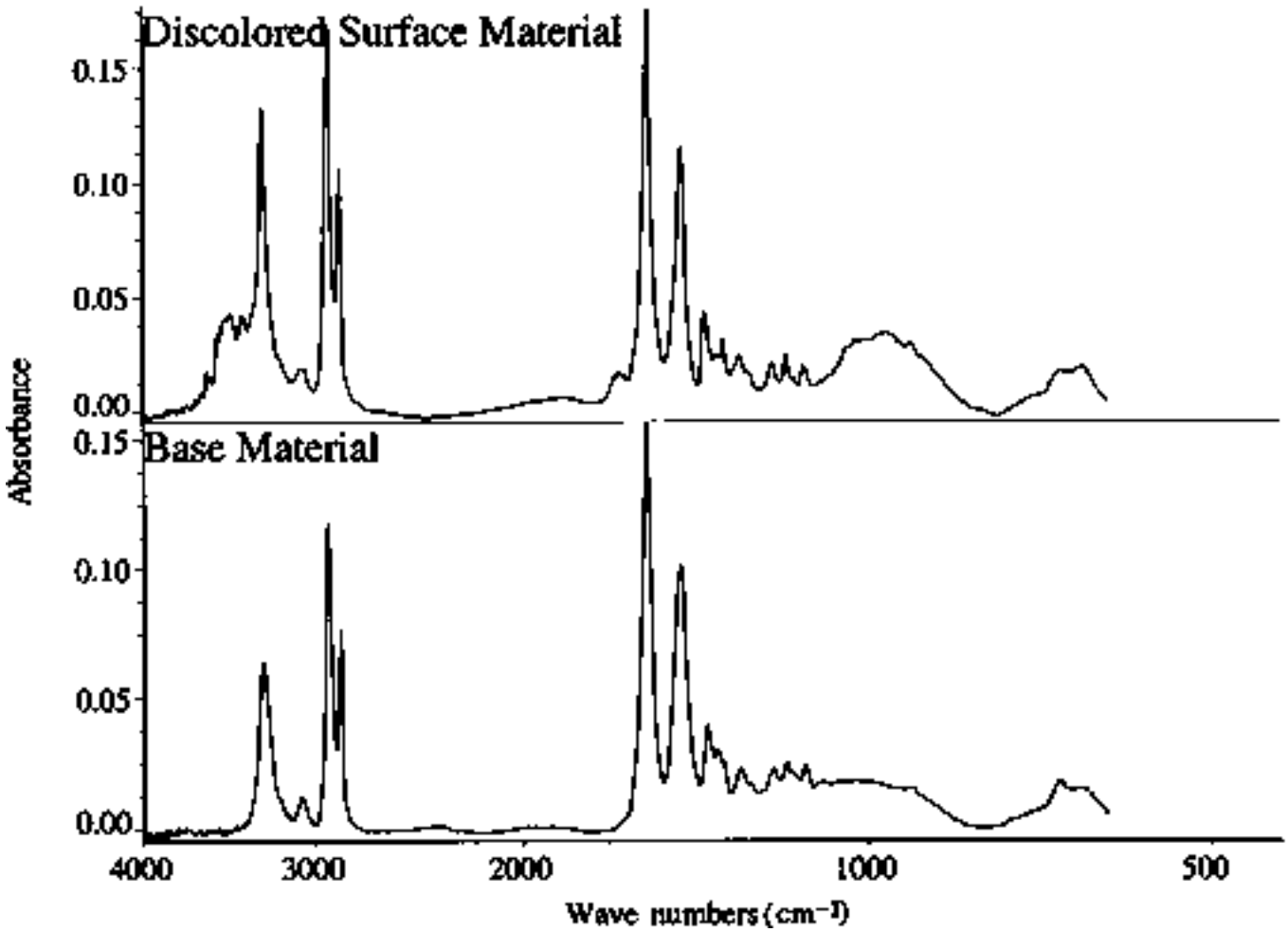


Fig. 36 Fourier transform infrared spectral comparison showing absorption bands associated with hydrolysis at 3350 cm^{-1} and oxidation at 1720 cm^{-1} in the results obtained on the discolored surface

Comparative TGA of the base material and the surface material also showed a significant difference. In particular, the results obtained on the surface material showed a lower temperature corresponding to the onset of polymer decomposition. This is illustrated in Fig. 37.

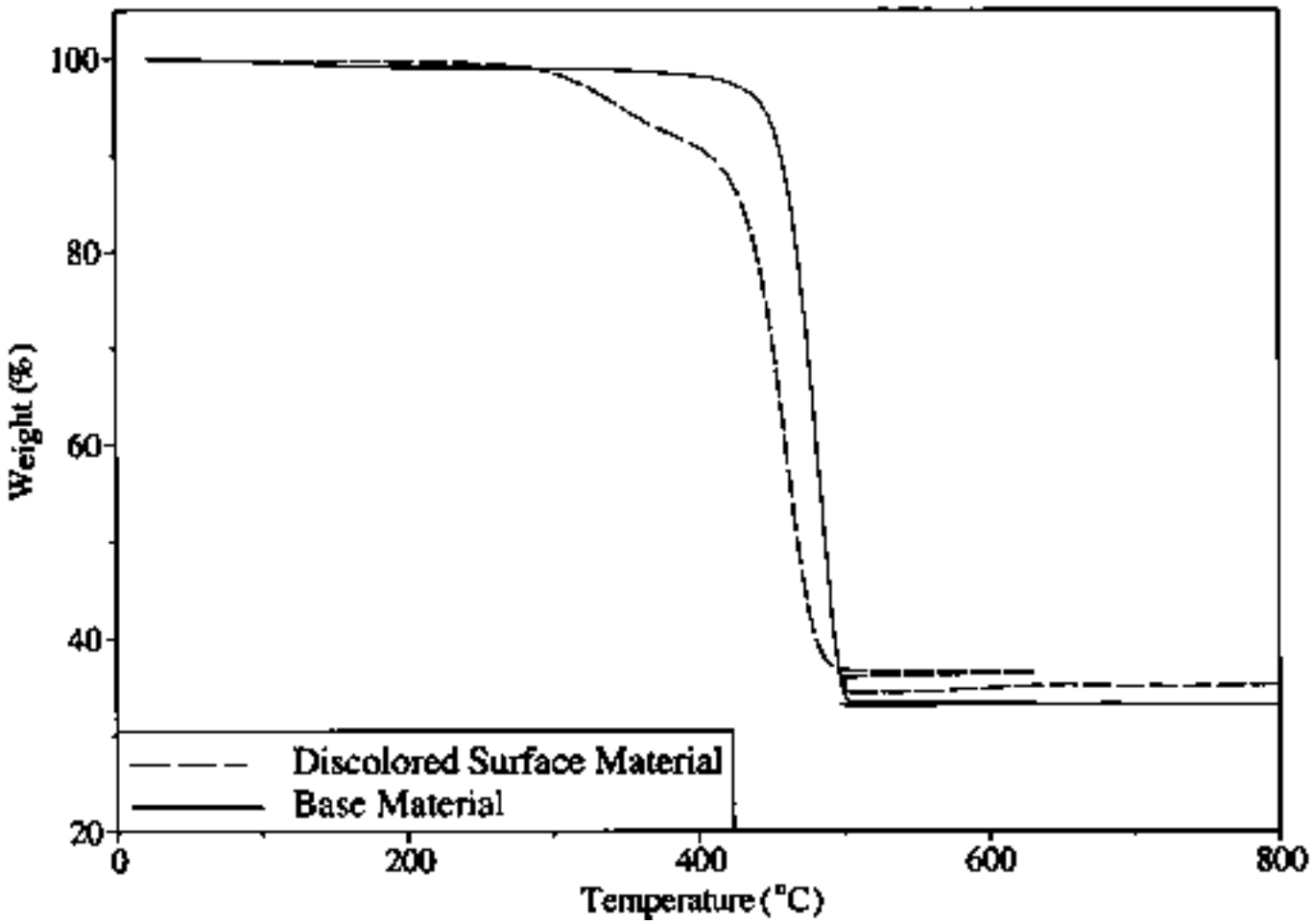


Fig. 37 Thermogravimetric analysis weight-loss profile comparison showing a reduction in the thermal stability of the discolored surface material relative to the base material

Conclusions. It was the conclusion of the evaluation that the filter component failed as a result of molecular degradation caused by the service conditions. Specifically, the part material had undergone severe chemical attack, including oxidation and hydrolysis, through contact with silver chloride. The source of the silver chloride was not established, but one potential source was photographic silver recovery.

Example 14: Failure of a PC Switch Housing. A housing used in conjunction with an electrical switch failed shortly after being placed into service. A relatively high failure rate had been encountered, corresponding to a recent production lot of the housings, and the failed part was representative of the problem. The housing had been injection molded from a commercially available, medium-viscosity grade of PC, formulated with an ultraviolet stabilizer. In addition to the PC housing, the design of the switch included an external protective zinc component installed with a snap-fit and two retained copper press-fit contact inserts. Control parts representing an earlier production lot were available for reference purposes.

Tests and Results. A visual examination of the submitted housing revealed massive cracking within the base of the part, including the retaining tabs securing the contacts. The fractures were primarily located adjacent to the copper contacts. Gray streaks, commonly referred to as splay, were also apparent on the PC housing, as shown in Fig. 38. Splay is often associated with molecular degradation, caused by insufficient drying or exposure to excessive heat, from the molding process. The visual examination also revealed that the contacts corresponding to the failed housing retaining tabs extended significantly, relative to contacts in nonfailed areas. This suggested a high level of interference stress between the contact and the tab. The fracture surface was further inspected using an optical stereomicroscope. The fracture surface showed no evidence of ductility, as would be evident in the form of stress whitening or permanent deformation. An oily residue was evident covering the crack surface.

Fig. 38 A view of the housing showing gray streaks characteristic of splay

The crack surface was further examined via SEM. The fracture surface exhibited multiple apparent crack origins and classic brittle fracture features, including hackle marks, river markings, and Wallner lines. A representative area on the fracture surface is shown in Fig. 39. No evidence of ductility, which would be apparent as stretched fibrils, was found. Overall, the observed features were indicative of brittle fracture associated with the exertion of stresses below the yield point of the material, over an extended period of time, by a creep mechanism.

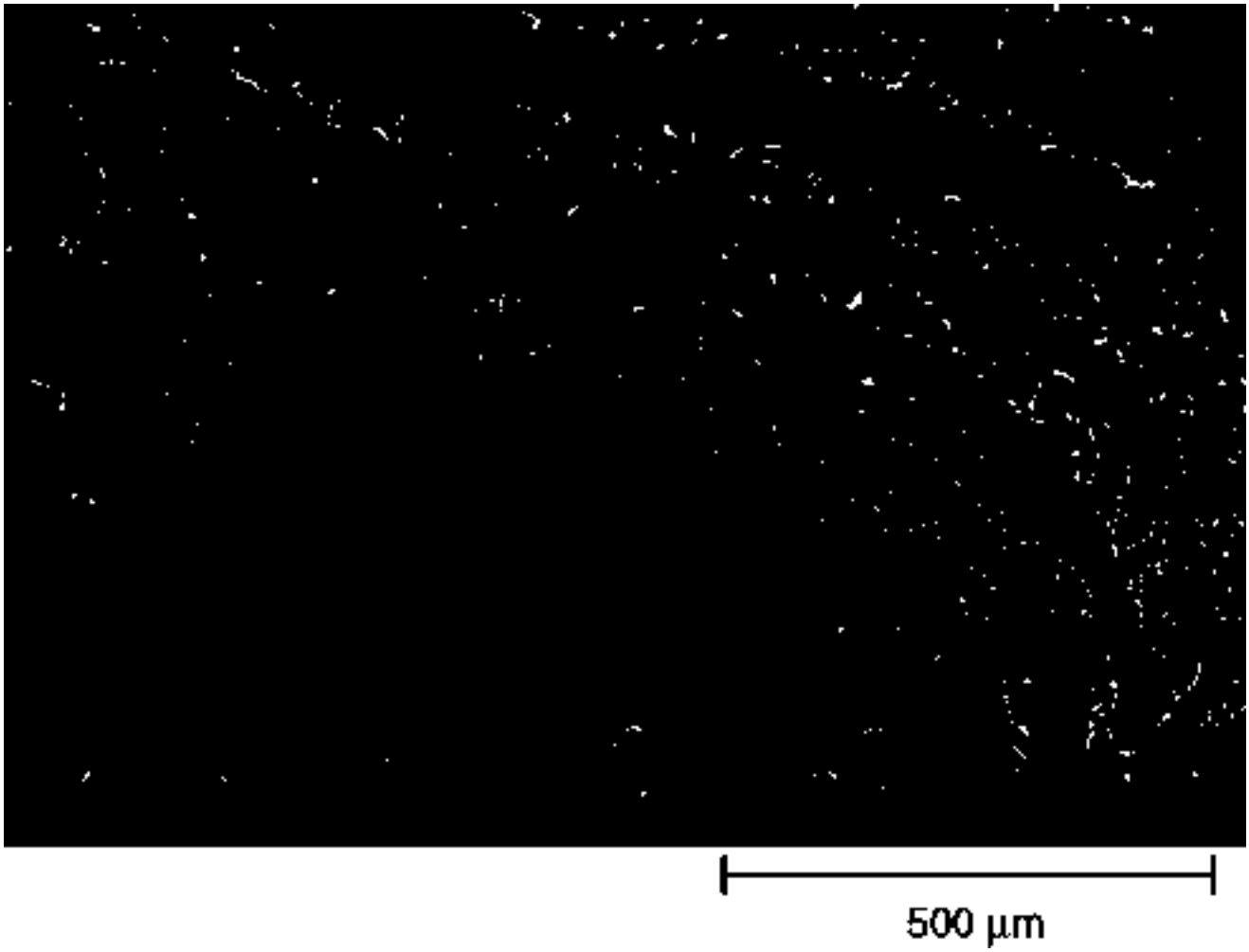


Fig. 39 Scanning electron image showing characteristic brittle fracture features on the housing crack surface. 100×

The housing base material was analyzed using micro-FTIR in the ATR mode, and the resulting spectrum contained absorption bands characteristic of PC. The results produced an excellent match with a spectrum obtained on a reference part, without evidence of contamination. The oily residue found on the part, including the fracture surface, was also analyzed. The obtained spectrum was characteristic of an aliphatic hydrocarbon-based oil, with no signs of aromatic hydrocarbons or other chemicals known to produce stress cracking in PC resins.

The housing material was also analyzed using DSC. The DSC thermogram showed a single transition at 141 °C (286 °F), associated with the glass transition temperature (T_g) of the material. This temperature was somewhat lower than expected for a PC resin, which usually undergoes this transition closer to 150 °C (300 °F). This difference was thought to be an indication of potential molecular degradation.

Thermomechanical analysis (TMA) was used to evaluate the failed retaining tab material, using an expansion probe. The TMA results confirmed the relatively low T_g, in particular, with a comparison to a reference part. A comparison showing this is presented in Fig. 40. No evidence was found in the results to indicate molded-in residual stress.

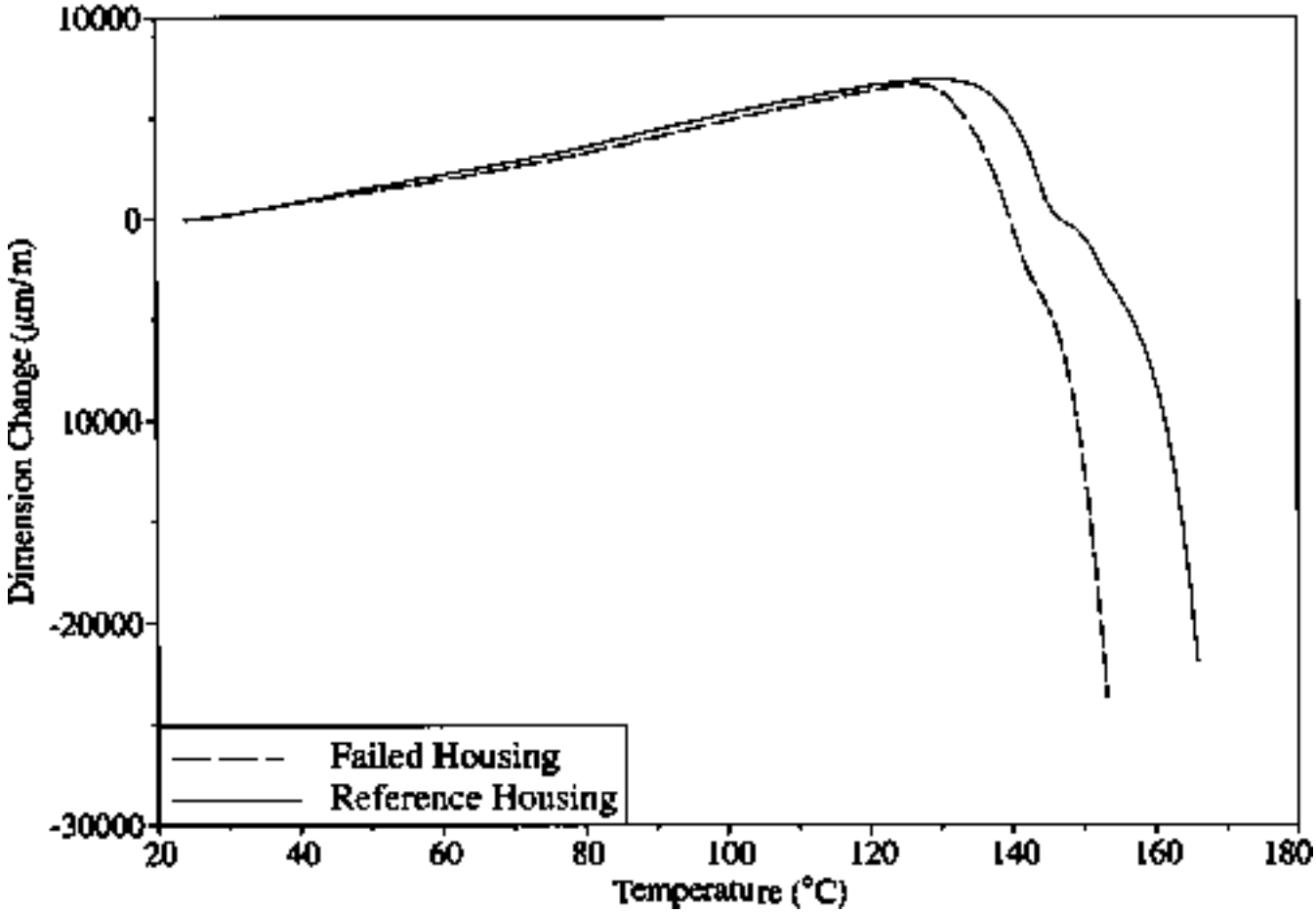


Fig. 40 The TMA results obtained on the failed and reference parts. The results exhibit differences corresponding to a reduction in the glass transition of the failed material.

Melt flow testing of the housing samples showed the submitted reference part to have a MFR of 39.7 g/10 min, compared with 78.1 g/10 min for the failed components. The nominal value for the resin used to produce the housing was 9 to 12 g/10 min. This indicated not only severe molecular degradation within the failed housing material but also within the reference parts. The most likely source of the degradation was the molding process. This degradation was consistent with the presence of splay observed on the part as well as the reduced T_g.

Conclusions. It was the conclusion of the evaluation that the switch housings failed via brittle fracture, likely through a creep mechanism. The failure was caused by severe embrittlement of the housing resin associated with massive molecular degradation produced during the molding process. A potential contributing factor was the design of the part, which produced significant interference stresses between the contact and the mating retaining tab.

Example 15: Failure of Nylon Hinges. A production lot of mechanical hinges had failed during incoming quality-control testing. The hinges were used in an automotive application and had cracked during routine actuation testing. Similar parts had been through complete prototype evaluations without failure. However, a change in part supplier had taken place between the approval of the prototype parts and the receipt of the first lot of production parts. The mechanical hinges were specified to be injection molded from an impact-modified, 13% glass-fiber-reinforced nylon 6/6 resin. A resin substitution was suspected, corresponding to the supplier change. Samples representing the failed components and the original prototype parts were available for the failure investigation.

Tests and Results. A visual examination of the failed parts confirmed catastrophic cracking within the mechanical hinge in an area that would be under the highest level of stress during actuation. The failures did not show signs of macroductility, which would be apparent in the form of stress whitening and permanent deformation. The fracture surfaces of the failed parts were further inspected via SEM. While the presence of glass-reinforcing fibers can render a plastic resin inherently more brittle, a certain level of ductility is still expected at the 13% glass level. This ductility is often only apparent at high magnification and only between the individual glass fibers. However, the failed hinge components did not exhibit any signs of ductility even at high magnification, with the fracture surface showing only brittle features. A laboratory failure was created on one of the prototype parts by overloading the component. Examination of the fracture surface using SEM showed the normally anticipated level of ductility, as indicated by the overlapping, rose-petal morphology. The crack surfaces of both the failed part and the laboratory fracture are shown in Fig. 41.

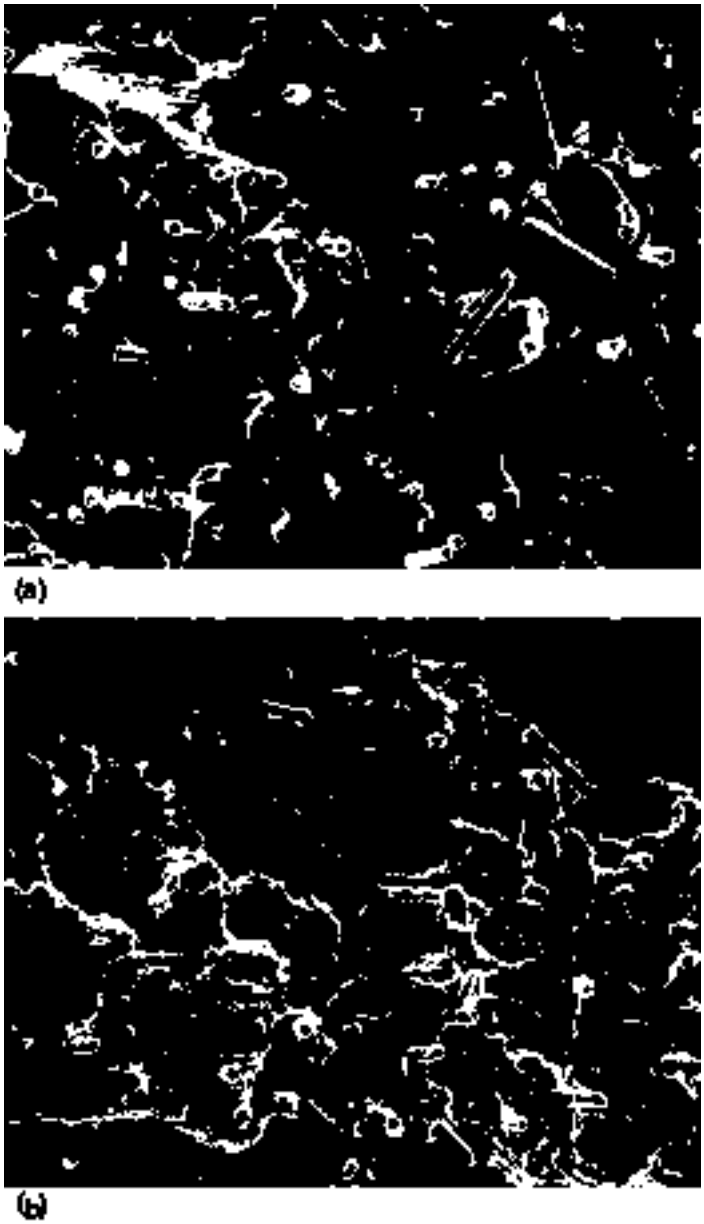


Fig. 41 Scanning electron images showing (a) brittle fracture features on the failed hinge and (b) ductile fracture features on the laboratory fracture. 118×

Analysis of the failed components and the corresponding molding resin via micro-FTIR produced results characteristic of a nylon resin. The molding resin and failed parts generated generally similar results. However, a distinct difference was apparent in that the spectra obtained on the failed parts showed an additional absorption band at approximately 1740 cm^{-1}, indicative of partial oxidative degradation of the resin. A spectral comparison illustrating this is presented in Fig. 42. Because the parts had not yet been in service, this degradation was thought to have occurred during the molding process.

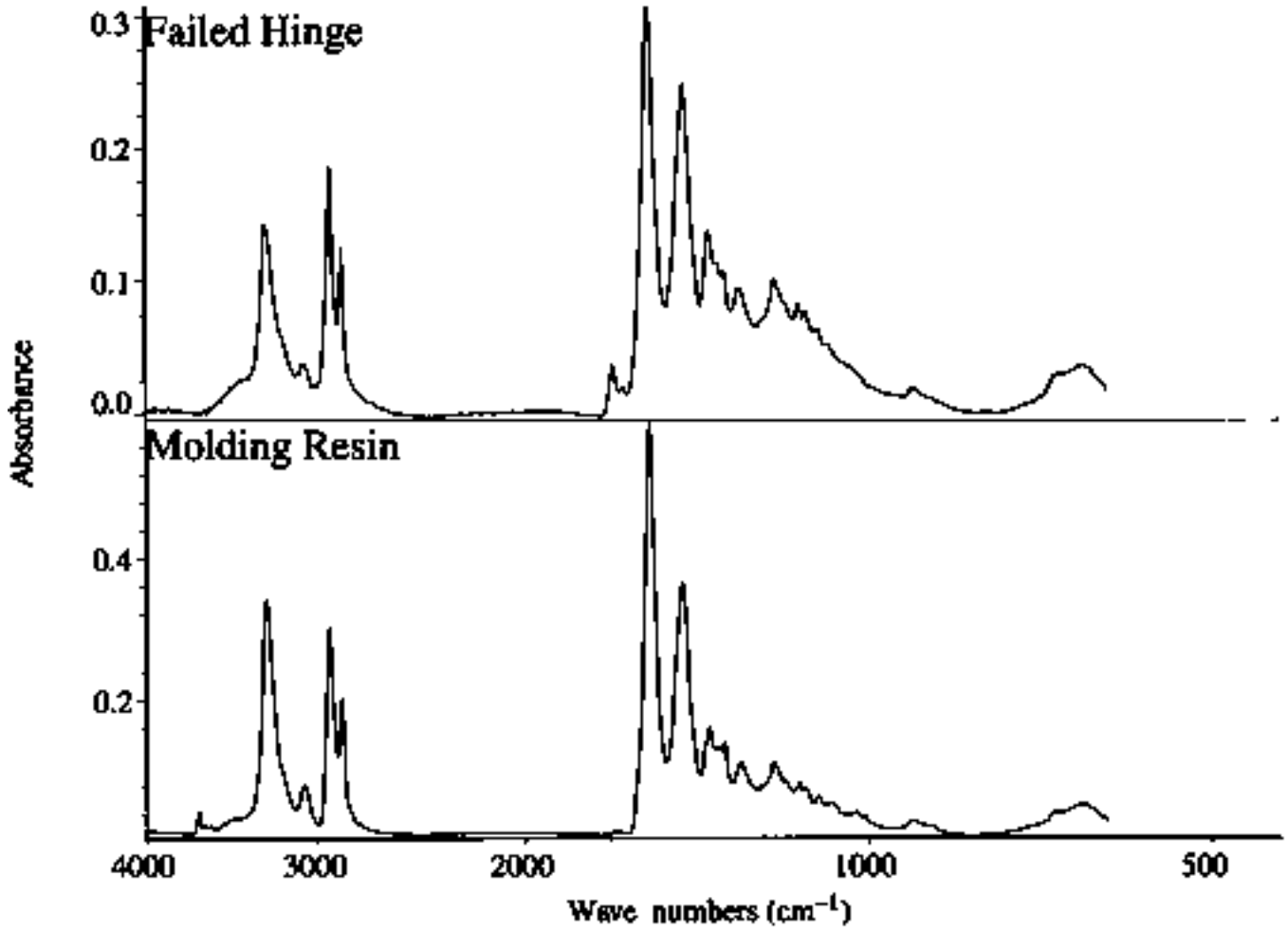


Fig. 42 Fourier transform infrared spectral comparison showing absorption bands at 1740 cm^{-1}, characteristic of oxidation within the results obtained on the failed parts

The failed parts were further tested using DSC. The obtained DSC results showed a melting point of 263 °C (505 °F), consistent with a nylon 6/6 resin. The molding resin was also analyzed via DSC, and a comparison of the results further indicated degradation of the molded nylon resin. This was apparent by a noted reduction in the heat of fusion in the results representing the failed parts.

The failed parts and the prototype parts were also analyzed using conventional thermogravimetric analysis (TGA), and both analyses produced results indicative of a nylon resin containing approximately 13% glass-fiber reinforcement. Further testing was performed using TGA in the high-resolution mode. This analysis was conducted in order to assess the level of impact-modifying rubber resin. The weight loss associated with the rubber was observed as a shoulder on the high-temperature side of the weight loss representing the nylon resin. This weight loss was particularly evident in the derivative curve. Because the weight losses could not be totally resolved, an absolute level of rubber could not be determined. However, a comparison of the results allowed a determination of the relative level of the impact modifiers in the two materials. This comparison showed a distinctly higher level of impact modifier in the prototype part material, relative to the failed part material.

Conclusions. It was the conclusion of this evaluation that the hinge assemblies failed through brittle fracture associated with stress overload during the actuation of the parts. The failed part material was found to be degraded, as indicated by both the FTIR and DSC analysis results. This degradation likely occurred either during the compounding of the resin or during the actual molding of the parts. A significant factor in the hinge failures is the conversion to a different grade of resin to produce the failed production parts as compared to the prototype parts. While both resins produced results characteristic of a 13% glass-fiber-reinforced, impact-modified nylon 6/6, the failed part material contained a significantly lower level of rubber. This decrease in rubber content rendered the parts less impact resistant and subsequently lowered the ductility of the molded hinge assemblies.

References

1. J.A. Jansen, Conducting a Plastic Component Failure Investigation: Examples from the Appliance Industry, *International Appliance Technology Conference,* March 2002, p 2

2. J.A. Jansen, Plastic Component Failure Analysis, *Adv. Mater. Process.,* May 2001, p 56, 58, 59

3. A.T. Riga and E.A. Collins, Analysis of Structure, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 825, 826

4. J. Scheirs, *Compositional and Failure Analysis of Polymers,* John Wiley & Sons, 2000, p 109, 138, 153, 393, 415

5. M.P. Sepe, *Thermal Analysis of Polymers,* RAPRA Technology, Shawbury, U.K., 1997, p 3, 4, 8, 17, 19, 22, 24, 33

6. L.C. Roy Oberholtzer, General Design Considerations, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 21

7. "Polymer Characterization: Laboratory Techniques and Analysis," Noyes Publications, 1996, p 15

8. S.B. Driscoll, Physical, Chemical, and Thermal Analysis of Thermoplastic Resins, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 533

9. S. Turner, Mechanical Testing, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 545

10. M. Ezrin, Plastics Analysis: The Engineer's Resource for Troubleshooting Product and Process Problems and for Competitive Analysis, *Plast. Eng.,* Feb 2002, p 45, 46

Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Selected References

- W. Brostow and R.D. Corneliussen, *Failure of Plastics,* Hanser Publishers, 1986
- T.R. Crompton, *Practical Polymer Analysis,* Plenum Press, 1993
- T.R. Crompton, *Manual of Plastics Analysis,* Plenum Press, 1998
- M. Ezrin, *Plastics Failure Guide,* Hanser Publishers, 1996
- *G.E. Engineering Thermoplastics Design Guide,* G.E. Plastics, 1997
- J.W. Gooch, *Analysis and Deformulation of Polymeric Materials,* Plenum Press, 1997
- J. Moalli, Ed., *Plastics Failure: Analysis and Prevention,* Plastics Design Library, 2001
- T.A. Osswald and G. Menges, *Materials Science of Polymers for Engineers,* Hanser Publishers, 1995

- R.C. Portney, Ed., *Medical Plastics: Degradation Resistance and Failure Analysis,* Plastic Design Library, 1998
- B.C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy,* CRC Press, 1996
- E.A. Turi, Ed., *Thermal Characterization of Polymeric Materials,* Academic Press, Inc., 1981
- D. Wright, *Failure of Plastics and Rubber Products,* RAPRA Technology, Shawbury, U.K., 2001

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Introduction

THE STRESSES acting on a component may cause unacceptable deformation, cracking, or fracture. By definition, after a stress related failure has occurred, it is a given that the stresses reached a critical level, at least at one critical location. Or it may be that fracture did not occur. It should be remembered that some parts are designed to fracture under certain circumstances. Not doing so may constitute the failure. Shear pins, for example, are commonly used to protect valuable equipment by acting as mechanical fuses. Stress directors in helicopter gearing are also designed so that if gear tooth fracture occurs, only half the tooth breaks off. This leaves the other half of the gear tooth to carry the load, simultaneously creating sufficient vibration to warn the pilot of imminent failure but allowing him sufficient flying time to land.

In any case, the goal of failure analysis is to understand a situation sufficiently well to determine if changes in design, manufacture, or other areas are needed. When confronted with a cracked, fractured, or deformed component, the failure analyst will usually seek to answer some basic questions:

- Were the loads and stresses encountered by the part at the level anticipated during design? Or did some unexpected condition(s) contribute to the failure?
- Was the material in the area of the cracking or deformation capable of meeting the conditions anticipated during design? Was there some deficiency or discontinuity that contributed to the failure, or was there a local stress raiser at the critical location? Was this taken into account by the designer?

Of course, the root causes of unacceptable stress concentrators are varied. They may occur from design (e.g., changes in section), manufacturing (e.g., surface condition deleterious inclusions or unacceptable porosity), and/or service exposure (e.g., corrosion and wear). The critical point is that if the causes (i.e., stresses and associated loads) contributing to the failure are not sufficiently understood, then it may be most difficult to prevent it from happening again. For example, consider a part with a fillet radius that is somewhat sharper than indicated on a drawing, and that the part fails at the fillet radius. Did the part fail because the fillet radius was out of design tolerance, or did it fail because of some other application or manufacturing factor(s)? A safety factor was probably used when the drawing was created with a maximum assumed load, but if the part was loaded well above the original design load, it could have failed whether the fillet radius was correct or not.

There are many reasons why a failure may occur, and a complete picture is required to gain insights on a situation that is potentially complex (and apparently unanticipated as well because, by definition, failures are unanticipated). The root-cause circumstances of a failure are therefore potentially subtle, complex, and/or synergistic. In this context, the analyst must be aware of not only the details of the component and the stress analysis, but also the “big picture.” It is common for the materials failure analyst to identify the fracture origin and any material discontinuity at the origin. But what is the root cause? For example, in a welded structure, the joints are by their very nature discontinuities and thus likely origins in fractures. However, did the welds meet the conditions for “fitness-of-purpose” requirements? Were the weld codes and standards applicable for the actual conditions?

A part is generally expected to fracture at the maximum stress location unless there is a material “deficiency” at some other location. However, if the deficiency is an expected material deviation that is properly anticipated during design, then the material condition may not be the root cause. In this case, a design review may be initiated with a more extensive analysis of loads and stresses. This design review also may find that none of the expected stresses are sufficient to explain the condition of the component (i.e., the designer correctly accommodated the anticipated stresses). In this case, the source of the additional stress may be unanticipated loads and possibly, inappropriate use. In this regard, the possibilities are many, such as operator overloading, unusual resonance vibration, or unexpected stress introduced as a result of assembly. The critical point is that

the analyst must look at the overall situation, including the loads and the distribution of stresses within the configuration. Is there an additional load/stress that was not included in the analysis, which changes the stress distribution? Or is the stress analysis incorrect?

In many cases, however, stress analysis alone is not sufficient to explain how a fracture occurred. The failure analyst must also apply the principles of fracture mechanics to understand how a crack developed and grew. The issue in this portion of the failure analysis is to determine if the expected stresses would have been sufficient to initiate, propagate, and grow a crack at the specific location and under the conditions for which the crack actually developed. The most insidious failures are those in which one would not expect fracture from consideration of the stress level alone, but cracking and ultimate fracture nevertheless do occur, often due to a combination of localized stress concentration, fluctuating loads, and/or aggressive environment.

This article describes concepts and tools that can be used by the failure analyst to address the above questions and additional queries that naturally follow. Only an overview and review of some of the basics is possible in an article such as this. The reader is referred to some of the many excellent texts (see the "Selected References" for this article) for further study. The perspective of this article is that of the failure analyst who is confronted with a problem; that is, a failed part, and asked to understand how it happened and how to keep it from happening again.

Throughout the material herein, it also should be kept in mind that the application of stress analysis and fracture mechanics tools is somewhat different than it is in design. The goal of the designer is to prevent failures from occurring by assuring that stresses in service do not reach a critical level. Uncertainty and unknowns in analysis, manufacture, service, and material conditions are typically addressed by applying so-called "factors of safety."

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

Analysis of Applied Stresses

Stress analysis provides a road map for the failure analyst. If fracture initiates in a low stress area, several working hypotheses are suggested. The part may not have been used as intended. It may have suffered mechanical or environmental damage prior to use or the presence of a large material flaw may have predisposed it to failure by creating a point or volume in which the actual stress exceeds the strength of the material. Additionally, the microstructure may vary from location to location. On the other hand, if failure occurs in an area of high stress, overload may also be suspected. Of course the possibility of improper design cannot be discounted. Knowledge of the stress distribution can point the failure analyst to the crack origin, especially if the crack surface is nearly featureless. It can also help check for consistency with the fracture features.

In the conventional approach to design and failure analysis, an elastic stress analysis of the component is necessary. Stress analysis is initially performed in the design stage of a development project. Frequently, principal stresses and maximum shear stresses are important to the designer because both are used in common failure expressions to calculate maximum load capability (e.g. multiaxial yield criteria). One failure mode theory is the Rankine (or maximum normal stress) criterion, where the maximum principal stress is used. The other is the Tresca theory based on the maximum shear stress. The Rankine criterion holds that inelastic deformation at a point begins when the maximum normal stress reaches the yield strength of the material. Tresca's theory states that yield at a point begins when the maximum shear stress equals the shear stress in a tension specimen when it begins to yield. Another common failure model is the von Mises criterion, which is based on a determination of distortion energy at yield and recognizes that in a constant-volume isotropic material, hydrostatic stresses cannot cause failure by yield.

An octahedral shear stress criterion provides a mathematically identical expression to the von Mises criterion. Several other criteria have been proposed (e.g., one based on maximum strain). Additionally, some criteria take into account polar behavior (unequal tensile and compressive failure stresses), anisotropic behavior and nonconstancy of volume during deformation. The historically early models (Rankine, Tresca, von Mises, and octahedral shear) were phenomenological and as proposed could be applied to either failure by yield or failure by fracture. They have also been applied to cyclic loading. In practice, the Tresca and von Mises criteria are most often used today to predict yield in isotropic constant-volume materials. The Rankine model is not often used because of its inability to predict behavior when one principal stress is negative and another positive. When used to predict failure of materials that are not very ductile, none of the

classic models are able to predict behavior very well in all four quadrants of biaxial stress space. A maximum normal stress criterion is sometimes used but again poorly predicts behavior when principal stresses are of opposite sign. Contemporary models to predict plastic deformation utilize the "flow" (i.e., yield) surface via the plastic potential equation. Because any linear model to predict flow (e.g., the Tresca criterion) is discontinuous in principal stress space, it cannot predict plastic strains for several loading conditions (e.g., balanced biaxial tension). As a result, modeling of plastic flow typically utilizes the nonlinear distortion energy criterion or its anisotropic equivalent, the Hill criterion (see the appendix in the article "Mechanisms and Appearances of Ductile and Brittle Fracture in Metals" in this Volume).

This section describes the underlying fundamentals and the relevance and necessity of performing proper stress analysis in conducting a failure analysis. Both plane stress and plane strain are explored, and instances in which seemingly appropriate two-dimensional (2D) analyses lead to erroneous results are presented. Three-dimensional (3D) stress transformations are discussed, and a procedure to obtain the principal stresses and associated directions as well as the maximum shear stresses and the planes on which they act is described. Separation of the stress tensor into hydrostatic and deviatoric components also is examined.

The method or tools of stress analysis may vary. In some instances of simple geometry and loading, stress analysis is performed by hand with well-known formulae. In other cases the material may not behave linearly, the geometry and/or loading may be complex, or time-varying boundary data may be present. These require the use of numerical procedures such as finite-element analysis (FEA) or boundary-element analysis (BEA). While FEA discretizes the entire volume into small elements (normally tetrahedra or hexahedra), BEA only discretizes the boundary. Both methods, however, employ continuum material models and produce only approximations. In general, accuracy improves as the number of elements increases. Convergence techniques are generally employed to ensure that a given solution is "correct." Program output is the stress state in a given component or assembly for a given combination of loading and boundary conditions. Although this stress state is unique to a specific coordinate system, most commercially available software packages can extract principal stresses as well as maximum shear stresses. A procedure for calculating the planes on which the principal stresses and maximum shear stresses act is described in the section "Transformation of Stresses" in this article.

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Fundamentals of Stress Analysis

Most of the conventional failure criteria are stress based. As such, the stress analysis of a geometrical component is important for design, reliability analysis, and failure analysis. This stress analysis is usually based on the mechanics of materials approach but could also use the theory of elasticity approach in some cases. There are many references in the literature that describe the state of stress for common geometries (Ref 1, 2, 3, 4), so in many instances it is not necessary to do original work but rather to know how to apply the equations that describe the state of stress.

In this section, issues related to the determination and use of stress are detailed. Stress is defined, and a procedure to deal with stress by determining maximum values through stress transformation is described. The stress analysis equations of typical component geometries are given, and some of the implications of the stress analysis relative to failure in components are also discussed. The component stresses determined by analysis usually lie in directions defined by the shape of the component geometry, for example, along major axes or perpendicular to major axes. These stresses may not be the maximum ones, so they must then be examined to find the actual maxima to be used in a failure criterion. These maxima are determined by taking stress transformations to find maximum normal stresses (principal stresses) and maximum shear stresses as well as by some other special stress definitions.

The section "Defects and Imperfections" in this article introduces the stress analysis of bodies containing cracklike imperfections and the topic of fracture mechanics. The fracture-mechanics approach is based on the analysis of the crack tip stress and strain field. The initial approach proposed by George Irwin, the father of fracture mechanics, used the stress-intensity factor, K, to describe the magnitude of a field with a unique stress distribution and to relate the boundary loading to the crack tip stresses. Irwin showed that as long as the region defined by the K field was small relative to the dimensions of a component, the stress field defined by the K field could be transferred from one body to another body, considerably easing the solution of problems. Results from simple laboratory geometries could be applied to manufactured components. As such, the fracture mechanics approach is an important part of stress analysis at the tips of sharp cracks or discontinuities. This is briefly discussed in the section "Fracture Mechanics" in this article.

Definitions

Stress Definition. A simple definition of stress is as a force per unit area. In a geometry that has a uniform stress distribution, such as a bar in simple tension, a force, *F*, applied along the axis of the bar that has area, *A*, has uniform stress, , where:

$$= F/A \qquad \text{(Eq 1)}$$

See Fig. 1(a). If the force is perpendicular to the area as shown, the stress, , is defined as an average normal stress. In a similar way a body may have a force, *V*, that is parallel to the area, *A*. The stress that is generated by this force is defined as a shear stress and usually given the symbol, !. Then:

$$! = V/A \qquad \text{(Eq 2)}$$

is the average shear stress (see Fig. 1b).

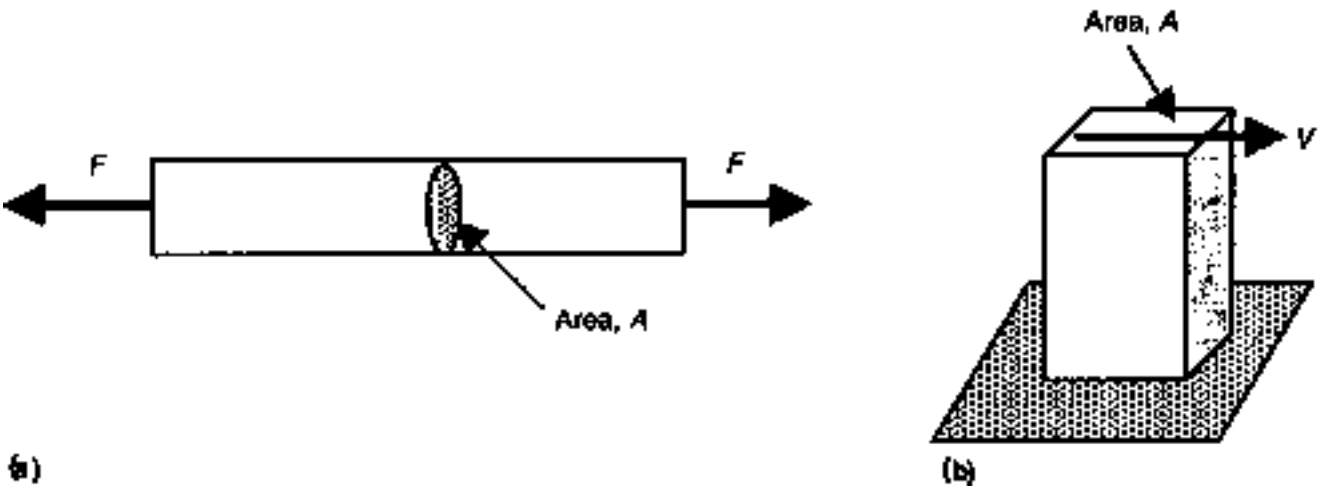


Fig. 1 Definition of (a) average normal stress and (b) average shear stress. *F*, force (load); *V*, force parallel to area.

For most geometrical configurations, the stress is not uniform but varies as a function of position in the body. A more general definition of the stress is then given as follows. An arbitrary body having forces F_i applied (such that the body is in equilibrium) is cut, revealing an internal surface (Fig. 2). The internal surface has a distribution of force that is needed for equilibrium of the cut section. Over an increment of area, " *A*, an increment of force, " *F*, is required. Then the stress vector, *S*, is defined as:

$$S = \lim_{\Delta A \to 0} \left(\frac{\Delta F}{\Delta A} \right) \qquad \text{(Eq 3)}$$

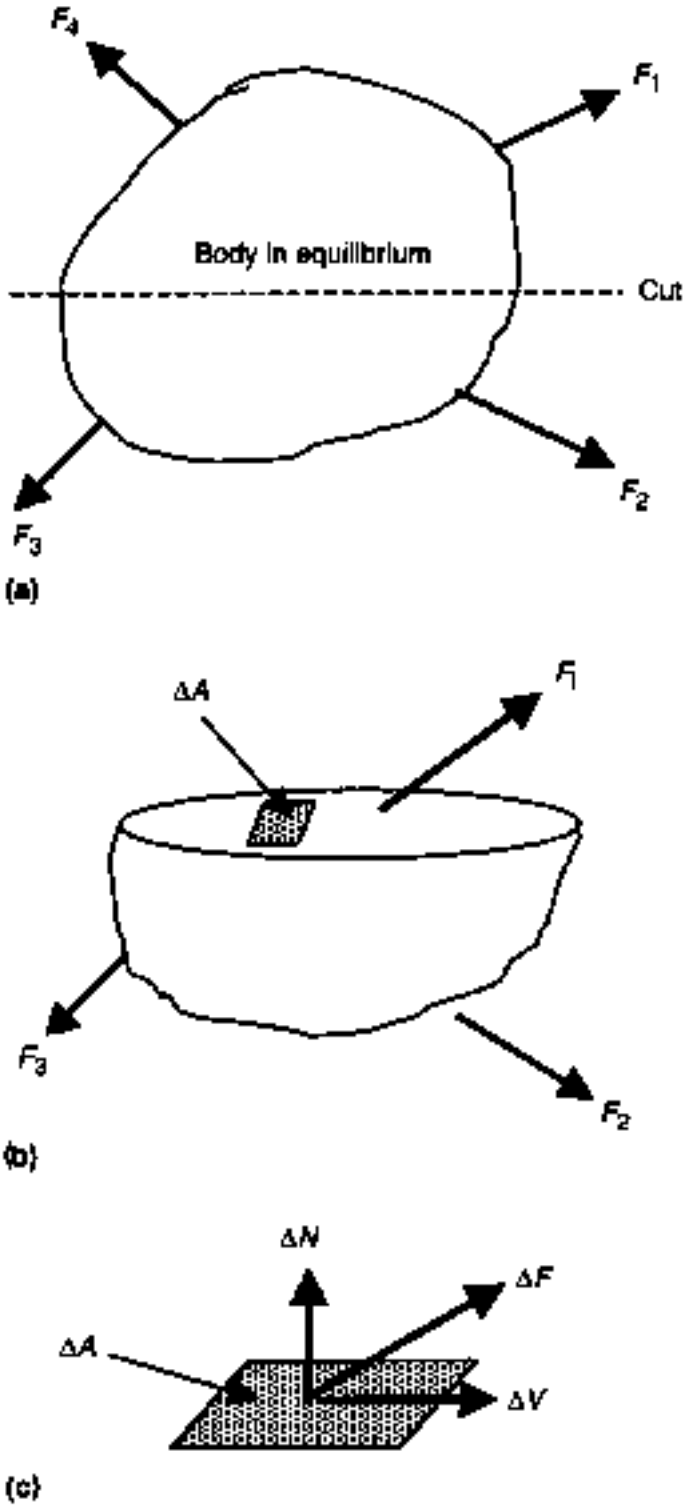


Fig. 2 Stress as a function of position in a body. (a) Cut from a body in equilibrium. (b) Internal force, F_i, on cut surface. (c) Components of F on A

If the force " F is resolved into mutually perpendicular components such that " N is a component normal to the area, " A, and " V is parallel to " A, then the normal component of stress, , and shear component of stress, !, are defined as:

$$s = \lim_{\Delta A \to 0} \left(\frac{\Delta N}{\Delta A} \right) \qquad \text{(Eq 4)}$$

and

$$t = \lim_{\Delta A \to 0} \left(\frac{\Delta V}{\Delta A} \right) \qquad \text{(Eq 5)}$$

The force " *F* and its components are functions of their position on the cut surface. The possibility that the stresses and ! are functions of position on the surface must also be permitted (as in the stress field around a stress concentrator or a stress state in bending or torsion). In addition, if the surface is cut in a different orientation, the forces are different and the stresses are different as well. Therefore the stresses, described with reference to a coordinate system are functions of position in a body as well as the orientation of the area over which the stress is defined and the coordinate system used to determine stresses.

The stress definitions given by Eq 4 and 5 is generated as a result of forces applied to a surface of the body. These stresses derived from surface forces are often labeled as *primary stresses.* Stresses can also be generated as a result of body forces and strain mismatches. Stresses coming from strain mismatch are often labeled *secondary stresses.* Common secondary stresses that are encountered are residual stresses and thermal stresses.

Residual stresses are important in failure analysis. They are different from the applied stresses in that when the forces are released, the stresses created by applied loads go to zero, but the residual stresses remain locked in the body. The way to treat residual stresses has been investigated and is still a matter of concern. Some methods to deal with it are discussed later.

Stress Components. The description of stress at a point in three dimensions requires the labeling of stresses that act on three mutually perpendicular planes and in three mutually perpendicular directions. Most generally, there are nine components of stress at a point. The labels can be given for the type of coordinate axes chosen, for example, Cartesian, polar, or spherical. For defining stress components, the labels may be given in Cartesian coordinates where the mutually perpendicular axes are labeled *x*, *y*, and *z*. Keeping the stresses defined the same as previously, for normal stress and ! for shear stress, the nine stress components are defined with the subscripts referring to the coordinate directions. For example, $_{xx}$ is the normal stress acting on a plane perpendicular to the *x* coordinate direction and acting in the *x* direction. Similarly, $!_{xy}$ is the stress acting on a plane normal to *x* and acting in the *y* direction. Using this nomenclature, nine stress components can be defined by:

$$\begin{bmatrix} s_{xx} & t_{xy} & t_{xy} \\ t_{yx} & s_{yy} & t_{yz} \\ t_{zx} & t_{zy} & s_{zz} \end{bmatrix}$$

These nine components are said to comprise the stress tensor. They are shown in Fig. 3 on a stress element.

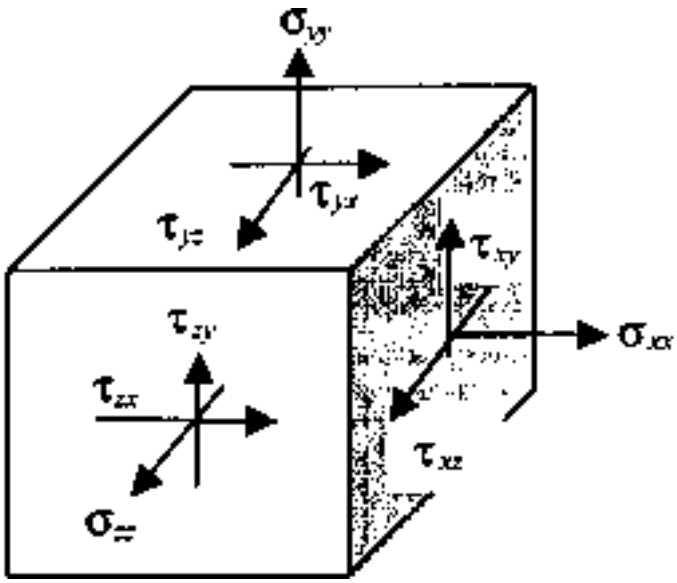


Fig. 3 Three-dimensional stress components

It is usual to drop one of the subscripts on the normal stress components because they always act on the plane that has the normal in the same direction that the stress acts. This gives repeated indexes. Therefore, $_{xx}$ can be written $_{x}$, and the other normal stresses can be written likewise. Also, a simple application of equilibrium (moment summation) shows that the shear stress components, $!_{xy}$ and $!_{yx}$ are equal. This applies as well to the other shear stress components with similar but transposed subscripts. The stress tensor is said then to be symmetric and therefore has only six independent components. The symmetric stress tensor can be written as:

$$\begin{bmatrix} s_x & t_{xy} & t_{xy} \\ t_{xy} & s_y & t_{yz} \\ t_{xz} & t_{yz} & s_z \end{bmatrix}$$

Stress Transformations

Internal forces as well as those on internal areas are functions of the orientation of the forces and areas. The redefining of stress components from one orientation to another is accomplished by an operation referred to as *stress transformation.* In a component design, reliability analysis, or failure analysis, the use of the stress transformation is essential so that the highest and most damaging stresses can be determined and applied in the analysis. In most stress analyses, the formulas that are developed from mechanics of materials principles or from theory of elasticity are in directions that correspond to the geometrical construction of the body. For example, Cartesian coordinates can be taken along an axis of a bar and perpendicular to the axis. For round shapes such as pipes or pressure vessels, the radial, circumferential, and axial directions can be chosen. These are taken for convenience in the analysis of the stresses in these components. However, finding the stresses in the directions corresponding to the body dimensions does not guarantee that they are the maximum and most damaging stresses. The ability to transfer stresses to another orientation (reference frame) is essential in determining these maximum stresses. The methods of stress transformations are discussed first, and then the determination of stress maxima is discussed.

To define stress transformation, it is easier to start with a two-dimensional case. One common way to define a two-dimensional (2-D) stress state is with a plane stress assumption. Plane stress assumes that all stresses in the third direction, *z*, are zero so that:

$$_z = 0 \qquad !_{xz} = 0 \qquad !_{yz} = 0 \qquad \text{(Eq 6)}$$

Also, the plane stress assumption requires the nonzero stresses to be only functions of *x* and *y*.

The stress transformation technique can be illustrated easily in two dimensions by using the principles of statics. In two dimensions, only three of the six stress components are used: $_x$, $_y$, $!_{xy}$. These can be shown on a 2D stress element, Fig. 4a, where *x* is a horizontal component and *y* a vertical component. The element is cut at an angle, #, and new axes defined as *x*$and *y*$are in the direction defined by a rotation of the axes through the angle # (Fig. 4b). A new stress element defined by the cut surface takes the shape of a triangle defined by one face perpendicular to *x*, a second face perpendicular to *y*, and the third perpendicular to the *x*$ The element is taken to have a unit thickness such that the area of each face can be defined by # and the area of either the *x* or *y* face as shown in Fig. 4(c). The stress components on the face perpendicular to the *x*$direction are labeled $_{x\$}$and $!_{x\$\$}$

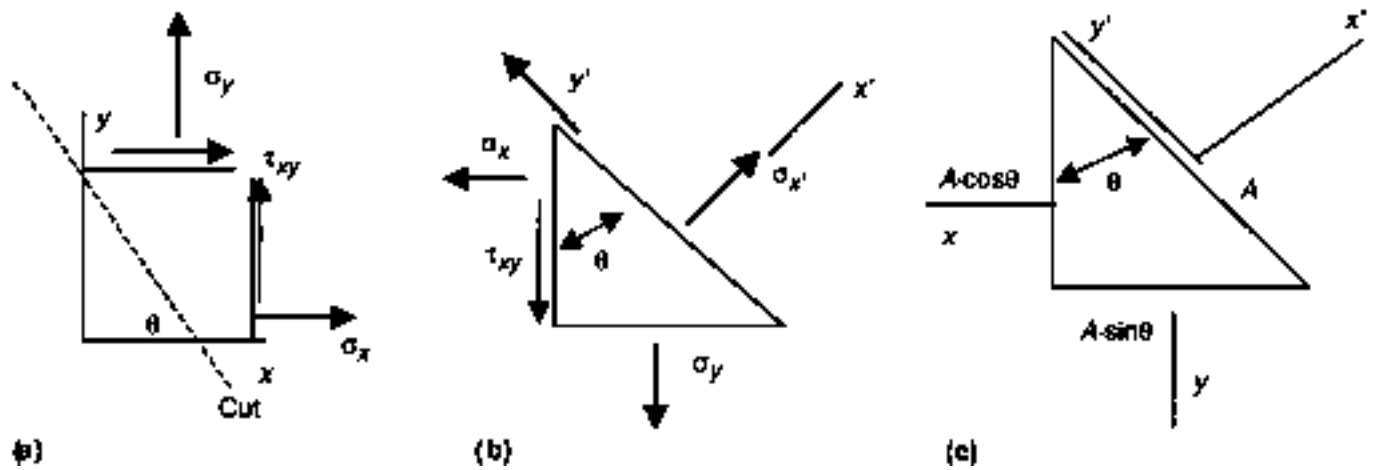


Fig. 4 Stress transformation in two dimensions. (a) Two dimensional stress at a point with a cut angle ! . (b) Stresses along rotated axes, *x*", *y*". (c) Area values on cut element

Static equilibrium can be applied to forces, not stresses. Therefore, the stresses have to be converted to force by multiplying by an appropriate area. In Fig. 4(c) the areas of the primed faces are related to the original faces by the angle #. If *A* is the area normal to *x*$ then *A* · cos# is the area normal to the *x* axis and *A* · sin# is the area normal to the *y* axis.

The force in the $x\$$ direction is $_{x\$}A$. This can be used to define equilibrium in the $x\$$ and $y\$$ directions by using the statics equilibrium equations:

$$\%F_{x\$} = 0 \qquad \%F_{y\$} = 0 \qquad \text{(Eq 7)}$$

Equilibrium in the $x\$$ direction gives:

$$\%F_{x\$} = 0 = {}_{x\$}(A) - {}_{x}(A\cos\#)\cos\# - {}_{y}(A\sin\#)\sin\# - !_{xy}(A\cos\#)\sin\# - !_{xy}(A\sin\#)\cos\# \qquad \text{(Eq 8)}$$

Solving for $_{x\$}$ gives:

$$_{x\$} = {}_{x}\cos^2\# - {}_{y}\sin^2\# - 2!_{xy}\cos\#\sin\# \qquad \text{(Eq 9)}$$

Theses are usually transformed to double angle format by:

$$\cos^2\# = (1 + \cos 2\#)/2$$

$$\sin^2\# = (1 - \cos 2\#)/2 \qquad \text{(Eq 10)}$$

$$2!_{xy}\cos\#\sin\# = \sin 2\#$$

so that

$$\sigma_{x'} = \frac{\sigma_x + \sigma_y}{2} + \frac{\sigma_x - \sigma_y}{2}\cos 2\theta + \tau_{xy}\sin 2\theta \qquad \text{(Eq 11)}$$

Writing the sum of the forces in the y direction, an expression can be developed for $!_{x\$y\$}$

$$\tau_{x'y'} = -\frac{\sigma_x - \sigma_y}{2}\sin 2\theta + \tau_{xy}\cos 2\theta \qquad \text{(Eq 12)}$$

For a cut face with perpendicular direction, $y\$$ the same equilibrium can be analyzed or # simply can be replaced by 90° + # in Eq 11 to get $_{y\$}$

$$\begin{aligned}\sigma_{y'} &= \frac{\sigma_x + \sigma_y}{2} - \frac{\sigma_x - \sigma_y}{2}\cos(2\theta + 180^\circ) \\ &\quad + \tau_{xy}\sin(2\theta + 180^\circ) \qquad \text{(Eq 13)} \\ &= \frac{\sigma_x + \sigma_y}{2} + \frac{\sigma_x - \sigma_y}{2}\cos 2\theta - \tau_{xy}\sin 2\theta\end{aligned}$$

Equations 11 to 13 are the 2D stress transformations. In a more complex way, a three-dimensional (3D) stress transformation equation can be defined for an arbitrary direction $x\$$ in 3D space. If the axis $x\$$ makes angles with the original x, y, and z axes defined by direction cosines l, m, and n where l is the cosine of the angle between x and $x\$$ m is the cosine of the angle between y and $x\$$ and n is the cosine of the angle between z and $x\$$ then the stresses on the plane normal to $x\$$ would be:

$$_{x\$} = {}_{x}l^2 + {}_{y}m^2 + {}_{z}n^2 + 2!_{xy}lm + 2!_{xz}ln + 2!_{yz}mn \qquad \text{(Eq 14)}$$

The 3D stress transformation is complex as are most stress analyses in three dimensions. Therefore, in most analyses an attempt is made to define conditions to achieve a 2D stress state such as plane stress. Therefore, most of the discussion will concentrate on 2D stress operations. However, it is essential to recognize those cases for which a 2D solution does not obtain the maximum shear stress in the body or the plane on which it acts (see the section “Three-Dimensional Stress Analysis” in this article).

One of the simple tools that is used to perform 2D stress transformation is the construction of Mohr's circle. With the use of Mohr's circle, a stress transformation or a determination of maximum stresses can be made with relative ease and without the memorization of all the stress transformation equations. Squaring and adding the 2D transformation equations, Eq 11 and 12, gives:

$$\left(\sigma_{x'} - \frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{x'y'}^2 = \left(\frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{xy}^2 \qquad \text{(Eq 15)}$$

It was noted that Eq 15 is the equation of a circle with axes and ! and centered on the axis with:

$$\text{center} = \left(\frac{\sigma_x + \sigma_y}{2}\right) \qquad \text{(Eq 16)}$$

and radius, R, given by:

$$R^2 = \left(\frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{xy}^2 \qquad \text{(Eq 17)}$$

The result is that stress transformations can be performed by using the geometric principles of a circle. For example, if stresses are known at a point, they are plotted on a figure that has horizontal axis and vertical axis ! so that ($_x$, !$_{xy}$) is the coordinate point at horizontal position $_x$ and vertical position !$_{xy}$ and so that ($_y$, !$_{xy}$) is the coordinate that corresponds to $_y$ and !$_{xy}$. Since the center of the circle is on the axis, this can be easily found by Eq 16. The radius of the circle is given by Eq 17. From this, the entire circle can be drawn. This is illustrated by the schematic Mohr's circle in Fig. 5.

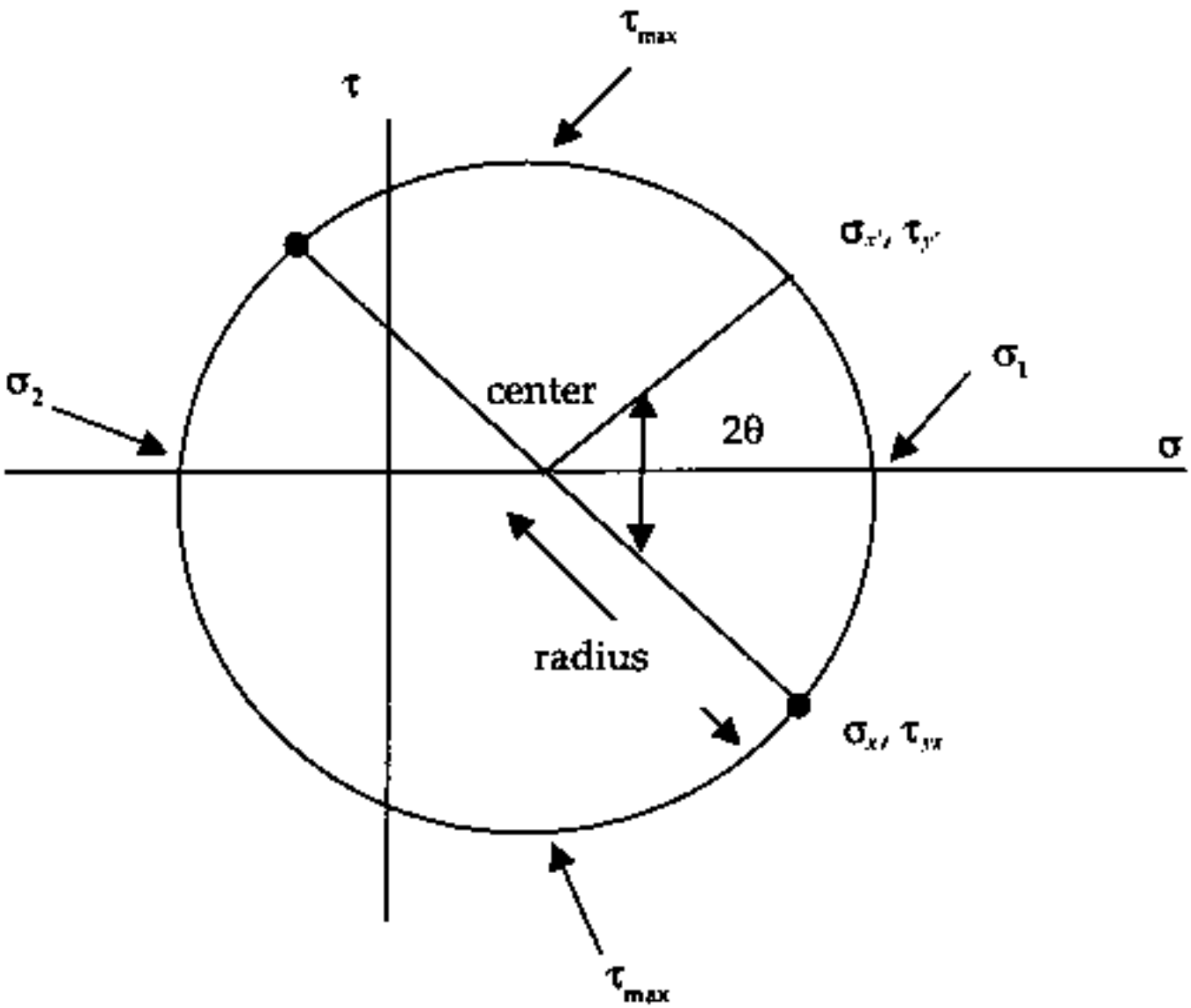


Fig. 5 Mohr's circle for two-dimensional stress transformation

Then for a rotation of axes by # in the stress element, the position on the Mohr's circle must go through a rotation of 2# in the same direction but around the circumference of the circle. Every point on the circle corresponds to a possible stress pair. Note that a rotation of 180° of the circle corresponds to a rotation of the stress element by 90° so that $_x$ is transformed to $_y$ as shown in Fig. 5.

Principal Stresses. The purpose of a stress transformation is mainly to find the stresses that can be used in a failure criterion. These would be either the largest magnitude stresses in any direction or the magnitude of stresses on a weak plane. For the former, the extreme stresses can be found by taking Eq 11 and applying the calculus principle $d\ _{x\$}/d\# = 0$. That is, principal stresses are normal (perpendicular) stresses on planes for which the shear stresses are zero. The values of principal stresses are given by:

$$\sigma_1, \sigma_2 = \frac{\sigma_x + \sigma_y}{2} \pm \sqrt{\left(\frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{xy}^2} \qquad \text{(Eq 18)}$$

where the positive radical gives a maximum stress labeled $_1$ and the negative radical a minimum stress $_2$. Notice that the positions of the extreme stresses are the points on the Mohr's circle where the circle crosses the axis (horizontal axis) (Fig. 5).

The right side of the circle corresponds to a maximum normal stress and the left side to a minimum stress. These stresses as called principal stresses and are often labeled $_1$ and $_2$. Notice that the extremal stress given by Eq 18 is equivalent to taking the center ± the radius of the Mohr's circle. In this way the extreme values of stresses can be easily found by simply using the Mohr's circle.

The maximum shear stress, $!_{max}$, can also be found by taking $!_{x\$\$}$ from Eq 12 and using the calculus principal $d!_{x\$\$}/d\# = 0$. This results in:

$$\tau_{max} = \pm\sqrt{\left(\frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{xy}^2} \qquad \text{(Eq 19)}$$

The value of $!_{max}$ is the radius of the Mohr's circle, and the corresponding points on the Mohr's circle are at the top and bottom of the circle as shown in Fig. 5. The maximum shear stress could also be written as:

$$\tau_{max} = \frac{\sigma_1 - \sigma_2}{2} \qquad \text{(Eq 20)}$$

Note that the maximum normal stresses are defined at a position where the shear stresses are zero. A plane with zero shear stress is defined as a principal stress plane, that is, a plane having only normal stresses. This can be extended to three dimensions. Given a symmetric stress tensor as defined previously, a transformation is desired to a plane that results in the stress tensor having only normal stress components; all shear components are zero. This would be given by:

$$\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}$$

The transformation results in a cubic equation in given by:

$$^3 - I_1 \ ^2 + I_2 \ - I_3 = 0 \qquad \text{(Eq 21)}$$

where

$$\begin{aligned} I_1 &= \sigma_x + \sigma_y + \sigma_z \\ I_2 &= \sigma_x\sigma_y + \sigma_y\sigma_z + \sigma_z\sigma_x - \tau_{xy}^2 - \tau_{xz}^2 - \tau_{yz}^2 \\ I_3 &= \sigma_x\sigma_y\sigma_z + 2\tau_{xy}\tau_{xz}\tau_{yz} - \sigma_x\tau_{yz}^2 \\ &\quad - \sigma_y\tau_{xz}^2 - \sigma_z\tau_{xy}^2 \end{aligned} \qquad \text{(Eq 22)}$$

The values of the coefficients, I_1, I_2, and I_3 are called the stress invariants and are independent of the coordinate system used to describe the stress state. However, here they are used to solve the stress cubic given in Eq 21 so that three principal stresses result: $_1$, $_2$, and $_3$. In three dimensions, the maximum shear stresses can be found in the same way to the 2D stresses, that is as one-half the difference of any two principal stresses.

If stresses $_1$, $_2$, and $_3$ are ordered from maximum to minimum, then the absolute maximum shear stress in three dimensions is:

$$\tau_{max(abs)} = \frac{\sigma_1 - \sigma_3}{2} \qquad \text{(Eq 23)}$$

The transformation of stress in three dimensions can be made with equations like Eq 14. For this transformation, a circle like the Mohr's circle cannot be defined. There is thus, no simple graphical way to transform the stresses or determine $_1$, $_2$, $_3$, and maximum shear stresses for a general 3D system of stresses. However, once $_1$, $_2$, and $_3$ are known from the solution of the cubic equation (Eq 21), it is possible to draw three circles (Fig. 6) having centers, *C*, at

$$C_1 = \frac{\sigma_1 + \sigma_2}{2}; \quad C_2 = \frac{\sigma_2 + \sigma_3}{2}; \quad C_3 = \frac{\sigma_1 + \sigma_3}{2} \qquad \text{(Eq 24)}$$

and radii

$$R_1 = \frac{\sigma_1 - \sigma_2}{2} \quad R_2 = \frac{\sigma_2 - \sigma_3}{2} \quad R_3 = \frac{\sigma_1 - \sigma_3}{2} \qquad \text{(Eq 25)}$$

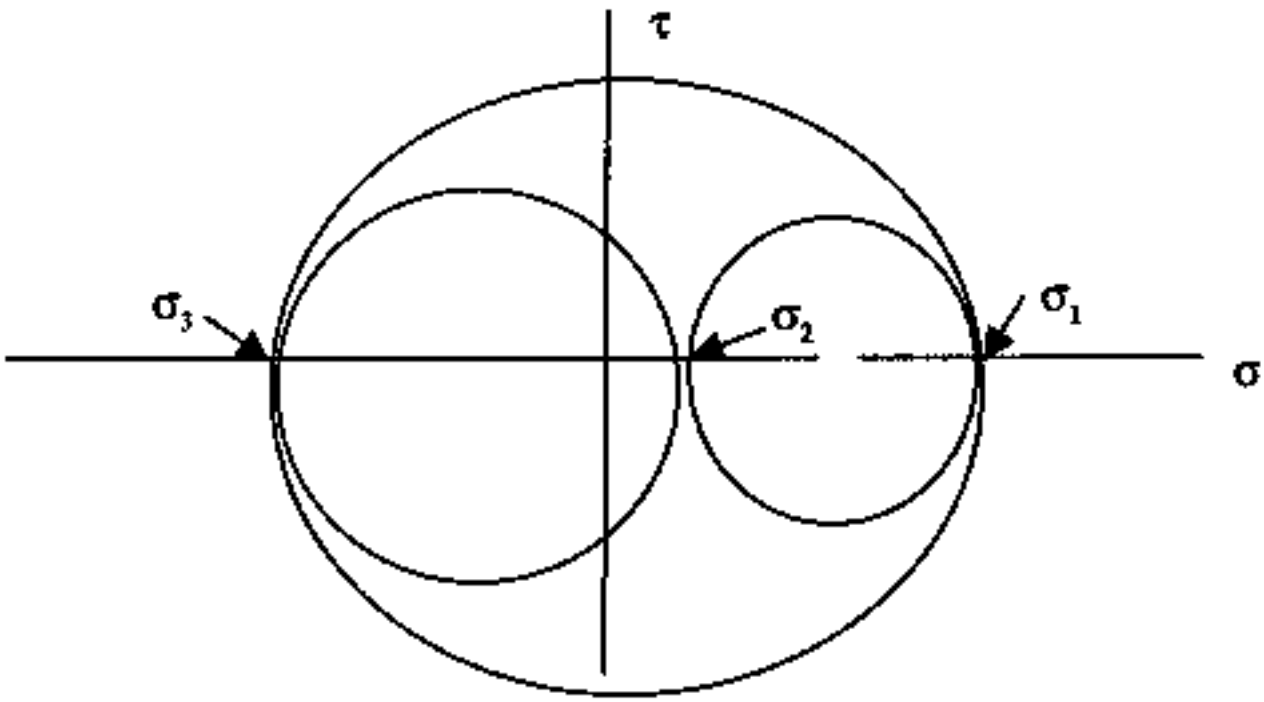


Fig. 6 Special three-dimensional Mohr's circle defined with principal stresses

These circles show the three principal stresses and the corresponding maximum shear stresses. Using them makes it is easy to see that the absolute maximum shear stress is given by Eq 23. They show that all of the principal stresses occur on planes that have zero shear stress. These three-dimensional (3D) Mohr's circles cannot be used for a general stress transformation but can be used for a particular transformation. If in one direction the principal plane is held fixed, then the circle that does not contain that point is the one for the other directions. A rotation of this circle gives all of the stresses that can be obtained from a rotation of the 3D element, holding the axes fixed in the perpendicular plane. For example, if the *z* axis is chosen for $_3$ direction, then a rotation of points on the circle containing $_1$ and $_2$ would define all of the stress states in the planes perpendicular to *z* that are the *x-y* plane for Cartesian coordinates.

There are cases in which a Mohr's circle construction can be used for a case of 3D stresses. Those cases are ones in which only *one shear stress* acts in the body. That is, if the stress tensor has the general form:

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & 0 \\ \sigma_{yx} & \sigma_{yy} & 0 \\ 0 & 0 & \sigma_{zz} \end{bmatrix}$$

In this case, $_{zz}$ is a principal stress. Because the directions of the principal stresses are mutually orthogonal, the other two principal stresses must lie in the *xy* plane and can therefore be determined from the two-dimensional case considering $_{xx}$, $_{yy}$, and !$_{xy}$.

This case clearly illustrates the dangers of considering only the applied loads in determining the maximum shear stress in the body. Consider a case of balanced biaxial tension ($_{xx}$ = $_{yy}$) and consider only the two-dimensional (2D) case. It is incorrectly concluded that there is no shear stress in the body because the radius of the Mohr's circle in the *xy* plane has been reduced to a point. However, when the 3D case is considered, $_{zz}$ is located on the -! axes, and now there are two circles having equal radius given by $\frac{1}{2}$ ($_{xx}$ - $_{zz}$) and $\frac{1}{2}$ ($_{yy}$ - $_{zz}$) and a third circle of zero radius given by $\frac{1}{2}$ ($_{xx}$ - $_{yy}$) so that there are now four planes on which the same maximum shear stress exists.

One plane of interest for defining stresses is called the octahedral plane (or {1, 1, 1} plane in crystallographic terminology). It is defined as the plane with normal *l* = *m* = *n*; that is, the normal to the plane makes equal angles to the *x*, *y*, and *z* axes. On this plane, there is a normal stress given by:

$$\sigma_{oct} = \frac{\sigma_x + \sigma_y + \sigma_z}{3} = \frac{\sigma_1 + \sigma_2 + \sigma_3}{3} \qquad \text{(Eq 26)}$$

and a shear stress given by:

$$\tau_{oct} = \tfrac{1}{3}\sqrt{(\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2 + 6(\tau_{xy}^2 + \tau_{xy}^2 + \tau_{xy}^2)}$$
$$= \tfrac{1}{3}\sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2} \qquad \text{(Eq 27)}$$

The shear stress on the octahedral plane is often given in the form of an equivalent or effective stress, $_e$, given by:

$$\sigma_e = \frac{1}{\sqrt{2}}\sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2}$$
$$= \frac{3}{\sqrt{2}}\tau_{oct} \qquad \text{(Eq 28)}$$

This effective stress and the von Mises stress have equal magnitude in this case, and along with the maximum shear stress is used to describe the yield condition in ductile materials. Alternatively, the effective stress could be described in terms of the Tresca criterion:

$$\sigma_e = \frac{\sigma_1 - \sigma_3}{2}$$

Other effective stresses can be defined for polar and/or anisotropic materials.

Three-Dimensional Stress Analysis. Although many practical stress states can be analyzed in two dimensions, the more complex 3D case is now examined. Complicated loading and/or geometry may induce a stress state in which none of the stress components is zero. For example, normal and tangential surface tractions as occur in combined rolling and sliding problems often generate a local 3D stress state, and failures are often associated with the attendant sliding friction. Notches, which are frequent crack initiation sites, when subjected to external loads often generate triaxial stress states and locally embrittle the material, predisposing it to failure. Thus, manipulation of the entire stress tensor is necessary for analysis.

As noted previously, Mohr's circle is a graphical representation of the 2D stress transformation equations. The cases described previously are convenient to show how Mohr's circle can be used to obtain the principal stresses and the planes of maximum shear stress. Three pairs of principal stresses define six planes of relative maximum shear stress (the sign of a shear stress does not matter). For purposes of failure analysis, it is important to know that the largest of these stresses has been obtained. A simple plane stress example illustrates the point. Consider a case of in-plane balanced biaxial tension of magnitude *S*; how does one determine the magnitude of the maximum shear stress and the plane(s) on which it exists? Treating this as a 2D problem would be incorrect, since every planar stress transformation leads to the same stress tensor, namely biaxial tension. Mohr's circle is a point. By improperly neglecting the third principal stress, even though it is zero, one would conclude that no shear stresses are present, when, in fact, $!_{max}$ is *S*/2 and acts on two perpendicular planes.

Consider next the case of a torsional stress, !, combined with an axial stress, $_1$. The 2D stress tensor is:

$$\begin{bmatrix} \sigma_1 & \tau \\ \tau & 0 \end{bmatrix}$$

Mohr's circle correctly predicts the maximum shear stresses to be:

$$\tau_{max} = \pm\sqrt{\left(\frac{\sigma_1}{2}\right)^2 + \tau^2}$$

However, if a second direct stress, $_2$, is added such that the 2D stress tensor becomes:

$$\begin{bmatrix} \sigma_1 & \tau \\ \tau & \sigma_2 \end{bmatrix}$$

then Mohr's circle may or may not give the correct maximum shear stress. It can be easily shown that when $_1$ and $_2$ are both positive or both negative, the incorrect maximum shear stress is incorrectly calculated assuming a two dimensional problem. This is because the third normal principal stress (which happens to have zero magnitude) has been neglected.

Next consider plane strain. If &z = 0, then plane strain conditions exist wherever $'_{xz} = '_{yz} = 0$. Plane deformation, in which $u_z = 0$ (where u_z is displacement in *z*-direction) and both u_x and u_y are independent of *z*, also leads to plane-strain conditions. It is approximated in the case of a long prismatic cylinder whose lateral-surface (parallel to the *z*-axis) boundary conditions are uniform (hence independent of *z*) along the length (*z*-axis) and are applied parallel to the *x*-*y* plane. Appropriate end conditions are $u_z = 0$, $!_{xz} = 0$ and $!_{yz} = 0$.

Suppose that a linearly elastic material has Poisson's ratio, (, of 0.5, and that $_x = _y$. Then by Hooke's law $_z = _y = _x$, that is:

$$\varepsilon_z = 0 = \frac{1}{E}\left[\sigma_z - \nu\left(\sigma_x + \sigma_y\right)\right]$$

Therefore, the hydrostatic stress ($_n$) is given by:

$$\sigma_n = \frac{\sigma_x + \sigma_y + \sigma_z}{3} = \frac{3\sigma}{3}$$

This describes a hydrostatic stress state, that is, the stress tensor, **T**, is defined as:

$$\mathbf{T} = \begin{bmatrix} \sigma_h & 0 & 0 \\ 0 & \sigma_h & 0 \\ 0 & 0 & \sigma_h \end{bmatrix}$$

Even if Poisson's ratio is unequal to 0.5, a hydrostatic stress whose magnitude is:

$$\sigma_h = \frac{(1+\nu)\left(\sigma_x + \sigma_y\right)}{3}$$

still exists. The effect of the hydrostatic stress is to raise the stress level at which yield occurs, causing it to approach the fracture stress. In a constant-volume material (metals, Ceramics), the hydrostatic stress cannot cause yield, but a hydrostatic stress does increase the magnitude of the stresses created by the applied loads at which yield occurs. In a nonconstant volume material (polymers), the hydrostatic stress affects yield behavior, so that a von Mises, octahedral shear, or Tresca-type criterion is inaccurate.

In application, consideration of the plane strain state of stress is significant in understanding macroscale brittle fracture surface orientation in tensile loading and the difference in fracture toughness for plane stress and plane strain loading (see the section “Fracture Mechanics” in later paragraphs). The difference in plane-stress and plane-strain loading is also important in understanding differences in fracture surface orientation and neck shape during macroscale ductile fracture of plate versus cylindrical tensile components (see the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals” in this Volume).

Loading relatively thick components in tension or bending and which contain notches, cracks, or similar stress concentrators parallel to the thickness direction, taken as the *z* direction to be consistent with the notation above, results in plane-strain loading conditions. The material at the tip of the notch or crack (area of maximum stress concentration) is constrained by the adjacent material, which is not as highly stressed. Thus, & is constrained to be (close to) zero. As discussed in other sections of this article and other articles in this Volume, the fracture toughness observed under plane-strain conditions is a minimum and is considered a material property, K_{Ic}. The underlying explanation for this observation follows from the preceding paragraph. Plane strain conditions increase the hydrostatic stress and inhibit yielding.

Under purely hydrostatic loading, no shear stresses exist, as is easily shown by examination of the three Mohr's circles. The circles reduce to zero radius. *Hydrostatic stress* is sometimes referred to as *spherical stress.* (A stress state in which a hydrostatic stress exists is sometimes incorrectly referred to as a case of *triaxial stress* Triaxial loading indicates only that the stresses act in three directions in the body.) *Spherical stress* is perhaps a better term since it conveys the omnidirectional nature of the loading and suggests the absence of shear stresses. The three 3D Mohr's circles reduce to circles of zero radius; that is, they become a point in a manner analogous to the case of biaxial tension analyzed in two dimensions. The difference, however, is that in a spherical stress state, no shear stresses are present, while under biaxial tension or compression, the 2D analysis leads to the incorrect conclusion that shear stresses are absent. This only highlights the need for proper stress analysis.

Stress-Strain Relationships

Stress and strain in an isotropic body for elastic deformation are related by linear expressions that result from Hooke's law. These are given by:

$$\varepsilon_x = \frac{\sigma_x}{E} - \frac{\nu}{E}\left(\sigma_y + \sigma_z\right) \qquad \text{(Eq 29)}$$

$$\varepsilon_y = \frac{\sigma_y}{E} - \frac{\nu}{E}\left(\sigma_x + \sigma_z\right) \qquad \text{(Eq 30)}$$

$$\varepsilon_x = \frac{\sigma_z}{E} - \frac{\nu}{E}\left(\sigma_x + \sigma_y\right) \qquad \text{(Eq 31)}$$

$$\gamma_{xy} = \frac{\tau_{xy}}{G}; \quad \gamma_{xz} = \frac{\tau_{xz}}{G}; \quad \gamma_{yz} = \frac{\tau_{yz}}{G} \qquad \text{(Eq 32)}$$

where *E* is the elastic modulus (often called Young's modulus), (is Poisson's ratio, and *G* is the shear modulus (sometimes called modulus of rigidity). Equations 29 to 32 are the 3D forms of the stress-strain equations for materials that are homogeneous and isotropic. The analysis of the stress-strain relationships shows that there are only two independent constants. Then *E*, *G*, and (are related by:

$$E = \frac{G}{2(1+\nu)} \qquad \text{(Eq 33)}$$

The volume change in a material is given by:

$$\Delta V = \left(\varepsilon_x + \varepsilon_y + \varepsilon_z\right)V = \left[\frac{1-2\nu}{E}\left(\sigma_x + \sigma_y + \sigma_z\right)\right]V \qquad \text{(Eq 34)}$$

where *V* is the original volume and " *V* is the volume change.
For typical materials the Poisson's ratio, (, varies from 0 to 0.5. For (= 0, there is no lateral strain with axial strain. From Eq 34, it can be seen that the volume change gets smaller as (increases from zero to 0.5. Experimental evidence indicates that metallic materials when plastically strained have a Poisson ratio of 0.5 and therefore deform under constant-volume conditions. An equivalent statement is that they are incompressible. Alternatively, polymerics have higher compressibility and in many applications should not be considered constant volume materials. As previously indicated, constant volume corresponds to (= 0.5. The most usual values of (for elastic metals are between 0.25 and 0.35. If exact data are not available, a value of (= 0.3 is typically used.
Two-dimensional forms can also be given for special stress or strain cases. For example, only the stresses in the *x-y* plane may be nonzero for an in-plane stress. The corresponding stress strain relationships become:

$$\varepsilon_x = \frac{\sigma_x}{E} - \frac{\nu}{E}\sigma_y \qquad \text{(Eq 35)}$$

$$\varepsilon_y = \frac{\sigma_y}{E} - \frac{\nu}{E}\sigma_x \qquad \text{(Eq 36)}$$

$$\gamma_{xy} = \frac{\tau_{xy}}{G} \qquad \text{(Eq 37)}$$

These can be inverted to solve the stresses as a function of strain:

$$\sigma_x = \frac{E}{1-\nu^2}\left(\varepsilon_x + \nu\varepsilon_y\right) \qquad \text{(Eq 38)}$$

$$\sigma_y = \frac{E}{1-\nu^2}\left(\varepsilon_y + \nu\varepsilon_x\right) \qquad \text{(Eq 39)}$$

$$!_{xy} = G'_{xy} \qquad \text{(Eq 40)}$$

The 2D equations in Eq 35 to 40 are the representation of Hooke's law for a homogeneous and isotropic material. A homogeneous material does not vary in properties from point to point; that is, for a translation of the coordinate axes there is no difference in material properties. Isotropic materials do not vary for a rotation of the axes. That is, the elastic properties do not vary for a change in orientation. No polycrystalline material satisfies these conditions exactly; however, the added benefit that comes from an analysis that uses elastic relationships for nonhomogeneous or anisotropic materials often does not justify the added difficulty introduced into the problem. For example, a completely anisotropic material can have 21 elastic constants. Even an orthotropic material can have nine elastic constants. Trying to work with these relationships presents a problem of such difficulty that it can usually only be handled numerically. However, it should be pointed out that anisotropy is taken into account (as for example with an anisotropic yield criterion) and controlled when large-strain plastic forming operations are carried out with metallic sheet.
Most metallic materials conform to these linear stress-strain relationships during the initial part of their loading. Most stress analyses are based on this linear-elastic relationship. The requirement for elastic behavior is that the body returns to its original shape on unloading. If metals are loaded beyond a point called the elastic limit, the body does not return to its original shape with unloading. Metals are said to yield and the relationship between stress and strain typically may

become nonlinear. (Note: the elastic limit and the yield point or yield stress are *not* equivalent. A material may be, and almost always is, nonlinearly elastic.) Above the yield stress, material behavior is nonlinear, and plastic strain results from increased stress. The volume of the material is assumed to be constant and Eq 34 shows that (approaches 0.5 for a constant-volume material. The relationship between stress and strain in the plasticity region often takes a nonlinear form, often parabolic in shape (i.e., $y = ax^n$), expressed as: =)&^{n}, where K is the constant strength coefficient and n is the constant strain hardening exponent. It is important to note that this equation is of fundamental importance in discussion of plastic forming behavior and that the stress, and strain are an instantaneous values, not nominal values.

Finally, it is worthwhile to reiterate the previous discussion of constraints on deformation along the z axis. Equations 35 to 40 are 2D equations, but the deformation along the z axis depends on the constraint imposed by the thickness of the material in the z-direction. When the part is thick, a lateral constraint is imposed on deformation along the z axis. When plane-strain conditions occur (i.e., & ~ 0), a stress component develops along the z axis:

$$_Z = ((\ _X + \ _Y) \quad \text{(plane strain)} \qquad \text{(Eq 41a)}$$

In contrast, if the section is thin along the z-axis, then the lateral constraint is reduced and deformation occurs along all major axes. In this case, the condition is plane stress, where (as the name suggests), stress is zero in one direction:

$$_Z = 0 \quad \text{(plane stress)} \qquad \text{(Eq 41b)}$$

Other Components of Stress

Effect of Residual Stresses. Residual stresses are locked into the body and are not released when external forces are released. In that way they are always present. However, in the absence of external forces, the body must be in equilibrium and hence have a zero net force. This gives the residual stress pattern a character that is called self-equilibrating. That is, for a set of residual stresses in tension, there must also be a similar set of residual stress in compression so that the net force caused by them is zero. For this reason the effect of the residual stresses cannot be accounted for correctly by simply adding the maximum value of residual stress to the primary stresses, those that come from the applied forces. Just as applied forces create complex internal states of stress, locked-in residual stresses may also be complex, and 3D stresses are vector quantities. Residual stresses must be added vectorially to applied stress.

The ways for handling residual stresses are varied but not always well understood. In the most elementary and conservative manner, it is sometimes assumed that the residual stresses are at a value of the yield strength and that they add algebraically (in a given direction) to the stresses from applied loads. In that way the combined applied and residual stresses are very severe, and the analysis of them is very conservative. Other ways for handling residual stresses are to measure the residual stress pattern and to add this measured pattern to the existing stress pattern that comes from applied forces.

There is not always a good way to measure this residual stress pattern. Among the ways to measure residual stresses is hole drilling with strain gages, a destructive procedure. X-ray diffraction, a surface measurement, and neutron diffraction are not always readily available. Because residual stresses come from strain mismatch, they are easily disturbed by such things as plastic deformation or geometrical change such as the introduction of a cracklike defect. So even with a measured residual stress pattern, it is not clear what will happen to the residual stress pattern during the life of a component.

Contact (Hertzian) Stresses. Sometimes neglected by designers, contact stresses can initiate failure, especially in the presence of friction. Many finite-element analysis (FEA) practitioners assume that component failures occur at locations far removed from the loading and, for the sake of convenience, apply either point or line loads to their models. This generally yields acceptable remote stress fields, but stresses local to the load may be seriously in error if that load is applied to a single node or is improperly distributed over an assumed area.

The theoretical area of contact between two nonconforming surfaces is frequently quite small. The interaction between these surfaces is often described as either “point” or “line” contact. Common examples of “point” contact are mating helical gears, cams and crowned followers, ball bearings and their races, and train wheels and rails. If the mating parts can be considered semi-infinite and if material behavior is linearly elastic, then the local stress state can be described by Hertzian theory as long as the contacting surfaces can be modeled as quadratic functions of those spatial coordinates defining the surfaces. Even in the absence of friction, the resulting 3D stress state is quite complex, and although the local stress state is compressive, that is, the principal stresses beneath the load are negative, large subsurface shear stresses, which can serve as crack initiation sites, exist beneath the load. A thorough analysis should be performed, especially if the contact stresses are the relevant stresses.

On the other hand, “line” contact, which occurs in cylindrical rollers and similar apparatuses, frequently can be considered to be a case of plane strain. Errors of FEA, such as load application over zero area, that is, applying a line load, may again yield acceptable far-field results, but local stresses will be overestimated. Uniform distribution of surface pressure will underestimate them. Again, if the contact stresses are responsible for failure, for example, pitting or spalling, proper load distribution is important to the failure analysis.

Consider radial loading of a cylindrical pin in a hole. First, look at the case in which significant clearance exists between the pin and the hole. In that event, a near-Hertzian stress state may exist beneath the pin, given sufficient material between the hole and an edge. However, if the hole is located near an edge and the loading is perpendicular (or nearly so) to that edge, then a large tensile stress exists at that edge beneath the hole. The magnitude of this tensile stress is strongly influenced by the stress state below the contact. Now examine the other extreme: little or no clearance between the pin and hole. In this event, Hertzian theory cannot be employed to develop the near-field subsurface stress state. Due to better load distribution, the edge tensile stress now will be significantly lower, as will the maximum subsurface shear stress. These two cases define the envelope in which pin/hole contact occurs. Differences between the two edge tensile stresses could approach 100%. Again this serves to highlight the need for proper stress analysis.
For additional information on complex stresses in components, see the article "Overview of Mechanical Properties and Testing for Design" in *Mechanical Testing and Evaluation,* Volume 8 of *ASM Handbook.*

References cited in this section

1. J.E. Shigley and C.R. Mischke, *Mechanical Engineering Design,* 5th ed., McGraw-Hill, 1989
2. W.K. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, Inc., 1997
3. W.C. Young, *Roark's Formulas for Stress and Strain,* 6th ed., McGraw-Hill, 1989
4. A. Blake, *Practical Stress Analysis in Engineering Design,* 2nd ed., Marcel Dekker, 1990

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Stress Analysis Of Common Geometries

The stress analysis of common geometries can be made with either a mechanics of material approach or a theory of elasticity approach. The mechanics of materials approach follows four steps to arrive at a solution; these steps are taken in a prescribed sequence. The mechanics of materials always proceeds as follows:

1. Define the geometry of deformation.
2. Define strain from the geometry of deformation.
3. Define stress from this strain through the stress-strain relationships.
4. Relate the stresses to the boundary loadings through equilibrium.

Following these four steps always leads to a solution. In the procedure, the first step is the most difficult, defining the geometry of deformation. This has been done in the past for most common geometries by applied mechanics theoreticians who studied deformation models. In developing the mechanics of materials solutions here, the author uses these classical deformation patterns invents rather than new ones.
The theory of elasticity method requires that a number of conditions be satisfied simultaneously, namely, the body must satisfy internally a set of equilibrium equations and a set of compatibility equations. The equilibrium equations ensure that all parts of the body are in equilibrium under the stress field, and the compatibility equations ensure that the displacements are single-valued and possible. The compatibility equations derive from strain-displacement equations. The theory of elasticity always gives an exact set of equations for the solution to a problem. However, the equations are often a partial differential equation that cannot be easily solved and/or which must be solved numerically.
Therefore, the common elements that are presented in the following sections for stress analysis follow the mechanics of materials method for solution. Examples presented include the tensile geometry, torsion geometry, bending and beam shear, pin shear, thin-walled pressure vessels, curved beams, and springs.

Simple Shaft or Bar

A summary in Fig. 7 shows simplified, 2D free-body diagrams illustrating the orientations of the normal-stress and shear-stress systems at any internal point in a shaft loaded in pure tension, torsion, and compression. The single-overload fracture behavior of both ductile and brittle materials also is illustrated for each type of load.

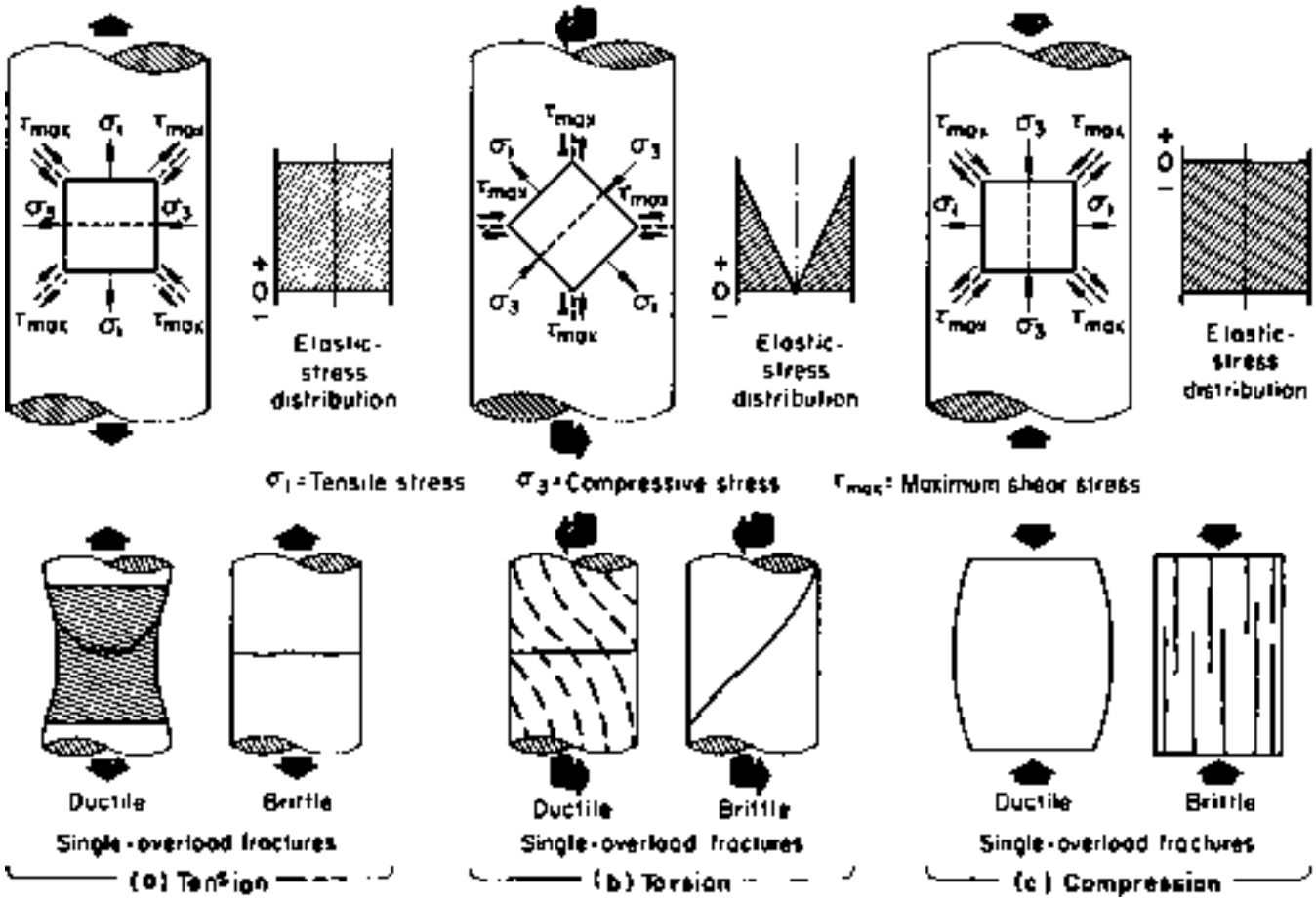


Fig. 7 Free-body diagrams showing orientation of normal stresses and shear stresses in a shaft and the single-overload fracture behavior of ductile and brittle materials. (a) Under simple tension. (b) Under torsion. (c) Under compression loading

Tensile Bars. The tensile bar has an axis, *x*, that goes through the centroid of a cross-sectional area. The area is often constant along the *x* axis but need not be. Tensile bars also are often round but need not be. Any shape of a tensile bar is permissible, but it must have a cross-sectional area with a centroid that lies on the load line. Assume that the load, *P*, is directed along the axis, that is, through the centroid of the area. Off-center loads, those not along a centroid, result in more complex loading and would not be analyzed as simple axial loading.

Step 1 in the mechanics of materials assumes that the deformation mode of the bar is uniform elongation, that is, the strain is averaged over the entire length (*L*) of the body. Step 2 gives the strain as the stretch, *, divided by the total length of the bar, *L*. The only nonzero stress is $_x$, which lies along the *x* axis so that step 3 gives the stress from the stress-strain equation as $_x = E\&$. The equilibrium condition from step 4 gives the relationship between the stress, , and the external loading, *P*, as load equals the stress times area (*A*). These steps result in following equations:

$$\sigma = \frac{P}{A} \qquad \sigma = E\varepsilon \qquad \varepsilon = \frac{\delta}{L}$$
$$\delta = \int \frac{Pdx}{AE} = \frac{PL}{AE} \qquad \text{(Eq 42)}$$

The strain in the above equation should be indicated as the strain in the direction of load (e.g., &)

The second result for the deflection, *, is correct when the load, *P*; area, *A*; and modulus, *E* do not vary along the *x*-axis. For simple geometries like the tension bar, there is only one stress component. As discussed in preceding sections, there can be transverse strains even though no stress is applied in these directions.

Under tension loading, the tensile stresses ($_1$ in Fig. 7) are longitudinal, whereas the compressive stress components ($_3$ in Fig. 7) are transverse. The stresses created by tensile loading of cylindrical or approximately square specimens cause considerable deformation (elongation and necking) before fracture, which originates near the center of the shaft within a

region perpendicular to the load and propagates toward the surface, ending with a shear lip. However, in a brittle material, the fracture surface from a tensile overload is roughly perpendicular to the direction of tensile stress and involves little or no macroscale visible permanent deformation. The elastic stress distribution in pure tension loading, in the absence of a stress concentration, is uniform across the section. Thus, fracture in this case can originate at any point within the highly stressed volume. However, once the specimen necks, the stress distribution is no longer constant across the cross section, and fracture is predicted to initiate either along the centerline or at the surface, depending on specimen geometry and material properties (strain handling and strain-rate hardening). See the article "Mechanisms and Appearances of Ductile and Brittle Fracture in Metals."

Torsion Bars. A bar is loaded with a torsional load if the loading is a couple, called torque, *T*, directed along the axis of the bar. The torsional loading of a round bar can be analyzed easily by the mechanics of materials method. It requires the same four steps previously listed in the same sequence required for all mechanics of materials solutions. The torsion of a round bar results in a state of stress described as pure shear. For the deformation mode, the assumption is made that the bar twists with no planar distortion and for the round bar, that the plane sections do not warp; that is, the plane sections remain plane. If the torsional bar is viewed as a roll of round coins, then the bar twists with one coin rotating slightly relative to the one next to it. As each coin rotates, the entire roll has a uniform shear strain that is linear in radius *r*. The coins do not distort but retain their original shape as they rotate about the axis. Therefore the strain is zero at the axis of the bar and reaches a maximum at the external surface Fig. 8.

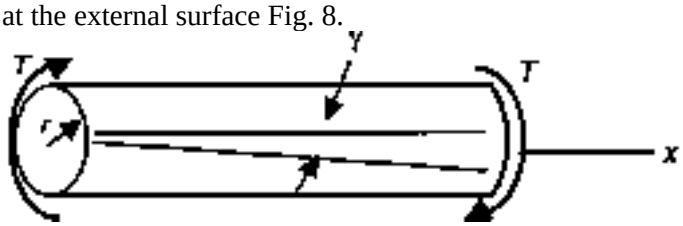


Fig. 8 Round bar in torsion

Because the strain is linear with *r* and the stress strain equation gives only a shear stress, the stress is also linear in *r*, (Eq 37). The evaluation of step 4 gives the relationship between the stress and the applied load, which here is a torque *T*. This results in a shear stress that is related to the torque by:

$$\tau = \frac{Tr}{J} \qquad \text{(Eq 43)}$$

where *T* is the applied torque and *J* is the polar moment of inertia. $J = (+/2)r^4$ for a round bar. The round bar does not need to be constant in size along its length. For this stress the strain is:

$$\gamma = \frac{\tau}{G} = \frac{Tr}{JG} \qquad \text{(Eq 44)}$$

The rate of twist is:

$$\frac{d\theta}{dx} = \frac{\gamma}{r} = \frac{T}{JG} \qquad \text{(Eq 45)}$$

so the total twist, #, can be found by:

$$\theta = \int_0^L \frac{T dx}{JG} = \frac{TL}{JG} \qquad \text{(Eq 46)}$$

where the second part of the equation is correct when all of the quantities are constant along the bar, *T*, *G* and *J* are not functions of *x*.

If the bar is not round, there is warping; that is, the bar displaces out of the plane of the cross-section in the *x* direction. The simple mechanics-of-materials method used for the round bar does not work, and some more advanced method like theory of elasticity can be used. For a common shape like a bar with a rectangular cross section (Fig. 9), the membrane analogy gives a method from which the torsional stress can be calculated. Since this stress is not a simple function of the applied torque, the solution is often given in parametric format. For example with a torsion bar that has a rectangular cross section $a \times b$, the maximum stress occurs at the middle of the long side and is given as:

$$\tau = \beta \frac{T}{ab^2} \qquad \text{(Eq 47)}$$

where , is a parameter that depends on the ratio of a/b. For $a = b$, a square cross section, , = 4.81. For $a = 10b$, a long thin rectangular cross section, , = 3.21.

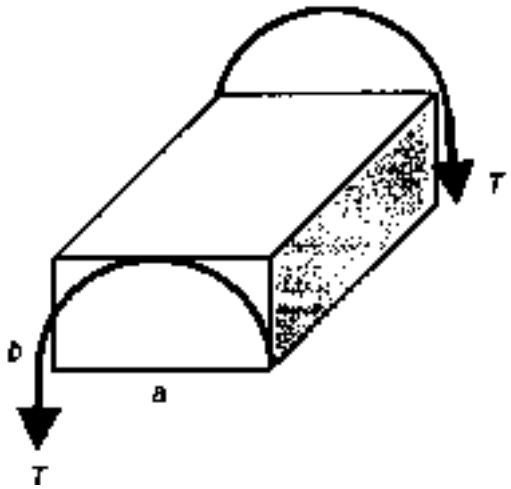


Fig. 9 Rectangular bar in torsion

In a ductile material loaded to failure in torsion, shear stresses cause considerable deformation before fracture. However, this deformation may not be usually obvious, because the shape of the shaft has not been changed. The distortion will be obvious if there were axial grooves or lines on the shaft before twisting or if the metal is hot etched to reveal grain-flow twisting or if a mounted microstructural section is examined. If a shaft loaded in torsion is assumed to consist of an infinite number of infinitely thin disks that slip slightly with respect to each other under the torsional stress, visualization of deformation is simplified. Torsional single-overload fracture of a ductile material usually occurs on the transverse plane, perpendicular to the axis of the shaft. In pure torsion, the final-fracture region is at the center of the shaft. Slight bending will cause it to be off-center. Additionally, final off-axis fracture is expected if the crack does not grow radially in a uniform manner with respect to the circumference. Since the latter is relatively uncommon, off-axis final fracture locations are common.

A brittle material in pure torsion will again fracture perpendicular to the tensile-stress component, which is now 45° to the shaft axis. The resulting fracture surfaces usually have the shape of a spiral. The elastic stress distribution in pure torsion is maximum at the surface and zero at the center of the shaft. Thus, in pure torsion, fracture normally originates at the surface (in non-surface-treated components), which is the region of highest stress. When surface-treated components are loaded in torsion, the material strength profile may cause subsurface crack initiation.

Compression. When a cylindrical shaft is loaded in axial compression (Fig. 7c), the stress in the axial direction is compressive ($_3$ in Fig. 7c) and tensile in the transverse direction ($_1$ in Fig. 7c). The maximum shear stresses are 45° to the shaft axis, as they are during axial tension loading.

In a ductile material overloaded in compression, shear stresses cause considerable deformation and possibly the creation of some cracking but usually do not result in total fracture. The shaft is shortened and bulges laterally. A brittle material loaded in pure compression, if it does not buckle, again will fracture perpendicular to the maximum tensile-stress component. Because the tensile stress is transverse, the direction of brittle fracture is parallel to the shaft axis. If fracture occurs, it will likely be in the longitudinal direction, because compression loading increases the shaft diameter and stretches the metal at the circumference.

Beam Bending

Beam Analysis: Internal Forces. A member is usually treated with a beam analysis if the loads or applied couples are transverse to the beam length. The tensile bar has forces along its axis, and the torsion bar has a couple along its axis. However, a beam has forces and couples that are perpendicular to its axis. The tension and torsion bars are most often loaded with constant forces or ones that change in discrete steps. The beam has internal forces that usually vary as a function of position along the axis of the beam. An internal force balance indicates that the beam is subjected to a shear force, *V*, an axial force and internal and moment. These generate two stress components: a normal stress, , acting along the length from the moment (*M*), and a shear stress, !, which acts on both longitudinal and transverse planes from the shear force.

Critical points in the beam correspond to the points where the shear force and bending moment are highest in magnitude, positive or negative. The sign convention for the internal loading in a beam is illustrated in Fig. 10. The positive configuration for *V* has it in the following directions, up on the left and down on the right. The positive configuration for *M* has it pointing up on both sides. The opposite directions correspond to negative values.

Fig. 10 Sign convention for shear (*V*) and bending moment (*M*)

The sign convention for the beam can be used to keep track of the stress directions; however, it is easier for the simple beam to use a physical sense to tell the direction of the stress.

The beam can be loaded with a distributed load that can be a function of position, discrete forces that are transverse to the beam axis, and moments that are transverse to the beam axis. Typical beam loads are shown in Fig. 11.

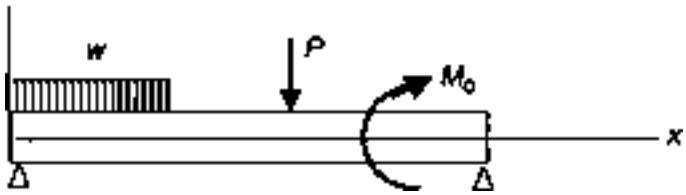


Fig. 11 Typical loading on a straight beam

For the beam, the horizontal and vertical axes *x* and *y* can be taken from the left end of the beam. The loads shown in Fig. 11 are *w*, a distributed load that may be a function of position, $w = w(x)$; *P*, a discrete force; and M_o, a discrete moment. The relationships between these applied loads and internal forces, *V* and *M* of Fig. 10, are determined from equilibrium as:

$$w = -\frac{dV}{dx} \qquad V = \frac{dM}{dx} \qquad \text{(Eq 48)}$$

or,

$$V_2 - V_1 = -\int_{x_1}^{x_2} w\,dx \qquad M_2 - M_1 = \int_{x_1}^{x_2} V\,dx \qquad \text{(Eq 49)}$$

These relationships can be used to analytically determine the shear force and moment in a beam or to make a graphical determination of them using the shear and bending moment diagram approach. For the latter, the value of the shear force, *V*, is plotted as a function of position on the beam (using the area under the *w* curve as given by Eq 49) to accumulate continuous change in *V*. The applied load, *P*, causes a sudden, discontinuous jump change in *V*. The *M* diagram uses the area under the *V* diagram to give the continuous change in *M* and a discrete moment to cause a jump change in *M*. The advantage of the diagram approach is that it is easy to determine the maximum magnitudes of the values of *V* and *M*. These values are used in the stress equations to determine maximum values of and !.

Beam Analysis: Stresses. The stress equations are developed from the mechanics of material methods. For this method to work, the loads must be applied in a plane of symmetry of the beam. This works for many geometries. For example, rectangular and circular cross sections as well as U-shaped and T-shaped beams have a plane of symmetry and can be analyzed by the simple mechanics of materials method. L-shaped or C-shaped beams do not fulfill this symmetry requirement and must be analyzed by other means.

The bending moment, *M*, gives a normal stress $_z$; the shear force, *V*, gives a shear stress $!_{zy} = !_{yz}$. If all applied loads are parallel to each other, the subscripts sometimes are dropped. (Note, however, that retention of subscripts minimizes errors in determining the net stress acting in a given direction as the combination of axial stress due to pressure and bending stress in a pressure vessel, or in a beam subjected to a torque.) The normal stress, , is analyzed by developing a geometry of deformation that assumes that there is a neutral axis in the beam over which there is a reversal of the stress and strain signs. On one side of the neutral axis the stress is positive and tension exists; on the other side, stress is negative and compression exists. The sections that were originally plane remain plane and rotate about this neutral axis without distortion of the cross section. This geometry of deformation is used to define a strain given by:

$$\varepsilon = -\frac{y}{\rho} \qquad \text{(Eq 50)}$$

where - is a radius of the beam curvature. This makes the strain linear in *y*, the distance from the neutral axis. Here the strain is negative for a positive value of *y* and positive for a negative value of *y*. The stress is then also linear in *y*.

Writing the equilibrium equations to relate the internal stress to the boundary load gives the position of the neutral axis and the relationship between stress and applied loading. Taking $\%F_z = 0$ shows that the neutral axis coincides with the centroidal axis of the cross-sectional area of the beam; that is, the centroidal axis of the cross section that is perpendicular

to the beam axis, the one with a normal orientation in the x direction in Fig. 11. Taking %$M_z = M$, the moment that comes from the applied loading gives the relationship between the stress, , and M. It is:

$$\sigma = -\frac{My}{I} \qquad \text{(Eq 51)}$$

where M is the bending moment at the point, usually $M = M(x)$, I is the moment of inertia, and y is a position in the cross section taken in a perpendicular direction from the neutral axis. Here the stress is negative or in compression for a positive value of M and y. Since M is a function of x and a function of y, the stress, , varies from point to point in the beam. It is often desirable to find maximum values of this stress. To find the maximum value of stress, a maximum M must be obtained as well as a maximum value of y. The maximum M comes from the beam analysis, for example, from a moment diagram, and the maximum y is at the outer edge of the beam, a distance farthest from the neutral axis. This is given the label c. Then:

$$\sigma_{max} = \frac{M_{max}c}{I} \qquad \text{(Eq 52)}$$

If the beam has two planes of symmetry, then the maximum stress occurs at $y = \pm c$, one sign giving maximum tension and the other maximum compression. This stress is sometimes called the outer fiber stress; it is the one used in determining design limits. For some beams, a quantity called section modulus, Z, is calculated where $Z = I/c$ so that at the outer edges:

$$\sigma = \frac{M}{Z} \qquad \text{(Eq 53)}$$

The beam shear stress is calculated from the shear force, V, using:

$$\tau = \frac{VQ}{It} \qquad \text{(Eq 54)}$$

When ! is determined at some point y from the neutral axis, then Q is the first moment of the area from a free surface, usually the top or bottom of the beam to the point y and taken about the neutral axis, I is the moment of inertia of the entire cross section, and t is the beam thickness at the point y. The thickness has to take all of the length that a cut through the section would encounter at y. For example, if the beam has two legs, as in a U beam, then t is the combined thickness of both sides.

As an example of using Eq 53 and 54 to calculate and ! in a beam that has a rectangular section $b \times h$, where b is the thickness of the section and h is the height, the stresses:

where the normal stresses, , are taken at the outer edges and the shear stresses, !, are at the neutral axis. (For beam stresses, ! are usually very small compared with .) The above ! is the maximum shear stress from the beam shear load and for a x-y oriented element. However, for a failure criterion based on shear stress, the maximum ! does not occur in the x-y orientation. At the top or bottom of the beam, the large usually dominates the stress magnitudes in the beam, and this stress is in the x direction. At 45° to the x direction, a plane of maximum shear stress occurs and the shear stress on that plane is /2. This can be found by applying a stress transformation as would be done in a Mohr's circle analysis like that illustrated in Fig. 5. For most beam geometries, from bending is at least an order of magnitude greater than ! from beam shear. Thus the shear stress associated with bending stresses, that is, the stress found by rotating the stress element, is the one that is used for developing safety against failure by yield.

Curved Beams. The above stress analysis of a beam is for a beam that is essentially straight, that is, the axis of the beam, x, must be straight. Some elements are beamlike in loading but have a curved axis. Examples might be chain links, loading hooks, and machine frames. For the curved beam to be analyzed by the mechanics of materials method, the two curved surfaces of the beam must be concentric circles. The beam is then specified by the circle dimensions, an inner radius, r_i, and an outer radius, r_o. The curved beam does not have its neutral axis on the centroid as a straight beam does. Therefore, both the centroidal and neutral axes must be calculated. The centroidal axis is given the symbol R and the neutral axis the symbol r_n. The main parameters of the curved beam are shown in Fig. 12.

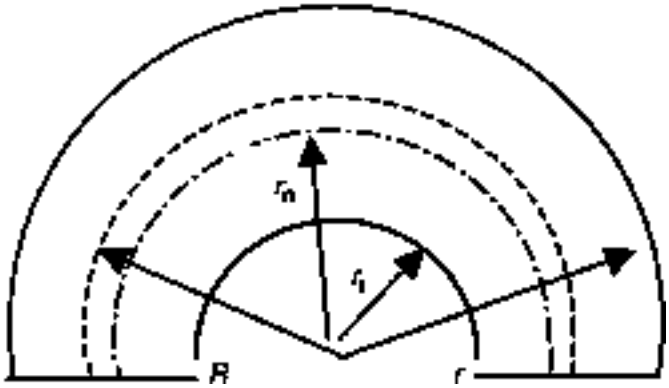


Fig. 12 Definitions of distances for a curved beam

To find r_n, the following expression can be evaluated:

$$r_n = \frac{A}{\int_A \frac{dA}{r}} \qquad \text{(Eq 56)}$$

although for most common cross-sectional shapes the value of r_n is tabulated. An important parameter is the difference between the centroidal radius and the neutral axis radius. This is called an eccentricity, e, where $e = R - r_n$. The stress, , for a curved beam loaded by end moments, M, is given by:

$$\sigma = \frac{My}{A\,er} \qquad \text{(Eq 57)}$$

The distance y is taken from the neutral axis. If the moment is positive then y is positive going toward the center of the circles. The stress becomes maximum at the inner radius, r_i. For example, at the inner radius $y = r_n - r_i$ and at the outside radius $y = r_n - r_o$. The latter is a negative number and reflects a negative bending stress at the outer radius. The stress in Eq 57 is for a pure moment, M, applied at the end of the curved beam. Often a force is applied. Fig. 13 shows the two possible cases. For an applied end force the stress must include the axial component of stress as well so that:

$$\sigma = \frac{My}{A\,er} + \frac{F}{A} \qquad \text{(Eq 58)}$$

where F is the applied force. In the case of Fig. 13, the axial force adds to the tensile bending stress at the inside radius and subtracts from the magnitude of the bending compression stress at the outside radius. The moment of an applied force is taken relative to the centroidal axis of the beam and not the neutral axis.

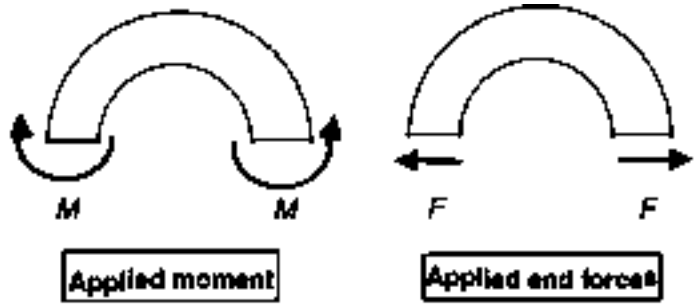


Fig. 13 Different ways of loading a curved beam

Other Geometries

Thin-Walled Pressure Vessels. A pressure vessel with a nominal radius, R, that is at least 10 times the thickness, t, is considered to be a thin-walled pressure vessel. The vessel is subjected to only internal pressure, p. The internal pressure is the gage pressure and is not absolute. The effect of the wall thickness is to control the magnitude of the radial stress. In thin-walled vessels, this radial stress is assumed to be zero, and the hoop stress is assumed to be constant across the cross section. The shear stress $_{\#r}$ is zero due to symmetry. Then, the only stresses present in the section due to an internal pressure are the hoop stress, $_{\#\#}$ and an axial stress, $_{zz}$.

The magnitudes of the axial and hoop stress are obtained by statics principles from equating the force of the internal pressure with the force of the stresses in the walls of the thin vessel. The pressure vessels can be cylindrical or spherical in shape. For the cylindrical vessel these stresses are:

$$\sigma_\theta = \frac{pr}{t} \qquad \sigma_z = \frac{pr}{2t} \qquad \text{(Eq 59)}$$

and the axial stress depends on having a capped end for the pressure vessel. A thin-walled pipe can also be treated in the same way but would not have a capped end. In that case the axial stress is zero. If the vessel is spherical, the stresses are equal in every direction

$$\sigma_{sph} = \frac{pr}{2t} \qquad \text{(Eq 60)}$$

The largest normal stress for a thin-walled pressure vessel is the hoop stress on a cylindrical vessel. The axial stress is half of the hoop stress. All of the stresses in a pressure vessel are principal stresses. Mohr's circle can be sketched (or transformation equations applied) to show that the maximum shear stress occurs on a plane rotated 45° around the axial direction and has a magnitude:

$$\tau_{max} = \frac{\sigma_\theta}{2} \qquad \text{(Eq 61)}$$

Therefore, at the macroscale, a fracture on a hoop plane is considered to be brittle fracture, and fracture on an inclined plane through the wall is considered to be ductile fracture. Macroscale brittle fracture (and therefore with no transverse thinning) is sometimes described as “thick-lipped” fracture. (An example is in the article “Fracture Appearance and Mechanisms of Deformation and Fracture).

Pin-Loaded Clevis. One geometry that is often analyzed is the pin-loaded clevis or pin-loaded plate. An example of a pin and clevis is shown in Fig. 14. This could be used for loading two bodies that must be coupled. The analysis can be applied to the pin as well as the clevis. For the pin, the load is a shear load.

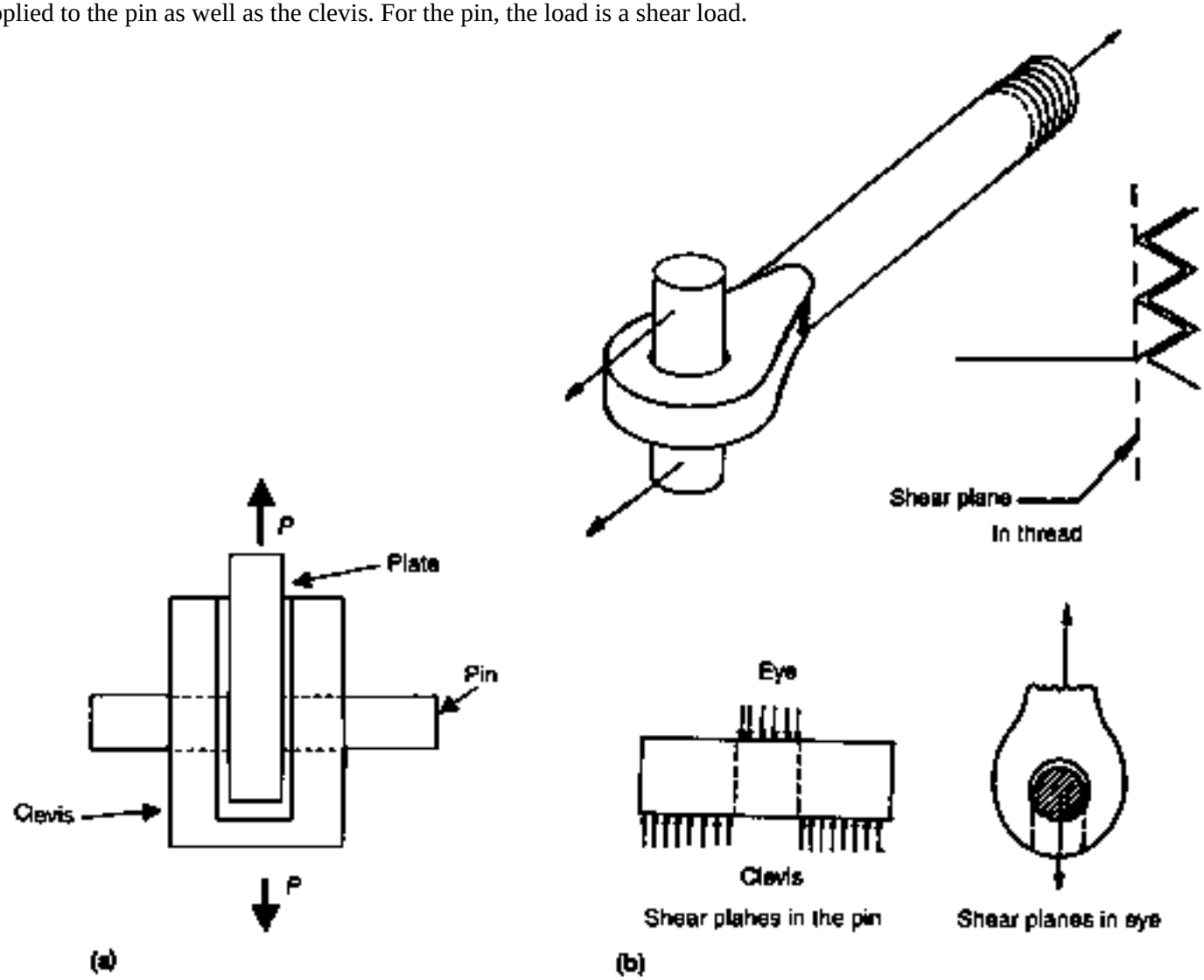


Fig. 14 Pin and clevis loading. (a) Pin going through loaded clevis and plate. (b) Shear planes in a clevis and pin connected to a threaded eye bolt

The pin often is in double shear as it is in this case. That means that it is stressed on both sides of the center. Then a load P is applied to the middle of the pin, and the internal shear force $V = P/2$ is resisted on either side of the central load. Given a pin diameter, d, and the pin shear stress, which is usually taken as the average stress given by:

$$\tau = \frac{V}{A} = \frac{P/2}{(\pi/4)d^2} = \frac{2P}{\pi d^2} \qquad \text{(Eq 62)}$$

The pin could be analyzed as a beam to get a maximum shear stress at the neutral axis but that is not commonly done. The clevis or plate with a pinhole can also be analyzed. The stress from the pinhole loading is not large enough to cause a problem unless the pinhole is too close to an edge, in which case it could be critical. The pinhole resists a shear load over the area that goes from the middle of the pinhole to the plate edge along a distance d (Fig. 15). That distance could go from the middle of the pinhole to the surface of the clevis as shown on the right; however, it is often advised to take d from the top of the pinhole to the edge of the clevis as shown on the left, here labeled d_{min}. That provides a conservative calculation of the stress. If the d_{min} dimension is about the same as the pin diameter, the design is usually okay. The stress then is:

$$\sigma = \frac{P}{2td_{min}} \qquad \text{(Eq 63)}$$

where t is the thickness of the plate. The stress concentration associated with the hole is not taken to influence the area defined by d_{min}. Here P is the total load applied to the one side of the clevis, as shown in Fig. 15, and not the load applied to the pin, as shown in Fig. 14.

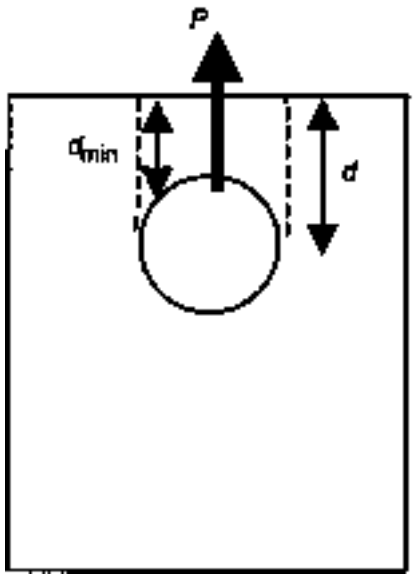


Fig. 15 Distance *d* for stress evaluation of a pin and clevis

Spring Stresses. Spring stress analysis can be done for two types of springs, the linear helical spring and the helical torsion spring. The linear helical spring has a coil of wire and is wound in a round helical shape. It is loaded along the axis of the spring (Fig. 16). In this case the spring wire is round as well. In Fig. 16, D is the diameter of the coil taken at midwire, d is the wire diameter for a round wire spring, and C is the spring index (D/d).

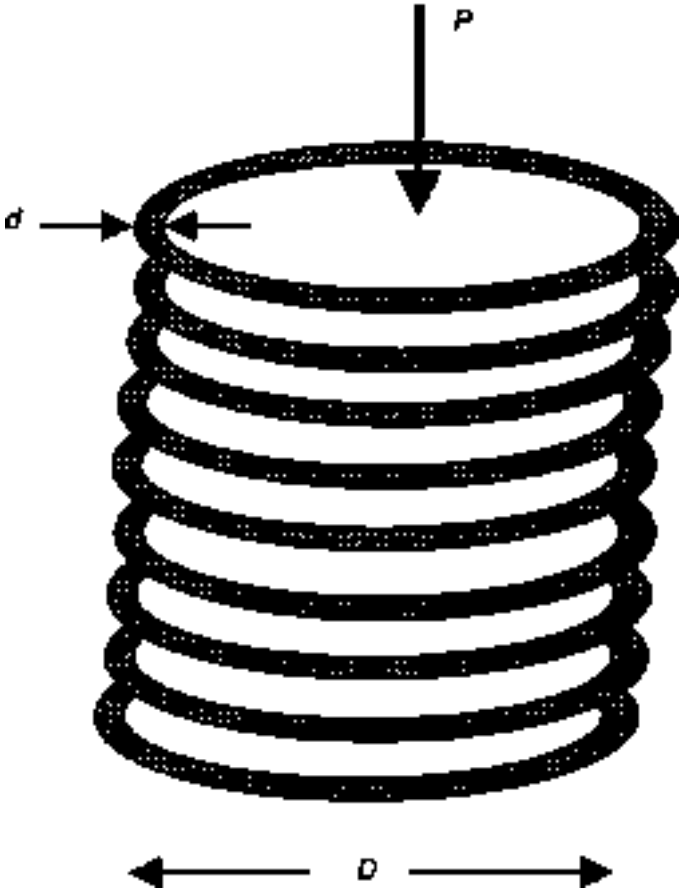


Fig. 16 Linear helical spring parameters

The only stress present in the linear spring is a shear stress, created primarily by the torsional load on the spring wire. Given force F applied to the spring along the axis and in the center of the coil diameter, D, the stress is given as:

$$\tau = K_s \frac{8FD}{\pi d^3} \qquad \text{(Eq 64)}$$

where $K_s = 1 + (0.5/C)$.

A helical torsion spring (rattrap spring) has a helical coil but it is loaded in a direction normal to the spring axis. This type of loading causes normal stresses resulting from a bending moment. The load is usually applied on an extension wire so that the applied moment is $M = FL$, where F is the force and L is the distance from the force application point, on the extension, to the center of the spring. The analysis is like that of a curved beam in bending. The regular bending formula of Eq 55 is used to calculate the bending stress, but a factor K_b is applied to the stress to account for the curvature. The resulting stress equation for the torsional spring is:

$$\sigma = K_b \frac{32M}{\pi d^3} \qquad \text{(Eq 65)}$$

where K_b takes different values for the inside of the beam versus the outside. On the inside:

$$K_{bi} = \frac{4C^2 - C - 1}{4C(C-1)} \qquad \text{(Eq 66)}$$

and on the outside

$$K_{bo} = \frac{4C^2 + C - 1}{4C(C+1)} \qquad \text{(Eq 67)}$$

and again, $C = D/d$.

Stress Concentration. In many structural shapes there are holes, corners and other geometric changes that alter the stress distribution. The result of such disturbances is a localized stress concentration. These usually result in a stress, $_{max}$, which is greater than the nominal stress, $_{nom}$, where $_{nom}$ is calculated from the application of the stress analysis formulas. This maximum stress, $_{max}$, is determined from the nominal stress, $_{nom}$, by multiplying it with a stress concentration factor, k_t. Therefore:

$$_{max} = k_t \;_{nom} \qquad \text{(Eq 68)}$$

Qualitatively, stress concentration factors increase as the difference in section at the discontinuity increases and as the fillet radius between the two sections decreases. It is also important to recognize that stress concentration is created not only by axial loading but also by bending and twisting loads and that the magnitude of the stress concentration factor depends on the type of loading.

The nominal stress from the stress formulas may be calculated based on either the net section or the reduced section and is indicated on each stress concentration factor chart; that is, for a plate with a hole, either the nominal area discounting the presence of the hole or the net section after the hole area is removed is used.

Standard references for stress concentration factors (Ref 2, 3) typically only provide values for the maximum stress created by the discontinuity, usually in a graphical format. Fig. 17 illustrates how a hole in a plate causes a stress concentration locally at the edge of the hole.

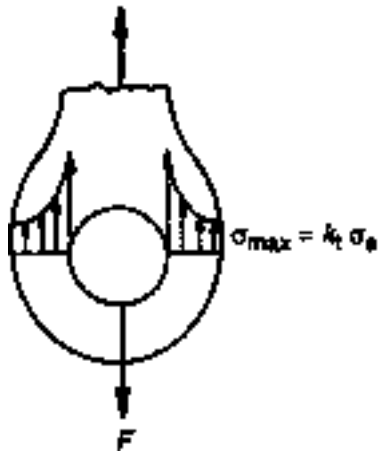


Fig. 17 Stress concentration at eyebolt. k_t, stress concentration factor; #$_a$, nominal (average) stress

A complete stress analysis often indicates that other stresses may be present and that the magnitude of those stresses depends on location and distance around the discontinuity. In some instances these stresses may be compressive. An example is the stress state associated with a circular hole in a wide plate subjected to an axial load. The stress concentration factor for such a geometry is taken as 3 (infinitely wide plate) or the actual correction for a finite width plate. However, it is important to recognize that the complete description of the stress state around the hole shows that a compressive stress is present. A stress concentration of 3 exists on a plane perpendicular to the applied load, but this factor changes to -1 on a plane parallel to the applied load at the top and bottom of the hole. The plate is not expected to fail at this location for uniaxial loading, and therefore the compressive stress is not considered in the design. However, if the plate is loaded biaxially, the situation changes. If the plate is loaded biaxially with balanced forces, the net stress concentration factor, obtained by superposition, is 2, not 3, and this stress concentration exists at four locations rather than two.

References cited in this section

2. W.K. Pilkey, *Peterson's Stress Concentration Factors*, 2nd ed., John Wiley & Sons, Inc., 1997

3. W.C. Young, *Roark's Formulas for Stress and Strain*, 6th ed., McGraw-Hill, 1989

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

Application of Stress Analysis

As noted at the beginning of this section, stress analysis provides a road map for the failure analyst. Stress analysis is an important and usually necessary part of the procedure to obtain a root cause for multiple reasons, including determining:

- Safe operating loads on a component in the original design
- Whether accurate and sufficiently precise stress analysis was done in the initial design in the event of failure
- From the macroscale examination of the fracture surface of a fractured component, whether the location of crack initiation is consistent with that expected based on proper fabrication/assembly and use of the component. If it is not, there are suggestions that the part may have been used for something other than its intended purpose or, alternatively, a metallurgical or geometric imperfection may be present in the material that caused the crack initiation site to be moved from the predicted location based on the nominal stress analysis.

Stress Analysis in Failure

Failure Criteria. Using a stress analyses for failure analysis requires a definition of the material behavior that constitutes failure. There can be many definitions for failure; yielding, fracture and buckling to name a few. A conventional design usually picks a yield criterion or a fracture criterion based on stress. For the former, the conventional yield criteria for metallic materials are based on the maximum shear stress approach (Tresca) or the maximum distortion energy approach (von Mises). For the former, Tresca, the maximum shear stress as given by Eq 23 is compared with the maximum shear stress in a tensile test. The failure stress for ductile materials is usually taken as the 0.2% offset yield strength, $_{ys}$, but in some cases (including brittle materials) it could be ultimate tensile strength, $_{uts}$. For design, a safety factor is usually chosen to determine a design stress from the failure stress. As an example, the yield strength with a safety factor of 2 or an ultimate tensile strength with a safety factor of 3 could be chosen. The maximum shear stress in a tensile test is /2 at 45° to the loading direction. Therefore having the maximum shear stress in a component equal half of the yield strength would be an example of a failure criterion based on the Tresca criterion.

The maximum distortion theory (von Mises) uses the effective stress of Eq 28. For a tensile test $_e$ equals the yield strength. Therefore a general failure criterion, $_e = _{ys}$ is usually the von Mises failure criterion. The Tresca and von Mises criteria do not give the same result except for a few special loading conditions, pure tension being one of them. Nevertheless, they are always within 15% of giving the same result and the Tresca criterion is more conservative. Both the isotropic von Mises and Tresca criteria are used for design purposes for ductile materials. (If the material is highly anisotropic, the common criterion is due to Hill, not discussed here). For the more brittle materials, the simplest criterion is a maximum normal stress and is often used but often does not accurately predict failure when uniaxial loading is compressive or for multiaxial loading when principal stresses are of opposite sign. This would be based on the principal stress, $_1$, reaching a limit, for example, a value of failure stress. For some brittle metallic materials and for most polymeric materials, a better representation of experimental data is obtained by using a criterion that includes both normal and shear stresses (that is, the normal stress on the plane of maximum shear stress affects the permissible shear stress for no yield).

For materials that behave in a brittle manner at the macroscale (due to constraint created by geometry or for an inherently brittle material), it is good to consider the effect of defects in the material. In that case, the fracture mechanics approach presented in the next section is usually a better way to assess failure potential than a maximum principal stress criterion.

For the preceding criteria it is assumed that the material behavior is even, that is, the failure occurs at the same load magnitude in tension loading as it does in compression loading. This assumption is acceptable for many metals. For some other materials it is not a good criterion. (A discussion of failure in uneven materials is given in many textbooks as the Coulomb-Mohr criterion. It is not discussed here; however, this criterion is important in reference to polymeric materials.)

Fracture Orientation. The orientation of the fracture surface also depends on the type of behavior. For a material that behaves in a macroscale brittle manner, the failure is perpendicular to the maximum normal direction. A macroscale ductile fracture is assumed to occur on a plane of maximum shear stress. The maximum shear stress often occurs on a 45° plane between that of the maximum principal stress and the minimum principal stress. Since the minimum principal stress is often the zero stress of a free surface, the fracture plane is often oriented at 45° to a free surface. (However, for plane strain deformation of plate geometries, this angle is changed to approximately 55° for uniaxial loading of isotropic material.) Mixed-mode fractures result when constraint is incomplete so that fracture occurs on both planes of maximum normal stress and maximum shear stress. This causes shear lips in platelike structures and the cup and cone fracture of a tensile specimen. Any fracture that causes a 45° ridge to a free surface is usually caused by a ductile fracture mode. On the other hand, a fracture surface that is flat is not always an indication of a brittle fracture. A microscale ductile fracture may occur with relatively flat fracture. That is, the macroscale appearance is brittle (with little transverse contraction) but the fracture mechanism (microscale) is ductile (i.e., microvoid coalescence).

Defects and Imperfections. An imperfection in a body may cause failure at a relatively low nominal stress, and if this stress is below the design stress, the imperfection can be considered a defect. Failures predicted by some of the preceding criteria even when stress concentration is taken into account, are often nonconservative. The fracture mechanics approach was developed to account for the effect of defects on failure for a body containing cracklike imperfections. This approach

is discussed in the next section and not described here. However, the question of whether an imperfection in the body or material constitutes a cracklike defect that should be analyzed by the fracture mechanics approach is one that needs some discussion. This is extremely important because a traditional approach to increase safety of a design is to increase the section to decrease the nominal stress. However, if the material contains a cracklike imperfection and the section thickness is increased, the design becomes less safe, not more safe.

It is not uncommon that many of the nondestructive methods used to find defects are unable to precisely distinguish the type of defect and how damaging it would be. Assuming the imperfection is cracklike and applying the fracture mechanics approach is the most conservative way to do a safety analysis, but it is necessary if cracklike imperfections are present.

In any stress analysis used in design, there are two inputs: material properties and loading conditions. Failure occurs when the loading conditions cause material strength, ductility, or toughness properties to be exceeded. The question then is how to handle the presence of imperfections. Two types of imperfections can exist in materials: geometric imperfections created by manufacturing operations or abuse, and material imperfections. Geometric imperfections include nicks and gouges, and in some cases, surface roughness created by machining. Material imperfections include deleterious microstructures (inherent brittleness due to mechanical/thermal processing errors, too large a grain size, precipitate-free zones, micro-and macroscale segregation, too high an inclusion content, shrinkage porosity in castings, etc.). Fracture mechanics analysis is a macroscale analysis and therefore properly used to analyze the effect of macroscale imperfections and thus primarily geometric imperfections. One example where a material imperfection might be used in a fracture mechanics analysis would be a macro-scale region of porosity (although this is not standard procedure in a fracture mechanics based analysis). Alternatively, material (microscale) imperfections cause a change in apparent mechanical properties and therefore are not used in determining the effect of loading conditions (and therefore as stress intensity input in a fracture mechanics analysis) but rather as input of the strength, toughness, or ductility of the material.

Critical Locations. Finding critical locations for possible failure and applying the analysis to these areas is an important consideration in the application of the stress-based failure analysis. To apply the ductile failure criteria, the regions with the highest stresses must be found. These are not always the largest tensile stress as in the first principal stress but the stresses that are used in the ductile failure criteria, such as the largest shear stress or the largest value of effective stress. With an analysis that involves the mechanics of materials presented previously, the critical locations can be determined by considering the nature and stress distribution that are found from each loading:

- Tensile loading is assumed to create a uniform distribution of stress.
- Torsion-bar loading has the highest stress in outer regions.
- Bending beam has the highest stress where the moment is highest and at a maximum distance from the neutral axis.

These are important and should be part of finding the most critical stress region. If failure is determined not to initiate at these locations, the implication is that a material imperfection (geometric or material) has moved the initiation site.

In addition, the highest stresses that result for the mechanics of materials formulas may not be the highest stresses in the body. These are only stresses in the major directions of the body. A stress transformation to find principal stresses and maximum shear stresses should be done. For ductile materials, shear stress or difference between stress components may be more important than the absolute maximum normal stress value.

An example of finding critical locations is shown in Fig. 18. A round beam has an applied torque, axial load, and transverse load. The stresses that would result from these loads are axial normal stress, bending normal stress, torsional shear stress, and beam shear stress. The location of maximum normal stress and maximum shear stress is then found by considering the location where the net stresses are maximum. The axial stress is independent of location, and the stress tensor associated with this stress is given by:

$$\sigma_{axial} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_{axial} \end{bmatrix}$$

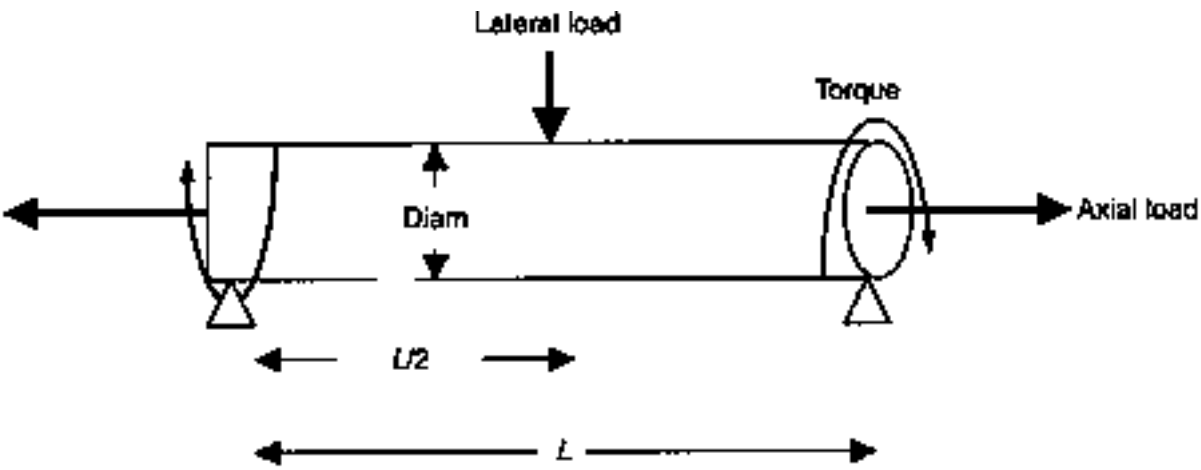


Fig. 18 Example of a round beam with multiple loading

The bending stresses vary with position and are given by:

$$\sigma_{bend} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_{zz,b} \end{bmatrix}$$

$$\tau_{bend} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \tau_b \\ 0 & \tau_b & 0 \end{bmatrix}$$

The torque applied to the section does not vary with position and is given by:

$$\tau_{torque} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \tau_{torque} \\ 0 & \tau_{torque} & 0 \end{bmatrix}$$

Therefore, examination of these stresses indicates that:

- The axial stress will add to the variable bending stress.
- The shear stress created by the torque will add to the variable shear stress created by the bending.

Clearly, then, the maximum net axial stress will occur on the bottom of the beam where the beam bending stress is tensile and will have a magnitude of the sum of these two stresses. Also, the maximum shear stress will occur along the neutral axis where the maximum shear stress from bending exists and will have a magnitude of the sum of the shear stress from the torque plus that from the bending. (In most cases, the shear stress from the bending is small in relation to that from the torque and is often neglected in an analysis. For solid prismatic sections, it is $\frac{4}{3}V/A$ and for solid circular sections it is $\frac{3}{2}V/A$, where V is the nominal shear force in the beam. However, for hollow tubes, the maximum shear at the neutral axis increases to $2V/A$, and therefore should not be so quickly neglected.

Examination of the aforementioned stress tensors then indicate that the critical locations to examine are (a) where the net shear stress is a maximum and (b) where the net axial stress is at a maximum. Location (a) is along the neutral axis and (b) is at the location of the maximum bending moment. Ductile fracture for (a) on the plane of maximum shear stress will occur on either a longitudinal or transverse plane, while brittle fracture at this location will occur on a plane at 45° to the maximum shear stress. Brittle fracture due to the net axial stress will occur on a plane normal to the axis, and ductile fracture will occur on a plane at 45° to the axis.

If the lateral load were not at the midlength or if several sets of lateral loading were involved, a shear and moment diagram construction could identify the critical locations for maxima. Then again, the top and bottom of the beam should be analyzed for the flexural stress and the midplane for shear. Other components can be analyzed in a similar way to find critical points. When the geometry or loading pattern becomes so complex that a simple application of the mechanics of

materials formulas cannot be made or does not reveal the critical locations, then a numerical analysis can be made and critical stress locations found.

Limit Analysis. An alternate method for stress analysis is a limit analysis. Here the approach is to analyze a structure with the assumption that is has reached a plastic limit. The calculations for this can often be much easier then than the elastic stress analysis described in the section “Stress Analysis of Common Geometries” in this article.

The limit analysis is done with two limit theorems, an upper bound theorem and a lower bound theorem. The upper bound theorem chooses a plastic collapse mechanism and balances external work on the structure against the internal energy dissipated by the collapse mechanism. Different collapse mechanisms can be assumed, and the one giving the lowest collapse load would be closest to being correct. The lower bound theorem uses an equilibrium analysis while requiring that no stress in the structure is above a limit stress like yield strength. The lower bound analysis can use different stress distributions, especially regarding the points that are at a limit stress. The analysis that gives the highest limit load would be the closest to being correct. If the lower and upper bound approaches give the same result, an exact limit solution is reached. If they differ, they bracket the exact solution, which may not be exactly determined.

The basic assumption on most limit analyses is that the body is undergoing ductile failure. In those cases the residual stresses do not have much influence. Under a large strain field as that imposed in a limit analysis, the strain mismatches that caused residual stresses tend to be washed out and the residual stresses disappear. Therefore in limit analysis, secondary stresses like limit residual stresses are often ignored.

Strain rate effects in safety analysis can have a confusing set of effects. For a ductile material, a fast loading rate would raise the strength of the material, possibly resulting in a higher failure load. However, if a brittle mechanism of failure may occur, a faster loading rate may lower the failure load (e.g., a higher DBTT with a higher strain rate). This is especially true for a structure with a notch or cracklike defect. In addition, the faster loading rate may change the mechanism of failure from a ductile one to a brittle one, and a failure prediction based on ductile behavior may not be correct. Therefore, it is important to consider the effect of a fast loading rate on the failure of a structure on a case by case basis and not use general principles to try to predict the effect of the loading rate. The evaluation requires knowledge of the failure mechanisms, for example, fracture or yield, material property data under rapid loading, and the loading rate (see ASM Handbook, Volume 8 for the regimes of strain rates). A fast loading rate may alter the stress analysis by introducing inertial effects that influence the stress distributions. A loading to failure that is in the millisecond range usually does not result in inertial effects but rates an order of magnitude faster than that which could introduce inertial effects.

Examples

The failure analyst also should realize that many low- and medium- technology products are designed and manufactured without ever having an engineer perform any sort of stress analysis. This obviously makes the failure analysis more challenging. Fortunately, there is usually someone who is able to help determine where the maximum stress should have been. In some cases, this is quite helpful and all that is necessary for the scope of the investigation (as a properly performed stress analysis may be out of the range of the time and budgetary constraints for the failure analysis project). In addition, if analysts gain an understanding of the basic types of loading (tension, compression, bending, and torsion), and study the expected crack appearances for each of these four loading types in cylindrical and prismatic geometries, they will be able to characterize the stresses of fracture regardless of the presence or absence of a correct or incorrect stress analysis.

It also is not always necessary for the failure analyst to personally become involved in a complex stress analysis. Those services may require extensive training, and the failure analyst should be careful not to perform work outside of his or her area of expertise. However, the failure analyst must be able to appreciate the implications of the analysis. It is more important for the failure analyst to be able to examine a failed part and to know whether the failure location and, in the case of fracture, the fracture surface orientation and initiation site are consistent with those expected in proper use of the component. Such stress analysis is qualitative or semiquantitative and may be key to identifying cause for failure. Two examples are cited, one simpler than the other, to illustrate this point.

Example 1: Stresses in a Screwdriver Used as a Pry-Bar. Normal use of the screwdriver imposes a torsional state of stress on the shank of the screwdriver. If the screwdriver is used as a pry-bar, the state of stress is that created by bending. Identification of root cause of a broken screwdriver clearly resides in recognizing indications on the fracture surface that tell whether the screwdriver failed during expected usage (torsional stresses), or while being used for other than its intended purpose (bending). The identification can be made because of difference in orientation of the fracture surface relative to the shaft. The screwdriver could still fail when serving its intended and proper use, but the implications of the failure are totally different in that case.

Example 2: Stresses in a Pressurized Cylindrical Section. As a second example, consider an internally pressurized cylindrical section (see the section “Thin-Walled Pressure Vessels” in this article for a more complete discussion of the stress state in such a section). The stresses present in this (assumed to be thin-walled) section are a tensile stress acting along the axis of the cylinder and a tensile circumferential stress. The magnitude of the circumferential stress is twice as

large as the axial stress. Therefore, failure, if it occurs, is assumed to be caused by the circumferential stress. However, the vessel must be supported in some way, so it is likely that bending stresses are also present in the vessel. The longitudinal stress created by the bending adds to the longitudinal stress created by the pressure and must exceed the longitudinal stress created by the pressure before it can change the plane on which the maximum stress acts. The latter could occur for example by the presence of a nick or gouge in the material in a location of high bending stress. There are then four possible ways in which fracture could occur: ductile or brittle fracture due to the internal pressure; or ductile or brittle fracture due to the internal pressure plus the bending stress. Identification of root cause clearly rests in part on being able to examine the fracture surface to determine whether bending loads were involved in the failure and whether the fracture was macroscale ductile or brittle.

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Fracture Mechanics

In general, there are two types of conditions that may lead to structural failure:

- Net-section instability where the overall structural cross section can no longer support the applied load
- The critical flaw size (a_c) is exceeded by some preexisting discontinuity or when subcritical cracking mechanisms (e.g., fatigue, stress-corrosion cracking, or creep) cause the crack to reach critical size.

Failures due to net-section instability typically occur when a damage process such as corrosion or wear reduces the thickness of a structural section. This type of failure can be evaluated by traditional stress analysis or finite element analyses. These methods can also help identify the likely areas of stress concentration where cracking may have initiated.
However, stress analyses by traditional methods do not easily account for crack propagation from preexisting cracks or sharp discontinuities in the material. Basic elasticity calculations show that both stress and strain become astronomical at a discontinuity such as a crack, far exceeding any recognized property levels that might offer some sort of limitation. Two situations are then possible:

- The crack reaches a critical length so that instability occurs and cataclysmic fracture occurs. The fracture is macroscale brittle (microscale ductile or brittle).
- The crack blunts, redistributing the stress state, with continued loading creating a tear zone (and sharpened crack-tip radius) in front of the crack. In steels, this tear zone can then cause the critical crack length to be exceeded such that unstable cleavage fracture occurs or unstable microscale ductile fracture is induced. Which event occurs depends on the temperature and the loading rate, but in either event, crack propagation is unstable (i.e., does not require an increasing load after creation of the tear zone). See the discussion below and in the article "Mechanisms and Appearances of Ductile and Brittle Fracture in Metals."

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Fracture Mechanics Concepts

The analysis of macroscale brittle fracture induced because of the presence of cracklike imperfections in a material is the basis of fracture mechanics.
In a simplistic overview, fracture mechanics says that there is a triangle of variables that must be considered in the analysis of fracture. This is in contrast to traditional design in which there are only two components of the design criteria—material properties and applied loading conditions. For linear elastic fracture mechanics (LEFM) analysis, the three parameters involved are the stress created by the applied load ( ), the stress

intensity (K), and the crack and component geometry. The K created by the applied loads is then related to the stress and crack length via:

$$K = \sigma\left(\sqrt{\pi a}\right)\left[f\left(a/W\right)\right]$$

where a is the crack length and $f(a/W)$ is a geometry correction factor tabulated in handbooks, which applies to a given component and crack geometry. Common examples include an edge-crack in a plate, a center crack, a double edge crack, and so forth. Fracture is assumed to occur when the $K_{applied}$ created by the applied loads exceeds a material property K_c. It is important to recognize that the geometry correction factor is not a stress concentration factor.

Fracture can be macroscale brittle, macroscale ductile, or macroscale mixed ductile and brittle. The attraction of fracture mechanics for fracture analysis is that so long as the volume of material described by the stress intensity factor is small with respect to the dimensions of the component and the size of the plastic zone at the crack tip is small with respect to the K-zone, the description of the stress state in the vicinity of the defect can be assumed to be elastic (LEFM analysis) and the resulting stress state and stress intensity factor is the same, no matter what the component geometry. Hence, the "one parameter" (K) description of the fracture process of fracture exists. Results from simpler laboratory test specimens can be transferred one-to-one to a full scale engineering component of different geometry. As shown below, there is a plastic zone at the tip of the crack. If the size of this zone is large with respect to the K-field, linear elastic analysis may not be appropriate. In this case, the more recently developed field of elastic-plastic fracture mechanics (EPFM) must be utilized in which some other material property (such as the J-integral) replaces the stress intensity factor to describe the "toughness" of the material. Analysis by LEFM is easier to apply (fracture control is described by a one-parameter analysis, the critical value of K). Additionally, from a practical standpoint, more material data is available in the literature for LEFM fracture than for EPFM fracture. If EPFM analysis is used, a two-parameter approach is necessary. The following section focuses on LEFM analysis, with some mention of EPFM analysis.

In conclusion, it is important to recognize that in the application of fracture mechanics analysis to a failure analysis, the material is treated as a black box. Analysis is based only on macroscale parameters of nominal stress, crack length, and component geometry. (This should not be interpreted to imply that microstructural parameters do not affect the value of K at fracture. Clearly, fracture toughness depends strongly on the specific microstructure present in the material and also on environmental conditions.)

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

An Introduction to LEFM

Assume the presence of a cracklike imperfection in a material as shown in Fig. 19. Irwin showed that the stress state around the imperfection could be described two dimensionally, where:

- On the plane of the crack, the only existent stress is $_{yy}$
- There is a singularity in the stress field as r approaches zero; that is, the stresses approach infinity.

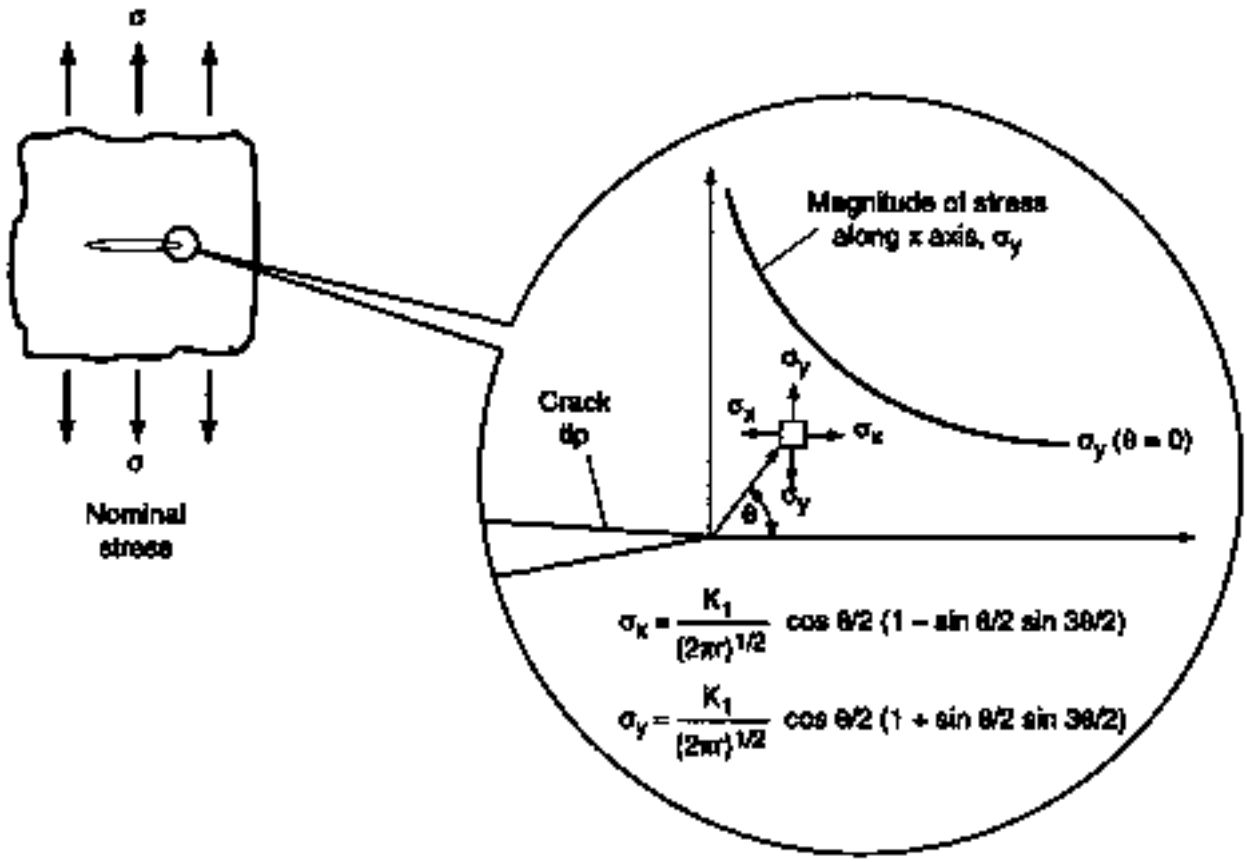


Fig. 19 Distribution of stresses near the tip of a through thickness crack in a plate

On the plane of the crack (# = 0 in Fig. 19), the only existent stress is $_{yy}$:

$$\sigma_{yy} = \frac{K}{\sqrt{2\pi r}}\left[f(\theta)\right] \qquad K = \sigma\left(\sqrt{\pi a}\right)\left[f(a/W)\right]$$

where a = crack length and W is the width of the plate containing the cracklike imperfection.

If the material is not inherently brittle, the crack tip stress $_{yy}$ reaches the yield stress (YS) at some distance in front of the crack tip, causing a plastic zone to be created. There are various ways to estimate the size and shape of the plastic zone (i.e., based on use of the yield equations in the previous section. However, the simplest approach is to assume that yield occurs at a distance in front of the crack tip:

$$r = \left(\frac{1}{2\pi}\right)\left(\frac{K_c}{\mathrm{YS}}\right)^2$$

The Irwin analysis was 2D. If the body is sufficiently thick (see subsequent paragraphs), the state of stress becomes 3D by the addition of a transverse stress, $_{zz}$, that acts in the thickness direction of the body and that has a magnitude given by:

$$_{zz} = \cdot\,(\,_{xx} + \,_{yy})$$

Stated differently, the state of stress at the crack tip changes from plane stress to plane strain, with the transverse stress $_{zz}$ preventing flow in the transverse direction. Then, application of yield equations shows that the plastic zone is reduced in size, being approximately one third as large as for the plane stress case. Irwin also observed that the critical value of K at fracture then becomes independent of section thickness and that the lateral contraction (due to $_{yy}$ at the crack tip) reaches a minimum value (i.e., the strain is 2D, plane strain). Since the value of K at fracture no longer depends on section thickness, it is considered to be a material property and described not as K_c, but as the plane-strain fracture toughness K_{Ic}. Determination of material property toughness is covered by several ASTM specifications, E-399 specifying the determination of K_{Ic}.

For quantitative evaluation in either design or in failure analysis, the material toughness at fracture must be known. Tabulated values in the literature are for plane strain conditions. If the section thickness to be used in the design or that of a failed component does not create sufficient constraint to develop plane strain conditions, the fracture toughness at fracture is unknown but higher than the tabulated plane strain value. For purposes of design, the required section thickness to obtain plane strain conditions can be estimated (for a through-thickness crack loaded in opening mode) according to:

$$B \geq 2.5\left(\frac{K_{Ic}}{S}\right)^2$$

where B = section thickness and S = design stress, typically the yield stress decreased by a factor of safety. This criterion is conservative in that plane strain values will be obtained at this section thickness but may be also obtained for smaller section thicknesses. For application to failure analysis of a fractured component, examination of the fracture surface will indicate whether all of the fracture surface is flat and normal to the applied loads or whether the fracture is mixed mode and contains shear lips. If shear lips are present, constraint was incomplete so that the value of K at the onset of fracture is greater than K_{Ic}.

This equation can also be used for purposes of inspection and determination of suitability for service. Assuming that K_{Ic} and the yield strength of the material is known, the equation can also be used where the crack length, a, replaces the section thickness B, allowing the estimation of critical crack lengths. Such estimates predict critical crack lengths at room temperature of a few thousands of an inch for a temper-embrittled steel to several inches for hot rolled and or properly quenched and tempered steels. Similar examination of the toughness of high-strength aluminum alloys indicate that unless these materials are specially tempered to maximize fracture toughness (e.g., the T73 temper), critical section thicknesses may be less than $\frac{1}{4}$ of an inch.

The general trend in all cases (of properly processed material is that the fracture toughness tends to decrease as the strength level is increased for materials typically used as structural or machine components. In steels, the fracture toughness does increase initially with carbon content for low (lower than that of the structural grades) carbon contents. However, the specific variation is quite sensitive to the microstructure present. Additionally, fracture toughness can be highly anisotropic in fabricated components. In rolled plate, fracture toughness is highest when cracks propagate in the long transverse direction and lowest when cracks propagate parallel to the rolling direction. Similar results are obtained for other deformation patterns with low fracture toughness associated with the largest principal fabrication strain in the body. This anisotropy is due to both banding (microscale solute segregation, which may result in microstructure constituent banding as for pearlite distribution in hot rolled steel) and fibering created by inclusion shape, distribution, and morphology. Some success in reducing fracture toughness anisotropy due to fibering and raising the overall toughness level has been achieved in the steel industry via ladle calcium treatments.

The analysis therefore must be modified to include a small zone of plasticity at the crack tip. The modified stress distribution is shown schematically in Fig. 20. The small plastic enclave is called a plastic zone, and its extent directly ahead of the crack is denoted by R_P.

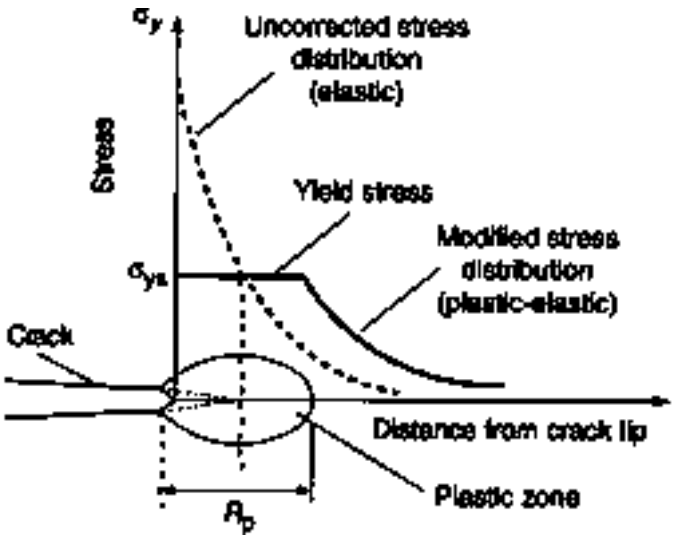


Fig. 20 Stress distribution ahead of crack in which small-scale plasticity is included

The stress field at the crack tip can be characterized in terms of the stress intensity (K), where the plastic zone size depends on both K and $_{ys}$:

$$R_p = C\left(\frac{K}{\sigma_y}\right)^2$$

The constant, *C*, depends on various factors including thickness and deformation characteristics of the material. It usually has a value between 1/+ and $\frac{1}{3}$+. This equation is only valid for linear elastic bodies; that is, when R_P is much smaller than the crack size (i.e., R_p < 10% of the crack size) and other physical dimensions of the body (Ref 5). As *K* increases (e.g., by increasing the load) R_p also increases, and the material in the plastic enclave becomes more severely strained. This plastic enclave is very important in metals and polymers since it is the region in which energy is absorbed and is responsible for the relatively high fracture toughness of metals compared to all other materials.

As a result of the direct relationship between *K* and the strains at the crack tip, a critical value of *K* can be defined for when fracture takes place. When the region of plastic deformation around a crack is small compared to the size of the crack (as is often true for large structures and high-strength materials), the magnitude of the stress field (in tension) around the crack is thus related to the stress-intensity factor, *K*, as follows:

$$K = Y\sigma\sqrt{\pi a}$$

where *Y* is a geometric factor, is the gross stress across the fracture plane, and *a* is the crack length. In this way, linear elastic analysis of small-scale yielding can be used to define a unique factor, *K*, that is proportional to the local crack-tip stress field outside the small crack-tip plastic zone. The stress-intensity factor can have a simple relation to applied stress and crack length, or the relation can involve complex geometry factors for complex loading, various configurations of real structural components, or variations in crack shapes. For example, the geometric factor, *Y*, is typically on the order of 1 for a tension-loaded plate with a through-thickness crack (Fig. 21). Geometric factors (expressed as a flaw shape parameter, *Q*) for surface cracks and embedded cracks are also shown in Fig. 21 (Ref 6). Specific formulas for *K* depend on the load/crack geometry, and numerous solutions for *K* can be found in Ref 7 and in the article "Summary of Stress-Intensity Factors" in *Fatigue and Fracture*, Volume 19 of *ASM Handbook* (Ref 8).

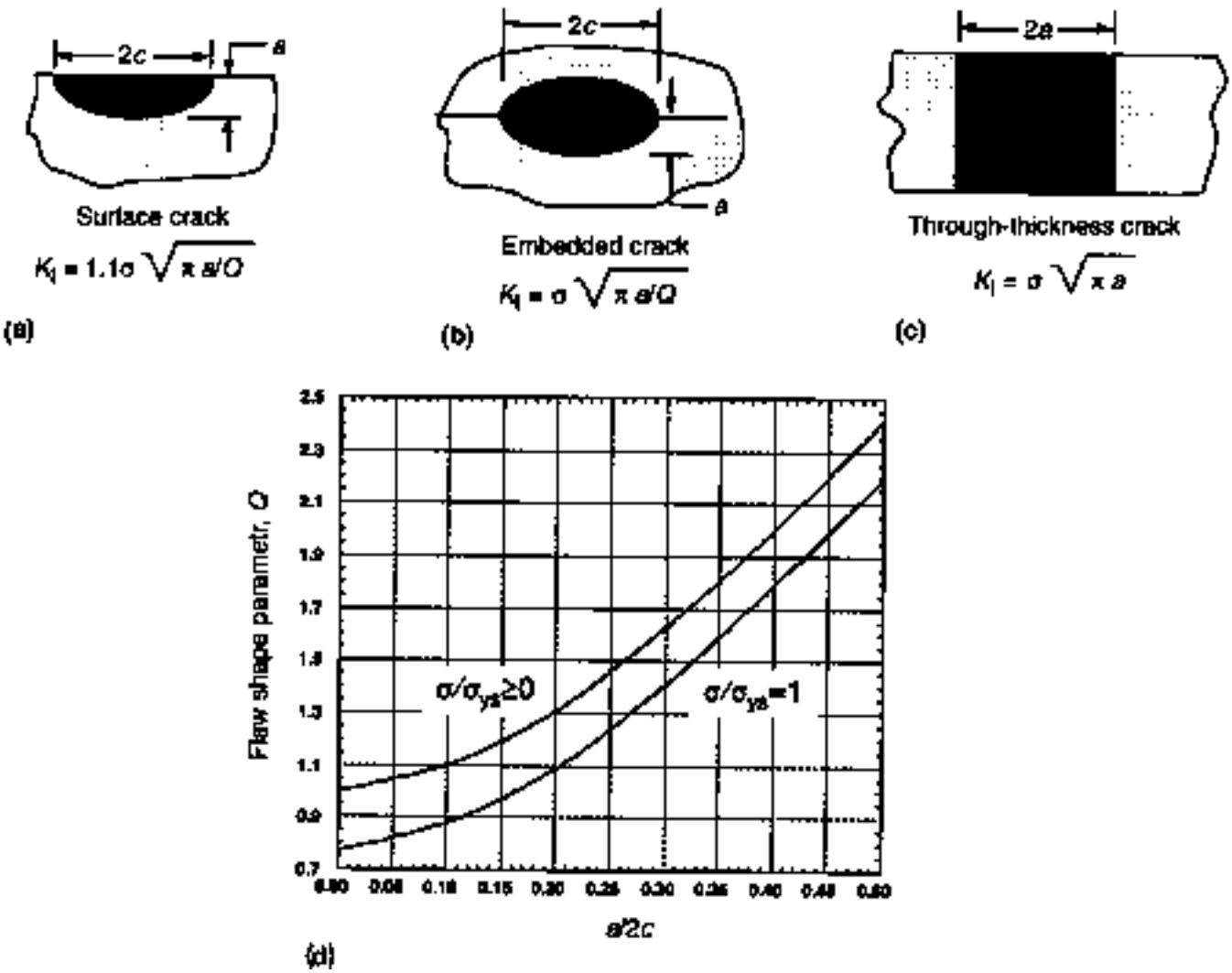


Fig. 21 Stress intensity factors (in tension, k_I) for various crack geometries. (a) Surface crack. (b) Embedded crack. (c) Through-thickness crack. (d) Flaw shape parameter (Q). Source: Ref 6

These concepts provide a basis for defining a critical stress-intensity factor (K_c) defined as:

$$K_c = Y\sigma_f\sqrt{\pi a_c}$$

where $_f$ is the fracture stress occurring with a critical crack size, a_c. The critical stress intensity, also known as fracture toughness (K_c), is the value of K that results in rapid, unstable fracture. The value of K_c depends not only on the material being considered but also on the thickness of the material in which the crack is found. Plastic deformation is the most important source of toughness for most structural materials, and anything that impedes plastic deformation will reduce the toughness. As previously noted, toughness is manifested primarily through plastic deformation at the crack tip; anything that impedes crack-tip plastic deformation will reduce the toughness. The plastic deformation might be visible in a test specimen as a non-linear displacement curve with rising load (although this could correspond to crack opening due to *stable* ductile tearing.) However, the deformation may well not be visible on the specimen surface, but it can be positively identified fractographically on a microscale.

The macroscopic deformation of the material also influences toughness. If the material is ductile enough or thin, it tends to contract (neck) parallel to the crack plane. This is indicative of plane stress conditions (Fig. 22b); that is, the stress along the crack plane is zero (Eq 41b). However, if the section size is sufficiently thick (or if the material is more brittle), the lateral contraction is opposed by the bulk of the thick material, and an additional stress component develops parallel to the crack plane (i.e., plane strain conditions, Eq 41a). Under plane-strain conditions, plastic deformation at the crack tip is severely limited. Lateral contraction is constrained along the crack plane, and the fracture surface is very flat with little or no macroscopic deformation perpendicular to the crack plane (Fig. 22a).

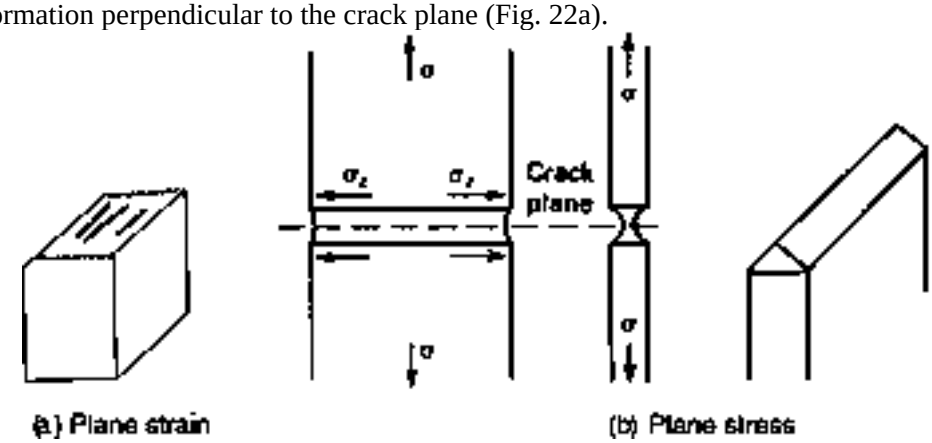


Fig. 22 Deformation and fracture surface for (a) plane strain and (b) plane stress. In plane strain, the lateral constraint of a thick section develops a stress along the crack plane (#$_z$, Eq 41aa).

The value of K_c thus depends not only on the ductility of the material but also on the state of stress. If the component is thin or sufficiently ductile, there is contraction along the crack plane, and the work of lateral deformation contributes to toughness. As the thickness of the section increases, the critical stress intensity is reduced until it reaches a minimum value for plane-strain fracture toughness (K_{Ic}) (Fig. 23). In the plane-strain state, a material is at its lowest point of resistance to unstable fracture, and the mode of fracture is brittle without any appreciable macroscopic plastic deformation. The plane-strain fracture toughness (K_{Ic}) is a minimum value and an inherent material property. Valid linear-elastics conditions for determination of K_{Ic} require a thickness / $2.5(K_{Ic}/\ _y)^2$

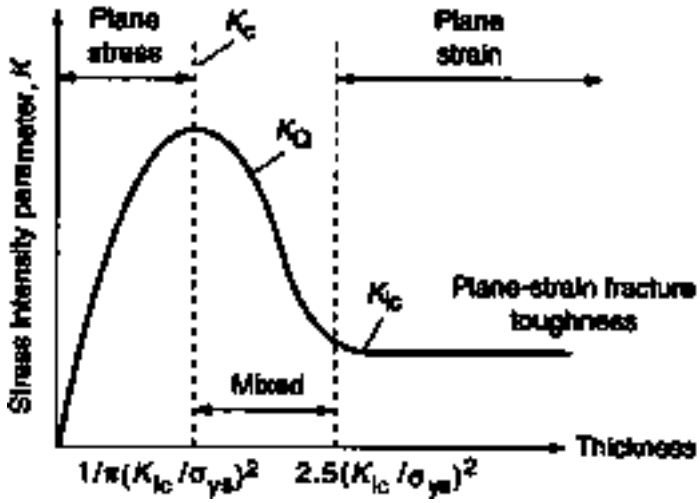


Fig. 23 Effect of thickness on state of stress and fracture toughness at the crack tip. Source: Ref 5

In general, when the specimen thickness and the in-plane dimensions near the crack are large enough relative to the size of the plastic zone, then the value of *K* at which growth begins is a constant and the plane-strain fracture toughness can be determined. If not, elastic-plastic fracture mechanics (EPFM) is required.

Linear elastic fracture mechanics (LEFM) is a useful tool in failure analysis, because many structural failures occur by subcritical crack growth mechanism (e.g., fatigue, stress corrosion, and creep) until a critical crack (a_c) is reached. In this regard, fracture mechanics is an effective tool for evaluating critical flaw size (a_c) that leads to rapid unstable fracture and can help answer questions during a failure analysis such as (Ref 6):

- Where should one look for the transition from subcritical crack growth to unstable rapid fracture?
- What was the load on the component at the time of failure?
- Was the correct material used and manufacturing/processing sound?
- Was the part designed properly?
- Did the environment influence the failure?

Linear elastic fracture mechanics can be applied to interpretation of subcritical crack growth, as briefly described in the next section. Care must be taken in applying LEFM to the onset of unstable (rapid) fracture of ductile materials, because the plastic component of fracture resistance may be significant. When plastic deformation is more excessive, the errors due to the assumption of elastic behavior become too large to be acceptable or realistic. In this case, elastic-plastic fracture mechanics (EPFM) is required.

When the crack plane is subjected to mixed modes (tension, shear, torsion) of loading, linear superposition is used to determine crack-tip stress fields (see Ref 9 for examples). However, failure criterion is written in terms of an effective *K*, which for modes I and II would be: $K_{eff}^2 = K_I^2 + K_{II}^2$. Many applications are for crack growth by opening (Mode I or Mode 1). When stress is applied in the crack plane, stress-intensity values are termed as K_{2c} (shear or sliding), or as K_{3c} (tearing or torsion along the crack plane).

References cited in this section

5. S. Antolovich and B. Antolovich, An Introduction to Fracture Mechanics, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 371–380

6. H. Nelson, Fractography in the Failure Analysis of Engineering Structures, seminar presented at ASM Annual Meeting, 10 Sept 1997

7. H. Tada, P.C. Paris, and G. Irwin, *The Stress Analysis of Cracks Handbook,* 3rd ed., American Society of Mechanical Engineers, 2000

8. A. Liu, Summary of Stress-Intensity Factors, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 980–1000

9. A. Liu, Chapter 7, Fracture Mechanics for Mixed Crack Tip Displacement Modes, *Structural Life Assessment Methods*, ASM International, 1998

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Subcritical Fracture Mechanics

Fracture mechanics analysis dealing with cracking (as opposed to fracture) is sometimes referred to as subcritical fracture mechanics. The final fracture is by overload. In subcritical cracking, the amount of plastic deformation during cracking is very small, and so LEFM concepts are typically valid for crack-growth analysis. An exception is creep cracking, for which EPFM concepts must be used. Creep crack growth is discussed in Ref 10.

Fatigue. In the case of cyclic loading, subcritical crack growth from fatigue loading, *da*/*dN*, is characterized by the stress-intensity range, " $K = K_{max} - K_{min}$. During a load cycle, *K* varies from a minimum and maximum:

$$K_{min} = Y\sigma_{min}\sqrt{\pi a}$$

$$K_{max} = Y\sigma_{max}\sqrt{\pi a}$$

where the stress ratio is $R = K_{min}/K_{max}$. The crack growth rate, *da*/*dN*, depends on " *K*, and *R* and must be provided by a test of the material. Crack growth rates versus " *K* typically have three distinct regions of behavior, as discussed in more detail in the article “Fatigue Failures” in this Volume. Region 1 is the stage of early crack growth with a threshold value (" K_{th}) below which cracks rates become diminishingly small. If values of " *K* are above " K_{th}, the crack size grows and thus causes the stress intensity to increase at the crack tip. The crack growth rates accelerate during stage 1 until they reach stage 2, where rates are described by the Paris equation:

$$da/dN = C(\text{"}\ K)^n$$

where *C* and *n* are empirical constants, which depend on the stress ratio. In practice, fracture mechanics testing is typically done at a small positive *R* value (*n*, *R* ~ 0.1), while service conditions range from -1 to +1 for *R*. Because the Paris equation covers only one *R* value, an alternate form is:

$$\frac{da}{dN} = \frac{C}{(1-R)^m}\Delta K^n = C\frac{\Delta K^m}{(1-R)^m}\Delta K^{n-m}$$

This is known as the Walker equation. Many other equations are used, but it should be pointed out that the equations are phenomenological curve-fitting equations. Conditions, especially environments, can alter the shape of the curves. For example, Fig. 24 shows three cases of environmental effects:

- *True corrosion fatigue* (TCF): Fig. 24(a). Material-environmental combination where the threshold value is lowered and the chemical reaction is enhanced by the plastic behavior at the crack tip under fatigue loading. The chemical reaction is not influenced by the static stress at the crack tip.
- *Stress corrosion fatigue* (SCF): Fig. 24(b). Material-environmental combination where the chemical reaction is strongly influenced by the static stress at the crack tip (i.e., stress corrosion is observed) but is not influenced by the plastic behavior at the crack tip as a result of fatigue loading.
- *TCF combined with SCF:* Fig. 24(c)

The latter is the most common.

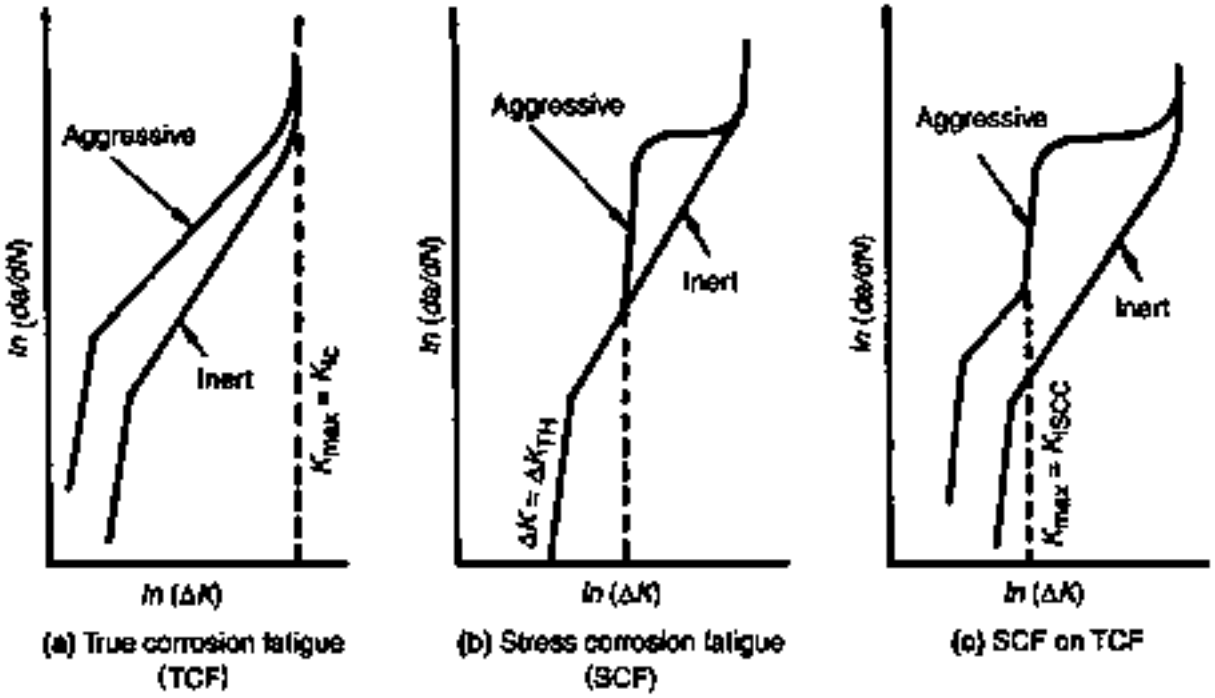


Fig. 24 Cases of environmental effects on fatigue crack growth. (a) True corrosion fatigue. (b) Stress-corrosion fatigue. (c) Stress-corrosion fatigue on true corrosion fatigue

Additionally, it is necessary to recognize that the Paris constants are in general anisotropic and for the same reasons that fracture toughness is anisotropic; that is, they depend on the direction of crack growth in mechanical/thermally processed material because of banding and segregation. Many references to " *K* - *da*/*dN* curves have superimposed comments about sensitivity of crack growth rates to loading condition variables (i.e., the *R* value), to the environment and to the microstructure. These references indicate high sensitivity of the microstructure in regions I and III but low sensitivity in region II (the Paris law region). This must be carefully interpreted. It is true that alloy compositions (e.g., the range of austenitic stainless steels) all have about the same Paris constants and that all steels, pearlitic and quench and tempered, have approximately the same Paris constants. This is of considerable benefit when back calculations are done using experimental determination of critical overload crack lengths and fatigue striation spacing, since constants may not be available for the specific alloy. However, as just noted, the Paris constants can be very sensitive to solute and microstructure banding as well as inclusion morphology and distribution. Thus, in a sense, there is also strong microstructure sensitivity in the Paris law region of the curve.

Stress-Corrosion Cracking. In the case of stress-corrosion cracking, the rate of growth will depend on *K* (similar to fatigue), but the rate of growth depends directly on time (*da*/*dt*) under a static stress. The curve of *da*/*dt* versus *K* typically exhibits three distinct regions (Fig. 25) in a fashion similar to fatigue. Stage 1 is the threshold region, where *da*/*dt* depends strongly on *K*. Stage 2 is the steady-state region where crack growth is only slightly dependent on *K*. Stage 3 is the *K*-dependent region of fast fracture. The threshold value is best determined by performing tests on cracked specimens and by measuring the time to fracture. A plot of the stress intensity applied as a function of time to failure will show an asymptote (Fig. 26). For combined stress-corrosion and fatigue cracking (Fig. 24c), one obtains *da*/*dN* from a cyclic-loading test in the appropriate environment.

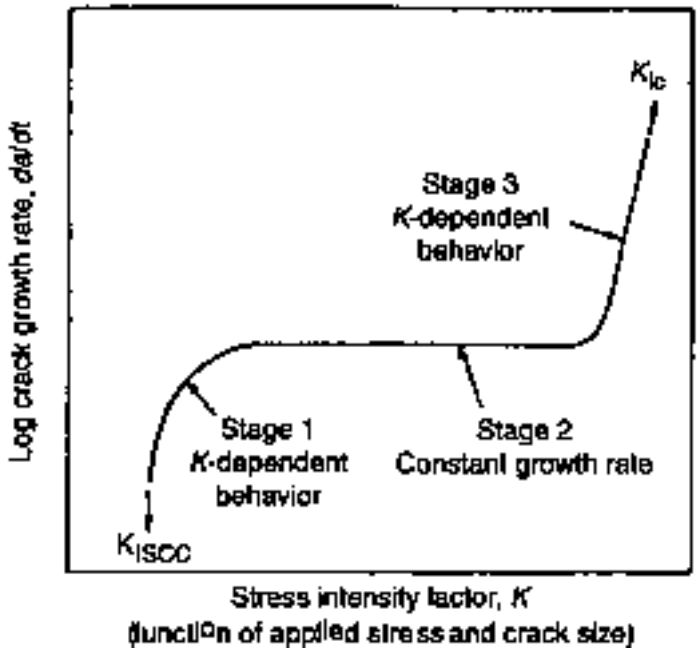


Fig. 25 Stages of stress-corrosion crack growth for a component under static loading in an aggressive environment

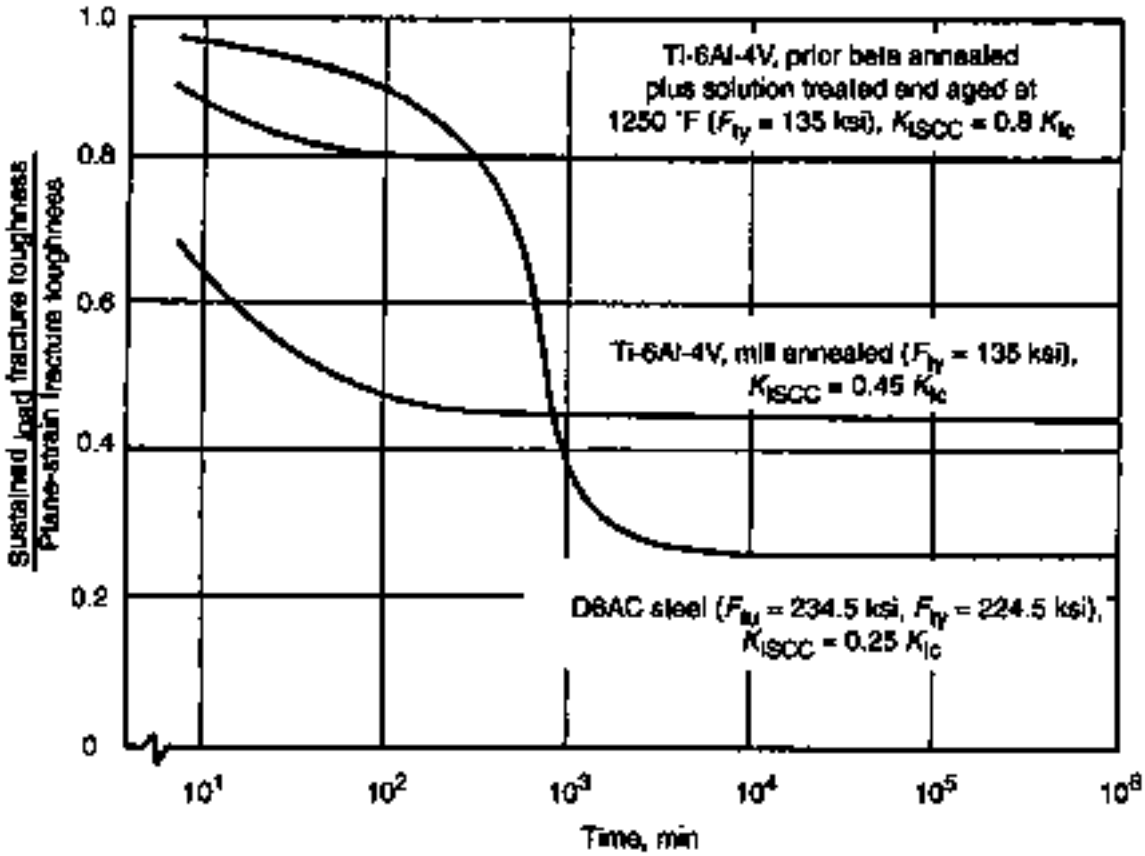


Fig. 26 Effect of environment (3.5% salt solution) on time at sustained load on fracture toughness for titanium alloys and steel. Source: Ref 10

Reference cited in this section

10. A. Liu, Chapter 6, Environmentally Assisted Crack Growth, *Structural Life Assessment Methods*, ASM International, 1998

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Elastic-Plastic Fracture Mechanics

Different approaches are used for evaluating materials that fracture (or behave) more plastically than they behave in LEFM. This requires the application of EPFM or nonlinear elastic fracture mechanics. Elastic-plastic fracture mechanics is still based on elasticity concepts, but it must also account for the nonlinear portion of the stress-strain curve in the region of plastic deformation. Hence, the term *nonlinear fracture mechanics* is used. It is described only briefly here with more details contained in Ref 11.

The *J*-integral and the crack-tip opening displacement (CTOD) methods are most often used to characterize fracture at the crack tip in EPFM. As plastic deformation is involved, fracture resistance in EPFM can be characterized logically and expressed in units of energy per unit area. For example, a general (and simplistic) form of the *J*-integral may be given as follows (Ref 12):

$$J = J_{el} + J_{pl}$$

where the elastic portion of the fracture-resistance energy (J_{el}) is:

$$J_{el} = \frac{K_I^2\left(1 - \nu^2\right)}{E}$$

and the plastic portion of the fracture-resistance energy (J_{pl}) is:

$$J_{pl} = \frac{\eta A_{pl}}{B_N B_0}$$

where 0 is factor based on the type of specimen, A_{pl} is the area under the load against displacement plot, B_N is the specimen thickness, and B_0 is the uncracked portion of the specimen (ligament). These equations apply to specific specimen geometries and load-crack situations as described in the ASTM test standards for fracture toughness testing (Ref 12, 13, 14). These expressions are given only as an example to show the link between *J* and *K* and to further show how the *J*-integral is expressed in two parts: the elastic and the plastic components. If the LEFM toughness is expressed in terms of an energy criterion, it equals K_c^2/E so that the dimension will be energy per unit area.

Due to the manner in which *J* is measured, the fracture energy often appears to increase when the fracture is in progress. For low-toughness materials, this increase is generally small and can then be ignored. For high-toughness material, the increase can be considerable. In this case, the fracture may first be stable, and a further increase in load is necessary to cause an uncontrollable (unstable) fracture. Because of historical development, *J* is usually referred to as the *J*-integral, and it can be expressed in the form of an integral. The *J*-integral is useful for computation of geometry factors, but for fracture mechanic applications, *J* can be derived without the integral.

References cited in this section

11. A. Saxena, *Nonlinear Fracture Mechanics for Engineers,* CRC Press, Boca Raton, FL, 1998

12. "Test Method for J-Integral Characterization of Fracture Toughness," E 1737-96, *Annual Book of ASTM Standards,* ASTM

13. "Standard Test Method for Crack Tip Opening Displacement (CTOD) Fracture Toughness Measurement," E 1290-93, *Annual Book of ASTM Standards,* ASTM

14. “Test Method for JIC, A Measure of Fracture Toughness,” E 813-89, *Annual Book of ASTM Standards,* ASTM

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Applications

The strength of a structural component decreases with crack size, and calculations of residual strength (rather than the original design strength as a function of crack size can be done using LEFM or EPFM. In most cases, the residual strength analysis is performed with LEFM and is based on the stress-intensity factor, K, and the toughness (the plane-strain fracture toughness, K_{Ic}, or the plane-stress fracture toughness, K_c), as discussed in more detail in the article “The Practice of Damage Tolerance Analysis” in *Fatigue and Fracture,* Volume 19 of *ASM Handbook* (Ref 15).

Residual strength analysis on the basis of EPFM and plastic fracture mechanics is discussed in the article “Residual Strength of Metal Structures” in *Fatigue and Fracture,* Volume 19 of *ASM Handbook* (Ref 16). One method is the failure assessment diagram, which represents the whole gamut of fractures from brittle to fully plastic. The failure assessment diagram and the application of J-integral and CTOD data are discussed in more detail in the article “Failure Assessment Diagrams” in this Volume.

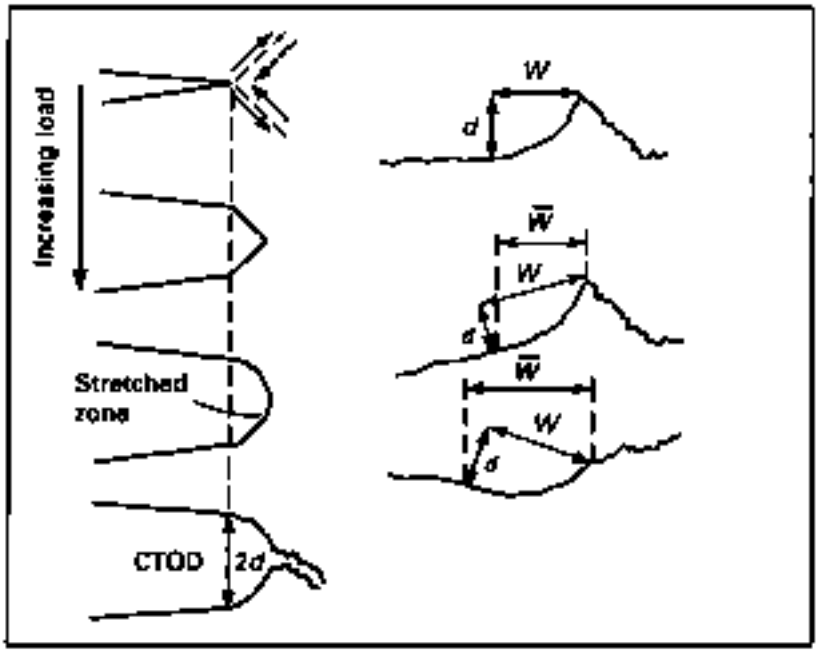
Subcritical fracture mechanics also permits calculation of the crack growth rates. Fracture control can be exercised, but only if both the residual-strength diagram and the crack-growth curve are available. If either is missing, nothing worthwhile can be said about the fracture safety of the structure. For example, if a crack is detected, one may be able to show that it can be sustained under the prevailing service loading. However, because the crack may be much longer the next day, it is not possible to keep the structure in operation unless it is known how long it will take for the crack to grow to a size that cannot be sustained. The crack-growth curve must be known as well.

Use of fracture mechanics in failure analysis can help in the evaluation of conditions leading to fracture and in fact may be critical to the determination of whether an imperfection is a defect (i.e., causative of failure). From the amount of crack growth (crack size at fracture), known stresses, and growth-rate properties, a reasonable insight can always be obtained regarding the question of misuse; for example, continuous overloading. The time to failure and final crack size are determined using fracture mechanics, as discussed in the following paragraphs. When results are not in accordance with observations, the analysis can be repeated to determine how much higher (or lower) the stresses would have had to have been to produce the cracking time and extent of cracking observed.

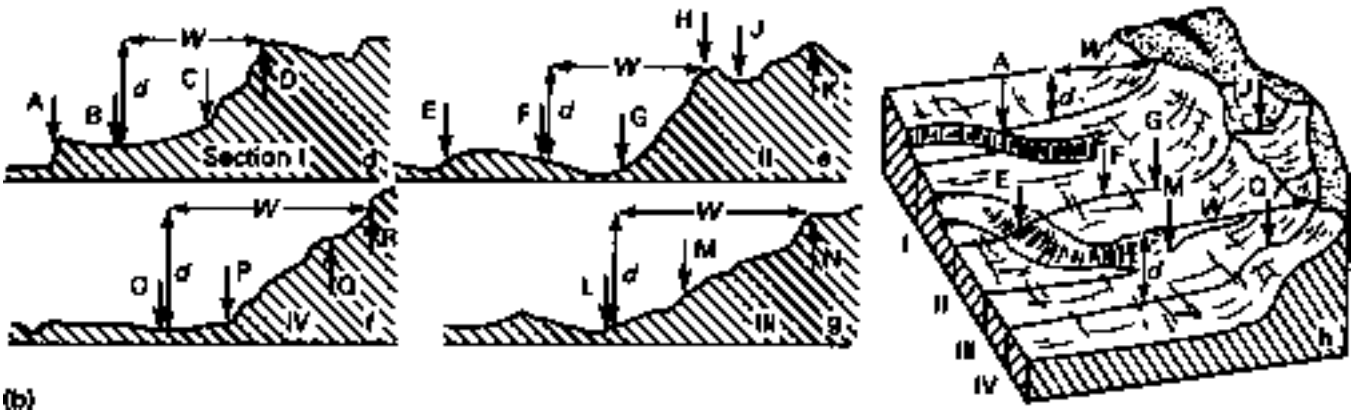
Fractography generally provides the crack size at which final fracture took place. Because fracture occurs in accordance with either LEFM or EPFM conditions, equations are available for using the fractographic results in various ways. If the operation stresses are known, equations can be used to calculate the toughness, which, if compared to the anticipated toughness, will show whether or not the material had a toughness higher or lower than expected. If the material specification indicated a toughness value, the calculation will show whether or not the material was in accordance with the specification. For more discussion, see the appendix in “Fracture Appearance and Mechanisms of Deformation and Fracture.”

Conversely, if the fracture stress is unknown, the fracture mechanics equations can be used to determine the stress at the time of fracture. In this case, a reasonably accurate (sometimes estimated) toughness value must be used. With the stress at fracture known, it can be determined whether the fracture took place due to an excessively high load or whether in general the operating stress levels were too high.

This information can often be corroborated by measuring the size of the stretched zone from the fractographs, provided stereomicroscopy is used (Fig. 27) (Ref 17). The size of the stretched zone is a measure of the CTOD at the time of fracture. It can be shown that the CTOD at fracture is related to the toughness as CTOD = $K_c^2/E\ _y$ or CTOD = $J/\ _y$.



(a)



(b)

Fig. 27 Measurement of the stretched zone in a fractured specimen. (a) Crack-tip blunting and the stretched zone. (b) Topography of stretched zone. Source: Ref 17

Therefore, the toughness can be calculated from the measured stretched-zone size. Figure 28 (Ref 18) shows that careful measurement can give reliable toughness values. With the toughness known, fracture stresses can be calculated as previously discussed.

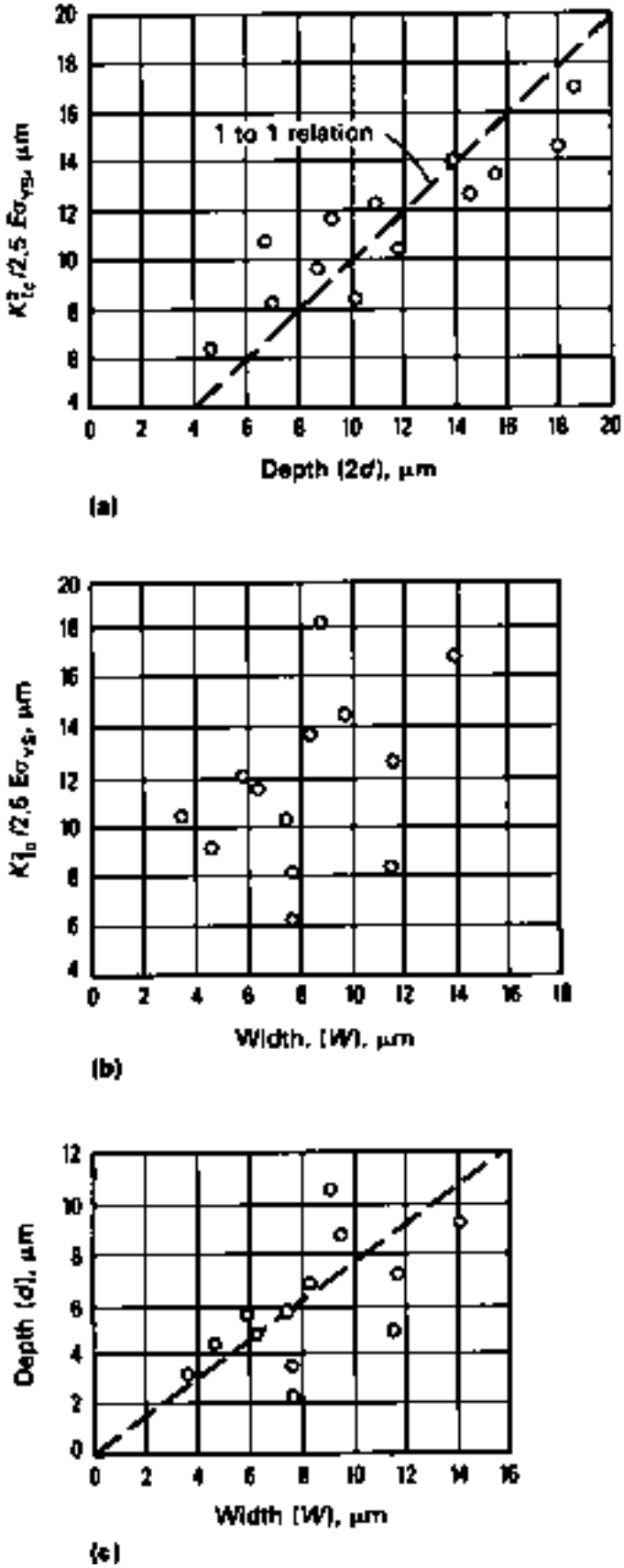


Fig. 28 Correlation between crack-tip opening displacement (CTOD) and toughness. (a) Stretched-zone depth versus CTOD. (b) Stretched-zone width versus CTOD. (c) Stretched-zone width versus depth. Source: Ref 18

Crack Growth. Fatigue-striation counts (if striations are present and if they can be accurately measured) generally provide a reasonable account of the crack-growth rates and crack-growth curve. If the crack-growth-rate behavior of the material and the stresses are known, the stress intensity can be calculated and a comparison can be made between actual and

anticipated properties for a conclusion about the adequacy of the material. Conversely, if the stresses are not known, the measured rates and rate properties can be used to estimate what the acting stresses were. This procedure at least provides the magnitude of the stresses.

Any change of loading or change in environment during the cracking process is likely to leave its mark. Such changes produce a change in crack-growth rate, which is usually associated with a change in topography (macro-scale surface roughness and at micro-scale a discontinuous change in striation spacing). Because the roughness affects the reflection of light, a change in roughness will appear as a line (*beach mark*). Any beach mark is an indication of a change in circumstances during cracking. Beach marks are also created by corrosion at the crack tip when the material is in the unloaded condition for a period of time. Therefore, the presence of a beach mark may only indicate that the material was unloaded for some period of time, not that the stress intensity range changed. If inclusions or second phase particles are present in the microstructure, the striation spacing is also changed as the crack front interacts with these particles. The striation spacing thus is locally affected by the particle, not by a change in stress intensity range. The beach marks on fatigue-crack surfaces are well known. Similar marks may occur on a stress-corrosion crack surface due to changes in loading or environment.

Initial Defects. If fractography shows the presence of initial defects (mechanical or metallurgical), crack-growth analysis can be used, starting from the initial defect. Again, depending on the objective, information can be obtained on stresses (known material properties) or material properties (known stresses). One can thus obtain information on whether the design stresses were too high or whether the material was inadequate.

Residual, Contact, and Thermal Stresses. In general, the stress intensities due to different stress systems are additive as long as all stress systems are of the same mode, for example, mode I. The crack tip stress field due to combined bending and tension is the sum of those due to tension and bending. However, the effective stress intensity factor is then given by:

$$K_{eff}^2 = K_I^2 + K_L^2$$

Residual stresses due to welding, cold deformation, surface treatments, or heat treatments give rise to a stress intensity. If this stress intensity is of the same mode as that due to external loads, the two are additive. Although this is very useful in principle, the stress intensity due to residual stresses usually cannot be determined simply because the residual stresses are unknown.

For a fracture problem, it may be assumed that local yielding at the crack tip is so extensive that residual stresses will be annihilated and can be ignored for the fracture analysis. In the case of fatigue-crack growth, the residual stress is static as opposed to the cyclic applied stress. As such, the effect of the residual stress may be interpreted as a simple change in *R*-ratio. It can be accounted for if the magnitude of the residual stress is known. In the case of sustained-load stress corrosion, the stress intensities are additive. Again, the residual stress can be accounted for if its magnitude is known.

Residual stresses occur only when plastic deformation occurs very locally. On release of elastic strains of the surrounding material, the plastically (permanently) deformed region does not fit in the surroundings. The surrounding elastic material will force it to fit and exert stresses to do so. These stresses are called residual stresses because they occur without the presence of an external load. The surrounding elastic material will essentially return to zero strain, which forces the permanently deformed material nearly back to zero strain. Just as the original plastic strain required yielding, the return to zero strain will require reverse yielding. This leads to the conclusion that residual stresses always equal the yield stress. However, if the material is at elevated temperature where the yield is lower, the residual stress may be equal to the lower yield stress only.

These guidelines can be used to evaluate the effect of residual stress on *K*. It should be noted, however, that *K* is an elastic parameter, and if the residual stress equals the yield, the applicability of *K* is doubtful. Accounting for residual stresses in EPFM is equally debatable. However, application of this rule can at least lead to an estimate of the effect. Cautious engineering judgment is certainly required. The same arguments hold more or less for thermal stresses and contact stresses. Stress intensities are additive, but the problem is in knowing what to add.

Example 3: The Application of Fracture Mechanics to Fractography (Ref 6). In this example, the material is quenched-and-tempered 4340 steel with a yield strength ($_{ys}$) of 160 ksi and a fracture toughness (K_{Ic}) of 80 ksi$\sqrt{in}$ On a failed component made from this material, fractography is used to determine values for critical flaw size (a_c) and 2*c* (as defined in Fig. 21a) for surface cracks. For a surface flaw, the fracture mechanics relationships are:

$$K_{Ic} = 1.1\sigma_f\sqrt{\pi a_c/Q}$$

$$\sigma_f = K_{Ic}/1.1\sqrt{\pi a_c/Q}$$

If a_c = 0.35 in., $a/2c$ = 0.35/0.7 = 0.5, and Q = 2.2, then:

$$\sigma_f = 80/1.1\sqrt{\pi 0.35/2.2} \qquad \sigma_f = 103 \text{ ksi}$$

In this case, fracture stress is as would be expected for a correctly designed structure (that is, the safety factor (SF) is 1.5). Failure occurred because of extensive service-induced subcritical crack growth.

If a_c = 0.17 in., $a/2c$ = 0.17/0.34 = 0.5, and Q = 2.2, then $_f$ = 148 ksi. In this case, fracture stress is well over the expected design stress—thus, the component was overstressed. Failure occurred either because of a design error or an operating error.

If a_c = 0.17 in., $a/2c$ = 0.17/0.34 = 0.5; Q = 2.2; and one knows that $_f$ can be no greater than 100 ksi, then $K_{Ic} = 54 \text{ ksi}\sqrt{\text{in.}}$ In this case, the fracture toughness of the material is well below the expected value of 80 ksi $\sqrt{in}$ Failure occurred because the wrong material was used or the material was processed incorrectly.

If a_c = 0.17 in.; $a/2c$ = 0.17/1.7 = 0.1; and Q = 1.05, then $_f$ = 102 ksi. In this case, fracture stress is well below the expected design stress. The configuration of the crack (large width, 2c, to its depth, a) suggests a failure due to a machining mark, surface scratch, or impact ding.

Note that these sample calculations are simple and are intended to show several different situations. However, the material is steel, and temperature is another important factor. There is a possibility that temperature at the time of failure was less than room temperature, and high strain rates could shift the DBTT. These calculations assume room temperature conditions and no strain rate effects on DBTT.

References cited in this section

6. H. Nelson, Fractography in the Failure Analysis of Engineering Structures, seminar presented at ASM Annual Meeting, 10 Sept 1997

15. D. Broek, The Practice of Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 420–426

16. D. Broek, Residual Strength of Metal Structures, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 427–433

17. D. Broek, Failure Analysis and Fracture Mechanics, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 56

18. D. Broek, Correlation Between Fracture Toughness and Stretched Zone Size, *International Congress on Fracture,* Vol III, 1973, p 422

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Acknowledgments

Portions of this article were also adapted from Ref 5, 6, and 17.

References cited in this section

5. S. Antolovich and B. Antolovich, An Introduction to Fracture Mechanics, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 371–380

6. H. Nelson, Fractography in the Failure Analysis of Engineering Structures, seminar presented at ASM Annual Meeting, 10 Sept 1997

17. D. Broek, Failure Analysis and Fracture Mechanics, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 56

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Appendix: Stress Tensor Decomposition and Transformation of Stress

Stress Tensor Decomposition

Any stress state can be decomposed into spherical and deviatoric tensors. In many engineering materials, while the spherical stress tensor is associated with volumetric change only, the deviatoric stress tensor (sometimes called the stress deviator) is associated with a change in shape. In these materials, plastic deformation is associated with a change of shape. Yield, therefore, occurs incompressibly, implying that Poisson's ratio during plastic deformation, ($_p$, is very nearly 0.5.

Decomposition of the stress tensor is straightforward and uncomplicated. If **T** represents the stress state at a given point, then **T** = $\mathbf{T}_h$ + $\mathbf{T}_d$, where $\mathbf{T}_h$ and $\mathbf{T}_d$ are the hydrostatic and deviatoric stress tensors, respectively. Mathematically:

$$\mathbf{T} = \begin{bmatrix} \sigma_h & 0 & 0 \\ 0 & \sigma_h & 0 \\ 0 & 0 & \sigma_h \end{bmatrix} + \begin{bmatrix} \sigma_x - \sigma_h & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \sigma_y - \sigma_h & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \sigma_z - \sigma_h \end{bmatrix}$$

in which

$$\sigma_h = \frac{\sigma_x + \sigma_y + \sigma_z}{3}$$

A yield criterion for ductile metals whose failure envelope is quite close to Tresca's is the distortion energy theory. Also known as the von Mises-Hencky theory, it can be shown that the von Mises effective stress, $_e$, is a function only of the stress deviator. Thus, spherical stresses play nearly no role in plastic deformation of these materials. A consequence of this is that if the normal stresses, $_x$, $_y$, and $_z$ are equal or nearly so, the spherical stress is relatively large and the effective stress is relatively small, reducing the tendency to yield.

Transformation of Stress

Ordered principal stresses, denoted as $_1$, $_2$, and $_3$, are the eigenvalues of the stress tensor, while the associated eigenvectors, $\hat{n}_1$, $\hat{n}_2$, $\hat{n}_3$, are unit normals to the planes on which the principal stresses act. Several commercially available software packages such as MathCAD have routines that easily extract both eigenvalues and eigenvectors. Other embedded capabilities include both numerical and symbolic manipulation of mathematical objects such as matrices, integrals, and differentials in addition to nonlinear equation solvers.

Maximum shear stresses (both positive and negative) and the planes on which they act can be easily determined once the principal stresses and directions are known. Because maximum shear stresses are oriented on planes rotated 45° with respect to the planes on which both $_1$ and $_3$ act, unit normals to these maximum shear stress planes $\hat{n}^1_s$ and $\hat{n}^2_s$ are found to be:

$$\hat{\mathbf{n}}_i^{(1)} = \frac{1}{\sqrt{2}}\left(\hat{\mathbf{n}}_1 + \hat{\mathbf{n}}_3\right)$$

$$\hat{\mathbf{n}}_i^{(2)} = \frac{1}{\sqrt{2}}\left(\hat{\mathbf{n}}_1 - \hat{\mathbf{n}}_3\right)$$

As many fractures originate on these planes this information has both predictive and diagnostic value. In the design stage possible failure sites and orientations can be predicted. In the event of failure, stress analysis can point to improper use or overload as contributing factors.

As a numerical example to determine principal stresses, maximum shear stresses, and the planes on which they act, consider the following stress tensor, **T**:

$$\mathbf{T} = \begin{bmatrix} 10 & 4 & 8 \\ 4 & 20 & 12 \\ 8 & 12 & -5 \end{bmatrix} \text{ksi}$$

The principal stresses and associated directions are found, using MathCAD, to be:

$$\sigma_1 = 27.531 \text{ ksi} \qquad \mathbf{n}_1 = \begin{pmatrix} -0.3737 \\ -0.8366 \\ -0.4005 \end{pmatrix}$$

$$\sigma_2 = 9.026 \text{ ksi} \qquad \hat{\mathbf{n}}_2 = \begin{pmatrix} 0.8844 \\ -0.4515 \\ 0.1181 \end{pmatrix}$$

$$\sigma_3 = -11.557 \text{ ksi} \qquad \hat{\mathbf{n}}_3 = \begin{pmatrix} 0.2797 \\ 0.3101 \\ -0.9086 \end{pmatrix}$$

Note that to maintain a right-handed coordinate system, $\hat{n}_1 \times \hat{n}_2 = \hat{n}_3$. Hence, normals to the planes of maximum shear stress are given by:

$$\hat{\mathbf{n}}_i^{(1)} = \begin{pmatrix} -0.066 \\ -0.372 \\ -0.926 \end{pmatrix} \text{ and } \hat{\mathbf{n}}_i^{(2)} = \begin{pmatrix} -0.462 \\ -0.811 \\ 0.359 \end{pmatrix}$$

Transformation of a three-dimensional (3D) stress state from one Cartesian coordinate system to another Cartesian coordinate system that has been rotated with respect to the first is governed by the following:

$$\mathbf{T\$} = \mathbf{ATA}^{\mathrm{T}}$$

in which **T** represents the original stress tensor and **T$**represents the stress state in the primed system. **A** is the matrix of direction cosines, and $\mathbf{A}^{\mathrm{T}}$ is its transpose. Each element of **A**, denoted a_{ij}, is the cosine of the angle between the *i*th axis of the primed system and the *j*th axis of the original, or unprimed system, $i, j = x, y, z$. Thus, if stresses are wanted in a coordinate system rotated 30° counterclockwise about the *y*-axis of a right-handed coordinate system, the matrix of direction cosines becomes:

$$\mathbf{A} = \begin{bmatrix} 0.866 & 0 & -0.500 \\ 0 & 1 & 0 \\ 0.500 & 0 & 0.866 \end{bmatrix}$$

The corresponding stress state, **T$**is then:

$$
\mathbf{T}' = \begin{bmatrix} 0.866 & 0 & -0.500 \\ 0 & 1 & 0 \\ 0.500 & 0 & 0.866 \end{bmatrix} \begin{bmatrix} 10 & 4 & 8 \\ 4 & 20 & 12 \\ 8 & 12 & -5 \end{bmatrix}
$$

$$
\times \begin{bmatrix} 0.866 & 0 & 0.500 \\ 0 & 1 & 0 \\ -0.500 & 0 & 0.866 \end{bmatrix}
$$

$$
= \begin{bmatrix} -0.678 & -2.536 & 10.495 \\ -2.536 & 20 & 12.392 \\ 10.495 & 12.392 & 5.678 \end{bmatrix}
$$

If another coordinate transformation (into the double primed system) is necessary, then a second matrix of direction cosines, **B**, is required. If **B** represents the orientation between the double primed system and the primed system, then the transformation equation becomes:

$$
\mathbf{T1} = \mathbf{BATA}^{T}\mathbf{B}^{T}
$$

The transformations **A** and **B** can be combined by defining **C** = **BA**. In that case

$$
\mathbf{T1} = \mathbf{CTC}^{T}
$$

which is in the same form as the single transformation.

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

References

1. J.E. Shigley and C.R. Mischke, *Mechanical Engineering Design,* 5th ed., McGraw-Hill, 1989

2. W.K. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, Inc., 1997

3. W.C. Young, *Roark's Formulas for Stress and Strain,* 6th ed., McGraw-Hill, 1989

4. A. Blake, *Practical Stress Analysis in Engineering Design,* 2nd ed., Marcel Dekker, 1990

5. S. Antolovich and B. Antolovich, An Introduction to Fracture Mechanics, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 371–380

6. H. Nelson, Fractography in the Failure Analysis of Engineering Structures, seminar presented at ASM Annual Meeting, 10 Sept 1997

7. H. Tada, P.C. Paris, and G. Irwin, *The Stress Analysis of Cracks Handbook,* 3rd ed., American Society of Mechanical Engineers, 2000

8. A. Liu, Summary of Stress-Intensity Factors, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 980–1000

9. A. Liu, Chapter 7, Fracture Mechanics for Mixed Crack Tip Displacement Modes, *Structural Life Assessment Methods,* ASM International, 1998

10. A. Liu, Chapter 6, Environmentally Assisted Crack Growth, *Structural Life Assessment Methods,* ASM International, 1998

11. A. Saxena, *Nonlinear Fracture Mechanics for Engineers,* CRC Press, Boca Raton, FL, 1998

12. "Test Method for J-Integral Characterization of Fracture Toughness," E 1737-96, *Annual Book of ASTM Standards,* ASTM

13. "Standard Test Method for Crack Tip Opening Displacement (CTOD) Fracture Toughness Measurement," E 1290-93, *Annual Book of ASTM Standards,* ASTM

14. "Test Method for JIC, A Measure of Fracture Toughness," E 813-89, *Annual Book of ASTM Standards,* ASTM

15. D. Broek, The Practice of Damage Tolerance Analysis, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 420–426

16. D. Broek, Residual Strength of Metal Structures, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 427–433

17. D. Broek, Failure Analysis and Fracture Mechanics, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 56

18. D. Broek, Correlation Between Fracture Toughness and Stretched Zone Size, *International Congress on Fracture,* Vol III, 1973, p 422

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

Selected References

Mechanics of materials

- F.P. Beer, E.R. Johnston, and J.T. Dewolf, *Mechanics of Materials,* 3rd ed., McGraw-Hill, 2001
- J.M. Gere and T.S. Timoshenko, *Mechanics of Materials,* 4th ed., PWS Publishing, 1997
- R.C. Hibbeler, *Mechanics of Materials,* 4th ed., Prentice Hall, Inc., 1999
- A. Higdon, S. Ohlsen, W. Stiles, J. Weese, and W. Riley, *Mechanics of Materials,* 4th ed., John Wiley & Sons, 1985

Elasticity

- I.S. Sokolnikoff, *Mathematical Theory of Elasticity,* 2nd ed., McGraw-Hill, 1956
- S.P. Timoshenko and J.N. Goodier, *Theory of Elasticity,* 3rd ed., McGraw-Hill, 1970

Fracture mechanics and fatigue

- T.L. Anderson, *Fracture Mechanics: Fundamentals and Applications,* 2nd ed., CRC Press, Boca Raton, FL, 1997
- J.M. Barsom and S.T. Rolfe, *Fracture and Fatigue Control in Structures: Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

- D. Broek, *Elementary Engineering Fracture Mechanics,* 4th ed., Martinus Nijhoff Publishers, 1987
- J.A. Collins, *Failure of Materials in Mechanical Design,* 2nd ed., John Wiley & Sons, 1992
- N.E. Dowling, *Mechanical Behavior of Materials,* 2nd ed., Prentice-Hall, 1999
- R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* 4th ed., John Wiley & Sons, 1995
- A. Saxena, *Nonlinear Fracture Mechanics for Engineers,* CRC Press, Boca Raton, FL, 1998

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Introduction

X-RAY DIFFRACTION (XRD) residual stress analysis is an essential tool for failure analysis. X-ray diffraction residual stress measurement methods are commonly applied to the successful completion of failure analyses and, in many cases, are the only viable ones for the acquisition of the required data. X-ray diffraction residual stress analysis methods can be used in (but are not necessarily limited to) situations where failures result from overloading, stress corrosion, fatigue, stress concentrations, or inappropriate manufacturing processes. X-ray diffraction residual stress analysis methods can also be applied to the evaluation of corrective measures and the optimization of production parameters for extending the service life of components.

This article focuses primarily on what the analyst should know about applying XRD residual stress measurement techniques to failure analysis. Discussions are extended to the description of ways in which XRD can be applied to the characterization of residual stresses in a component or assembly, and to the subsequent evaluation of corrective actions that alter the residual stress state of a component for the purposes of preventing, minimizing, or eradicating the contribution of residual stress to premature failures.

The article presents a practical approach to sample selection and specimen preparation, measurement location selection, and measurement depth selection; measurement validation is outlined as well. A number of case studies and examples are cited. Failures with root causes other than residual stress are beyond the scope of this article and are not discussed. The article also briefly summarizes the theory of XRD analysis and describes recent advances in equipment capability. Comprehensive, detailed treatments of XRD residual-stress measurement theory can be found in Ref 1, 2, 3, 4, 5, 6, 7, 8.

X-ray diffraction techniques have been applied as early as 1925 to the measurement of residual stress on a variety of materials (Ref 9). Considerable advances have since been made in detector speed and resolution (Ref 10, 11), data analysis and handling (due primarily to the advent of the personal computer), and equipment portability (Ref 12). The importance of the XRD method resides in its ability to measure residual and applied stress with high spatial resolution, speed, and excellent accuracy, and, in many cases, measurements can be performed nondestructively (Ref 13). The measurement of residual stress via XRD is generally limited to polycrystalline materials (Ref 6), that is, in materials with a grain structure (long-range ordering) as normally found in metals and their alloys as well as in polycrystalline ceramics (i.e., oxides, carbides, nitrides, etc.). The size of specimens that can be evaluated using XRD can vary widely from as small as the head of a pin to as large as a ship (Ref 12, 14), a building/structure (Ref 15, 16), or a bridge (Ref 17, 18). Line of sight for both the incident beam and the detector to the measurement location is required, and the incident x-ray beam must irradiate the volume of interest (Ref 8). The precision and accuracy of results obtained is generally a function of the instrument used, the material condition, the measurement technique used, and subsequent data analysis. Unless otherwise specified, data analysis is based on the assumption that a near-random grain orientation distribution is sampled in a homogeneous and isotropic material (Ref 3). Although a detailed discussion is

beyond the scope of this article, XRD techniques can also be used for the measurement of percent retained austenite (Ref 19), and other phases, pole figures (Ref 7), and other analyses too numerous to mention.

References cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods,* Elsevier, 1997

3. M.E. Hilley et al., "Residual Stress Measurement by X-Ray Diffraction" J784a, Society of Automotive Engineers, 1971

4. P.S. Prevey, X-Ray Diffraction Residual Stress Techniques, *Materials Characterization,* Vol 10, *ASM Handbook,* American Society for Metals, 1986, p 380–392

5. J. Lu, et al., *Handbook of Measurement of Residual Stress,* Fairmont Press, 1996

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* 2nd ed., Wiley-Interscience, 1974

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000

9. H.H. Lester, and R.H. Aborn, *Army Ordnance,* Vol 6, 1925–1926, p 129, 200, 283 and 364

10. C.O. Ruud, P.S. DiMascio, and D.J. Snoha, A Miniature Instrument for Residual Stress Measurement, *Advances in X-Ray Analysis,* Vol 27, Plenum, 1984

11. H. Berger, Sensing Stress, *Research/Penn State,* Vol 10 (No. 2), 1989

12. M.E. Brauss, G.V. Gorveatte, and J.F. Porter, Development of a Miniature X-Ray Diffraction Based Stress Analysis System Suitable for Use on Marine Structures, *Nondestructive Evaluation of Materials and Composites—SPIE,* Vol 2944, 1996

13. J.A. Pineault, and M.E. Brauss, Stress Mapping—A New Way of Tackling the Characterization of Residual Stress, *Exp. Tech.,* March 1995

14. J.A. Pineault, M.E. Brauss, and J.F. Porter, Characterization of Plastic Strain in HY-80 Using X-Ray Diffraction Techniques, *Proceedings of the Conference on Naval Applications of Material Technology,* (Halifax), D.R.E.A., 1993

15. M.G. Carfagno et al., *X-Ray Diffraction Measurement of Stresses in Post-Tensioning Tendons: Extending the Lifespan of Structures,* IABSE, Zurich, Switzerland, 1995

16. M.E. Brauss and J.A. Pineault, Residential Strain Measurements of Steel Structures, *NDE for the Energy Industry,* NDE-Vol 13, American Society of Mechanical Engineers, 1995

17. M.E. Brauss et al., "Deadload Stress Measurements on Brooklyn Bridge Wrought Iron Eye Bars and Truss Sections Using X-Ray Diffraction Techniques: Proceedings of the 14th Annual Meeting of the International Bridge Conference," Engineers' Society of Western PA, 1997

18. M.E. Brauss, J.A. Pineault, M. Belassel, and S.I. Teodoropol, Nondestructive, Quantitative Stress Characterization of Wire Rope and Steel Cables, *Proceedings of SPIE, Structural Materials Technology,* Vol 3400, Society of Photo-Optical Instrumentation Engineers, 1998

19. C.F. Jatczak et al., "Retained Austenite and Its Measurement by X-Ray Diffraction," SP-453, Society of Automotive Engineers, 1980

Residual Stress in Failures and XRD Analysis

The residual stresses present in a component can arise in almost every step of processing (Ref 1) and, in many cases, can play a key role in determining the effective service life of a component. When the failure analyst suspects residual stress as a contributing factor to a premature failure, such suspicions may be validated by experiment and by measurement. X-ray diffraction residual stress measurement analysis should be specified when damaged components:

- Experience static in service loads that are a significant fraction of the maximum allowable equivalent stress (stress concentrations, overloaded)
- Distort or form cracks without applied loads
- Are placed in corrosive environments (stress corrosion)
- Experience cyclic loading in service (mechanical or thermal fatigue)
- May have been subject to improper processing on manufacture (shot peening, grinding, milling, etc.)
- May have been subject to inappropriate heat treatment (stress relief, induction hardening, service temperature, thermal strains, etc.)

Reference cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

X-Ray Diffraction Theory and Residual Stress Measurement

The XRD technique uses the distance between crystallographic planes, that is, *d*-spacing, as a strain gage. This method can only be applied to crystalline, polycrystalline, and semicrystalline materials (Ref 1, 20), where thousands of grains are sampled in a typical measurement. When the material is in tension, the *d*-spacing increases in the direction of stress, and when the material is in compression, the *d*-spacing decreases. The presence of residual stresses in the material produces a shift in the XRD peak angular position that is directly measured by the detector (Ref 6). For a known x-ray wavelength, 2, and *n* equal to unity, the diffraction angle, 2#, is measured experimentally, and the *d*-spacing is then calculated using Bragg's law:

$$n2 = 2d \sin\#$$ (Eq 1)

Once the *d*-spacing is measured for unstressed (d_0) and stressed (d) conditions, the strain is calculated using the following relationship:

$$\&= (d - d_0)/d_0$$ (Eq 2)

When applying a plane-stress model, the unstressed lattice spacing, d_0, can be substituted with the *d*-spacing measured for the specimen of interest at 3 (psi) = 0; that is, the unstressed lattice spacing need not be known precisely for the material in question (Ref 3). To evaluate stresses in the direction normal to the specimen surface $_{33}$, the unstressed lattice spacing for the material in question must be known with an accuracy of better than 0.01%. An error in the estimation of the unstressed lattice spacing, d_0, as small as 0.1% can lead to large errors in the stress measurement results (Ref 1, 21).

For the $\sin^2 3$ method, where a number of *d*-spacings are measured, stresses are calculated from an equation derived from Hooke's law for isotropic, homogeneous, fine-grain materials:

$$\begin{aligned}\varepsilon_{\phi\psi} = {} & \tfrac{1}{2} S_2 \left(\sigma_{11}\cos^2\phi + \sigma_{12}\sin 2\phi + \sigma_{22}\sin^2\phi - \sigma_{33}\right) \\ & \times \sin^2\psi + \tfrac{1}{2} S_2 \sigma_{33} - S_1\left(\sigma_{11} + \sigma_{22} + \sigma_{33}\right) \\ & + \tfrac{1}{2} S_2 \left(\sigma_{13}\cos\phi + \sigma_{23}\sin\phi\right)\sin 2\psi\end{aligned}$$ (Eq 3)

where $\frac{1}{2}S_2$ and S_1 are the x-ray elastic constants of the material, variations of $_{ij}$ are the stress-tensor components, 4 is the angle of the direction of strain measurement with respect to the frame of reference, 3 is the angle subtended by the bisector of the incident and diffracted beam, with the specimen normal, and $\&_{43}$ is the strain for a given 43 orientation. Figure 1 is a schematic showing the reference axes and direction of measurement.

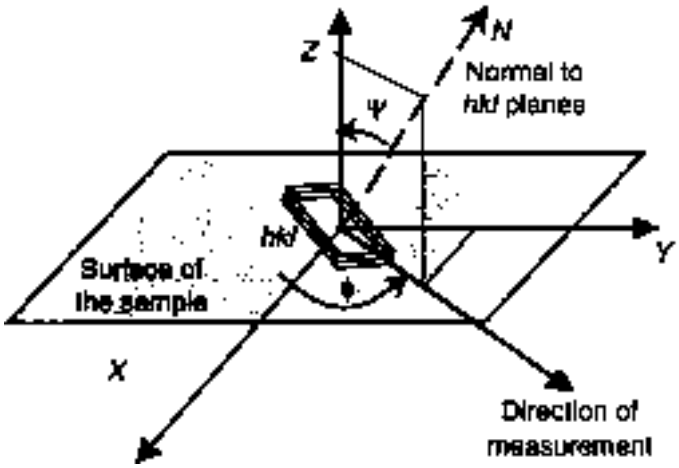


Fig. 1 Definition of the reference axes and the direction of measurement in XRD residual-stress analysis

Evaluation of the stress-tensor components, $_{ij}$, is generally straightforward and is normally carried out by plotting the measured *d*-spacing versus $\sin^2 3$, with careful selection of the measurement directions, 3 and 4. A variety of mathematical models and measurement approaches have been proposed to evaluate the stress-tensor components of interest (Ref 1, 2, 3, 4, 5, 6, 7, 8).

References cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation*, Springer-Verlag, 1987

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods*, Elsevier, 1997

3. M.E. Hilley et al., "Residual Stress Measurement by X-Ray Diffraction" J784a, Society of Automotive Engineers, 1971

4. P.S. Prevey, X-Ray Diffraction Residual Stress Techniques, *Materials Characterization*, Vol 10, *ASM Handbook*, American Society for Metals, 1986, p 380–392

5. J. Lu, et al., *Handbook of Measurement of Residual Stress,* Fairmont Press, 1996

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* 2nd ed., Wiley-Interscience, 1974

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000

20. C.O. Ruud, X-Ray Diffraction Methods for Process Monitoring and Quality Control, *Topics on Nondestructive Evaluation Series,* Vol 1, *Sensing for Materials Characterization Processing and Manufacturing,* ASNT, 1998

21. C.O. Ruud and K.J. Kozaczek, Errors Induced in Triaxial Stress Tensor Calculations Using Incorrect Lattice Parameters, *Proceedings of the 1994 SEM Spring Conference,* Society for Experimental Mechanics, June 1994

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Analysis of XRD Data

Residual-stress measurement *d* versus $\sin^2 3$ data can normally be categorized as follows: linear, elliptical with 3-splitting, and nonlinear with oscillatory behavior (Fig. 2). Data obtained from near-isotropic and homogeneous samples with no shear stress are linear and resemble that shown in Fig. 2(a). Elliptical data, shown in Fig. 2(b) and commonly defined as 3-splitting in *d*-spacing versus $\sin^2 3$ plots obtained via XRD, are generally evidence of shear stress and/or instrument misalignment (Ref 1). Data similar to that found in Fig. 2(c) are most often caused by a nonrandom crystallographic orientation of the coherently diffracting domains in the volume sampled by the incident x-ray beam, that is, material condition.

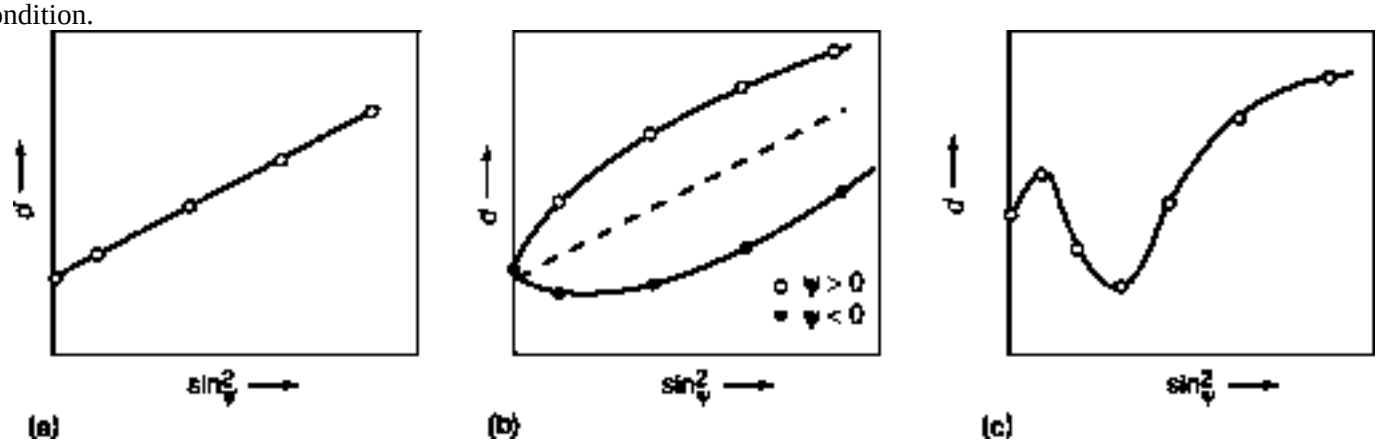


Fig. 2 Common types of *d*-spacing versus $\sin^2$\$ plots. (a) Linear: exhibiting no shear stress. (b) Elliptical: exhibiting \$-splitting due to shear stress. (c) Nonlinear: oscillatory behavior due to preferred crystallographic orientation. Source: Ref 1

Mechanical surface treatments with forces out of plane with the surface create shear stresses (and thus, 3-splitting). Shear stresses can typically be found in samples after grinding, turning, rolling, inclined shot peening, and honing processes. Shear stresses can also be created due to material heterogeneity with nonuniform microstrains. This is often seen in multiphase materials (Ref 2). When 3-splitting is observed, a nonlinear regression should be employed to fit the data points in *d*-spacing versus $\sin^2 3$ plots (Eq. 3). A triaxial analysis can also be employed (Ref 1, 2, 5). These analyses generally require the collection of many data points at both positive and negative 3 angles and in many 4 directions.

The examples shown in Fig. 3 demonstrate that an elliptical fit gives good results (Fig. 3a), whereas a linear fit using either the 3 > 0 or the 3 < 0 branch leads to incorrect results (Fig. 3b). In this case, the results from the two opening branches of the curve differ by approximately 54 ksi.

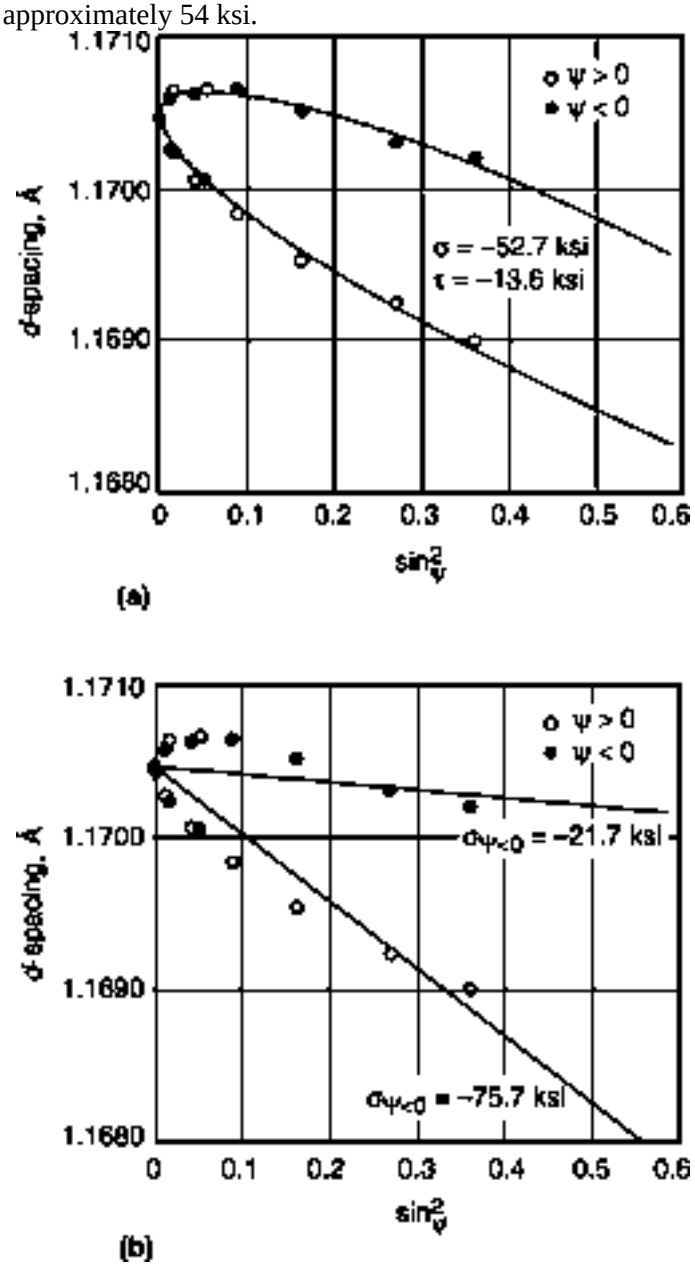


Fig. 3 $-splitting on steel using (a) elliptical fit and (b) linear fit for $ > 0 and $ < 0

With recent advances in equipment speed, the full XRD peak profile can be obtained quickly and easily. When using instrumentation having single-channel detectors, the full diffraction profile should always be collected. Equipment that allows the use of only two 3 tilts (typically, 3 = 0° and 3 = 45°) has limited applications. The use of only a few 3 tilts can lead to erroneous results; thus, the double-exposure method with 3 > 0 or 3 < 0 is only viable in special cases and when it is certain that the *d*-spacing versus $\sin^2 3$ plot is linear. This means that, to ensure accurate results on a wide variety of samples and measurement locations, many 3 tilts using both 3 > 0 and 3 < 0 and elliptical fitting methods should be applied. The use of true 3 or 5 (chi) geometry, sometimes referred to as 6 (omega) and 3, respectively, is recommended. The use of 5 goniometer geometry has advantages, primarily a slightly increased tilt range where the 3 tilt range is limited due to sample geometry constraints. However, when using the modified 5 (or side-inclination) geometry, errors may be introduced due to the 3 angle offset and therefore, should only be used when the specimen geometry eliminates the preferred methods (i.e., true 3 or 5). True-5 and modified-5 geometry are more sensitive to goniometer alignment, particularly with small goniometer radii; thus, care must be taken in sample and goniometer alignment.

References cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods*, Elsevier, 1997

5. J. Lu, et al., *Handbook of Measurement of Residual Stress*, Fairmont Press, 1996

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Instrument Calibration and Validation of Stress Measurements

A comprehensive approach to the measurement of residual stress using XRD should address instrument calibration and the validation of results. The five steps required to calibrate instrumentation and to validate stress measurement results are:

1. Alignment of XRD instrumentation
2. X-ray elastic constant determination
3. Surface condition evaluation
4. Collection parameter selection
5. Repeatability and reproducibility determination

Once all of the previous five steps have been completed successfully, residual-stress measurements via XRD techniques can be applied with a high level of confidence.
Alignment of XRD instrumentation is performed according to the ASTM E 915 standard (Ref 22). This procedure is also performed periodically to verify instrument conformity to specifications. The maximum acceptable stress is ±2 ksi (for steel), tensile or compressive. The verification of instrument alignment is usually performed on a stress-free powder of the material in which residual stress is to be measured. Ideally, the shear stress should also be on the same order of magnitude as the normal stress, that is ±2 ksi, although technically, this would exceed the ASTM E 915 specification. Excellent instrumental alignment can easily be achieved for a goniometer with a focal radius as small as 30 mm (1.2 in.) (Fig. 4).

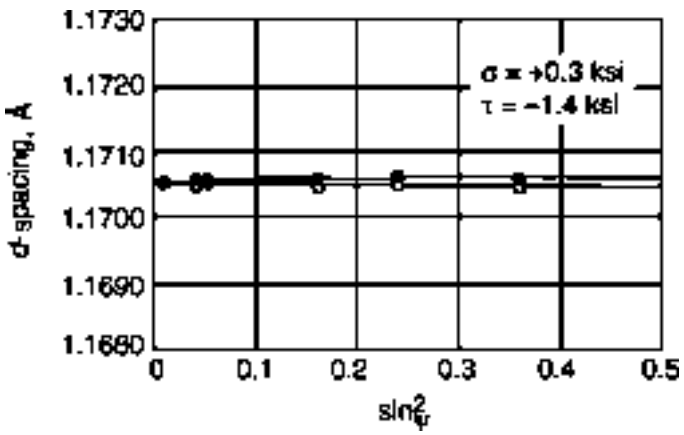


Fig. 4 Elliptical analysis on stress-free iron powder with a 30 mm (1.2 in.) focal radius goniometer

X-Ray Elastic Constant Determination (ASTM E 1426). For stress measurements using XRD techniques, one needs to determine the x-ray elastic constants (XECs). They are a function of the mechanical properties of the material and the crystallographic plane selected to perform the measurements. Once known, they can be used in subsequent measurements on the same alloy but can be affected by preferred orientation. In general, the XECs are determined experimentally or by calculation using theoretical methods. Theoretical models, such as those in Ref 23 and 24, are generally used when a suitable coupon cannot be obtained. The experimental method is usually the preferred method, particularly when the physical parameters of the material structure are not well known. The experimental method uses either a four-point bend or uniaxial tensile specimen, and ASTM E 1426 (Ref 25) is usually applied. In the case of four-point bend testing, a strain gage is attached to the opposite side and/or near the measurement location on the specimen. The applied strains are

recorded via a strain gage, and the stresses are measured by XRD at different load increments. The maximum applied stress is approximately 70% of the yield stress.

For example, the slope of the stress measured via XRD versus the applied stress for a ground steel coupon (Fig. 5) was used to calculate the S_2 parameter in Eq 3. S_1 is calculated for triaxial methods, and its determination is not described by ASTM E 1426 (Ref 25).

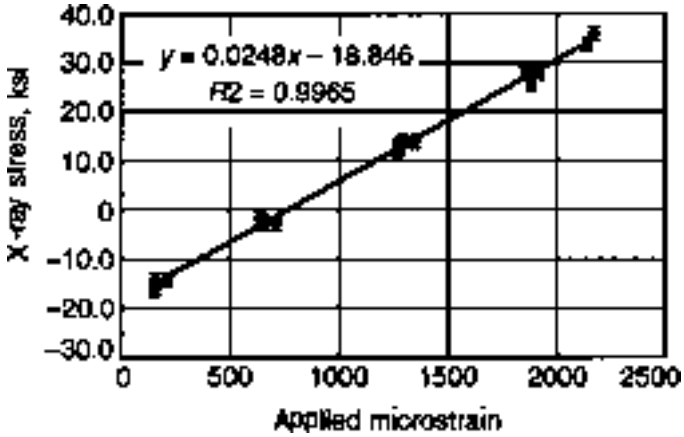


Fig. 5 Results obtained from a steel four-point bend specimen in round robin tests. The XEC (S_2) was determined to be 3.687 10^{-5} ksi^{-1}

Surface Condition Evaluation. The effect of factors such as surface condition (surface roughness, oxide layers, etc.) on XRD residual stress measurement results should be addressed prior to surface-stress measurements on real components.

For example (Ref 26), a steel specimen was machined in three regions, each with different machine tools, to introduce variations in surface roughness on the same specimen. The stress was then measured in the three regions via XRD techniques under independently monitored tensile loads. X-ray diffraction results were plotted versus the applied stress for locations where the average roughness (R_a) was 1, 3, and 6 7m (1 7m = 39.4 7in.). The linear least squares regression for each data set yields slopes of 1.01, 0.96, and 1.03 for 1, 3, and 6 7m R_a, respectively (Fig. 6a). All of these results are well within the experimental error (less than 5%), thus indicating that an R_a of up to 6 7m for these experimental conditions has no measurable effect on residual-stress results obtained via XRD.

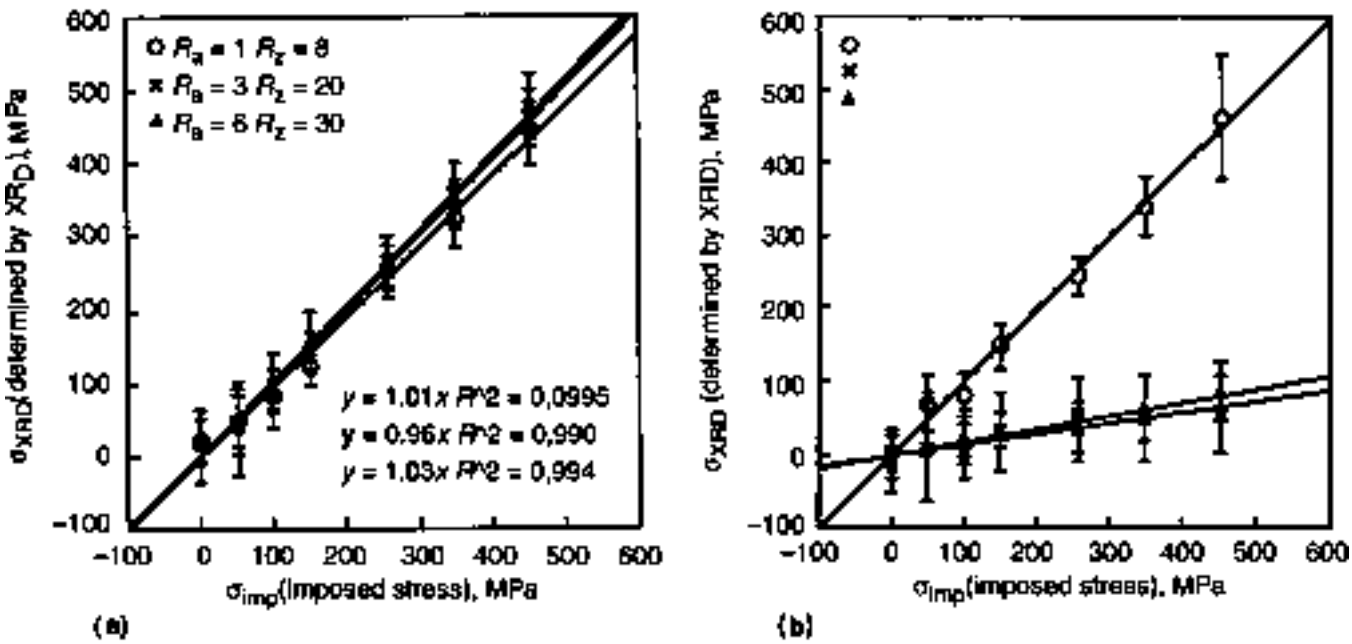


Fig. 6 X-ray diffraction stress versus applied stress for varying average roughness (R_a). (a) Samples with R_a of 1, 3, and 6 %m. (b) Samples with R_a of 1, 40, and 56 %m. (Source: Ref 26)

Another sample was prepared with regions machined to an R_a of 1, 40, and 56 7m. It was loaded in tension, and stresses were measured via XRD. In this case, the slopes were 0.9, 0.15, and 0.18 for 1, 40, and 56 7m R_a, respectively (Fig. 6b).

This demonstrates that the surface-stress measurements do not respond as well to the applied stress when the roughness is increased.

The response of stress measurements via XRD to the applied load for an R_a of up to 56 7m is plotted in Fig. 7. Here, it can be seen that the sensitivity to the applied load is reduced when the R_a is greater than 10 7m (0.0004 in.), in this particular case.

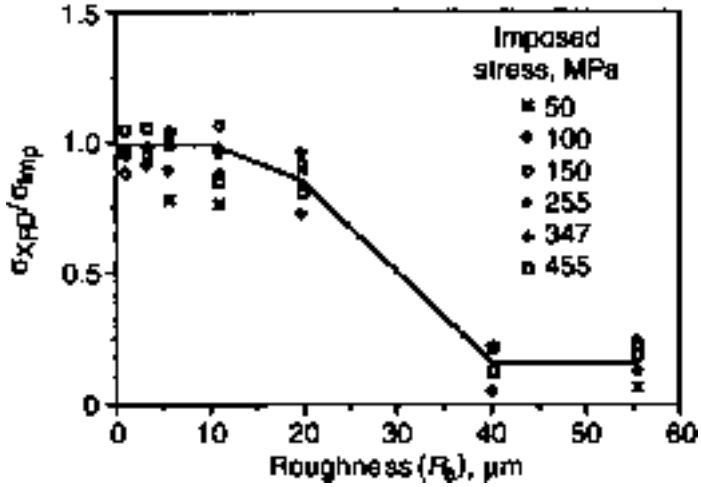


Fig. 7 Ratio of measured stress and applied stress for varying R_a. Source: Ref 26

When the surface R_a is less than the penetration depth of x-rays in the material, the measured stress results more accurately reflect the applied load (Fig. 8a). However, when the surface R_a is greater than the penetration depth of the x-rays, the measured stress reflects the applied load to a lesser degree (Fig. 8b).

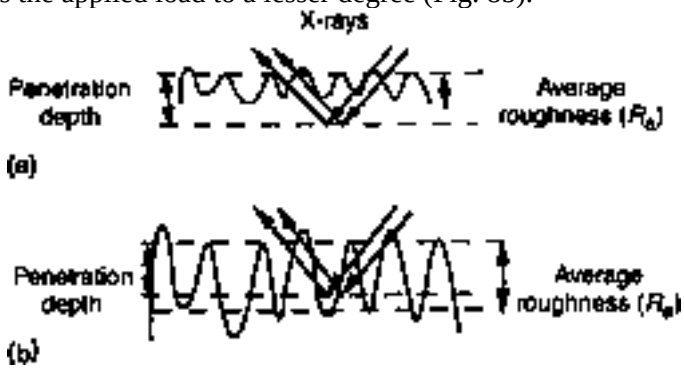


Fig. 8 Effect of surface R_a on XRD stress measurements. (a) X-ray penetration depth is greater than R_a. (b) X-ray penetration depth is less than R_a. Source: Ref 26

The depth of penetration of x-rays in a material is dependent on the wavelength, 2, (i.e., the energy) of the incident radiation and the mass absorption coefficient, 7, of the material (Ref 6). The mean depth of penetration can vary but is typically 10 to 20 7m for most metals.

Effect of Corrosion. The effect of surface corrosion on measurement response can be demonstrated by an example. A strain gage was attached to the surface of a steel wire on the inner side adjacent to the XRD stress-measurement location of interest. The XRD measurement location was visibly corroded. The wire was then incrementally loaded in tension. The stress was measured using XRD techniques, and the applied stress was monitored via the strain gage. The slope of the XRD stress versus applied stress was 0.889. The same steel wire was electropolished, and the experiment was repeated. The slope of the XRD stress versus applied stress was 0.997.

The results plotted in Fig. 9 indicate that, in this particular case, the electropolished surface (Fig. 9b) responds better than the corroded surface (Fig. 9a). Because the near-surface layers are often of interest to the failure analyst, electropolishing techniques and near-surface residual-stress measurements may be required.

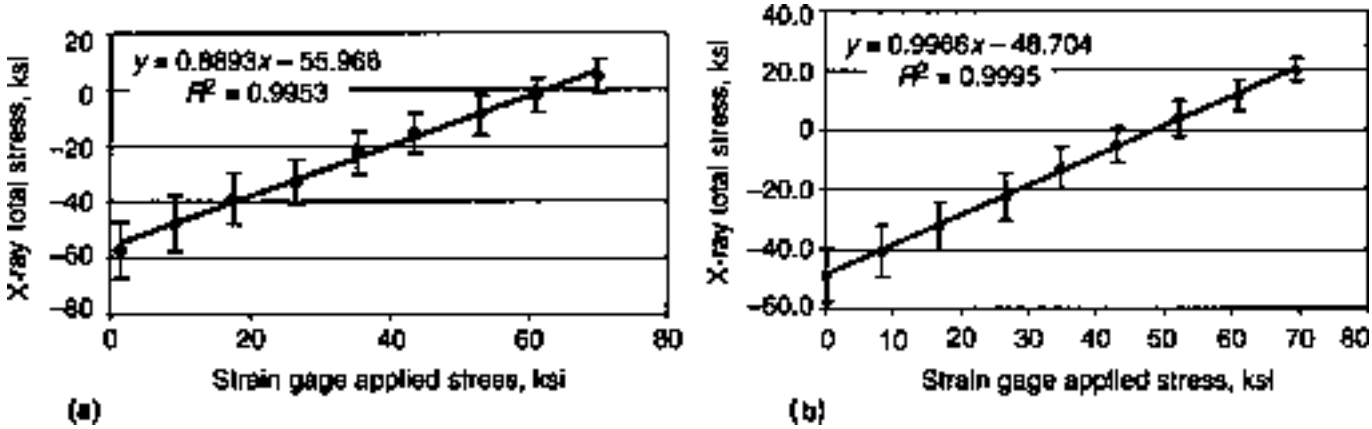


Fig. 9 X-ray diffraction stress versus applied stress on (a) as-received and (b) electropolished surfaces

Collection Parameter Selection. A number of factors can contribute to the introduction of random and/or systematic errors into residual-stress measurements performed via XRD techniques; thus, the data collection parameters should be carefully selected. Components can exhibit different microstructures that require the use of different collection parameters. In general, these collection parameters can be summarized as follows:

- Collection time
- Number of 3 tilts used for *d*-spacing versus $\sin^2 3$ plots
- X-ray diffraction peak position determination
- Effects of microstructure
- Effects of surface curvature and beam size

Optimization of the previous collection parameters should be addressed prior to proceeding with residual-stress measurements.

Collection Time. Low collection times can introduce random errors due to insufficient x-ray counting statistics (Ref 1, 7, 27). Random errors become apparent during repeatability tests; therefore, x-ray data collection times should be increased sufficiently until the desired repeatability is achieved. This technique works well for all detector types, particularly multichannel solid-state detectors that do not resolve individual counts. When using position-sensitive proportional counters, the number of counts can be used as a measure of counting statistics.

Number of 3 Tilts Used for d-Spacing Versus $\sin^2 3$ Plots. Improper selection of the number of 3 angles used can introduce systematic errors (see the section "Analysis of XRD Data" in this article). Systematic errors become apparent during reproducibility tests. In general, a minimum of five 3 tilts are required for a reasonable assessment of the shape of the *d*-spacing versus $\sin^2 3$ curve; however, it is recommended that more than five tilts be used as a general practice.

X-Ray Diffraction Peak Position Determination. Improper selection of the peak location method can also introduce systematic errors in XRD residual stress measurement results. Depending on the specimen under examination, the XRD $K8_1$-$K8_2$ doublet may or may not be resolved. A number of factors can affect how well the doublet is resolved, including material effects such as the dislocation density, the crystallite size in the sample, and instrumental effects such as slit size and defocusing (Ref 1). In practice, that is, in most cases, XRD patterns obtained from real-life components exhibit broadening due to cold-working effects such as machining, grinding, shot peening, and so on. If the doublet is not well resolved, the peak position can be determined by fitting the blended doublet as a whole. The position of the blended doublet can thus be determined via analysis of the full profile (e.g., gravity center methods), and the splitting of $K8_1$ and $K8_2$ lines is quasi-automatically averaged. The application of peak fitting methods based on the analysis of selected parts of the profile should only be used exceptionally when peak asymmetry exists (Ref 2).

If the $K8_1$-$K8_2$ doublet is reasonably well resolved, it is recommended that the Rachinger correction be applied (Ref 7). This method consists of eliminating $K8_2$, assuming that it has the same shape as $K8_1$ and its intensity is half. Once the $K8_2$ portion of the doublet has been eliminated, symmetric functions such as Voigt, Gaussian, Pearson VII, and full-profile methods work well (Ref 2). In contrast to the center-of-gravity method, the Gaussian offers the advantage that it may easily be extended to two or more profiles with their $K8_2$ peaks removed, using the Rachinger correction. Because the intensity ratio and the difference in wavelength, "2 , of the $K8_1$ and $K8_2$ lines is known for all x-ray tube targets, two functions (such as Pearson VII, Gaussian, modified Lorentzian, etc.) with the appropriate angular separation and relative intensity can be used to fit the doublet. In special material conditions or in the case of high-resolution experiments, it is recommended to use Pearson VII or the more physically based Voigt functions (Ref 2). The cross-correlation method is another method for peak location determination; however, this method is not applicable for textured materials (Ref 1).

Effects of Microstructure. All materials, at the single-crystal level, are anisotropic to some degree. If a nonrandom grain orientation distribution is sampled, the effects become apparent in either the shape and/or relative intensity of the diffraction peaks and in the *d*-spacing versus $\sin^2 3$ plots. Typically, large grain size or preferred orientation effects result in a nonlinear or oscillatory *d*-spacing versus $\sin^2 3$ plots (Fig. 2c). In the case of large grain-size effects, the peak shape generally changes drastically from 3-tilt to another 3-tilt, and corrective actions may include increasing the aperture size and/or oscillation of the 3 and/or 4 axes. In the case of preferred orientation, the peak shape generally remains constant with progressive broadening, but the peak intensity varies widely from 3-tilt to another 3-tilt. As a rule of thumb, an intensity ratio greater than 2.0 indicates there is a potential material condition problem. Corrective actions may include linearization (Ref 2) over the highest 3-tilt range possible and others (Ref 1, 28, 29).

Effects of Surface Curvature and Beam Size. For specimen curvature effects, the following rule of thumb should be adhered to: for cylindrical specimens of radius *R*, the maximum incident x-ray spot size to use is *R*/6 for 5% accuracy and *R*/4 for 10% accuracy in the hoop direction, and *R*/2.5 and *R*/2 for 5% and 10% accuracy, respectively, in the axial direction. In cases where the beam size cannot be made sufficiently small, corrections can be applied (Ref 5).

Repeatability and Reproducibility. To assess the effects of random and/or systematic errors in residual stress measurement results, the repeatability and reproducibility of such measurements may be determined. Repeatability is a measure of the consistency in measurement results collected sequentially without a change in experimental setup. Reproducibility is a measure of the consistency in measurement results when the experiment is setup again at a different time, with a different instrument and/or with a different operator.

The residual-stress measurement data found in Table 1 were collected on a laboratory standard. The repeatability test result average in Table 1 has a standard deviation of 0.7 ksi, and the reproducibility test result average has a standard deviation of 1.1 ksi. The slightly higher standard deviation in the reproducibility test versus the repeatability test is due to:

- A slightly different population of grains sampled
- Small variations in sample positioning and alignment
- The nonuniformity in surface residual stress of the specimen

Table 1 Repeatability and reproducibility of stress results

Measurement No.	**Stress, ksi**	
	Repeatability	**Reproducibility**
1	-62.8	-62.9
2	-62.9	-61.1
3	-63.2	-63.0
4	-62.6	-63.7
5	-61.4	-63.8
Average	**-62.6**	**-62.9**
Standard deviation	**0.7**	**1.1**

A similar repeatability and reproducibility test should be performed when a new material, material condition, or type of sample is to be analyzed using XRD.

References cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation*, Springer-Verlag, 1987

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods*, Elsevier, 1997

5. J. Lu, et al., *Handbook of Measurement of Residual Stress*, Fairmont Press, 1996

6. B.D. Cullity, *Elements of X-Ray Diffraction*, 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials*, 2nd ed., Wiley-Interscience, 1974

22. "Standard Test Method for Verifying the Alignment of X-Ray Diffraction Instrumentation for Residual Stress Measurement," E 915-90, ASTM

23. E. Kröner, *Z. Phys.*, Vol 151, 1958, p 504–518

24. E. Kröner, *J. Mech. Phys. Solids,* Vol 15, 1967, p 319–329

25. "Standard Test Method for Determining the Effective Elastic Parameter for X-Ray Diffraction Measurements of Residual Stress," E 1426-91, ASTM

26. V. Li, J.L. Ji, and G.E. Lebrun, "Surface Roughness on Stress Determination by the X-Ray Diffraction Technique," *Exp. Tech.,* Vol 19 (No. 2), March/April 1995, p 9–11

27. A.J.C. Wilson, *Br. J. Appl. Phys.,* Vol 16, 1965, p 665

28. J.A. Pineault and M.E. Brauss, Measuring Residual and Applied Stress Using X-Ray Diffraction on Materials with Preferred Orientation and Large Grain Size, *Advances in X-Ray Analysis,* Vol 36, 1993

29. C.M. Mitchell, Stress Measurement by X-Ray Diffractometry, U.S. Patent 4,561,062, 1985

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Sample Selection

In general, a statistically representative population of specimens should be chosen for analysis, so as to accurately sample the potential scatter within a population. In practice, the number of specimens that are representative of a population is related to how well the manufacturing processes are controlled. Such scatter (or lack of process control) in and of itself can be a contributing factor to a wide variability in the effective service-life-to-failure period for components manufactured with uncontrolled or poorly controlled processes. This can result in a drastically reduced predicted life for a given component, simply to accommodate the shortcomings of a small percentile of the component population.

If a normal distribution is used to predict the residual-stress distribution at a given location among a population of specimens, then a minimum number of five specimens should be chosen to adequately sample the potential distribution within the population. For processes where the standard deviation in the residual stress is less than the experimental error, it makes economic sense to reduce the number of specimens to represent the population. In the authors' experience, budget constraints often limit the sample population to two or three specimens; however, if variations in results at equivalent locations on equivalent specimens are larger than the experimental error, more specimens should be obtained for measurement.

Separate specimen populations should be selected to represent the manufacturing stage(s) of interest, so as to evaluate the effect of a given process on the stress state of the component. This can help identify the manufacturing process or production line that may be introducing an undesirable stress state in the component. Knowledge of the component service history can also help identify how undesirable stress states were introduced in service or how a failure may have occurred. Ideally, samples should be selected from real production populations to more accurately represent the true variance that can be expected within a population. Characterizing a sample population of only one using XRD residual-stress analysis can be risky, because it may inadequately represent the statistical distribution of residual stresses in the component population. Components that have already failed may be of limited use, because residual stresses may have changed or relaxed considerably upon failure. Ideally, nonfractured components with a similar history should be considered for analysis as well.

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Measurement Location Selection and Location Access

When a failure occurs, the failure analyst must define the damage type and the origin of failure. This may be relatively straightforward or difficult, depending on the type of component and/or structure in question. Generally, the point of origin of a crack can be easily located on a given component with standard nondestructive evaluation techniques (i.e., visual inspection, eddy current, fluorescent dye penetrant, ultrasonic, etc.). In more complicated structures, the failure analyst must identify the components and locations most vulnerable to cracking.

Empirical Observation Method for Measurement Location Selection. An empirical guideline for determining the locations where cracks are most likely to occur and thus, where residual-stress measurements should be concentrated, can be summarized as follows:

- The highest-stressed areas are the areas with the smallest thickness or cross section for a given applied load.
- Stress concentration geometry: undercuts, necks, radii, weld toes, fillets, etc.
- Locations of dynamic contact (fatigue contact)
- Locations exposed and susceptible to corrosion

Once a crack has already initiated or failure has already occurred, the location of interest for XRD measurement is in that area. It should be noted that in the vicinity of the crack, the residual stress is in part or completely relieved in the direction normal to the crack line, hence the importance of evaluating parts prior to crack initiation. Triaxial XRD stress measurements are preferable to determine the orientation and the magnitude of the stress field, when it is possible or practical. In cases where the accessibility around the location of interest is limited to one or two directions (either due to sample geometry, economic, or time constraints), the priority is given to stress measurements in the direction perpendicular to the crack line.

Finite-Element Method for Measurement Location Selection Finite-element methods are very useful to predict the stress distribution in a component under applied loads. This method requires:

- Building a two- or three-dimensional (2-D, 3-D) map of the component or structure
- Determining the boundary conditions
- Performing calculations in 2-D or 3-D, depending on the complexity of the loading
- Printing the 2-D or 3-D stress distribution results (components in different directions, plastic strains, equivalent strains and stresses, etc.)

The results identify high- and low-stressed areas and stress gradients. The locations with high priority for residual-stress measurements are the highest-tensile-stressed areas, that is, the areas that are most susceptible to localized yielding or cracking. At the boundaries where the stress gradient changes sign, shear stresses may be present; thus, triaxial stress measurements at these locations are recommended. For components of complex geometry where finite-element methods are not economically practical, the empirical observation method is preferable.

Stress-Mapping Method for Measurement Location Selection The introduction of stress-mapping techniques has allowed the rapid and precise characterization of entire component surfaces, including areas of interest such as steep stress gradients (as found in welds) and their associated tensile residual-stress maxima. In cases where the component geometry is such that stress measurements can be performed in the areas of interest (near the failure), stress mapping can be performed in both loaded and unloaded conditions. Stress-mapping techniques are routinely applied to welds, ground areas, shot-peened areas, rolled areas, and locally heat treated areas with temperature gradients. Subsequent measurements should be concentrated in areas identified by stress maps as the most tensile with higher resolution and/or with depth.

The stress-map display makes available to the failure analyst a complete and accurate visual analysis of the magnitude and distribution of residual stresses in components. The plots shown in Fig. 10, and 11 are examples of stress maps collected on components with surface-stress gradients. In Fig. 10, the tensile stresses detected in the center of the map (in the weld metal and adjacent heat-affected zone) may limit the service life of the saw blade (Ref 30). Similarly, in Fig. 11, the unpeened tensile stress region of a 316L stainless steel butt welded plate may be the location most susceptible to failure (Ref 31).

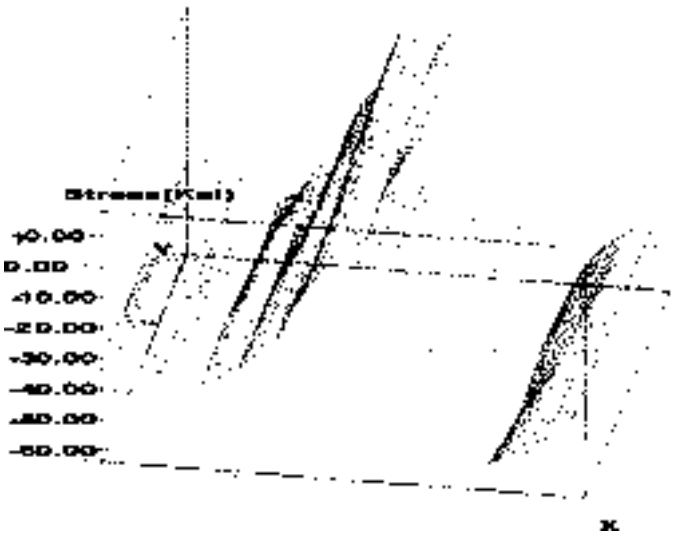


Fig. 10 Surface residual-stress map of resistance welded, heat treated, and ground steel saw blade. Source: Ref 30

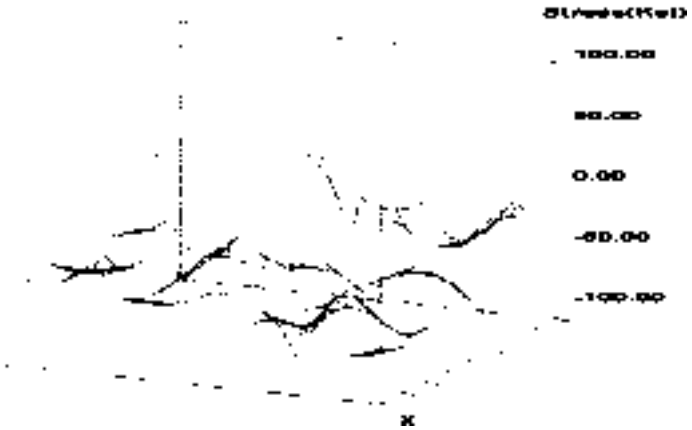


Fig. 11 Residual-stress map of welded 316L stainless steel plate. Source: Ref 31

Accessing Measurement Locations of Interest. Using XRD, stresses in areas as small as a fraction of a millimeter can be measured. Accessing measurement locations on flat surfaces is straightforward but impossible at locations such as the inside diameter of a 2 mm (0.08 in.) hole. X-ray diffraction residual-stress measurements at locations where access is a problem normally require sectioning of the component to access the location of interest and, in general, require the evaluation of applied stresses caused by the relaxation due to sectioning. The most common method to monitor the strain relaxation during sectioning is with the use of electrical resistance strain gages. The placement and orientation of strain-gage elements prior to sectioning requires careful consideration.

For example, when residual-stress measurements are required on the inner diameter of a small ring, and the diameter of the ring is too small for the goniometer to obtain line of sight to the inner-diameter measurement locations, strain gages should thus be placed on both the outer and inner surfaces, so as to measure strains in the direction(s) of interest. Both outer and inner gages should be as close to the stress measurement location as possible. If other directions are to be monitored, dual-element strain gages or rosettes can be used for this purpose. The relaxation is generally assumed to be elastic; however, in highly stressed materials, it can also be elasto-plastic. The strain monitoring should therefore be performed in real time, that is, continuous mode. Care must be taken to ensure that the chosen cutting method does not cold work or overheat the XRD stress measurement location of interest.

The following summarizes the steps for sectioning samples prior to XRD analysis:

- Orient strain-gage element(s) so as to measure strains in the direction parallel to XRD stress measurement direction(s).
- Use coolant with the cutting tool (saw blade, cutoff wheel, EDM wire, etc.) to keep the sample temperature low, thus minimizing localized thermal effects and preventing stress relaxation due to heat dissipation generated during sectioning.
- Always protect the gages from coolant contact (particularly if the coolant is conductive) with epoxy, lacquer, or equivalent coatings (see strain gage manufacturer's recommendations).

- Monitor (in real-time) the variation of the strains during sectioning.
- Look for strain peaks during sectioning, and compare the value to the yield stress of the material, because plastic deformation may occur.
- Use the final strains to calculate the stress relaxation correction due to sectioning. In general, the calculated stress values are subtracted (added with opposite sign).

When complicated sectioning is required, more advanced models may be required and applied (see Ref 8 for more details). Finite-element methods can also be useful in such cases.

References cited in this section

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation*, Vol 8, *ASM Handbook*, ASM International, 2000

30. M. Belassel, M.E. Brauss, and J.A. Pineault, "Residual Stress Characterization Using X-Ray Diffraction Techniques, Applications on Welds," American Society of Mechanical Engineers Conference (Atlanta), 2001

31. J.A. Pineault, M.E. Brauss, and J.S. Eckersley, Residual Stress Characterization of Welds Using X-Ray Diffraction Techniques, *Welding Mechanics and Design*, American Welding Society, 1996

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Selecting Measurement Directions and Depths

Selecting Measurement Directions. X-ray diffraction residual stress measurement directions can be selected as follows:

- For characterization of the complete stress tensor (triaxial analysis), the measurement of stresses in six directions is recommended (a minimum of three are required).
- Initial stresses generated during processing can be a good indication; for example, in the case of rolling, grinding, or turning processes, the directions parallel and perpendicular to the rolling, grinding, or turning direction are of interest.
- In specimens where induced stresses are omnidirectional, one direction or two orthogonal directions may be sufficient.
- If a crack has initiated in any direction, the highest priority measurements are in the direction perpendicular to the crack.
- When measurement directions are constrained by sample geometry, measurements at 0°, 45°, and 90° for any reference frame may be used to calculate principle and shear stresses using Mohr's circle for the biaxial case. Similar methods can be applied for the triaxial case as well.

Selecting Measurement Depths. A comprehensive residual-stress investigation using XRD is seldom limited to the surface, and thus, subsurface measurements are generally required. If subsurface locations must be measured nondestructively, neutron diffraction is recommended (Ref 5, 8). When surface and subsurface measurements are performed, stress gradients normal to the specimen surface can be characterized, thus exposing potentially beneficial or harmful subsurface residual stresses in the material. Using XRD, the residual stresses should be corrected for stress relaxation using the Moore-Evans (Ref 32) method, where the material removed is over the whole surface. The stress correction with this technique requires measurements in one direction for flat surfaces and two directions for cylindrical surfaces (hoop and axial). For more complicated specimen geometry, finite-element techniques can be used to correct for stress relaxation due to material

removal (Ref 3, 33). In general, the correction is within the experimental error of the residual stress measurement (i.e., is negligible) if the deepest layer removed is a small fraction of the total thickness of the component. When steep stress gradients normal to the surface exist in the component, the stress gradient correction should also be applied to collected data (Ref 3).

The number of subsurface residual stress measurements performed is not limited, and measurement depths are generally selected (as required) to view the full shape of the stress versus depth profile to depths where stress gradients tend to level off. The actual depths required to characterize a subsurface stress gradient are very process- and sample-dependent and should be determined on a case-by-case basis.

References cited in this section

3. M.E. Hilley et al., “Residual Stress Measurement by X-Ray Diffraction” J784a, Society of Automotive Engineers, 1971

5. J. Lu, et al., *Handbook of Measurement of Residual Stress,* Fairmont Press, 1996

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000

32. M.G. Moore and W.P. Evans, Mathematical Corrections in Removal Layers in X-Ray Diffraction Residual Stress Analysis, *SAE Trans.,* Vol 66, 1958, p 340–345

33. A. Constantinescu and P. Ballard, On the Reconstruction Formulae of Subsurface Residual Stresses After Matter Removal, *Proceedings of ICRS5* (Linkoping), 1997, p 703–708

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Specimen Preparation

Most mechanical and structural components are protected when used in aggressive or harsh environments to prevent or minimize material degradation mechanisms such as oxidation, wear, erosion, corrosion, and so on. Structures exposed to the elements are often coated with zinc or painted, whereas mechanical components used in a dynamic mode are often lubricated. Protective coatings can partially or completely attenuate the incident x-ray beam and must be removed prior to residual-stress measurement. Removal of coatings must be performed without modifying the surface or subsurface residual-stress state. Mechanical polishing or grinding modifies the residual stress at the surface and the subsurface and is not recommended. If mechanical surface preparation is required for other nondestructive testing techniques (such as eddy current or ultrasonic) to be used in parallel with XRD, the XRD analysis should be performed before the surface is additionally disturbed or cold worked.

It is important that the chemicals used to remove coatings never etch the material. If etching occurs, subsequent electropolishing below the etched layer may be required. To evaluate the appropriateness of the surface-preparation techniques applied and the surface condition of the component to XRD residual-stress measurements, a complete surface-condition evaluation (four-point bend test) must be performed on a coupon taken from the component in the as-prepared condition.

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Residual-Stress Effects on Components under Quasi-Static Loading (Ref 2)

The basis for a quantitative assessment of statically loaded components is dependent on the determination of the maximum allowable equivalent stress, $_e$, which is determined by the underlying effect chosen (i.e., the resulting normal stresses, shear stresses, deformations, etc.), its location, and the defined failure criterion (i.e., fracture, plastic deformation, etc.). Thus:

$$\sigma_e \leq \frac{R}{s} \qquad \text{(Eq 4)}$$

where R is the resistance to failure (and depends on the failure criterion), s is the safety factor, and the allowable equivalent stress, $_e$, is a function of the loading stresses and residual stresses. The sign and magnitude of the residual stress may increase or diminish the equivalent stress, and subsequently, surface and subsurface residual-stress gradients can modify their contributions locally. In cases where plastic deformation occurs in ductile materials, their effect may diminish, because residual stresses can change or relax substantially prior to and on failure and thus may or may not play a significant role in the failure (depending on the failure criterion). The effects of residual stresses are generally more dominant in brittle failures or when plastic deformation is the failure criterion (Ref 2).
Tensile residual stresses can significantly reduce fracture loads (Fig. 12), whereas compressive residual stresses generally act to increase the crack opening thresholds for given loading stresses (Fig. 13).

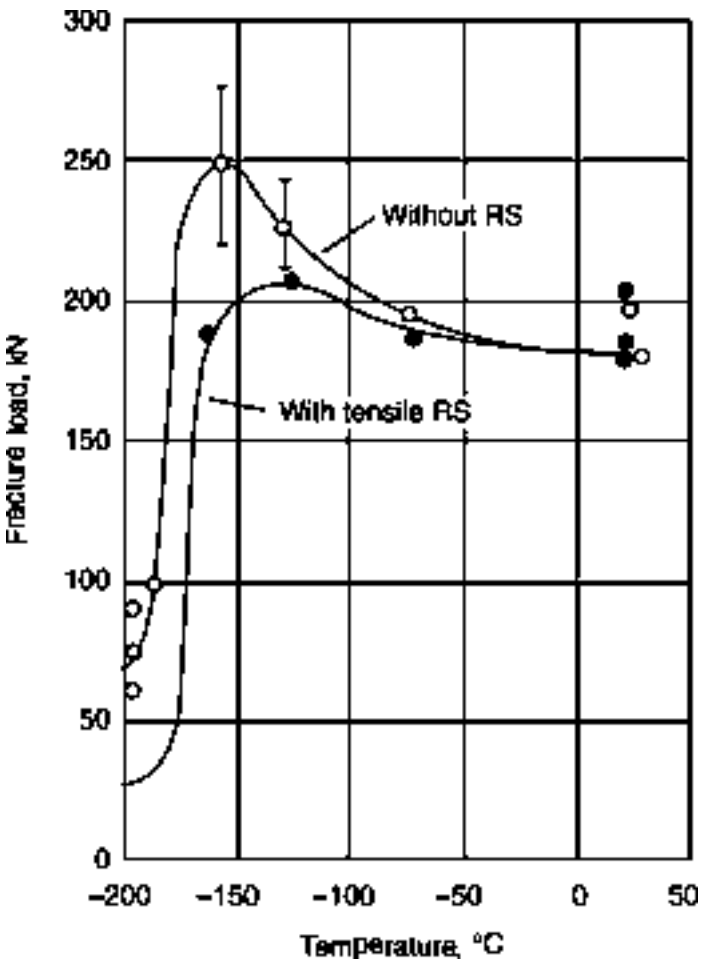


Fig. 12 Effect of tensile residual stress (RS) on fracture loads as a function of test temperature. Source: Ref 34

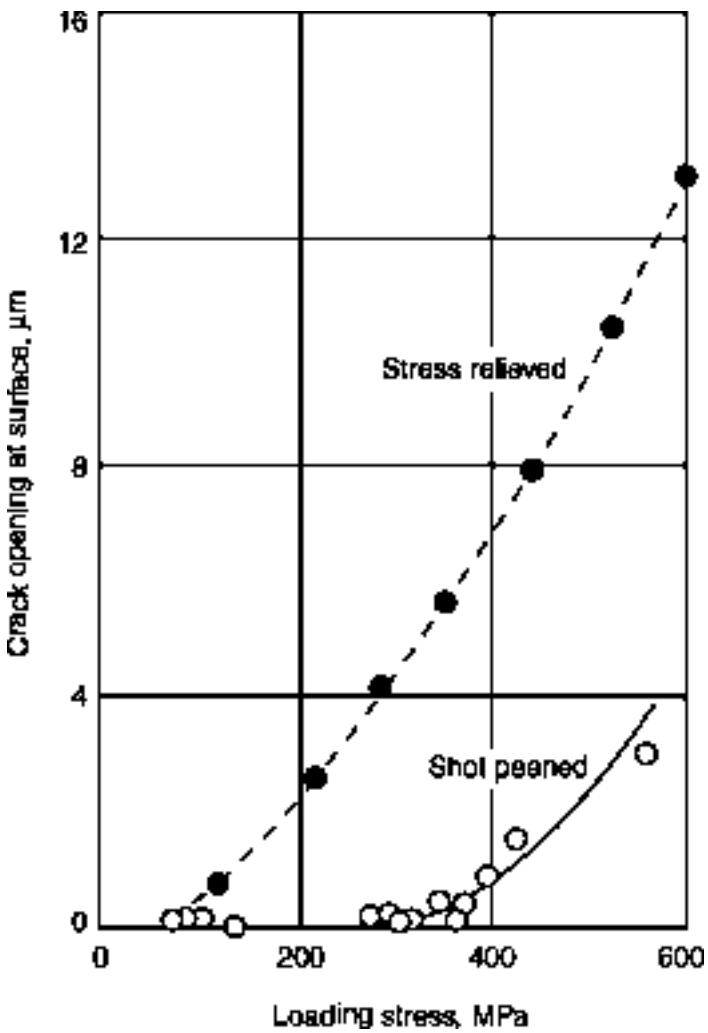


Fig. 13 Crack tip opening of a shot-peened and residual-stress-free Ti-6Al-4V specimen. Source: Ref 35

When applying linear elastic fracture mechanics, the residual-stress contributions to the stress-intensity factor can be calculated and used to predict whether crack growth and/or arrest occurs. It should be noted that the effect of residual stresses on crack growth rates tends to diminish with increasing fracture toughness. Residual stresses can also be taken into account when elastic-plastic fracture mechanics are used (Ref 2).

References cited in this section

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods*, Elsevier, 1997

34. C.L. Formby and J.R. Griffiths, *Proc. of Conf. on Residual Stresses in Welded Constructions and Their Effects*, The Welding Institute, 1987, p 359

35. J.E. Hack and G.R. Leverant, "Residual Stress Effects in Fatigue," STP 776, ASTM, 1982, p 204

Stress-Corrosion Cracking and Corrosion Fatigue (Ref 31)

Environmentally assisted cracking, also known as stress-corrosion cracking (SCC), is a major source of potential failures in the process industries, in pulp mills, in storage vessels, and even in aircraft. Quite often, SCC occurs in the heat-affected zone (HAZ) immediately adjacent to a weld, simply because the HAZ is left in a state of very high residual tensile stress as a result of the shrinkage and differential cooling occurring in most welds. Tensile stress (resulting from the superposition of residual and applied stresses) (Ref 61) is the main component of the SCC triangle; the other two are a susceptible metal and an environment that often needs to be only slightly corrosive to that metal. For instance, grade 316 stainless steel is essentially inert to the corrosive effect of common salt unless tensile stresses are present, when it becomes very sensitive to chloride-induced SCC.

There are a number of possible solutions to the SCC problem. The obvious one is to change the environment, but that is rarely possible. The next is to change the metal, but usually that is expensive and, if the component in question is already built, impractical. Thermal stress relieving is a partial solution at best, because, to completely relieve all the tensile stresses in a component, it is necessary for the heating to reach the annealing temperature that may change the material properties. In addition, annealing cannot be used to overcome any subsequently applied tensile service loads. Corrosion engineers have long recognized that an effective solution for the retardation or even prevention of SCC is the introduction of compressive stresses.

For example, it can be seen in Fig. 14 that the shot peening technique used had a significant effect on the stress state of the weld and parent material, as seen by the “step” or drop in residual stress near the center of the stress map. On the left-hand side, a typical weld stress map is observed, with tensile residual stresses in the weld and in the HAZ, then dropping off in the parent material. The right side of this map was the peened portion. Here, the characteristic profile is much more compressive (or less tensile) and smooth; however, tensile residual stresses still exist. This indicates that the peening process had the effect of reducing the tensile residual-stress field in the weld and HAZ and introducing a much more uniform compressive residual-stress level in the parent material. However, it was not sufficient to make the surface stresses in the HAZ entirely compressive. This would suggest that the postweld treatment could be changed or augmented to increase the compressive residual stress imparted on the weld and HAZ.

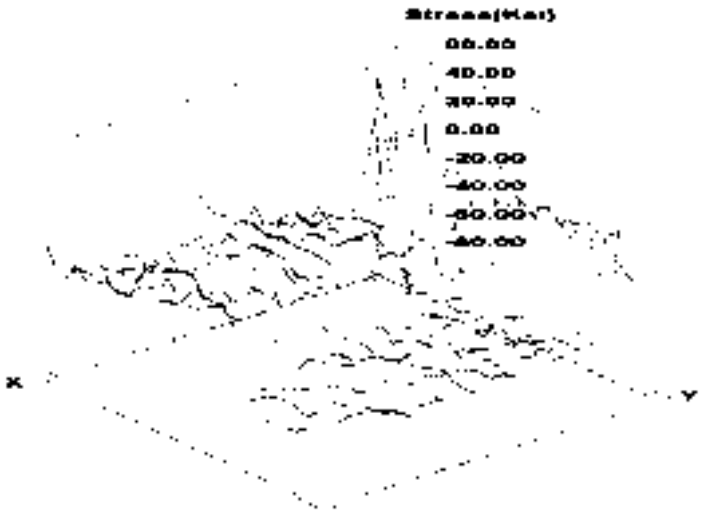


Fig. 14 X-ray diffraction residual-stress map showing the introduction of compressive surface residual stresses in the parent material and the reduction, but not elimination, of tensile residual stresses in the weld metal on the unmasked side from shot peening a nickel alloy weldment. Source: Ref 31

X-ray diffraction techniques can thus be used to characterize the stress state of components that may be susceptible to SCC either prior to or after they have been put into service. In the case of corrosion fatigue, compressive surface residual stress has a beneficial effect on lifetime and strength. In the case of data shown in Fig. 15 (Ref 36), it can be seen that XRD residual stress measurements can be used to compare the effects of grinding and shot peening on the surface and subsurface residual stress state.

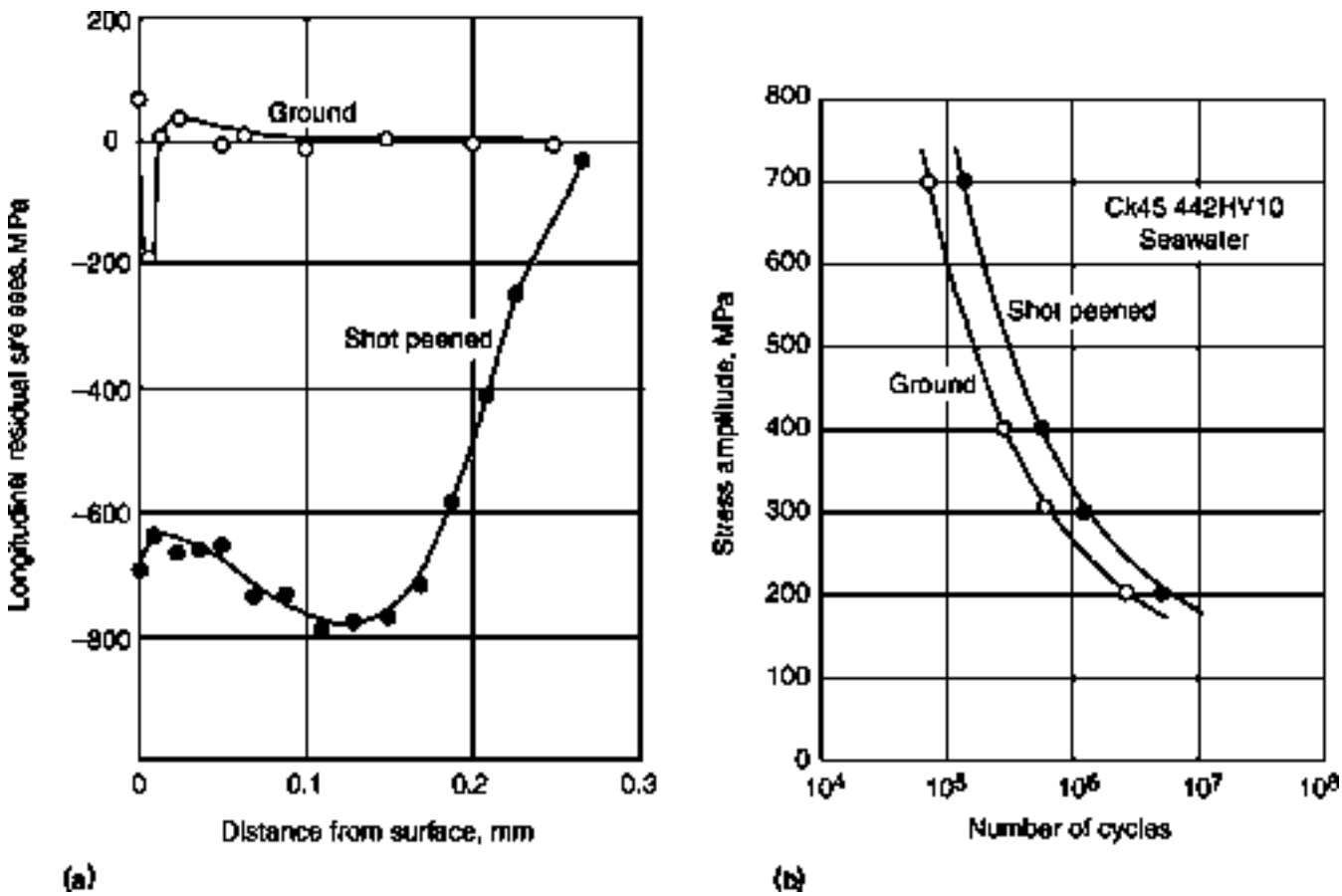


Fig. 15 Effects of grinding and shot peening on surface and subsurface residual stress in low-carbon (CK 45) steel tested in seawater. (a) Residual stress versus depth profiles. (b) Bending fatigue stress-number of cycles (*S-N*) curves. Source: Ref 36

References cited in this section

31. J.A. Pineault, M.E. Brauss, and J.S. Eckersley, Residual Stress Characterization of Welds Using X-Ray Diffraction Techniques, *Welding Mechanics and Design,* American Welding Society, 1996

36. R. Herzog, Dr. Ing. thesis, University Gh Kassel, 1996.

61. A.R. McIlree, C.O. Ruud, and M.E. Jacobs, The Residual Stress In and the Stress Corrosion Performance of Roller Expanded Inconel 600 Steam Generator Tubing, *International Conf. on Expanded and Rolled Joint Tech.,* Canadian Nuclear Society, 1993, p 139–148

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

The Importance of Residual Stress in Fatigue

The strong impact that surface and near-surface residual stresses have on the fatigue life of components underlies the importance of studying the effect of surface treatments and manufacturing processes (Ref 2, 37, 38). It is now known that if reliable fatigue-life estimates are to be made, it is necessary to characterize the residual-stress fields in test specimens and engineering components (Ref 39).

When a component undergoes cyclic loading, it can be susceptible to fatigue. The fatigue life for a given component is often characterized with a Wöhler diagram or *S-N* curve, where *S* is defined as the total stress range the material experiences during cyclic loading, and *N* is defined as the number of cycles to failure. Consider the simple case of constant amplitude cyclic loading (Fig. 16) (Ref 40). The peak-to-peak difference in the maximum and minimum stress is defined as the stress range. However, in the case of Fig. 16, the mean stress cannot be simply considered as the mean applied stress but must be considered as the superposition of the mean applied stress and the residual stress. Thus, a nonzero residual stress offsets the mean stress about which the stress amplitude cycles. This principle applies to variable amplitude cycling as well. Changes in the residual stress for a component in a given application (and thus, the mean total stress) can have the effect of displacing the *S-N* curve and subsequently changing the fatigue life.

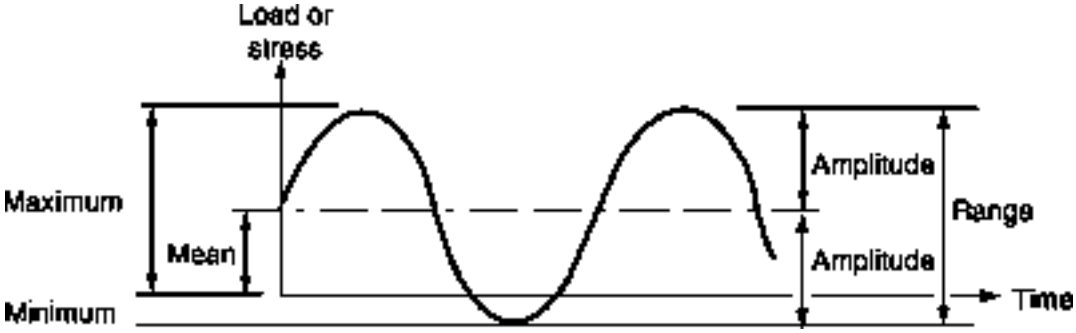


Fig. 16 Typical constant amplitude cyclic loading spectrum. Source: Ref 40

S-N curves are plotted in Fig. 17 for two identical gears of identical hardness, except one was double shot peened. The stress range of the fatigue limit of gear A is 1256 MPa, whereas an increase of 38% to 1710 MPa is observed for the double shot peened gear B (Ref 41).

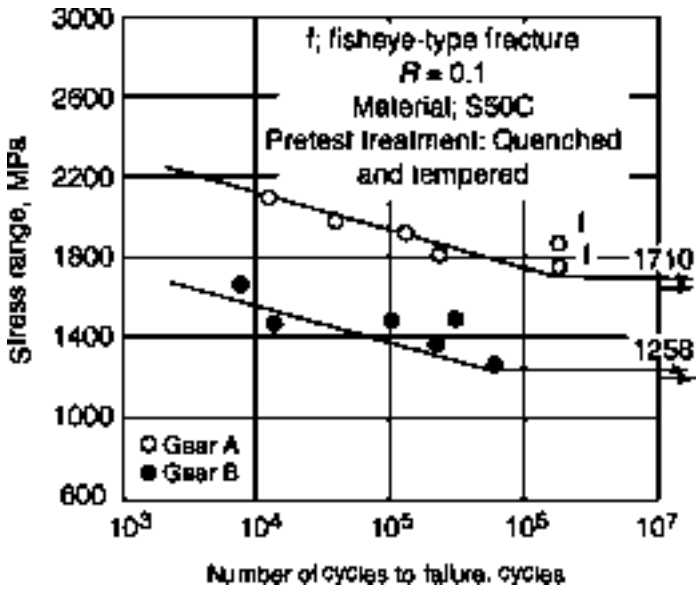


Fig. 17 *S-N* curves for as-hardened (gear A) and as-hardened plus double shot peened (gear B) gears. Source: Ref 41

Plastic strain, which is generally more significant when operating in cyclic fatigue, has a predominant influence in the number of cycles to failure (Ref 42). The relation between the plastic and the elastic strain can be summarized as follows: in high-life ranges, the plastic strain rapidly diminishes to negligible values, and the elastic strain range dominates. In the low-life range, the plastic component of strain dominates the material behavior.

The predominant driving force in fatigue (elastic or plastic) is thus determined by the in-service strain/stress range. In the case of short fatigue lives, plastic strain is more dominant than the elastic strain, so that low-cycle fatigue (LCF) life is controlled by the ductility of the material. At longer fatigue lives, as in the case of high-cycle fatigue (HCF), the elastic strain is more significant than the plastic strain, and the fatigue life is determined by fracture strength. Another difference between LCF and HCF failures is that LCF is characterized by multiple cracking in highly stressed areas, whereas HCF failures generally initiate at precise stress concentration sites, with cracks propagating from a single initiation (Ref 43).

In cases of LCF failures, the in-service stress exceeds the elastic limit of the material. Low-cycle fatigue is usually used to define the reverse elastoplastic loading mode (Ref 44, 45). Low-cycle fatigue is characterized by cumulative fatigue damage associated with cycles to failure of up to 10^4 to 10^5 cycles. Experience has also demonstrated that LCF life is sensitive to changes in stress (or strain). A 5 to 10% difference in stress can result in a 50% difference in life; thus, errors in the representation of total stress have a profound effect on life, if residual stress is not in the equation (Ref 46).

Example 1: Use of XRD to Assess Residual Stresses in Steel Springs. A steel spring used in an automotive application suddenly began to fail in the field. It was understood that "nothing had changed" in the fabrication process of these springs, yet the incidence of field failures suddenly increased dramatically. Fatigue tests, using springs fabricated prior to field failures, lasted 500,000 cycles to failure, whereas fatigue tests performed on springs fabricated after field failures lasted only 50,000 cycles to failure. It was discovered that the percent coverage of shot peening prior to and subsequent to the increase in failure incidence was much less than 100%, with a shot peening time of 12 min. Subsequently, a potential corrective action was introduced by the engineers: an increase in the time the spring was shot peened from 12 to 60 min. The residual-stress state of "as fabricated" springs in three conditions were thus evaluated: springs manufactured prior to failure incidence increase, 12 min peen; springs manufactured following failure incidence increase, 12 min peen; and 60 min peen.

The residual-stress measurement results, as seen in Fig. 18 indicate that something had indeed changed, namely the effective depth and magnitude of compressive residual stresses in the shot-peened springs when comparing before and after the increase in failure incidence. The cause of this change may have been either material or process related. Additionally, increasing the peening time from 12 to 60 min significantly increased the compressive residual-stress levels in the springs. In fact, the springs that were shot peened for 60 min were found to fail at much more than 500,000 cycles in fatigue tests. Thus, in this example, the suspicion that inadequate compressive residual stresses were imparted by the peening process was validated, and the validity and effect of the proposed corrective action was substantiated. In this case, an increased peening time resulted in an increased percent coverage and effective compressive stress level. This example demonstrates that once the source of the failure is understood and validated by experiment, corrective measures can be applied, verified, and subsequently monitored with confidence. This example also demonstrates that XRD can be used to explore quality-related issues. The residual-stress benchmarks established in this case study can be referenced by the manufacturer or the customer in future residual-stress measurement audits and/or incorporated into the blueprints for this component.

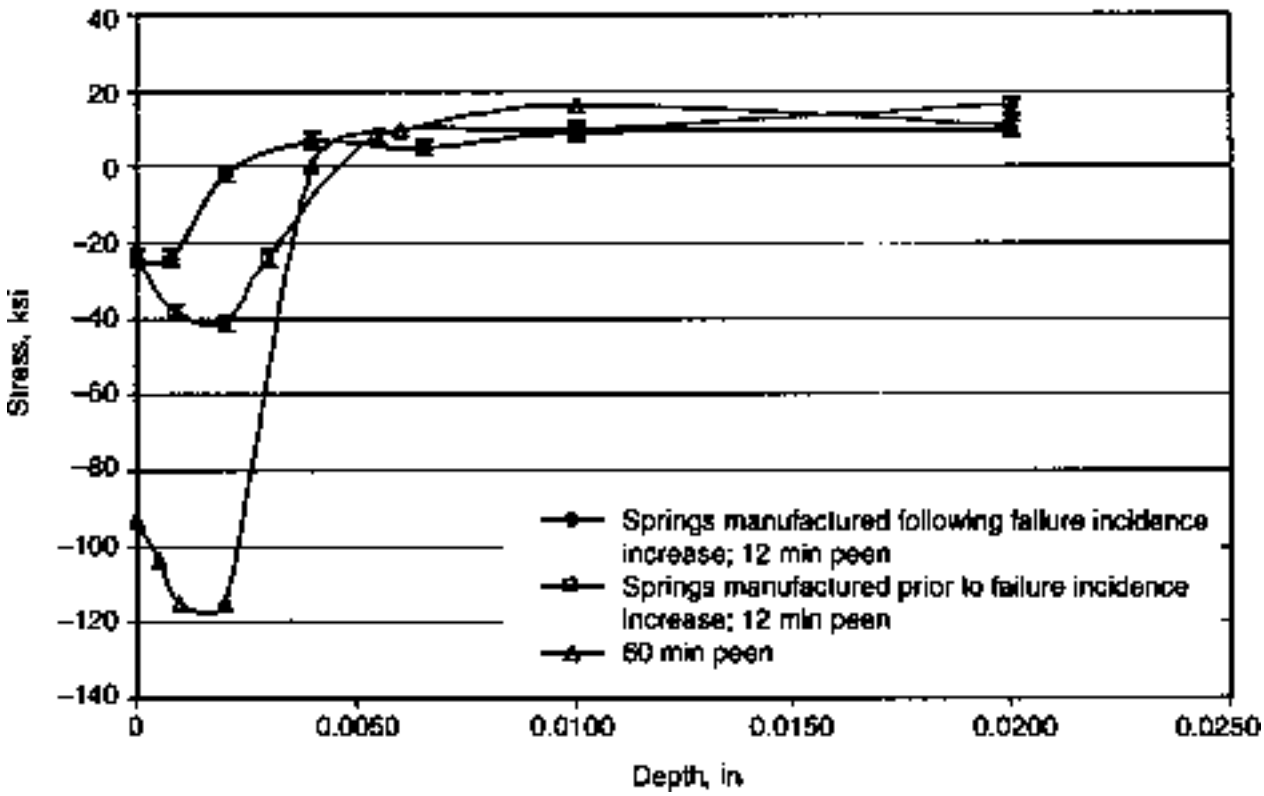


Fig. 18 Stress versus depth profiles for different steel coil springs

Example 2: Use of XRD to Assess the Effect of Mechanical Loading on Stress Relaxation of Machined and Shot-Peened Nickel-Base Alloy (Waspaloy). Experiments were conducted on rotating beam specimens at different loading levels. The specimen conditions are defined as shot peened and not shot peened/polished. Initial surface residual stresses were

determined using XRD techniques, and the specimens were thus divided into three groups: first level, second level, and third level (shot peened), where the third level was the most compressive.
Prior to failure, multiple cracks were observed at 500× magnification adjacent to the macroscopically visible main crack, indicating that damage was not localized in the middle of the gage section, that is, evidence of LCF (Fig. 19).

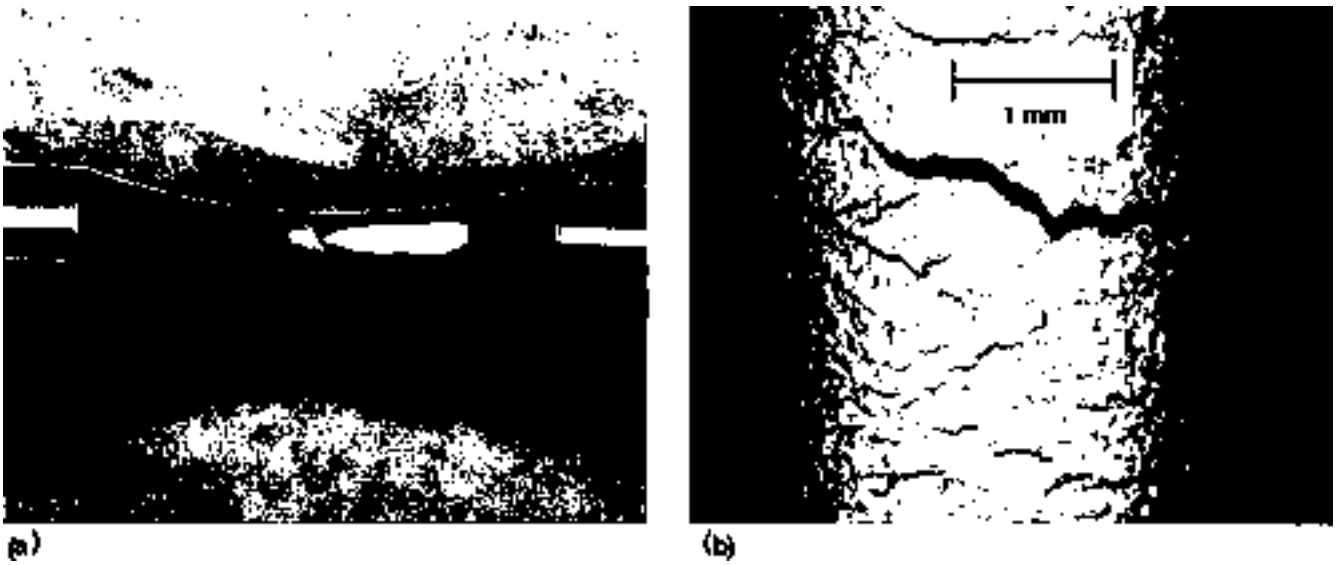


Fig. 19 Observation of failed nickel-base alloy (Waspaloy) specimen after rotating bend fatigue. (a) Macro view. (b) Micrograph. Source: Ref 43

S-N curves for all three stress levels can be seen in Fig. 20. Level 1 and level 2 curves are very similar; thus, the difference is negligible, according to the physical mechanism of failure. The number of cycles to failure for shot-peened specimens is longer at low applied bending moment and shorter at higher applied bending moment, as for non-shot-peened specimens. This can be explained by the effect of two parameters: the level of the compressive residual stress and the surface roughness. The roughness of the surface due to peening can act as a stress raiser and may diminish the fatigue life, but the high compressive stresses can counterbalance the effect of the roughness (Ref 47, 48). In addition, the benefit of shot peening at higher applied fatigue loads may be diminished due to the surface roughness and the potentially steeper evolution (i.e., lack of stability) of residual stresses when the yield strength of ductile materials, such as Waspaloy, are exceeded.

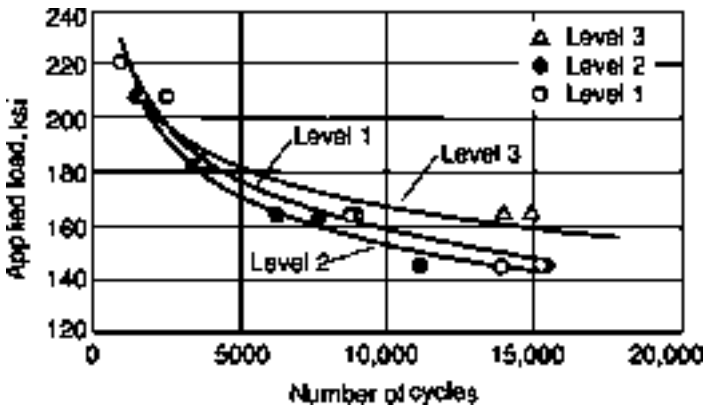


Fig. 20 Applied bending load and number of cycles to failure, *R* = -1, for Waspaloy specimen. Source: Ref 43

Residual stress relaxation is a consequence of micro- and/or macroplastic deformations and is important, whether the superposition of residual, loading, or mean stresses exceeds the monotomic or cyclic yield strength. The fatigue limit of high-strength materials is generally lower than the cyclic yield strength, whereas in low-strength materials, these values may nearly coincide. This explains why residual stresses do not relax in high-strength materials subjected to load amplitudes in the range of the fatigue limit (Ref 2).
In the case of materials such as Waspaloy, the evolution of residual stresses with LCF can be used as a mechanism for tracking the useful fatigue life remaining in the component. It has been shown, in some cases, that a linear relationship holds between the residual-stress relaxation rate and the residual stress amount of the starting condition (Ref 43, 45, 46, 49, 50).

Example 3: Using XRD to Track Residual- Stress Levels at Critical Locations on Turbine Engine Disks as They Accumulate Engine Cycles. The results shown in Fig. 21 indicate that there is a continual degradation of residual compressive stress with an increase in operational engine cycles. Based on the sensitivity of fatigue life with stress, the degradation of compressive residual stress from cold-working processes is an indication of increased “active” stress and could, therefore, be used as a measure of remaining fatigue life. For this to be realized, a trend must exist whereby the compressive residual stress, from shot peening or other cold-working processes, relaxes to a value below which the disk is at increased risk to crack. In other words, it is assumed that residual-stress degradation is a precursor to fatigue crack initiation in LCF mode where loads exceed the yield strength of the material (Ref 45). (The implications of the comparison in Fig. 21 of the “large spread—process not controlled” versus the “controlled process reduces spread” are discussed in the section “Sample Selection” in this article.)

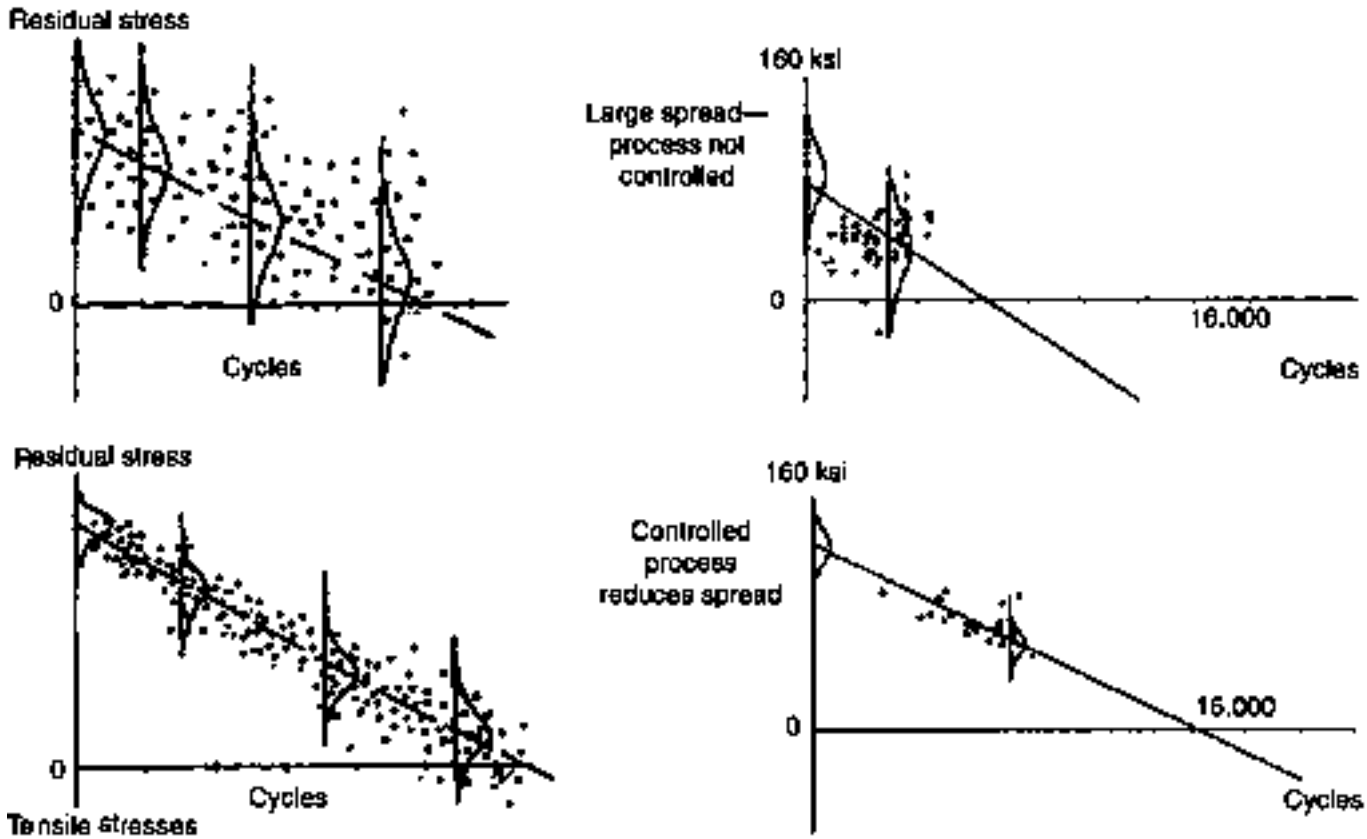


Fig. 21 Theoretical model versus XRD-generated experimental data plots of residual stress versus number of cycles. Source: Ref 43

Residual Stress and HCF. High-cycle fatigue may occur when the cyclic load range does not exceed the cyclic yield strength of the material but exceeds the fatigue limit (endurance limit) of the material (although many materials, such as aluminum, do not have well-defined fatigue limits). In components where no initial stresses are present, stresses may develop, and, if the stresses that develop are compressive, they may actually add to the life of the component. However, if stresses are induced by processing, they may “fade” at a rate that increases with cycling stress (Fig. 22) (Ref 51).

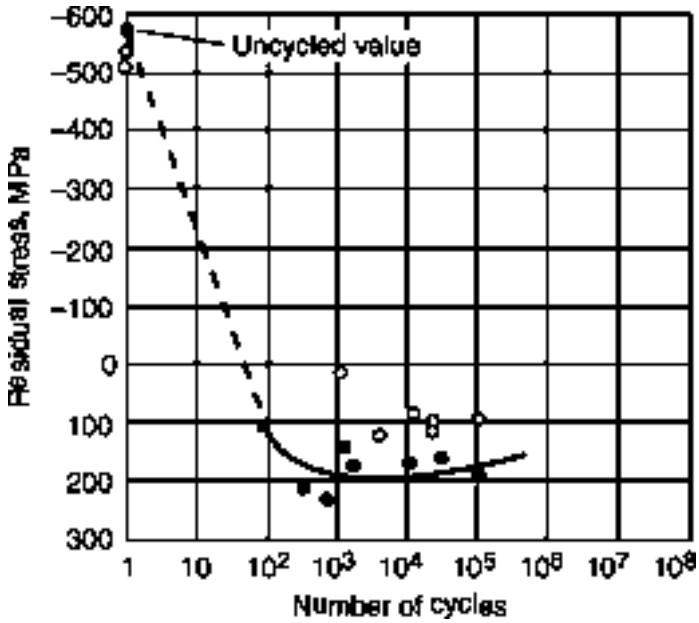


Fig. 22 Residual stresses in peened 1040 steel samples resulting from tension-tension fatigue (the two symbols represent two samples). Source: Ref 51

Other Considerations Regarding Residual Stress and Fatigue. There are cases where the influence of residual stress on fatigue life is negligible. For hardened and ground specimens, a clear positive influence of compressive stresses and a negative influence of tensile stresses are often exhibited. Compressive residual stresses impart the maximum benefit when the volumes with the highest compressive residual stress coincide with the highest loaded component volumes. This explains why the influence of near-surface residual stresses is more pronounced for bending fatigue than in the case of tension-compression loading. For example, in the case of a normalized material with thin-reaching machining residual stresses, they may have very little impact on the resulting tension-compression loading *S-N* curves. Positive residual-stress effects are also more pronounced if detrimental starting conditions, such as decarburization or oxidation, exist in the case of hardened steels. Thus, the stability of residual stresses is of central importance; that is, their effect is more pronounced when less relaxation occurs during fatigue loading (Ref 2).

References cited in this section

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods,* Elsevier, 1997

37. V.M. Faires, *Design of Machine Elements,* 4th ed., MacMillan, 1965

38. J.S. Eckersley, *Shot Peening Theory and Application,* IITT-International, 1989, p 241–255

39. J.F. Throop et al., “Residual Stress Effects in Fatigue,” STP 776, ASTM, 1982

40. H.S. Reedsmeyer, Life Assessment Under Variable Amplitude Loading, *Fatigue and Fracture Analysis of Ship Structures,* Fleet Technologies Limited, 1998

41. K. Matsui et al., An Increase in Fatigue Limit of a Gear by Compound Surface Treatment, *Proceedings of ICRS6* (Oxford), 2000, p 871–878

42. D. François, A. Pineau, and A. Zaoui, *Comportement Mécanique des Matériaux* (*Mechanical Behavior of Materials*), Vol 2, Hermes, Paris, 1993

43. M. Belassel, M.E. Brauss, and S.G. Berkley, Residual Stress Relaxation in Nickel Based Alloys Subjected to Low Cycle Fatigue, *Proceedings of ICRS6* (Oxford), 2000, p 144–151

44. C.M. Verpoort, and C. Gerdes, *Shot Peening Theory and Application,* IITT-International, 1989, p 11–70

45. S.G. Berkley, U.S. Patents 5,490,195 and 5,625,664

46. S.G. Berkley et al., “Residual Stress Measurement and Its Application to Achieve Predicted Full Life Potential of Low Cycle Fatigue (LCF) Limited Engine Disks,” Ninth International Symposium on Transport Phenomena and Dynamics of Rotating Machinery, ISROMAC-9, (Honolulu, HI), 2002

47. C.O. Nonga et al., The Influence of Residual Stresses and Surface Finish on the Fatigue of Metal Matrix Composites, *Proceedings of ICRS5* (Linkoping), 1997, p 95–100

48. *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996

49. G. Kuhn et al., Instability of Machining Residual Stresses in Differently Heat Treated Notched Parts of SAE 1045 during Cyclic Deformation, *Proceedings of ICRS3,* Elsevier, 1992, p 1294–1301

50. G. Kuhn, Dr. Ing. thesis, University Karlsruhe (TH), 1984

51. M. McClinton and J.B. Cohen, *Mater. Sci. Eng.,* Vol 56, 1982, p 259–263

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

The Effect of Manufacturing Processes on Residual Stress

Because of the enormous impact of manufacturing processes and loading history on resulting residual-stress states, they have to be taken into account in the case of failure analysis (Ref 2, 8). It is assumed, at least in principle, that in-service loads for a given component can be determined either by direct calculation, measurement, or finite-element modeling. X-ray diffraction can be used to measure the residual stress in as-manufactured components so as to enable the detection of stress states that may be potentially harmful to service life and to subsequently identify unfavorable process parameters.
For example, consider the effects of what are identified in Fig. 23 (Ref 3) as gentle, conventional, and abusive grinding. It is important to note, in the case of abusive grinding (where grinder burn is present), that the surface residual stress may be neutral (as in this example) or even compressive; however, the life-limiting tensile residual stresses may be at very shallow depths below the surface. Thus, failure analysts must consider surface and subsurface residual stresses in their analysis.

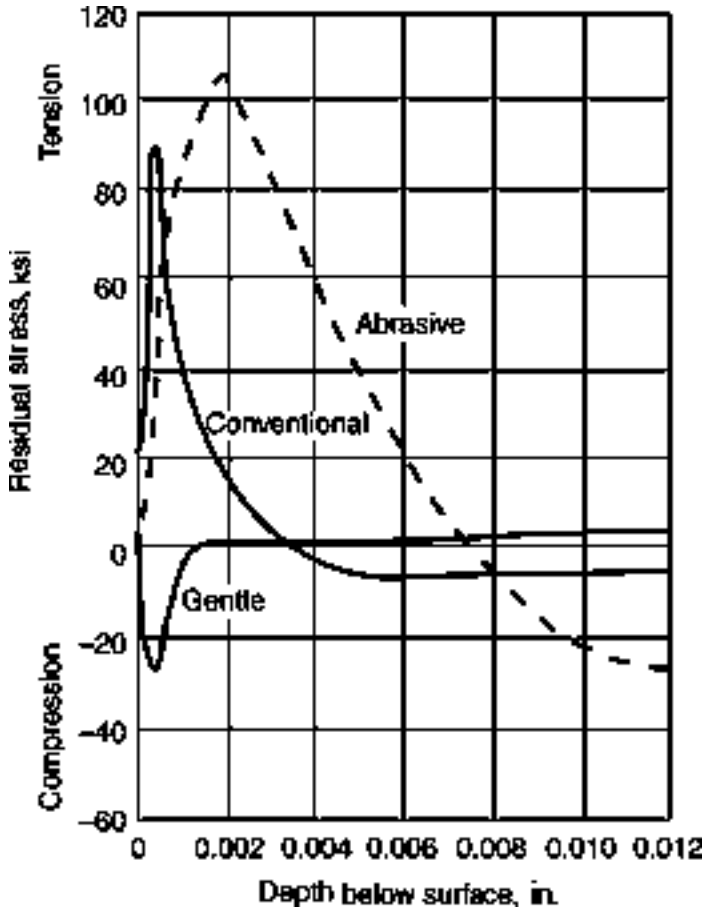


Fig. 23 Subsurface residual-stress distribution after grinding hardened steel (stress measured in the direction of grinding). Source: Ref 3

References cited in this section

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods,* Elsevier, 1997

3. M.E. Hilley et al., "Residual Stress Measurement by X-Ray Diffraction" J784a, Society of Automotive Engineers, 1971

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

The Characterization of Stress Gradients Using XRD

The XRD technique is generally applied to the measurement of residual stress in polycrystalline materials used in mechanical, electronic, and structural components. After machining, casting, shot peening, turning, heat treatment, and so on, the residual stress can vary significantly from the surface through the subsurface (Ref 2). These residual stresses must therefore be characterized at the surface and through the subsurface to evaluate the effects of the process and the stress gradients generated (Ref 52, 53). Changes in the process can then be evaluated via changes in the surface and subsurface stress state. To access subsurface measurement locations, material removal is required and can be performed using electro-polishing techniques. This technique is discussed in more detail in the section "Specimen Preparation" in this

article. Stress relaxation due to material removal and stress gradient corrections can subsequently be applied to stress versus depth results (Ref 3). Residual-stress characterization in both the surface and subsurface is a powerful and essential tool in failure analysis, process control, and optimization (Ref 6, 52).

The subsurface stress gradients shown in Fig. 24 reflect the effect of varying peening intensity on shot-peened components from the same population. The compressive residual stress is low at the surface for the unpeenend sample and high at the surface for the peened ones. Normally, samples such as these are subsequently exposed to fatigue testing in order to select the optimal process. This information can then be used to establish a benchmark for quality control and quality assurance. In cases where the process is adequately controlled, the surface stresses can be used as an indication of the presence and/or effectiveness of the shot peening in line with the subsurface stress profiles confirmed via periodic (hourly/daily/weekly) audits. Indirect information about the effect of the loading history on the residual-stress state of components can also be obtained after they have been placed into service.

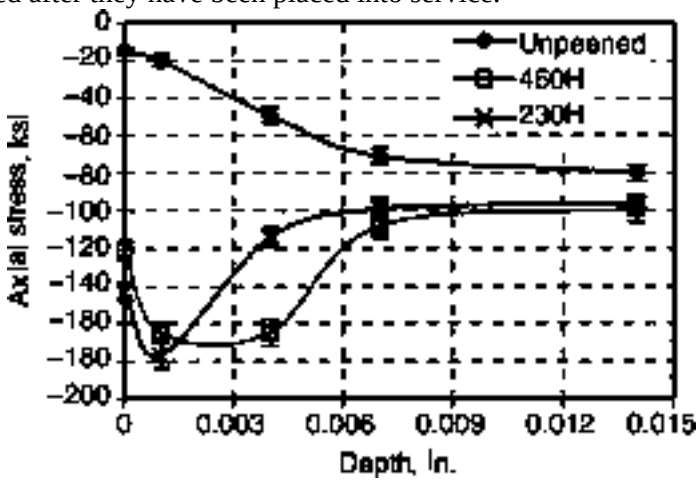


Fig. 24 Residual-stress profiles on unpeened and peened samples

Effects of In-Service Loads on Residual Stress. The residual stress versus depth plots in Fig. 25 indicate that the in-service loads have a profound effect on the near-surface stress state of the material in the gear. The service cycling introduced an increased compressive near-surface residual stress layer in the region of the pitch diameter. This compressive residual stress as a result of in-service cycling may, in fact, increase the resistance to fatigue of the gear at this location. This effect is, of course, localized and does not address the stress state of other locations, such as the root of the gear.

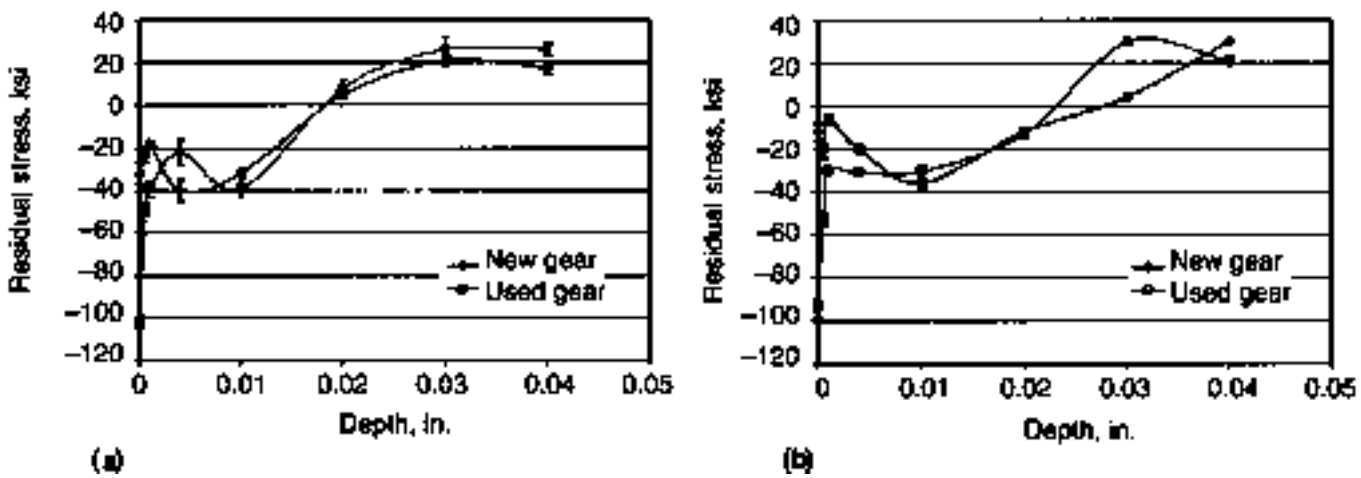


Fig. 25 Comparison of the residual stress on the tooth pitch diameter found in two different types of hardened steel gears in new and used conditions using XRD. (a) Pinion gears. (b) Sun gears

The residual stress versus depth plot in Fig. 26 indicates that the in-service loads have the effect of reducing the compressive near-surface stress state of the material in the spring. In this case, the service cycling introduced probably decreases the resistance of the spring to fatigue failure.

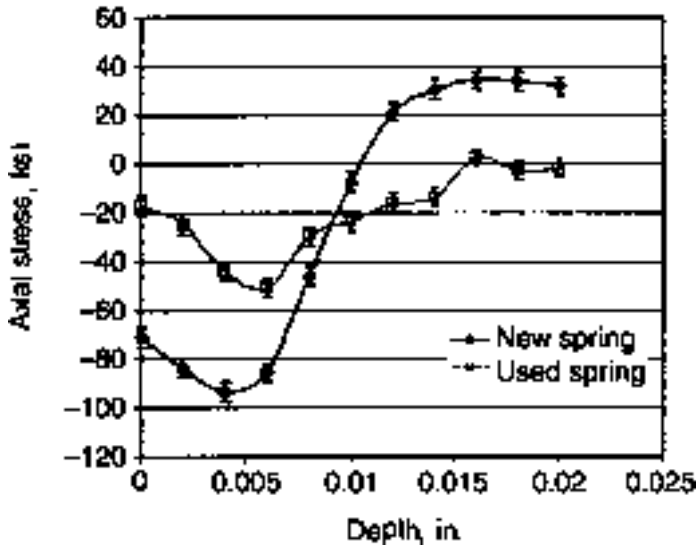


Fig. 26 Comparison of the residual stress on the inner diameter of shot-peened coil springs in new and used conditions using XRD

The examples shown in Fig. 25 and 26 illustrate the potential effects of cold working on manufactured components prior to and subsequent to being placed into service.

References cited in this section

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods,* Elsevier, 1997

3. M.E. Hilley et al., “Residual Stress Measurement by X-Ray Diffraction” J784a, Society of Automotive Engineers, 1971

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

52. J.A. Pineault, and M.E. Brauss, *Measuring Residual Stress Using X-Ray Diffraction on Shot Peened Components,* MAT-TEC, IITT-International, France, 1993

53. R. Fathallah et al., Effect of Shot Peening Parameters on Introduced Residual Stresses, *Proceedings of ICRS4* (Baltimore), 1994, p 340–346

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Effects of Heat Treatment on Residual Stresses

Heat treatment processes are also commonly applied to engineering components to obtain the desired microstructure, such as in the case of quenching and/or tempering. Heat treatments may also be applied that have as their purpose the effect of relaxing micro- and/or macroresidual stresses. X-ray diffraction can be used effectively in the characterization of these processes as well.

Example 4: Use of XRD to Assess the Effects of Heat Treatment on Residual Stress in Steel Coil Springs. Shot-peened steel coil springs were heat treated for 45 min through a range of temperatures to observe stress relaxation effects with various tempering temperatures. The surface residual macrostress was measured prior to and subsequent to heat treatment. The purpose of measuring residual stresses on the same springs before and after heat treatment was to minimize the effect of potential variations in the as-manufactured residual-stress state. The data plotted in Fig. 27 were obtained.

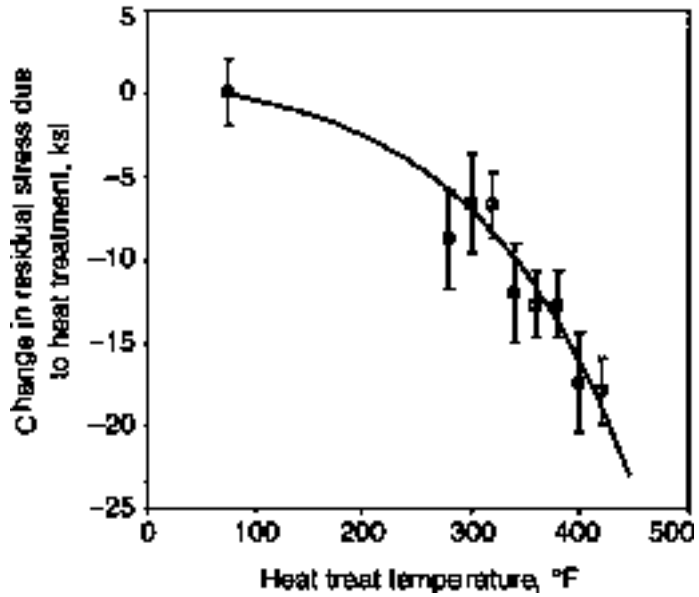


Fig. 27 Plot of the change in the compressive residual stress due to heat treatment

One coil spring was kept as a control, and the reduction in compressive residual stress due to heat treatment was plotted for the remaining springs that were subject to various heat treatment temperatures. A trend of increasingly relaxed residual stress was observed for increased heat treatment temperature.

Similar studies can be conducted on a variety of components to assess the change in residual macrostress due to heat treatment. The effect may be a positive influence, if life-limiting (undesirable) residual stresses are relaxed. Conversely, the impact of life-extending (desirable) residual stresses may be reduced.

Example 5: Use of XRD to Evaluate the Effect of Varying Heat Treatment Temperature on Residual Stress for Iron Alloys. As this example demonstrates, XRD can also be applied to the evaluation of both micro- and macrostress simultaneously. Trends in the relaxation of macrostresses are evaluated using the associated shifts in the atomic lattice spacing and thus the diffraction angle, whereas the distribution of microstresses can be correlated to the XRD peak width.

The residual stress with varying heat treatment temperature for two iron alloys is plotted in Fig. 28. The magnitude (or absolute value) of the residual stress for these specimens tends to decrease for a given increase in heat treatment temperature. The effect is pronounced for the AISI 01 samples, because they had a significant residual stress prior to heat treatment. The residual stresses in the AISI 1070 samples were lower, on average, prior to heat treatment; thus, there were lower-magnitude macrostresses to relax. The scatter in these data (beyond the quoted error bars) can be attributed to the slightly different residual-stress state prior to heat treatment in the various samples used to represent each heat treatment temperature.

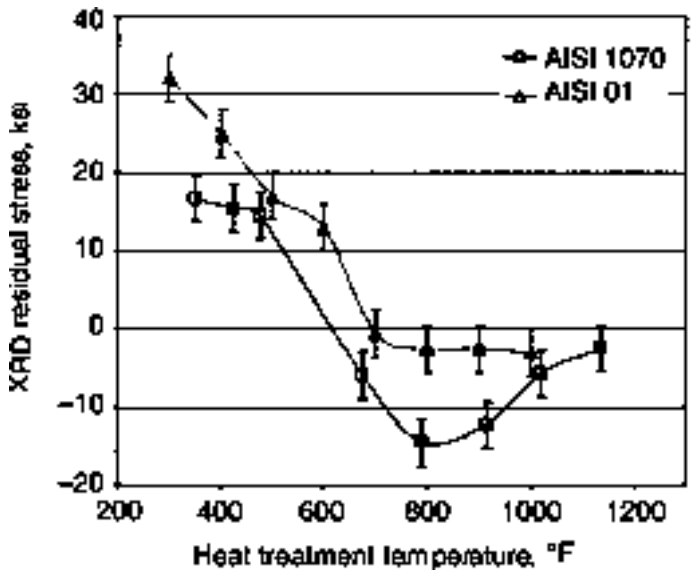


Fig. 28 X-ray diffraction residual stress versus heat treatment temperature for various iron alloys. Specimens were held at temperature for 1 h and furnace cooled.

Dislocations are typically introduced into materials by cold working. The formation of small crystallites (coherent domains) and the introduction of elastic microstrains both result in broadening of the XRD line profile (Ref 7). Line broadening is also a function of instrumental broadening, that is, the XRD instrument optics.

Despite trends in macrostress (Fig. 28), a trend of decreased hardness with increased heat treatment temperature is often observed in iron alloys sample sets (Fig. 29a). An associated decrease in the XRD peak breadth is also often observed (Fig. 29b). In this example, the range of heat treatment temperatures remained below the annealing temperature for each alloy; thus the instrumental and particle size contributions to the broadening remained more or less constant. If the heat treatment temperature exceeds the annealing temperature, the observed XRD peak breadth may be reduced either by recovery, recrystallization, or grain growth (Ref 6). Heat treatment with slow cooling has the effect of relieving or reducing microstresses as well as decreasing the dislocation density, thus decreasing the hardness.

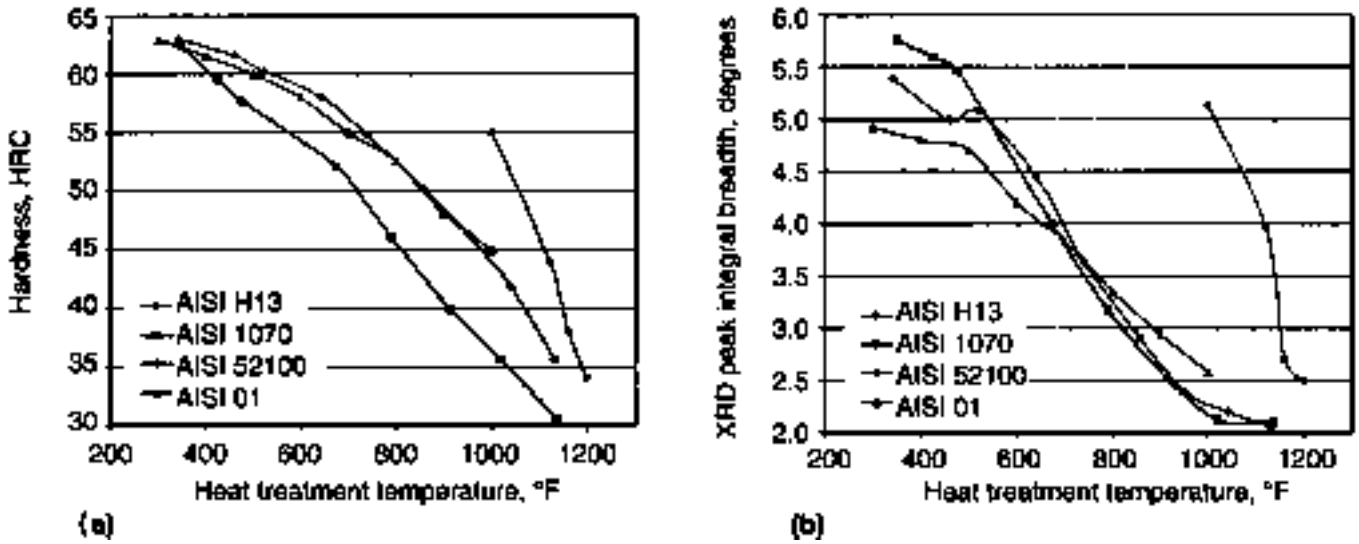


Fig. 29 Effect of heat treatment temperature on (a) hardness (HRC) and (b) XRD peak integral breadth.

The effects of these mechanisms can thus be observed in the XRD peak breadth. In general, this is used as a trending tool, and the data are most easily used in empirical form to assess the relative hardness, cold working, or dislocation density in a given material.

More detailed analyses of the source of XRD line broadening can be performed using Fourier space methods, such as the Warren-Averbach method (Ref 54), or real space methods, such as Voigt deconvolution (Ref 55). The XRD peak integral breadth and/or full width at half maximum is generally characterized as a function of depth with the use of electropolishing techniques in parallel with the measurement of residual macrostresses.

References cited in this section

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* 2nd ed., Wiley-Interscience, 1974

54. B.E. Warren, *X-Ray Diffraction,* Addison-Wesley, 1969

55. J.I. Langford, A Rapid Method for Analyzing the Breadths of Diffraction and Spectral Lines Using the Voigt Function, *J. Appl. Crystallogr.,* Vol 11, 1978

X-Ray Diffraction Stress Measurements in Multiphase Materials and Composites

In many materials, the presence of more than one phase is not uncommon. Materials that are subject to heat treatment and/or have alloying elements may be composed of two, three, or more phases. For composite materials, the failure analyst should consider all constituent phases. The diffraction conditions and measurement parameters may differ for each phase, however, the additional effort is often worthwhile for a complete analysis.

For example, carbon steel containing over 0.5% C and other elements added to increase toughness, improve corrosion resistance, and so on can contain ferrite (and/or martensite), austenite, cementite, and/or other phases. The volume fraction of each phase can vary with alloying content, heat treatment, and even cold working; therefore, the stresses have to be evaluated for all phases of significant volume fraction. The residual stress in each phase and the weighted average stress (based on the volume fraction of each phase) should be calculated, because the fracture resistance of the material may vary locally, depending on the phase under consideration. For example, untempered martensite is more brittle than austenite; thus, when both phases are present in a component, each may behave quite differently during cyclic loading. The residual stress in all phases present in the material should thus be measured and analyzed.

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

X-Ray Diffraction Stress Measurements in Locations of Stress Concentration

Stress concentrations and stress raisers can be important contributors to the initiation and/or propagation of cracks and subsequent component or structural failure. Stress concentrations are ubiquitous, found in locations associated with geometric discontinuities such as sharp corners, weld “toes,” notches, undercuts, machining marks, scratches, microcracks, and so on, and naturally, crack tips themselves. The degree of care taken in the estimation, modeling, and ultimate management of the various stress concentration contributors in a given structure or component should be dependent on the criticality of the component in the structure.

Typically, stress concentration factors (SCFs) are either found in publications such as *Peterson's Stress Concentration Factors* (Ref 56) or determined via computer-based modeling using various techniques. The application of SCFs obtained using either of these two approaches may involve assumptions regarding component geometry, the SCF uniformity, and, in some cases, estimates of the residual stresses present. Experimental determination of SCFs using XRD is often possible for specific geometric features by performing stress measurements at concentration locations under an incremental loading regime (Ref 12, 31). Data can also be collected to determine the variation of the SCF due to inhomogeniety of the stress concentration geometry, that is, where the radius of curvature is nonuniform.

It is important to measure residual stresses local and adjacent to stress concentrations using XRD, because residual stresses are additive to applied stresses and thus influence the ultimate effect of the total stress present in the region of stress concentration (Ref 57). It has been shown that surface residual stresses can significantly affect crack opening behavior (thus, crack growth behavior) even when the crack has grown well beyond the zone of residual stress (Ref 58). This indicates that the residual compressive stress present at the surface can strongly influence the surface crack tip opening displacement throughout the entire loading cycle (Ref 58).

Example 6: Use of XRD to Evaluate Cracks in Steel Cargo Tiedown Sockets. Cracks initiating from the tip of the cloverleaf pattern in steel cargo tiedown sockets were observed by the builder following installation aboard several cargo vessels in various stages of construction. A root-cause failure analysis was launched, and three possible mechanisms of failure were considered: overload failure, fatigue fracture, and environmentally assisted cracking (stress-corrosion cracking, SCC). Failure mechanisms of fatigue fracture and environmentally assisted cracking were eliminated, because no evidence of SCC or fatigue fracture was observed; thus, the failure analysts focused on the possibility of overload

failure. Further tests showed that the overload failure mode and the transition from ductile to brittle fracture were facilitated by the combination of high brittleness of a carbon-rich transformed martensite layer introduced by flame cutting, increased hardness due to the cold-working coining process, and high residual stresses created by the welding process (Ref 59).

Finite-element models (Fig. 30) were used to predict residual stresses in the areas of high stress concentration introduced during the manufacturing (welding) process and additional applied stresses introduced by the “sagging” loads inherent in the ship hull construction and ballast distribution. The high SCF thus had the effect of locally amplifying the manufacturing-induced welding residual stresses and the in-service dead load stresses.

Fig. 30 Finite-element model showing maximum stress concentration in cloverleaf radius. The highest stress concentration is in the small black area surrounded by white.

To validate the “overload failure” postulate, measurements were performed in the field on cargo ships with the cracking problem. A typical residual-stress profile collected on the tiedown sockets installed on ships can be seen in Fig. 31.

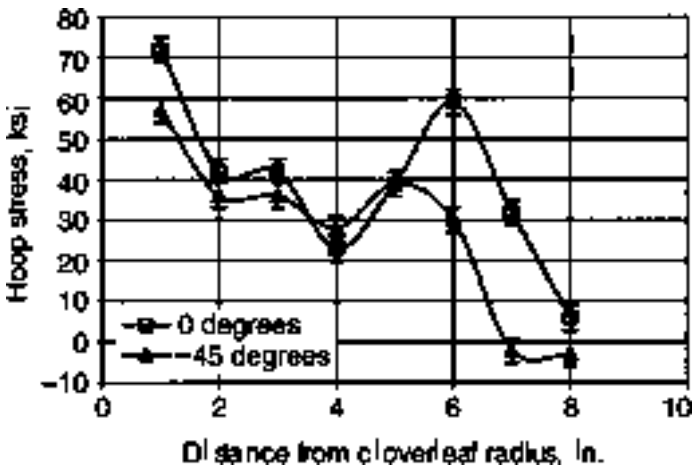


Fig. 31 Typical residual-stress profile as a function of distance from the maximum stress concentration in the radius of a tiedown socket

X-ray diffraction was thus used to validate the postulated failure mechanism and effectively demonstrated the existence of residual weld-induced stresses in installed tiedown sockets (Ref 60). Subsequently, failure analysts were able to recommend corrective measures: the removal of the brittle, carbon-rich transformed martensite layer introduced by flame cutting, which facilitated the overload failure; and the application of a localized stress-relief heat treatment process. X-ray diffraction residual-stress measurements were then performed on heat treated tiedown sockets to verify the effectiveness of the localized heat treatment process applied.

References cited in this section

12. M.E. Brauss, G.V. Gorveatte, and J.F. Porter, Development of a Miniature X-Ray Diffraction Based Stress Analysis System Suitable for Use on Marine Structures, *Nondestructive Evaluation of Materials and Composites—SPIE,* Vol 2944, 1996

31. J.A. Pineault, M.E. Brauss, and J.S. Eckersley, Residual Stress Characterization of Welds Using X-Ray Diffraction Techniques, *Welding Mechanics and Design,* American Welding Society, 1996

56. W.D. Pilkey et al., *Peterson's Stress Concentration Factors,* 2nd ed., Wiley-Interscience, 1997

57. M.E. Brauss, J.A. Pineault, and M.J. Vinarcik, Characterizing Residual Stresses Induced by Rolling in Crankshaft Fillets Using X-Ray Diffraction, *Proceedings of SEM Spring Conference,* Society for Experimental Mechanics, 1995

58. J.E. Hack and G.R. Leverant, "Influence of Compressive Residual Stress on the Crack Opening Behavior of Parts Through Fatigue Cracks," Residual Stress Effects in Fatigue: ASTM Special Technical Publication 776, ASTM PCN 04-776000-30, 1981, p 204–210

59. X.J. Zhang and M. Gaudett, Steel Failure Analysis Results in Stronger Deck Sockets, *Wavelengths—An Employee Digest of Events and Issues,* Jan/Feb 1999

60. M. Zoccola, Division Trio Receive Navy Superior Service Award, *Wavelengths—An Employee Digest of Events and Issues,* Aug 1999

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Acknowledgments

The authors would like to thank Debbie Aliya, Clayton O. Ruud, Stan Berkley, Robert Drake, and Fred Tonch for their helpful insight and careful review of the ideas presented in this article.

References

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods,* Elsevier, 1997

3. M.E. Hilley et al., "Residual Stress Measurement by X-Ray Diffraction" J784a, Society of Automotive Engineers, 1971

4. P.S. Prevey, X-Ray Diffraction Residual Stress Techniques, *Materials Characterization,* Vol 10, *ASM Handbook,* American Society for Metals, 1986, p 380–392

5. J. Lu, et al., *Handbook of Measurement of Residual Stress,* Fairmont Press, 1996

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* 2nd ed., Wiley-Interscience, 1974

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000

9. H.H. Lester, and R.H. Aborn, *Army Ordnance,* Vol 6, 1925–1926, p 129, 200, 283 and 364

10. C.O. Ruud, P.S. DiMascio, and D.J. Snoha, A Miniature Instrument for Residual Stress Measurement, *Advances in X-Ray Analysis,* Vol 27, Plenum, 1984

11. H. Berger, Sensing Stress, *Research/Penn State,* Vol 10 (No. 2), 1989

12. M.E. Brauss, G.V. Gorveatte, and J.F. Porter, Development of a Miniature X-Ray Diffraction Based Stress Analysis System Suitable for Use on Marine Structures, *Nondestructive Evaluation of Materials and Composites—SPIE,* Vol 2944, 1996

13. J.A. Pineault, and M.E. Brauss, Stress Mapping—A New Way of Tackling the Characterization of Residual Stress, *Exp. Tech.,* March 1995

14. J.A. Pineault, M.E. Brauss, and J.F. Porter, Characterization of Plastic Strain in HY-80 Using X-Ray Diffraction Techniques, *Proceedings of the Conference on Naval Applications of Material Technology,* (Halifax), D.R.E.A., 1993

15. M.G. Carfagno et al., *X-Ray Diffraction Measurement of Stresses in Post-Tensioning Tendons: Extending the Lifespan of Structures,* IABSE, Zurich, Switzerland, 1995

16. M.E. Brauss and J.A. Pineault, Residential Strain Measurements of Steel Structures, *NDE for the Energy Industry,* NDE-Vol 13, American Society of Mechanical Engineers, 1995

17. M.E. Brauss et al., "Deadload Stress Measurements on Brooklyn Bridge Wrought Iron Eye Bars and Truss Sections Using X-Ray Diffraction Techniques: Proceedings of the 14th Annual Meeting of the International Bridge Conference," Engineers' Society of Western PA, 1997

18. M.E. Brauss, J.A. Pineault, M. Belassel, and S.I. Teodoropol, Nondestructive, Quantitative Stress Characterization of Wire Rope and Steel Cables, *Proceedings of SPIE, Structural Materials Technology,* Vol 3400, Society of Photo-Optical Instrumentation Engineers, 1998

19. C.F. Jatczak et al., "Retained Austenite and Its Measurement by X-Ray Diffraction," SP-453, Society of Automotive Engineers, 1980

20. C.O. Ruud, X-Ray Diffraction Methods for Process Monitoring and Quality Control, *Topics on Nondestructive Evaluation Series,* Vol 1, *Sensing for Materials Characterization Processing and Manufacturing,* ASNT, 1998

21. C.O. Ruud and K.J. Kozaczek, Errors Induced in Triaxial Stress Tensor Calculations Using Incorrect Lattice Parameters, *Proceedings of the 1994 SEM Spring Conference,* Society for Experimental Mechanics, June 1994

22. "Standard Test Method for Verifying the Alignment of X-Ray Diffraction Instrumentation for Residual Stress Measurement," E 915-90, ASTM

23. E. Kröner, *Z. Phys.,* Vol 151, 1958, p 504–518

24. E. Kröner, *J. Mech. Phys. Solids,* Vol 15, 1967, p 319–329

25. "Standard Test Method for Determining the Effective Elastic Parameter for X-Ray Diffraction Measurements of Residual Stress," E 1426-91, ASTM

26. V. Li, J.L. Ji, and G.E. Lebrun, "Surface Roughness on Stress Determination by the X-Ray Diffraction Technique," *Exp. Tech.,* Vol 19 (No. 2), March/April 1995, p 9–11

27. A.J.C. Wilson, *Br. J. Appl. Phys.,* Vol 16, 1965, p 665

28. J.A. Pineault and M.E. Brauss, Measuring Residual and Applied Stress Using X-Ray Diffraction on Materials with Preferred Orientation and Large Grain Size, *Advances in X-Ray Analysis,* Vol 36, 1993

29. C.M. Mitchell, Stress Measurement by X-Ray Diffractometry, U.S. Patent 4,561,062, 1985

30. M. Belassel, M.E. Brauss, and J.A. Pineault, "Residual Stress Characterization Using X-Ray Diffraction Techniques, Applications on Welds," American Society of Mechanical Engineers Conference (Atlanta), 2001

31. J.A. Pineault, M.E. Brauss, and J.S. Eckersley, Residual Stress Characterization of Welds Using X-Ray Diffraction Techniques, *Welding Mechanics and Design,* American Welding Society, 1996

32. M.G. Moore and W.P. Evans, Mathematical Corrections in Removal Layers in X-Ray Diffraction Residual Stress Analysis, *SAE Trans.,* Vol 66, 1958, p 340–345

33. A. Constantinescu and P. Ballard, On the Reconstruction Formulae of Subsurface Residual Stresses After Matter Removal, *Proceedings of ICRS5* (Linkoping), 1997, p 703–708

34. C.L. Formby and J.R. Griffiths, *Proc. of Conf. on Residual Stresses in Welded Constructions and Their Effects,* The Welding Institute, 1987, p 359

35. J.E. Hack and G.R. Leverant, "Residual Stress Effects in Fatigue," STP 776, ASTM, 1982, p 204

36. R. Herzog, Dr. Ing. thesis, University Gh Kassel, 1996.

37. V.M. Faires, *Design of Machine Elements,* 4th ed., MacMillan, 1965

38. J.S. Eckersley, *Shot Peening Theory and Application,* IITT-International, 1989, p 241–255

39. J.F. Throop et al., "Residual Stress Effects in Fatigue," STP 776, ASTM, 1982

40. H.S. Reedsmeyer, Life Assessment Under Variable Amplitude Loading, *Fatigue and Fracture Analysis of Ship Structures,* Fleet Technologies Limited, 1998

41. K. Matsui et al., An Increase in Fatigue Limit of a Gear by Compound Surface Treatment, *Proceedings of ICRS6* (Oxford), 2000, p 871–878

42. D. François, A. Pineau, and A. Zaoui, *Comportement Mécanique des Matériaux* (*Mechanical Behavior of Materials*), Vol 2, Hermes, Paris, 1993

43. M. Belassel, M.E. Brauss, and S.G. Berkley, Residual Stress Relaxation in Nickel Based Alloys Subjected to Low Cycle Fatigue, *Proceedings of ICRS6* (Oxford), 2000, p 144–151

44. C.M. Verpoort, and C. Gerdes, *Shot Peening Theory and Application,* IITT-International, 1989, p 11–70

45. S.G. Berkley, U.S. Patents 5,490,195 and 5,625,664

46. S.G. Berkley et al., “Residual Stress Measurement and Its Application to Achieve Predicted Full Life Potential of Low Cycle Fatigue (LCF) Limited Engine Disks,” Ninth International Symposium on Transport Phenomena and Dynamics of Rotating Machinery, ISROMAC-9, (Honolulu, HI), 2002

47. C.O. Nonga et al., The Influence of Residual Stresses and Surface Finish on the Fatigue of Metal Matrix Composites, *Proceedings of ICRS5* (Linkoping), 1997, p 95–100

48. *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996

49. G. Kuhn et al., Instability of Machining Residual Stresses in Differently Heat Treated Notched Parts of SAE 1045 during Cyclic Deformation, *Proceedings of ICRS3,* Elsevier, 1992, p 1294–1301

50. G. Kuhn, Dr. Ing. thesis, University Karlsruhe (TH), 1984

51. M. McClinton and J.B. Cohen, *Mater. Sci. Eng.,* Vol 56, 1982, p 259–263

52. J.A. Pineault, and M.E. Brauss, *Measuring Residual Stress Using X-Ray Diffraction on Shot Peened Components,* MAT-TEC, IITT-International, France, 1993

53. R. Fathallah et al., Effect of Shot Peening Parameters on Introduced Residual Stresses, *Proceedings of ICRS4* (Baltimore), 1994, p 340–346

54. B.E. Warren, *X-Ray Diffraction,* Addison-Wesley, 1969

55. J.I. Langford, A Rapid Method for Analyzing the Breadths of Diffraction and Spectral Lines Using the Voigt Function, *J. Appl. Crystallogr.,* Vol 11, 1978

56. W.D. Pilkey et al., *Peterson's Stress Concentration Factors,* 2nd ed., Wiley-Interscience, 1997

57. M.E. Brauss, J.A. Pineault, and M.J. Vinarcik, Characterizing Residual Stresses Induced by Rolling in Crankshaft Fillets Using X-Ray Diffraction, *Proceedings of SEM Spring Conference,* Society for Experimental Mechanics, 1995

58. J.E. Hack and G.R. Leverant, “Influence of Compressive Residual Stress on the Crack Opening Behavior of Parts Through Fatigue Cracks,” Residual Stress Effects in Fatigue: ASTM Special Technical Publication 776, ASTM PCN 04-776000-30, 1981, p 204–210

59. X.J. Zhang and M. Gaudett, Steel Failure Analysis Results in Stronger Deck Sockets, *Wavelengths—An Employee Digest of Events and Issues,* Jan/Feb 1999

60. M. Zoccola, Division Trio Receive Navy Superior Service Award, *Wavelengths—An Employee Digest of Events and Issues,* Aug 1999

61. A.R. McIlree, C.O. Ruud, and M.E. Jacobs, The Residual Stress In and the Stress Corrosion Performance of Roller Expanded Inconel 600 Steam Generator Tubing, *International Conf. on Expanded and Rolled Joint Tech.,* Canadian Nuclear Society, 1993, p 139–148

Metallographic Techniques in Failure Analysis

George F. Vander Voort, Buehler Ltd.

Introduction

METALLOGRAPHIC EXAMINATION is one of the most important procedures used by metallurgists in failure analysis. Development of powerful electron metallographic instruments, such as the scanning electron microscope, has not diminished the importance of light microscopy. Basically, the light microscope is used to assess the nature of the microstructure and its influence on the failure mechanism. The purpose in using the light microscope may be twofold. One purpose may be to determine the relationship between the microstructure and the crack path (in failures involving fracture) and/or the nature of corrosion or wear damage. The second purpose is to determine whether processing or service conditions have produced undesirable microstructural conditions that have contributed to the failure, such as abnormalities due to material quality, fabrication, heat treatment, and service conditions. Examples are given in this article to demonstrate such analytical work.

Conducting a materials failure analysis, a common activity for many metallurgists, requires a carefully planned series of steps (Ref 1, 2) designed to arrive at the cause of the problem. Proper implementation of light microscopy is of critical importance in failure analysis, and this article focuses on the use of metallographic techniques and examinations using the light microscope (LM) in failure analysis. Metallographic examination typically should follow nondestructive and macroscopic examination procedures and should precede use of techniques of electron microscopy.

Examination of fractured components should begin with the low-power stereomicroscope. Hand-held magnifying lenses are still widely used to study fractures but mainly in the field. While the light microscope has limited value for direct observation of fracture surfaces, a great deal can be learned by indirect examination, that is, by examination of the fracture profile and secondary cracking.

Detailed observation of the fracture surface is best accomplished by use of the scanning electron microscope (SEM) or by examination of replicas with the transmission electron microscope (TEM). However, lack of access to a SEM or TEM should not be viewed as a crippling obstacle to performing failure analysis, because such work was done successfully prior to the development of these instruments. In many studies, such equipment is not needed, while in other cases, they are very important tools. In most cases, a more thorough job can be accomplished using such tools.

The LM, on the other hand, is a virtually indispensable tool for the failure analyst. The SEM and TEM find application in microscopy when the required magnification/resolution exceed the capabilities of the LM (about 1000×), but the first tool of choice is the LM for microstructural examination and identification. Electron microscopy and LM are complementary tools. Microstructural examination can be performed with the SEM over the same magnification range as the LM, but examination with the latter is more efficient. Contrast mechanisms for viewing microstructures are different for LM and SEM. Many microstructures, for example, tempered martensite, exhibit poor contrast in the SEM and are best viewed by light microscopy. When atomic number contrast or topographic contrast is strong, the SEM provides good structural images, particularly above 500× (Ref 3). Again, because of the limitations and advantages of each instrument, they are complementary rather than competitive tools. All studies of microstructures and fractures should begin at the lowest magnification level, the unaided human eye, and progress upward, first using the stereomicroscope for fractures and the LM for fracture path and microstructural studies, before using electron metallographic equipment.

References cited in this section

1. G.F. Vander Voort, *Met. Eng. Q.*, Vol 15, May 1975, p 31–36

2. G.F. Vander Voort, *Pr. Fail. Anal.*, Vol 1 (No. 2), April 2001, p 14–19, 38–46

3. G.F. Vander Voort, The SEM as a Metallographic Tool, *Applied Metallography*, Van Nostrand Reinhold Co., 1986, p 139–170

George F. Vander Voort, Buehler Ltd.

Examination of Fractures

Microfractography is a relatively new field; its roots can be traced to the light optical fractographs published by Zapffe and coworkers (Ref 4) beginning in the early 1940s, although a few studies of historical value predated their efforts. Zapffe's work, however, was almost exclusively confined to observation of cleavage facets on rather brittle, coarse-grained specimens. The technique, basically an interesting academic exercise, did stimulate interest in fracture examination as part of failure analysis. However, the depth-of-field limitation of the light microscope has restricted its use for such work. Aside from the published light optical fractographs made by Zapffe (see Ref 5 for a review of many of these), very few optical fractographs of metallic materials have been published by others. Microfractography gained momentum with the development of TEM replication methods and became commonplace after the commercial introduction of the SEM in approximately 1965.

A flat, brittle fracture can be examined with the light microscope by orienting the fracture perpendicularly to the optical axis. It is best to start with a low-power objective; long-working distance types are preferred. Focusing reveals the limitations of the method, because only part of the fracture is in focus at any setting. Thus, photographs reveal only a portion of the fracture in focus, depending on the coarseness and orientation of the fracture facets. Figure 1 shows an example of a brittle fracture in a low-carbon steel examined in this manner. Figure 1 also shows a LM image of the fracture profile, a LM image of a replica of the fracture, SEM images of the fracture and a replica of the fracture, and a TEM replica of the fracture. Although Zapffe used bright-field illumination for this work, dark-field illumination often produces superior results. Figure 2 illustrates the use of bright-field and dark-field illumination for viewing a brittle fracture in an Fe-Cr-Al alloy, plus a SEM fractograph of the same area. Dark-field illumination is better at collecting the light scattered from the fracture features; glare is reduced, and image contrast is improved.

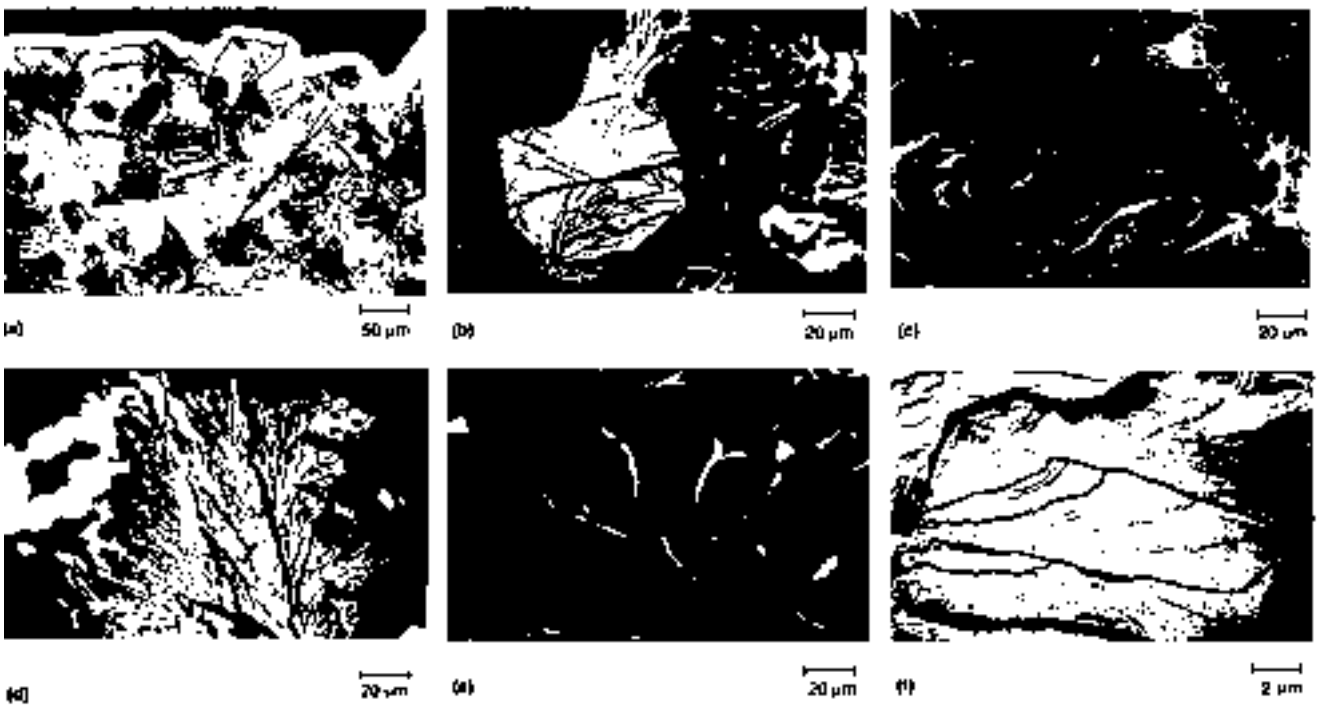


Fig. 1 Illustration of a cleavage fracture in a quenched and tempered low-carbon steel examined using three direct methods and three replication methods. (a) LM cross section

(nickel plated). Etched with Vilella's reagent. (b) LM fractrograph (direct). (c) SEM fractograph (direct). (d) LM replica. (e) SEM replica. (f) TEM replica

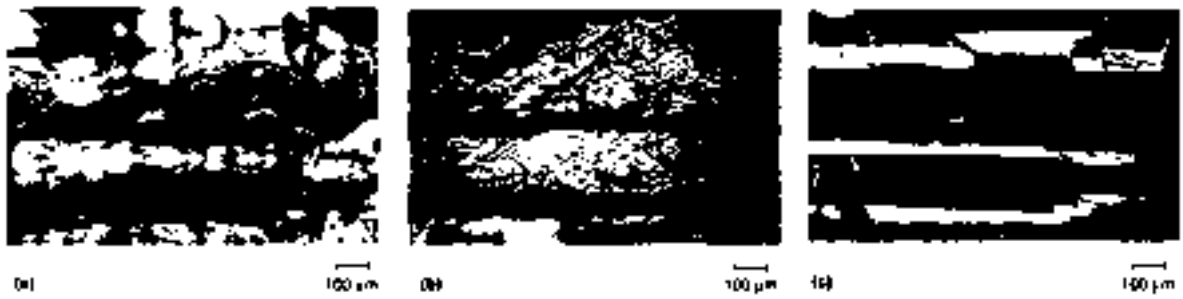


Fig. 2 Light microscope fractographs taken with (a) bright-field and (b) dark-field illumination compared to (c) a SEM secondary-electron image fractograph of the same area. Sample is an Fe-Al-Cr alloy.

Examination of fracture replicas with the light microscope (Ref 5, 6, 7) can extend the use of the method only to a limited extent, because the replica collapses slightly, producing less depth of field. Also, with a replica, the risk of damaging the objective is eliminated.

Considerable information concerning the fracture mode and the relationship of the microstructure to the fracture path can be obtained by LM examination of the profile of partially fractured (Ref 8, 9, 10) or completely fractured (Ref 11, 12, 13, 14, 15) polished metallographic specimens. Such examinations have been conducted for many years, long before the development of electron metallographic techniques, and continue to be used because of the value of the method. If the fracture has progressed to complete rupture, so that only one side of the fracture is to be examined, it may be best to nickel plate the fracture to enhance edge retention. This is not required if the crack has not separated the component into two pieces, or if a secondary crack is to be examined.

Figure 3(b) shows a nickel-plated brittle impact fracture of a low-carbon steel with a ferrite-pearlite microstructure. The crack consists of numerous connected straight-line segments in the ferrite phase. Several subsurface cleavage cracks are also present. In comparison, a ductile impact fracture in a quenched and tempered carbon plate steel is shown in Fig. 3(a). Note that the fracture surface exhibits a much rougher appearance due to the linking up of the microvoids. Below the fracture surface, spherical ruptures are frequently seen, which are also indicative of a ductile fracture mechanism.

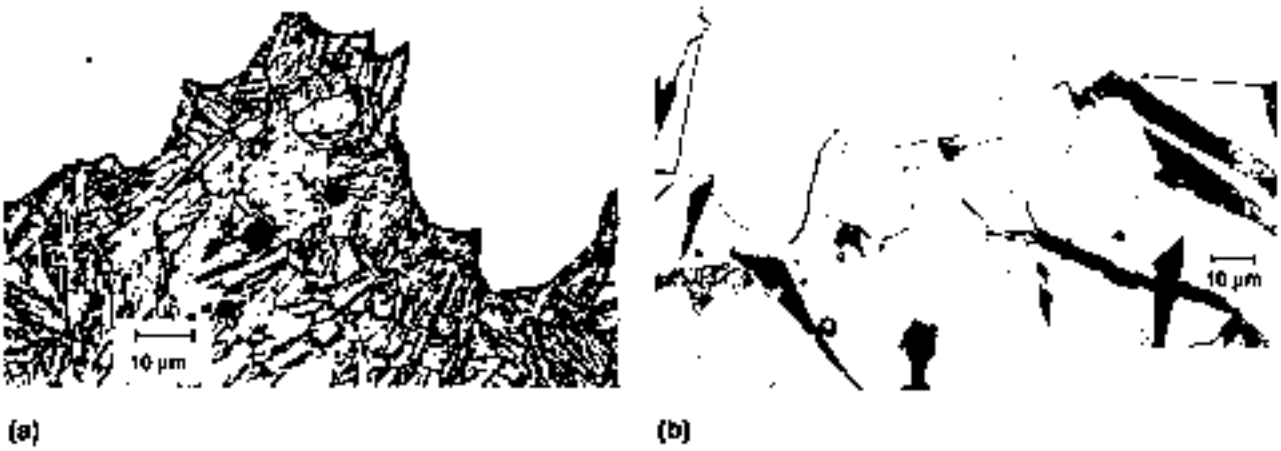


Fig. 3 Light micrographs of section profiles of (a) a nickel-plated ductile fracture and (b) a nickel-plated brittle fracture. Both are carbon steels etched with 2% nital.

Fatigue fractures can also be examined using fracture profiles, as illustrated in Fig. 4. This shows a low-cycle fatigue crack that has not propagated to final rupture. Note that the crack exhibits some minor branching and does not traverse the ferrite grains in a straight-line fashion, as with cleavage (Fig. 3b). Also, there are no spherical cavities near the crack, as observed with ductile fractures (Fig. 3a). The crack shows no preference for either the ferrite or pearlite during its growth, and the macroscopic appearance of the crack is rather flat, which is typical of fatigue cracks.

Fig. 4 Light micrograph of the path of a fatigue crack through a low-carbon steel specimen. Etched with 2% nital

Figure 5 shows partially broken and completely broken sensitized (649 °C, or 1200 °F, for 4 h) impact specimens of American Iron and Steel Institute (AISI) 304 stainless steel. Scanning electron microscope examination of the fracture face reveals extensive microvoid coalescence, that is, ductile rupture, although the impact strength (at -196 °C, or -320

°F) was only 40% of that of a nonsensitized sample. The partially broken sample reveals a fracture path that often follows the grain boundaries. Rupture cavities are observed behind and ahead of the crack; most are associated with coarse carbides. The fracture profile of the completely broken specimen reveals microvoid coalescence.

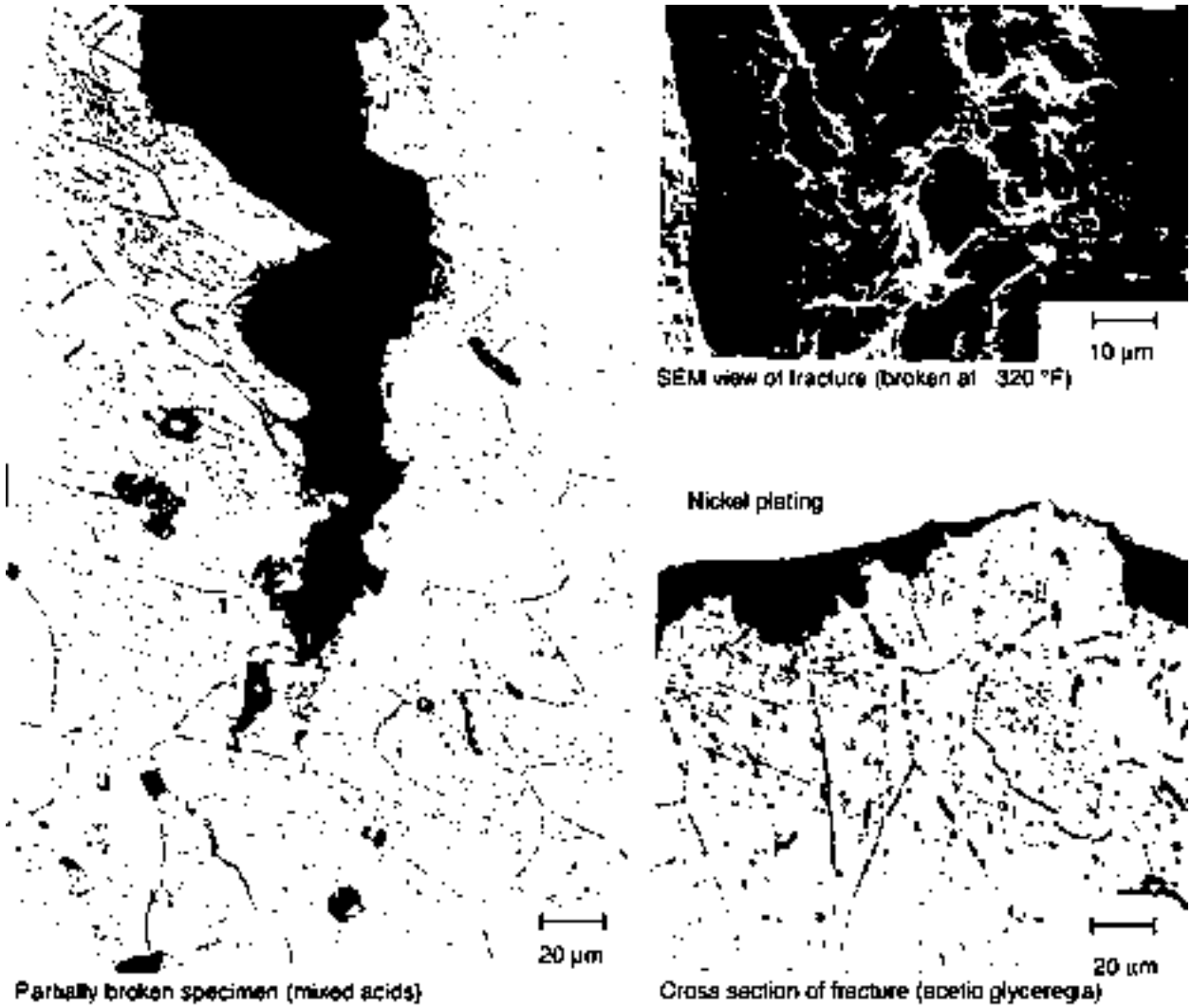


Fig. 5 Light micrographs of a partially broken and a completely broken specimen of sensitized (649 °C, or 1200 °F, for 4 h) AISI 304 austenitic stainless steel, and a SEM fractograph of the broken Charpy V-notch specimen

As a comparison to Fig. 5, Fig. 6 shows a similar fracture made in AISI 312 stainless steel weld metal that was aged at 816 °C (1500 °F) to transform the delta ferrite to sigma phase. In this case, the impact strength at room temperature was only 7% of that of an as-welded sample containing austenite and delta ferrite. Note that the fracture path is microscopically flatter and consists of numerous connected short straight-line segments. The SEM fractograph shows numerous small, flat fracture regions indicative of the embrittlement due to sigma. The high-magnification LM profile view of the completely fractured surface reveals the presence of extensive sigma along the crack path.

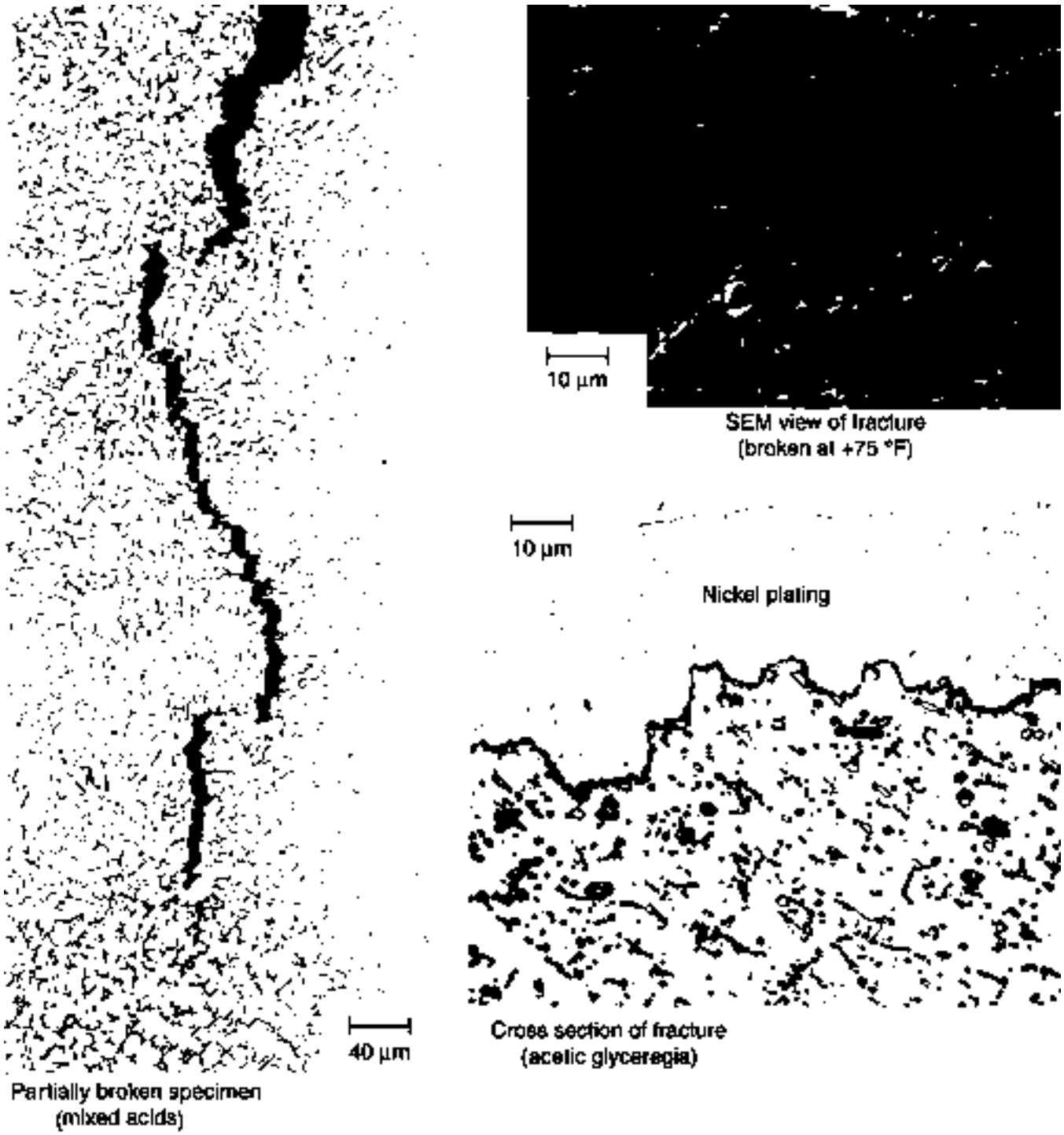


Fig. 6 Light micrographs of a partially broken and a completely broken specimen of AISI 312 stainless steel weld metal heat treated to transform the delta ferrite to sigma, and a SEM fractograph of the broken Charpy V-notch specimen

Examination of the crack path using cross sections is also very useful for study of fractures due to environmental problems. Figure 7 shows a stress-corrosion crack in a partially broken sample of solution-annealed AISI 304 stainless steel tested in boiling (151 °C, or 304 °F) magnesium chloride. The crack path is predominantly intergranular, but considerable transgranular fracture is also present. The SEM fractograph of the specimen clearly reveals the intergranular nature of the crack.

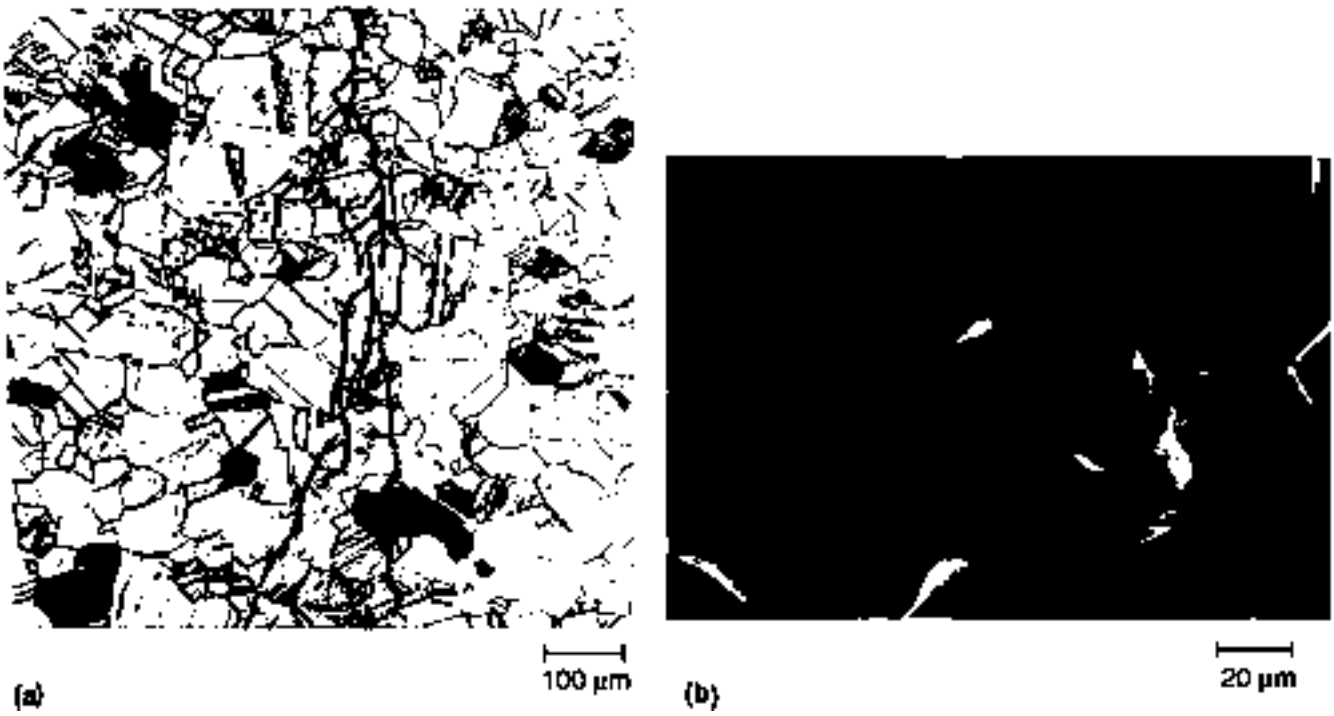


Fig. 7 Light micrograph of a cross section of (a) a partially broken specimen and (b) a SEM fractograph of a completely broken specimen of solution-annealed AISI 304 stainless steel after stress-corrosion crack testing in boiling (151 °C, or 304 °F) magnesium chloride

Fracture profile examination is also very useful in the study of failures due to liquid metal embrittlement (LME). Figure 8 shows the microstructure adjacent to a LME crack in a eutectoid steel where liquid copper has penetrated the grain boundaries at 1100 °C (2012 °F) while the sample was austenitic and under an applied tensile load. Light microscope examination reveals a discontinuous film of copper in the prior-austenite grain boundaries and an intergranular fracture path. Scanning electron microscope examination of the fracture also reveals the intergranular nature of the crack path.

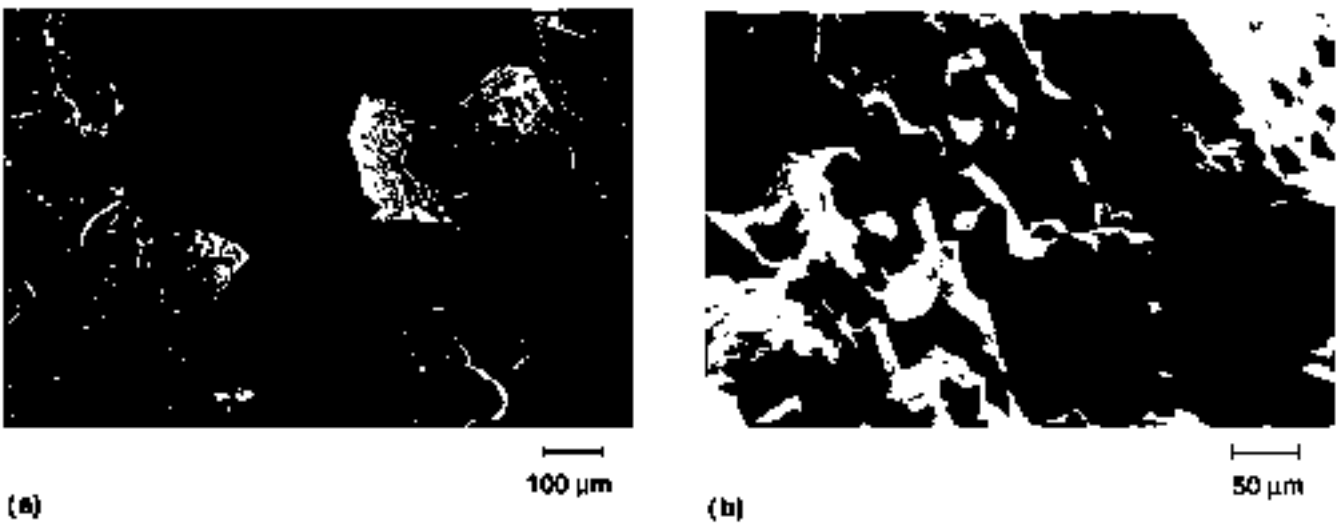


Fig. 8 Light micrograph of (a) a partially broken eutectoid carbon steel specimen embrittled by liquid copper at 1100 °C (2012 °F) (arrows point to grain-boundary copper penetration), and (b) SEM fractograph of the completely broken specimen

Surface detail can also be studied by LM using taper sections (Ref 16, 17, 18). This method has been used to study wear phenomena, surface coatings, fatigue damage, and other fine surface detail. In this method, the surface is sectioned at a slight angle to the surface. Polishing on this plane produces a magnified view of the structure in the vertical direction. The degree of magnification is defined by the cosecant of the sectioning angle; an angle of 5° 43$ produces a tenfold magnification.

Considerable progress has been made in applying the principles of quantitative metallography to the study of fractures (Ref 14, 15, 19, 20, 21, 22, 23). Much of this work has used measurements made on polished sections taken parallel to the crack-growth direction. This work provides new insight into fracture processes and should be useful in failure analysis, although its application to date has been limited mainly to research studies.

References cited in this section

4. C.A. Zapffe and M. Clogg, *Trans. ASM,* Vol 34, 1945, p 71–107

5. K. Kornfeld, *Met. Prog.,* Vol 77, Jan 1960, p 131–132

6. P.J.E. Forsyth and D.A. Ryder, *Metallurgia,* March 1961, p 117–124

7. K.R.L. Thompson and A.J. Sedriks, *J. Aust. Inst. Met.,* Vol 9, Nov 1964, p 269–271

8. H.C. Rogers, *Trans. AIME,* Vol 218, June 1960, p 498–506

9. C. Laird and G.C. Smith, *Philos. Mag.,* Vol 7, 1962, p 847–857

10. R.H. Van Stone and T.B. Box, "Use of Fractography and Sectioning Techniques to Study Fracture Mechanisms," STP 600, ASTM, 1976, p 5–29

11. W. Staehle et al., *Corrosion,* Vol 15, July 1959, p 51–59 (373t–381t)

12. D. Eylon and W.R. Kerr, "Fractographic and Metallographic Morphology of Fatigue Initiation Sites," STP 645, ASTM, 1978, p 235–248

13. W.R. Kerr et al., *Met. Trans.,* Vol 7A, Sept 1976, p 1477–1480

14. W.T. Shieh, *Met. Trans.,* Vol 5, May 1974, p 1069–1085

15. J.R. Pickens and J. Gurland, Metallographic Characterization of Fracture Surface Profiles on Sectioning Planes, *Proc. Fourth International Congress for Stereology,* NBS Spec. Publ. 431, 1976, p 269–272

16. E. Rabinowicz, *Met. Ind.* Vol 76, 3 Feb 1950, p 83–86

17. L.E. Samuels, *Metallurgia,* Vol 51, March 1955, p 161–162

18. M.H. Hurdus, "Taper Sectioning of Tubular Specimens and Its Application to Corrosion Oxide Film Examination," Report AERE-R9704, U.K. Atomic Energy Authority, Harwell, Oct 1980

19. S.M. El-Soudani, *Metallography,* Vol 11, July 1978, p 247–336

20. E.E. Underwood and E.A. Starke, Jr., "Quantitative Stereological Methods for Analyzing Important Features in Fatigue of Metals and Alloys," STP 675, ASTM, 1979, p 633–682

21. E.E. Underwood and S.B. Chakrabortty, "Quantitative Fractography of a Fatigued Ti-28V Alloy," STP 733, ASTM, 1981, p 337–354

22. M. Coster and J.L. Chermant, *Int. Met. Rev.,* Vol 28, 1983, p 228–250

23. E.E. Underwood, Quantitative Fractography, *Applied Metallography,* Van Nostrand Reinhold, 1986, p 101–122

George F. Vander Voort, Buehler Ltd.

Metallographic Specimen Preparation

Because the metallographer cannot predict in advance what the microstructural examination will reveal, specimen preparation must be perfect; otherwise, critical information can easily be lost. This basic truth has been proven over and over, yet is violated regularly. The preparation procedure and the prepared specimen must provide the following:

- Deformation induced by sectioning, grinding, and polishing must be removed or be shallow enough to be removed by the etchant.
- Coarse grinding scratches must be removed; even very fine polishing scratches may not be tolerable in examining failed parts.
- Pullout, pitting, cracking of hard particles, smear, and other preparation artifacts must be avoided.
- Relief (i.e., excessive surface-height variations between structural features of different hardness) must be minimized; otherwise, portions of the image are out of focus at high magnifications. Excessive relief invalidates image analysis measurements and is undesirable for wavelength-dispersive chemical analysis.
- The surface must be flat, particularly at edges (if they are of interest), or they cannot be examined. Edge preservation is of critical importance in failure studies, because many failures start at external surfaces.
- Coated or plated surfaces must be kept flat if they are to be examined, analyzed, measured, or photographed.
- An etchant should be used that reveals all of the structure at first. Later, it may be useful to use a selective etchant that reveals only the phase or constituent of interest, or at least produces strong contrast or color differences between two or more phases present, to improve the precision of microstructural measurements or to better reveal the relative presence of undesirable constituents or phases.

If these characteristics are met, then the true structure is revealed and can be interpreted, measured, analyzed, and recorded. The preparation method should be as simple as possible, should yield consistent, high-quality results in a minimum of time and cost, and must be reproducible.
Preparation of metallographic specimens (Ref 24) generally requires five major operations: sectioning, mounting (optional), grinding, polishing, and etching (optional).

Sectioning

It is certainly not uncommon in failure analysis to encounter large specimens. Indeed, the initial sectioning operation may be quite a challenge. Bulk samples for subsequent laboratory sectioning may be removed from larger pieces using methods such as core drilling, band- or hacksawing, flame cutting, or similar methods. Flame or torch cutting may be the only recourse in the field. If this is done, the torch-cut area must be well away from the area to be examined, because the heat from this operation severely alters the original microstructure for some distance from the cut. Subsequent cutting can be performed with laboratory devices that are much less damaging to the structure. Laboratory abrasive-wheel cutting is recommended to establish the desired plane of polish.
The most commonly used sectioning device in the metallographic laboratory is the abrasive cutoff machine. All abrasive-wheel sectioning should be performed wet. An ample flow of coolant, with an additive for corrosion protection and lubrication, should be directed uniformly into the cut. Wet cutting produces a smooth surface finish and, most importantly, guards against excessive surface damage caused by overheating. Figures 9(a) and (b) show the surface of quenched and tempered A2 tool steel (59 HRC) cut without using coolant. The cut surface was nickel plated for edge preservation. Figure 9(a) shows a light-etching surface zone extending to a depth of approximately 0.22 mm (0.009 in.), with a hardness of approximately 62.5 HRC. Beneath the light-etching surface zone is a region that was softer (53 to 56 HRC) and etches darker. The unaffected matrix is beneath this zone, off the edge of the micrograph. Figure 9(b) shows the extreme surface at high magnification. Incipient melting can be observed to a depth of approximately 10 7m. The light-etching zone contains untempered martensite, because the temperature in this region was high enough to reaustenitize the structure. The dark-etching zone beneath it saw temperatures below the lower critical but greater than the original tempering temperature. Abrasive wheels should be selected according to the manufacturer's recommendations.

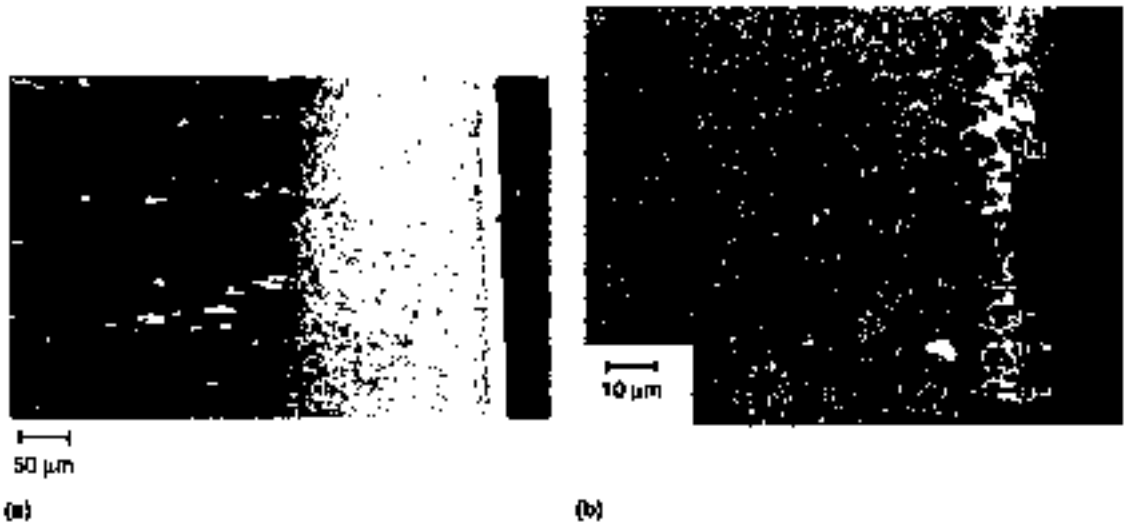


Fig. 9 Light micrographs of the surface (plated with nickel) of quenched and tempered A2 tool steel cut with an abrasive wheel without using coolant showing (a) a reaustenitized zone (light-etching area) and a back-tempered heat-affected zone (dark-etching area), and (b) details of incipient melting at the surface (arrows). Specimen etched with nital

Wheels consist of abrasive particles, chiefly alumina or silicon carbide (SiC), and filler in a binder material that may be a resin, rubber, or a mixture of resin and rubber. Alumina is the preferred abrasive for ferrous alloys, and SiC is the preferred abrasive for nonferrous metals and minerals. Wheels have different bond strengths and are recommended based on the suitability of their bond strength and abrasive type for the material to be sectioned. In general, as the hardness of a material increases, abrasives become dull more quickly, and the binder must be adjusted to release the abrasives when they become dull, so that fresh abrasive particles are available to maintain cutting speed and efficiency. Consequently, these wheels are called “consumable” wheels, because they wear away with use. If they do not wear at the proper rate, dull abrasives rub against the region being cut, generating heat and altering the existing true microstructure. If this heat becomes excessive, it can lead to grain or particle coarsening, softening or phase transformations, and, in extreme cases, burning or melting. Different materials have different sensitivities to this problem, but the need to balance the wheel break-down rate with the hardness of the piece being sectioned produces the various recommendations listed for cutting different materials and metals at different hardnesses, such as steels.

Precision saws are commonly used in metallographic preparation to section materials that are small, delicate, friable, extremely hard, or where the cut must be made as close as possible to a feature of interest, or where the cut width and material loss must be minimal. As the name implies, this type of saw is designed to make very precise cuts. They are smaller in size than the usual laboratory abrasive cutoff saw and use much smaller blades, typically from 7.6 to 20.3 cm (3 to 8 in.) in diameter. These blades can be of the nonconsumable type, made of copper-base alloys with diamond or cubic boron nitride abrasive bonded to the periphery of the blade, or they can be consumable blades using alumina or SiC abrasives with a rubber-based bond. Blades for the precision saws are much thinner than the abrasive wheels used in an abrasive cutter, and the load applied during cutting is much less. Consequently, less heat is generated during cutting, and damage depths are reduced. While pieces with a small section size that would normally be sectioned with an abrasive cutter can be cut with a precision saw, the cutting time is appreciably greater, but the depth of damage is much less. Precision saws are widely used for sectioning sintered carbides, ceramic materials, thermally sprayed coatings, printed circuit boards, electronic components, bone, teeth, and so on.

Mounting

The primary purpose of mounting metallographic specimens is for convenience in handling specimens of difficult shapes or sizes during the subsequent steps of metallographic preparation and examination. A secondary purpose is to protect and preserve extreme edges or surface defects during metallographic preparation. The method of mounting should in no way be injurious to the microstructure of the specimen. Pressure and heat are the most likely sources of injurious effects.

The most common mounting method uses a device, called a mounting press, to provide the required pressure and heat to encapsulate the specimen with a thermosetting or thermoplastic mounting material. Common thermosetting resins include phenolic, diallyl phthalate, and epoxy, while methyl methacrylate is the most commonly used thermoplastic mounting resin. Both thermosetting and thermoplastic materials require heat and pressure during the molding cycle, but, after curing, mounts made of thermoplastic resins must be cooled under pressure to at least 70 °C (158 °F), while mounts made of thermosetting materials may be ejected from the mold at the maximum molding temperature. However, cooling

thermosetting resins under pressure to near-ambient temperature before ejection significantly reduces shrinkage gap formation. Never rapidly cool a thermosetting resin mount with water after hot ejection from the molding temperature. This causes the metal to pull away from the resin, producing shrinkage gaps that promote poor edge retention (Fig. 10a, b) because of the different rates of thermal contraction.

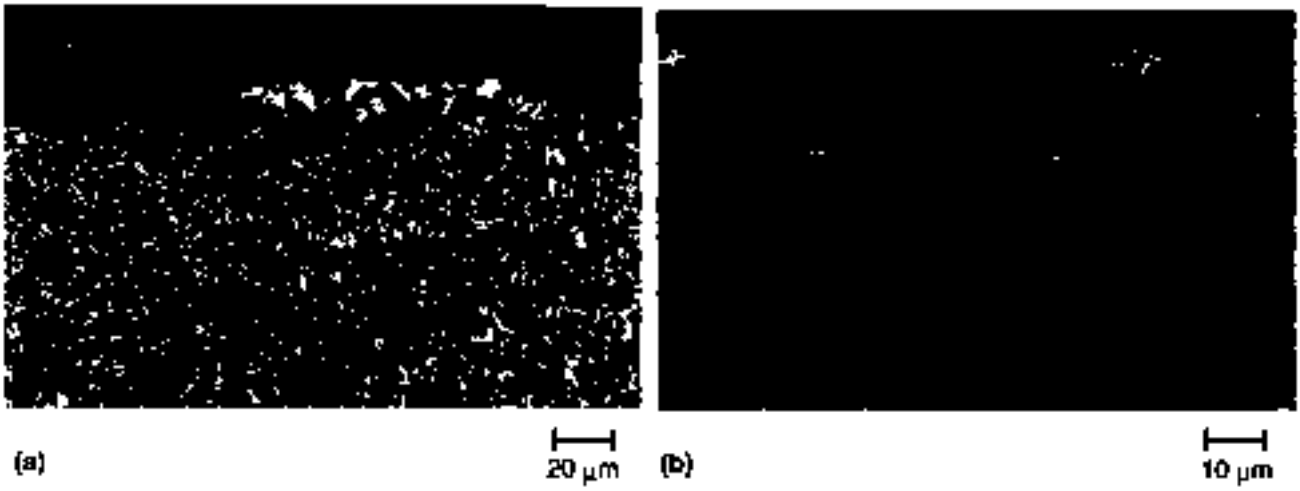


Fig. 10 Light micrographs of the surface of a carburized 8620 alloy steel specimen mounted in phenolic resin. Note the shrinkage gap (see arrows in a) that has reduced the edge flatness. In (b), taken at 1000×, decarburization at the surface has caused ferrite and pearlite to form, and this area is slightly out of focus. Specimen etched with nital

A thermosetting epoxy with special filler (Epomet, Buehler Ltd.) provides superior edge retention (Fig. 11a) compared to other resins (see Fig. 11(b) for a phenolic mount and Fig. 11(c) for methyl methacrylate) and is virtually unaffected by hot or boiling etchants, while phenolic resins are badly damaged. The thermoplastic resins, such as methyl methacrylate, produce a transparent mount, which is helpful when trying to grind to a specific feature, but provide poor edge retention (Fig. 11c). Electroless nickel plating is an effective method for improving edge retention, particularly for steels. However, if the area to be studied is white or with low contrast, it may be difficult to determine where the nickel plating ends and the surface begins, as shown in Fig. 11(d). Figure 12 shows an example of ion-nitrided hot work die steel with a brittle white-etching iron nitride surface layer that is quite visible when mounted in Epomet resin but would probably be very hard to detect if the surface had been plated with nickel.

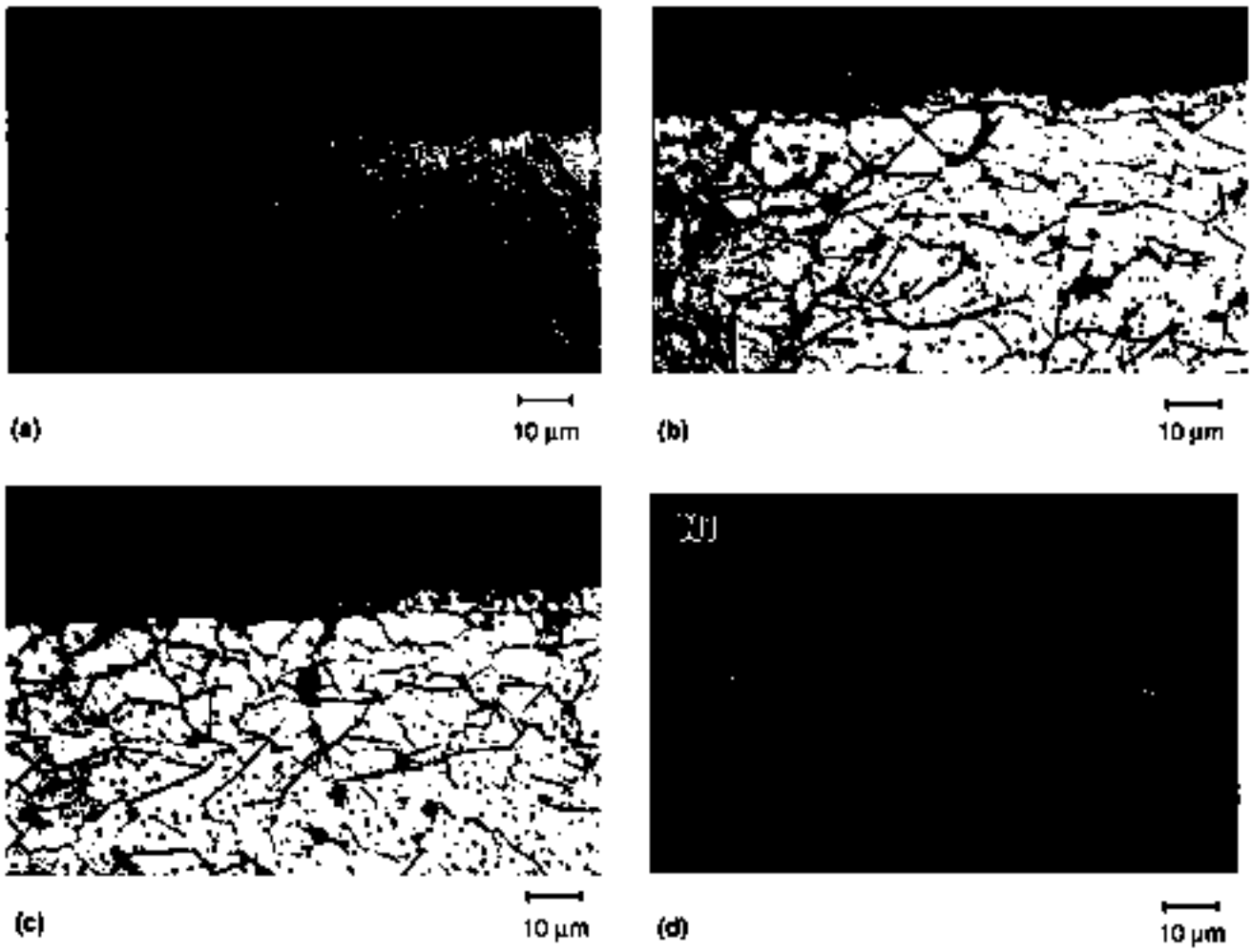


Fig. 11 Light micrographs of specimens of 1215 carbon steel that were salt bath nitrided and mounted in different resins. (a) Epomet thermosetting epoxy resin. (b) Phenolic thermosetting resin. (c) Methyl methacrylate thermoplastic resin. (d) Electroless nickel plated and mounted Epomet in epoxy resin (resin not in the field of view). All four specimens were prepared in the same holder and were etched with nital. The arrows point to the nitrided surface layer.

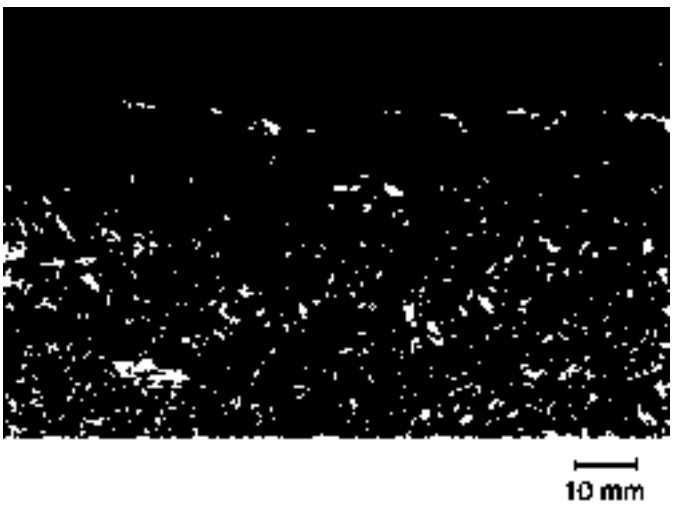


Fig. 12 Light micrograph of an ion-nitrided H13 tool steel specimen mounted in epoxy thermosetting resin (Epomet). The arrows point to a white-etching iron nitride layer at

the surface that probably would not have been observed if the specimen was nickel plated for edge protection. Specimen etched with nital

Shrinkage gaps between specimen and mount are a prime cause of loss of edge retention, as discussed subsequently. Besides this, abrasives can become lodged in the gap and fall out, causing contamination problems in a subsequent step. Further, liquids can seep out of the gaps, despite best efforts to dry the specimen carefully, and obscure the microstructural details at the edge, or, worse yet, drip onto the objective (in an inverted microscope), causing loss of image clarity or even damage. Figure 13 shows a large shrinkage gap between a phenolic mount and a piece of 6061-T6 aluminum etched with dilute aqueous hydrofluoric acid. Nomarski differential interference contrast (DIC) reveals the curvature at the edge of the specimen and water stains (arrows) along the edge of the specimen. Figure 14 shows a high-speed steel specimen in a phenolic mount, where a large shrinkage gap is present, and the etchant, Vilella's reagent, has seeped out and now obscures the edge detail (arrows).

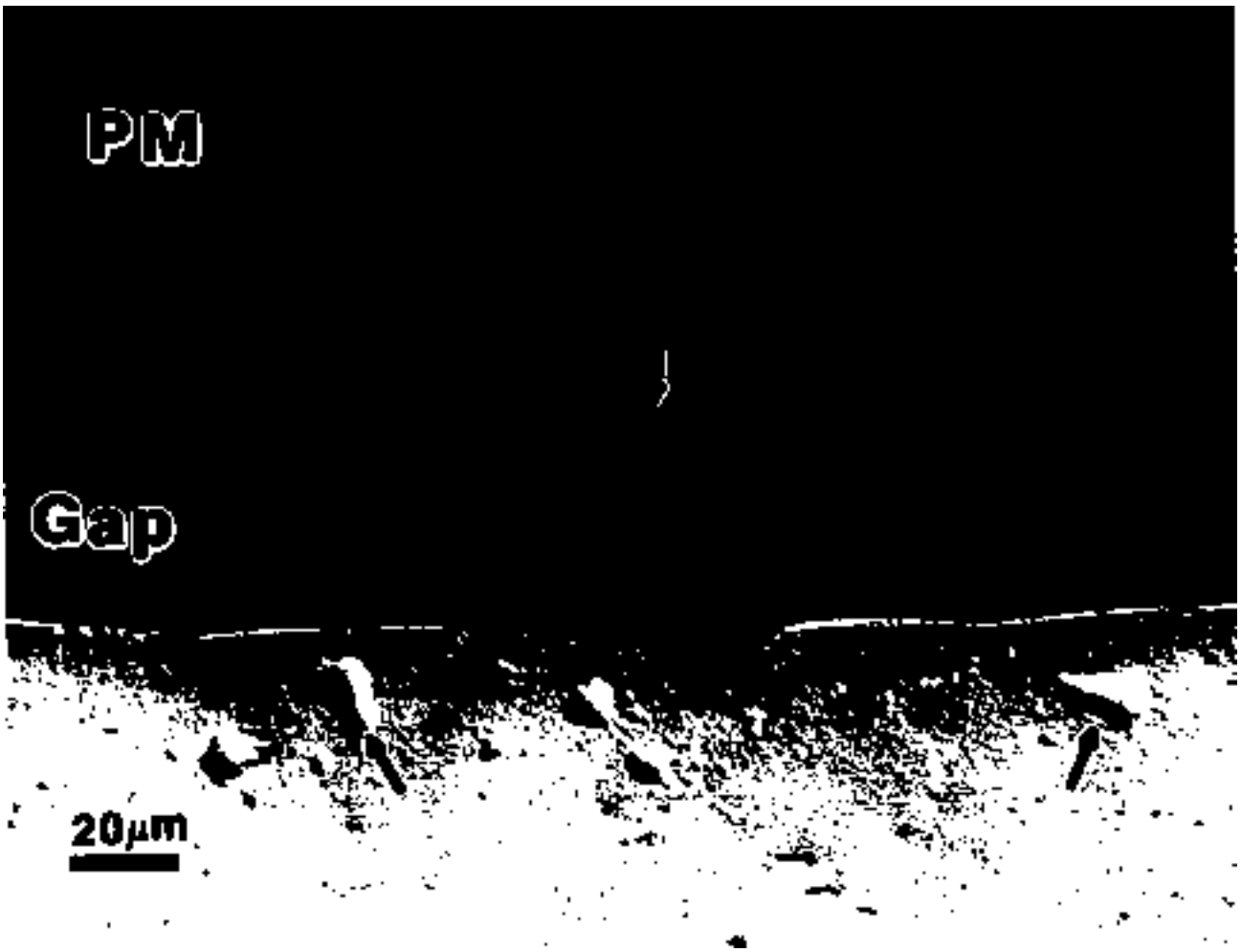


Fig. 13 Light micrograph showing a very large shrinkage gap between the phenolic resin mount (PM) and a specimen of 6061-T6 aluminum etched with aqueous 0.5% hydrofluoric acid. Note the metal flow at the specimen edge (revealed using Nomarski DIC illumination) and the water stains (arrows on the aluminum specimen).

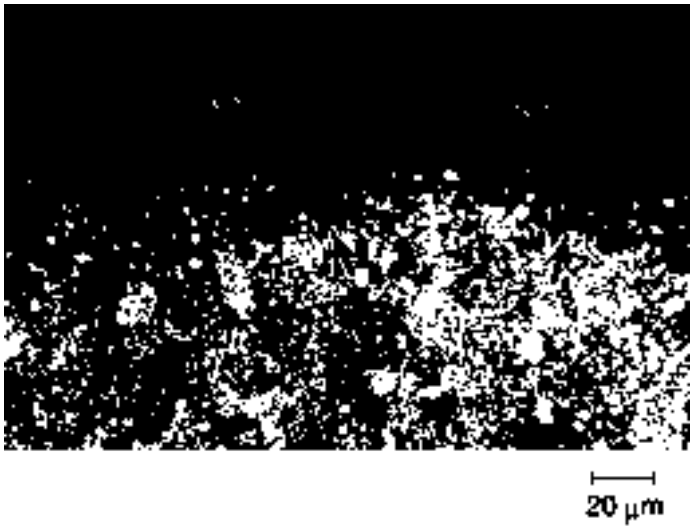


Fig. 14 Light micrograph showing stain (arrows pointing up) from the etchant (Vilella's reagent) that seeped from the shrinkage gap (wide arrows pointing down) between the phenolic resin mount and the specimen of M2 high-speed steel

An advantage of compression mounting is production of a mount of a predicable, convenient size and shape. Further, considerable information can be engraved on the backside—this is always more difficult with unmounted specimens. Manual (hand) polishing is simplified, because the specimens are easy to hold. Also, placing a number of mounted specimens in a holder for semi- or fully-automated grinding and polishing is easier with standard mounts than for unmounted specimens. Mounted specimens are easier on the grinding/polishing surfaces than unmounted specimens.

Cold-mounting materials require neither pressure nor external heat and are recommended for mounting specimens that are sensitive to heat and/or pressure. Acrylic resins are a widely used castable resin, due to their low cost and short curing time, but they are generally unsatisfactory for failure studies, because shrinkage is a problem with acrylics. Epoxy resins, although more expensive than acrylics, are commonly used in failure studies, because epoxy physically adheres to specimens and can be drawn into cracks and pores, particularly if a vacuum impregnation chamber is employed and a low viscosity epoxy is used. Epoxies are very suitable for mounting fragile or friable specimens and corrosion or oxidation specimens. Dyes or fluorescent agents may be added to epoxies for the study of porous specimens such as thermal spray coated specimens. Epoxy resins are much more useful in failure analysis work than acrylic resins.

Most epoxies are cured at room temperature, and curing times can vary from 2 to 20 h. Some can be cured at slightly elevated temperatures in less time, as long as the higher temperature does not adversely affect the specimen. Acrylics do generate considerable heat during curing, and this can be strongly influenced by the molding technique used. Castable epoxy resins generate much less heat during curing, but this can vary substantially. The amount of heat generated increases as the epoxy volume increases and as the curing time decreases.

Edge Preservation

Edge preservation is the classic metallographic problem in failure analysis work, and many “tricks” have been promoted (most pertaining to mounting but some to grinding and polishing) to enhance edge flatness. These methods include the use of backup material in the mount, the application of coatings to the surfaces before mounting, or the addition of a filler material to the mounting resin. Plating of a compatible metal on the surface to be protected (electroless nickel has been widely used) is generally considered to be the most effective procedure. However, image contrast at an interface between a specimen and the electroless nickel may be inadequate for certain evaluations. Figures 11(a) and (d) show the surface of a specimen of 1215 free-machining steel that was salt bath nitrided. One specimen was plated with electroless nickel; both were mounted in epoxy resin. It is hard to tell where the nitrided layer stops for the plated specimen (Fig. 11d), which exhibits poor image contrast between the nickel and the nitrided surface. This is not a problem for the nonplated specimen (Fig. 11a).

Introduction of new technology has greatly reduced edge preservation problems. Mounting presses that cool the specimen to near-ambient temperature under pressure produce much tighter mounts. Gaps that form between specimen and resin are a major contributor to edge rounding, as shown in Fig. 10. Staining from bleed-out at shrinkage gaps obscures edge detail, (Fig. 14). Use of semiautomatic and automatic grinding/polishing equipment, rather than manual (hand) preparation, increases surface flatness and edge retention. To achieve the best results, particularly with a 200 mm (8 in.) diameter platen and a 125 mm (5 in.) diameter holder, the position of the specimen holder relative to the platen should be adjusted

so that the outer edge of the specimen holder rotates out over the edge of the surface on the platen during grinding and polishing. This procedure can be used effectively with larger-diameter wheels, if the specimen holder diameter is large, relative to the platen. The use of “hard,” woven or nonwoven, napless surfaces for polishing with diamond abrasives (rather than softer cloths, such as canvas, billiard, and felt) maintains flatness. Rigid grinding discs (RGDs) yield surfaces with exceptional flatness. Final polishing with low-nap cloths for short times introduces very little rounding, compared to use of higher-nap, softer cloths.

These procedures produce better edge retention with all thermosetting and thermoplastic mounting materials. Nevertheless, there are still differences among the polymeric materials used for mounting. Thermosetting resins provide better edge retention than thermoplastic resins. Of the thermosetting resins, diallyl phthalate provides little improvement over the much-less-expensive phenolic compounds. The best results are obtained with an epoxy-based thermosetting resin that contains a special, hard filler material (Epomet). For comparison, Fig. 11 shows micrographs of a salt bath nitrided 1215 steel specimen mounted in a phenolic resin (Fig. 11b) and in methyl methacrylate (Fig. 11c) at 1000×. These specimens were prepared in the same specimen holder as those shown in Fig. 11(a) and (d), but neither displays acceptable edge retention at 1000×.

In the 1970s, very fine alumina spheres were mixed with liquid epoxy in an effort to improve edge retention. This is not a satisfactory procedure, because the particles are extremely hard (~2000 HV) and their grinding/polishing characteristics are incompatible with softer metals placed inside the mount. As a result, this product is no longer promoted for improving the edge retention of metallic specimens. Recently, a soft ceramic shot (~775 HV) was introduced that has grinding/polishing characteristics compatible with metallic specimens placed in the mount. Figure 15 shows an example of improving edge retention of annealed hot work die steel using soft ceramic shot in an epoxy mount.

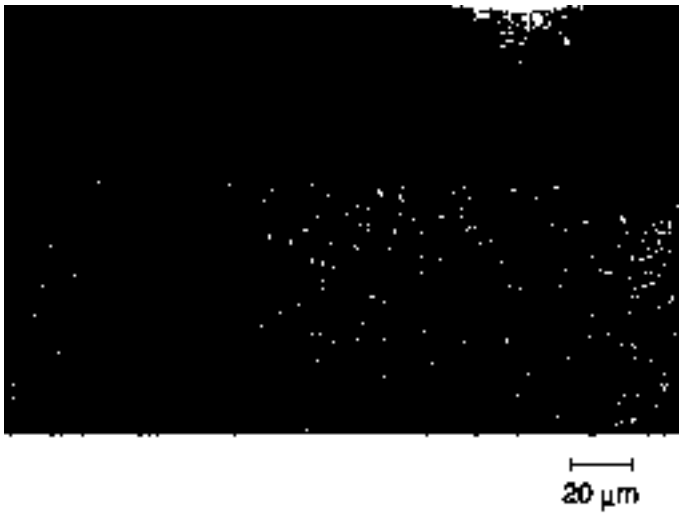


Fig. 15 Good edge retention obtained in a cast epoxy mount containing soft ceramic shot filler. (Note the round particles in the epoxy at the top.) The specimen is annealed H13 hot work die steel, and it was etched with picral.

Following are general guidelines for obtaining the best possible edge retention. All of these factors contribute to the overall success, although some are more critical than others:

- Properly mounted specimens yield better edge retention than unmounted specimens, because rounding is difficult, if not impossible, to prevent at a free edge. Hot compression mounts yield better edge preservation than castable resins.
- Electrolytic or electroless plating of the surface of interest provides excellent edge retention. If the compression mount is cooled too quickly after polymerization, the plating may be pulled away from the specimen, leaving a gap. When this happens, the plating is ineffective for edge retention.
- Thermoplastic compression mounting materials are less effective than thermosetting resins. The best thermosetting resin for edge retention is an epoxy-based resin (Epomet) containing a hard filler material.
- Do not hot eject a thermosetting resin after polymerization and cool it quickly to ambient (e.g., by cooling it in water), because a gap forms between specimen and mount due to the differences in thermal contraction rates. Fully automated mounting presses cool the mounted specimen to near-ambient temperature under pressure, and this greatly minimizes gap formation due to shrinkage.
- Automated grinding/polishing equipment produces flatter specimens than manual (hand) preparation.
- Use the central force mode (defined later in this article) with an automated grinder/polisher, because this method provides better flatness than individual pressure mode (defined later in this article).

- Orient the position of the smaller-diameter specimen holder so that, as it rotates, its periphery slightly overlaps the periphery of the larger-diameter platen.
- Use pressure-sensitive-adhesive (PSA)-backed SiC grinding paper (when SiC is used) rather than water on the platen and a peripheral holddown ring, and PSA-backed polishing cloths rather than stretched cloths.
- Metal-bonded or resin-bonded diamond grinding discs produce excellent flat surfaces for a wide variety of materials.
- Use hard, napless surfaces for rough polishing (until the final polishing step(s)) and fine polishing. Use a napless or a low- to medium-nap cloth, depending on the material being prepared, for the final step(s), and keep it brief.
- Rigid grinding disks (RGD) produce excellent flatness and edge retention and should be used whenever possible.

Grinding

Grinding should commence with the finest grit size that establishes an initially flat surface and removes the effects of sectioning within a few minutes. An abrasive grit size of 180 or 240 in the American National Standards Institute/Coated Abrasive Manufacturers' Institute (ANSI/CAMI) grading system (P180 or P280 in the FEPA, or Fédération Européenne des Fabricants de Produits Abrasifs system) is coarse enough to use on specimen surfaces sectioned by an abrasive cutoff wheel. Hacksawed, bandsawed, or other rough surfaces usually require abrasive grit sizes in the range of 120- to 180-grit (P120 to P180). Grinding must remove the damage created by sectioning (Fig. 16). If the initial grinding step does not remove this layer, the plane of polish may be within the zone of surface damage from cutting, and the true structure is not observed. The abrasive used for each succeeding grinding operation should be one or two grit sizes smaller than that used in the preceding step. A satisfactory grinding sequence might involve SiC papers with grit sizes of 240-, 320-, 400-, and 600-grit (P280, P400, P800, and P1200, respectively). This sequence is used in the “traditional” preparation approach.

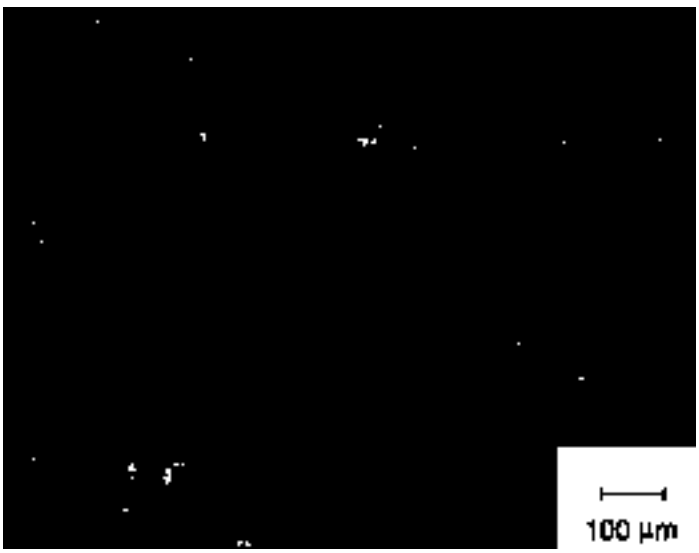


Fig. 16 Light micrograph showing cutting damage (arrows at left) and a burr at the corner of a specimen of commercial-purity titanium (ASTM F67, grade 2) etched with modified Weck's reagent and viewed with polarized light plus sensitive tint. The arrow along the top edge points to a surface layer containing mechanical twins.

As with abrasive-wheel sectioning, all grinding steps should be performed wet using water, provided that water has no adverse effects on any constituents of the microstructure. If water cannot be used during grinding, then some other non-aqueous coolant must be used, for example, kerosene or mineral spirits. Wet grinding minimizes specimen heating, prevents the abrasive from becoming loaded with metal removed from the specimen being prepared, and minimizes airborne metal-particle contamination and health problems.

Each grinding step, while producing damage itself, must remove the damage from the previous step. The depth of damage decreases with the abrasive size but so does the metal removal rate. For a given abrasive size, the depth of damage introduced is greater for soft materials than for hard materials. Grinding abrasive can become entrapped, or embedded, in the surface of specimens. This is especially true for soft, low-melting-point alloys ground using SiC paper. Embedding is more common with the finer-grit-size papers. Figure 17 illustrates embedding of SiC grinding paper abrasive in soft metals. The electron microprobe was used to study embedding of abrasives during grinding, and coating of the paper with candlewax or soap greatly reduced embedding (Ref 25).

Fig. 17 Light micrograph showing a SiC grinding-abrasive particle (arrow) lodged in a weldment in 6061-T6 aluminum etched with aqueous 0.5% hydrofluoric acid

For automated preparation using a multiple-specimen holder, the initial step is called planar grinding. This step must remove the damage from sectioning while establishing a common plane for all of the specimens in the holder, so that each specimen is affected equally in subsequent steps. Silicon carbide and alumina abrasive papers are commonly used for the planar grinding step and are very effective. Besides these papers, there are a number of other options available. One option is to planar grind the specimens with a conventional alumina grinding stone. This requires a special-purpose machine, because the stone must rotate at a high speed, / 1500 rpm, to cut effectively. The stone must be dressed regularly with a diamond tool to maintain flatness, and embedding of alumina abrasive in specimens can be a problem. However, the approach provides high removal rates.

Polishing

Polishing is the final stage in producing a deformation-free surface that is flat, scratch-free, and mirrorlike in appearance. Such a surface is necessary for subsequent metallographic interpretation, both qualitative and quantitative. The polishing technique used should not introduce extraneous structures such as disturbed metal (Fig. 18), pitting (Fig. 19), dragging out of inclusions, “comet tailing” (Fig. 20), staining (Fig. 21), relief (height differences between different constituents, or between holes and constituents) (Fig. 22), or embedding (Fig. 23). Polishing usually is conducted in several stages. Rough polishing is conducted with 9, 6, or 3 7m diamond abrasives charged onto napless or low-nap cloths. For hard materials, such as through-hardened steels, ceramics, and cemented carbides, two rough-polishing steps may be required. The initial rough-polishing step may be followed by polishing with 1 7m diamond on a napless, low-nap, or medium-nap cloth. A compatible lubricant should be used sparingly to prevent overheating or deformation of the surface. Intermediate polishing should be performed thoroughly, so that final polishing may be of minimal duration.

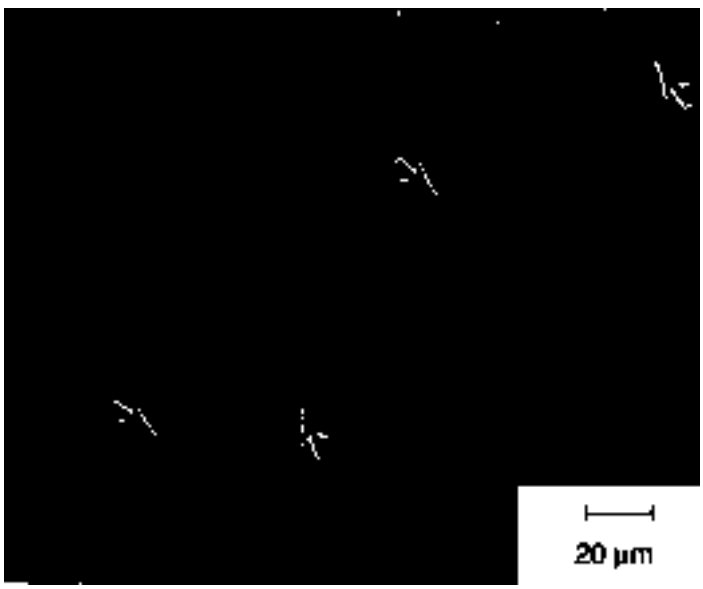


Fig. 18 Light micrograph showing residual damage (arrows) from preparation that was not removed by the procedure when this specimen of commercial-purity titanium was prepared. The specimen was etched with Kroll's reagent and photographed with Nomarski DIC illumination.

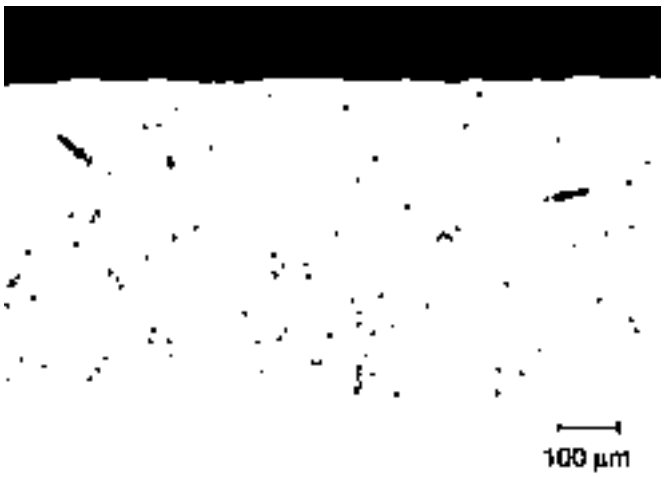


Fig. 19 Light micrograph showing pitting (arrows) after preparation of cold-drawn brass (Cu-20%Zn). Not etched

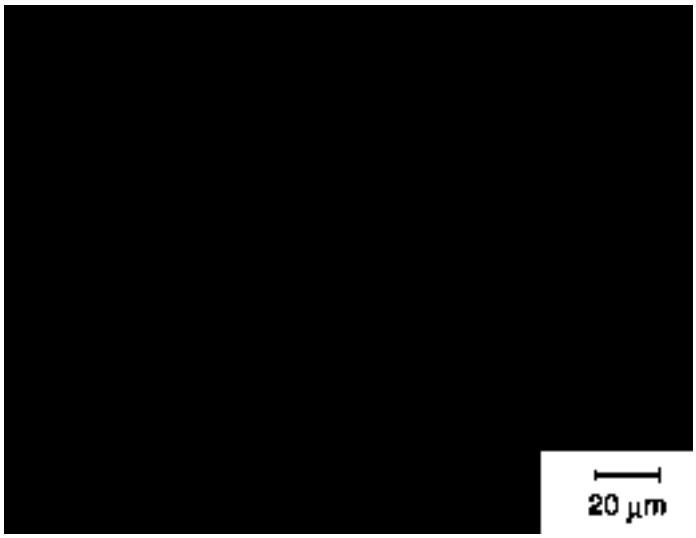


Fig. 20 Light micrograph illustrating “comet tails” emanating from hard nitrides on the surface of a prepared specimen of H13 tool steel. The specimen is unetched and viewed with Nomarski DIC.

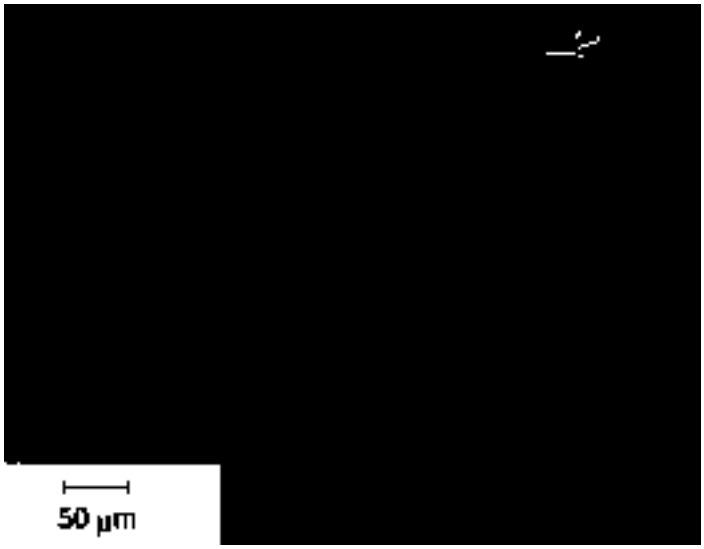


Fig. 21 Light micrograph illustrating staining (arrow) on the surface of a prepared specimen of Ti-6%Al-2%Sn-4%Zr-2%Mo. The specimen was not etched.

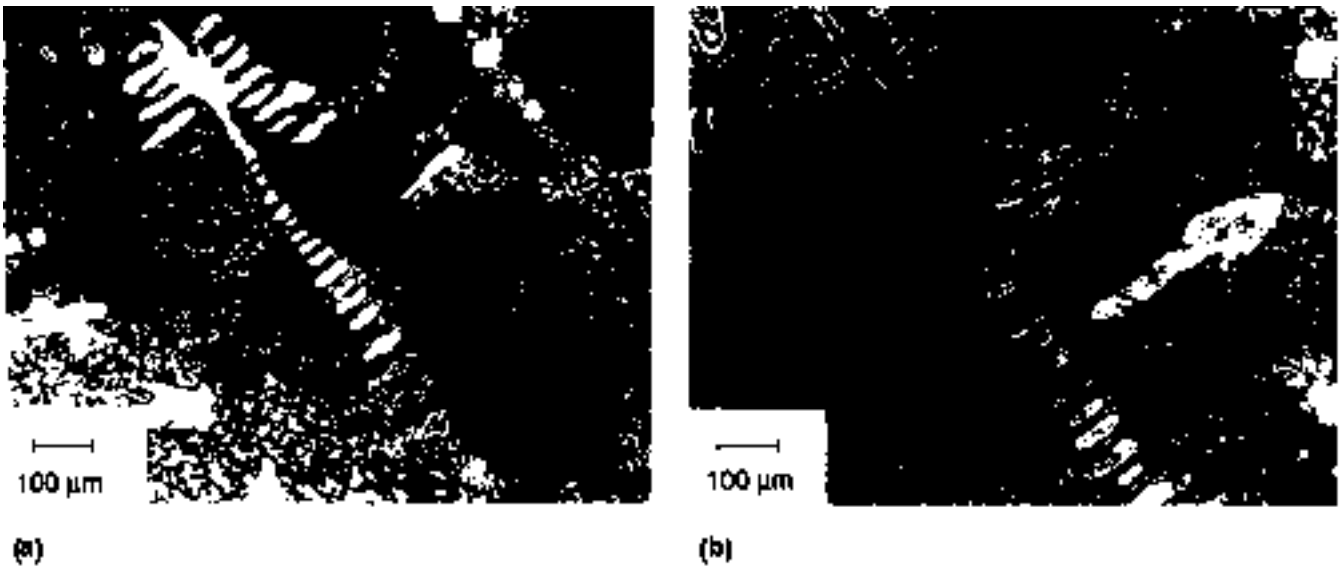


Fig. 22 Light micrographs depicting (a) excessive and (b) low relief around voids in a braze between an austenitic stainless steel and Monel. The specimen was etched with glyceregia.

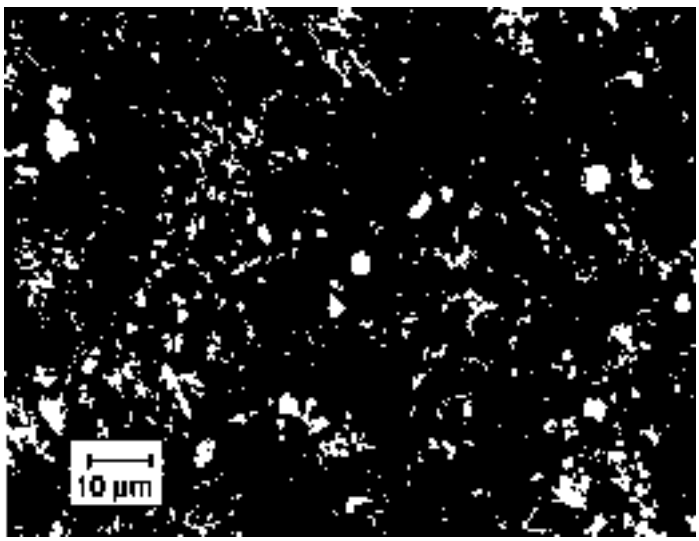


Fig. 23 Light micrograph showing 6 %m diamond (arrows) abrasive embedded in the surface of a partly prepared specimen of lead

Manual polishing, or hand polishing, is usually conducted using a rotating wheel, where the operator rotates the specimen in a circular path counter to the wheel rotation direction. To obtain the best possible surfaces, it is necessary to use another step, typically with a 0.05 7m alumina or colloidal silica abrasive. This step can be performed on a wide variety of cloths. In the traditional approach, a medium-nap synthetic suede cloth is used. While this is still a useful cloth for many materials, increasing use is being made of synthetic polyurethane pads. This type of cloth is recommended when edge retention must be maximized.

The requirements of a good polishing cloth include the ability to hold the abrasive media, long life, absence of any foreign material that may cause scratches, and absence of any processing chemical (such as dye or sizing) that may react with the specimen. Many cloths of different fabrics, weaves, or naps are available for metallographic polishing. Napless or low-nap cloths are recommended for rough polishing with diamond abrasive compounds. Napless, low-, medium-, and, occasionally, high-nap cloths are used for final polishing. This step should be brief to minimize relief.

Mechanical polishing can be automated to a high degree using a wide variety of devices ranging from relatively simple systems to rather sophisticated, minicomputer, or microprocessor-controlled devices. Units also vary in capacity from a single specimen to a half-dozen or more at a time. These systems can be used for all grinding and polishing steps. These devices enable the operator to prepare a large number of specimens per day, with a higher degree of quality than by hand polishing and at reduced consumable costs. Automatic polishing devices produce the best surface flatness and edge retention. There are two automated approaches for holding specimens. Central force uses a specimen holder, with each specimen held in place rigidly. The holder is pressed downward against the preparation surface, with the force applied to the entire holder. Central force yields the best edge retention and specimen flatness. If the results after etching are inadequate, the specimens must be placed back in the holder, and the entire preparation sequence must be repeated. Instead of doing this, most metallographers repeat the final step manually and then re-etch the specimen. Individual force machines have a holder that does not hold the specimen rigidly. Specimens are placed inside a hole cut into a holder, and a piston comes down and presses the specimen against the working surface. Thus, the specimens can be removed and examined easily during the preparation cycle without losing planarity. This provides convenience if a step must be repeated, but the method is limited to mounted specimens, usually round mounts, and edge retention is not as good as with a central force holder.

Polishing usually involves the use of one or more of the following abrasives: diamond, alumina, and amorphous silicon dioxide in colloidal suspension. For certain materials, cerium oxide, chromium oxide, magnesium oxide, or iron oxide may be used, although these are used infrequently. With the exception of diamond, these abrasives are normally suspended in distilled water, but if the metal to be polished is not compatible with water, other suspensions, such as ethylene glycol, alcohol, kerosene, or glycerol, may be required. The diamond abrasive should be extended only with the product recommended by the manufacturer. Most diamond pastes and suspensions are water-based products, and these are suitable for most materials. However, oil-based diamond suspensions are needed to prepare materials sensitive to water.

Over the past forty years, a general procedure has been developed that is quite successful for preparing most metals and alloys. This method is based on grinding with SiC waterproof papers through a series of grits, then rough polishing with one or more diamond abrasive sizes, followed by fine polishing with one or more alumina suspensions of different particle size. This procedure is called the “traditional” method and is described in Table 1, listing equivalent ANSI/CAMI and FEPA grit sizes for the SiC paper.

Table 1 The traditional method for preparing most metals and alloys

Surface	Abrasive/size	Load, N (lbf)	Speed, rpm/direction	Time, min
Waterproof PSA paper	120/P120-grit SiC, water cooled	27 (6)	240–300 Comp	Until plane
Waterproof PSA paper	240/P280-grit SiC, water cooled	27 (6)	240–300 Comp	1–2
Waterproof PSA paper	320/P400-grit SiC, water cooled	27 (6)	240–300 Comp	1–2
Waterproof PSA paper	400/P800-grit SiC, water cooled	27 (6)	240–300 Comp	1–2
Waterproof PSA paper	600/P1200-grit SiC, water cooled	27 (6)	240–300	1–2

			Comp	
Canvas	6 7m diamond paste with lubricant	27 (6)	120–150 Comp	2
Billiard or felt cloths	1 7m diamond paste with lubricant	27 (6)	120–150 Comp	2
Synthetic suede pad	Aqueous 0.3 7m 8-alumina slurry	27 (6)	120–150 Comp	2
Synthetic suede pad	Aqueous 0.05 7m ’ -alumina slurry	27 (6)	120–150 Comp	2

Note: Comp, complementary (platen and specimen holder both rotate in the same direction)

This procedure is used for manual or automated preparation, although manual control of the force applied to a specimen would not be very consistent. Complementary motion means that the specimen holder is rotated in the same direction as the platen and does not apply to manual preparation, because this cannot be done. In manual preparation, the specimen is held still in grinding, aside from moving between the edge and the center. In manual polishing, the specimen is rotated clockwise, against the counterclockwise wheel rotation direction. Some machines can be set so that the specimen holder rotates in the direction opposite to that of the platen, called “contra.” This provides a more aggressive action but was not adopted when the traditional approach was automated. The traditional method is not rigid, because other polishing cloths may be substituted, and one or more of the polishing steps might be omitted. Times and pressures could be varied, as well, to suit the needs of the work or the material being prepared. This is the “art” of metallography.

New concepts and new preparation materials have been introduced that enable metallographers to shorten the process while producing better, more consistent results. Much of this effort has centered on reducing or eliminating the use of SiC paper in the grinding steps. In all cases, an initial grinding step must be used, but there is a wide range of materials that can be chosen instead of SiC paper. There is nothing wrong with the use of SiC for the first step, except that it has a short life. If an automated device is used that holds a number of specimens rigidly (central force), then the first step must remove the sectioning damage on each specimen and bring all of the specimens in the holder to a common plane. This first step is often called planar grinding. Silicon carbide paper can be used for this step, although more than one sheet may be needed. Alternatively, the metallographer could use alumina paper, an alumina stone on a dedicated high-speed grinder, metal- or resin-bonded diamond discs, stainless steel mesh cloth (diamond is applied during use), RGDs (diamond is applied during use), or lapping platens of several types (diamond is applied and becomes embedded in the surface during use).

In contemporary preparation methods, one or more steps using diamond abrasives on napless surfaces usually follow planar grinding. Pressure-sensitive-adhesive-backed silk, nylon, or polyester cloths are widely used. These give good cutting rates, maintain flatness, and minimize relief. Silk cloths provide the best flatness and excellent surface finishes relative to the diamond size used. Thicker hard, woven cloths are more aggressive, give nearly as good a surface finish, similar excellent flatness, and longer life than silk cloths. Synthetic chemotextile pads give excellent flatness and are more aggressive than silk. They are excellent for retaining second-phase particles and inclusions. Diamond suspensions are very popular with automated polishers, because they can be added easily during polishing, although it is still best to charge the cloth initially with diamond paste of the same size to get polishing started quickly.

Final polishing could be performed with a very fine diamond size, such as 0.1 7m diamond, depending on the material, the metallographer's needs, and personal preferences. Otherwise, final polishing is performed with colloidal silica or with alumina slurries using napless or low- to medium-nap cloths. For some materials, such as titanium and zirconium alloys, an attack polishing solution is added to the abrasive slurry to enhance deformation and scratch removal and improve polarized light response. Contra rotation (head moves in the direction opposite to the platen) is preferred, because the slurry stays on the cloth better, although this does not work if the head rotates at a high number of revolutions per minute (>90). Examples of generic preparation practices for many metals and alloys are found in Tables 2, 3, 4.

Table 2 Generic four-step contemporary practice for many metals and alloys

Surface	**Abrasive/size**	**Load, N** **(lbf)**	**Speed,** **rpm/direction**	**Time, min**
Waterproof PSA paper	120/P120-, 180/P180-, or 240/P280-grit SiC, water cooled	27 (6)	240–300 Comp	Until plane

Silk cloth	9 7m polycrystalline diamond suspension	27 (6)	120–150 Comp	5
Synthetic woven cloth	3 7m polycrystalline diamond suspension	27 (6)	120–150 Comp	4
Synthetic short nap cloth	~0.05 7m colloidal silica or sol-gel alumina suspensions	27 (6)	120–150 Contra	2

Note: Comp, complementary (platen and speciman holder both rotate in same direction). Contra, platen and specimen holder rotate in opposite directions

Table 3 Four-step contemporary practice for steels using a rigid grinding disc

Surface	**Abrasive/size**	**Load, N (lbf)**	**Speed, rpm/direction**	**Time, min**
Waterproof PSA paper	120/P120-, 180/P180-, or 240/P280-grit SiC, water cooled	27 (6)	240–300 Comp	Until plane
Rigid grinding disk[(a)]	9 7m polycrystalline diamond suspension	27 (6)	120–150 Comp	5
Synthetic woven cloth	3 7m polycrystalline diamond suspension	27 (6)	120–150 Comp	4
Synthetic short nap cloth	~0.05 7m colloidal silica or sol-gel alumina suspensions	27 (6)	120–150 Contra	2

Note: Comp, complementary (platen and specimen holder both rotate in the same direction). Contra, platen and specimen holder rotate in opposite directions.
(a) Disk for metals >175 HV

Table 4 Four-step contemporary practice for nonferrous metals using a rigid grinding disc

Surface	**Abrasive/size**	**Load, N (lbf)**	**Speed, rpm/direction**	**Time, min**
Waterproof PSA paper	240/P280- or 320/P400-grit SiC, water cooled	22 (5)	240–300 Comp	Until plane
Rigid grinding disk[(a)]	6 7m polycrystalline diamond suspension	22 (5)	120–150 Comp	5
Synthetic woven cloth	3 7m polycrystalline diamond suspension	22 (5)	120–150 Comp	4
Synthetic short nap cloth	~0.05 7m colloidal silica or sol-gel alumina suspensions	22 (5)	120–150 Contra	2

Note: Comp, complementary (platen and specimen holder both rotate in the same direction). Contra, platen and specimen holder rotate in opposite directions.
(a) Disk for metals >50 HV
The starting SiC abrasive size is chosen based on the degree of surface roughness and depth of cutting damage and the hardness of the material. Never start with an abrasive size coarser than necessary to remove the cutting damage and

achieve planar conditions in a reasonable time. A 1 7m diamond step can be added for more difficult-to-prepare materials using a napless cloth and a similar approach as the third step, but a 3 min polish.

A similar scheme can be developed using RGDs. These discs are generally restricted to materials above a certain hardness level, such as 175 HV, although some softer materials can be prepared using them. The disc can also be used for the planar grinding step. An example of such a practice, applicable to nearly all steels (results are marginal for solution-annealed austenitic stainless steels), is given in Table 3.

The planar grinding step could also be performed using a 45 7m metal-bonded or a 30 7m resin-bonded diamond disc or with a RGD and 15 or 30 7m diamond, depending on the material. Rigid grinding discs contain no abrasive; they must be charged during use, and suspensions are the easiest way to do this. Polycrystalline diamond suspensions are favored over monocrystalline synthetic diamond suspensions for most metals and alloys due to their higher cutting rate. Again, a 1 7m diamond step can be added for difficult materials or to ensure generation of the required degree of perfection in the surface finish.

Rigid grinding discs designed for soft metals and alloys are used in a similar manner. These discs are quite versatile and can be used to prepare harder materials as well, although their wear rate is greater when used to prepare very hard materials. A generic four-step practice is given in Table 4 for soft metals and alloys.

The planar grinding step can be performed with the 30 7m resin-bonded diamond disc or with a second RGD and 15 or 30 7m diamond, depending on the metal or alloy. For some very difficult metals and alloys, a 1 7m diamond step on a synthetic woven cloth (similar to step 3 but for 3 min) could be added, and/or a brief vibratory polish (use the same cloths and abrasives as for step 4) may be needed to produce perfect publication-quality images.

Electrolytic Polishing

Electrolytic polishing can be used to prepare specimens with deformation-free surfaces. The technique offers reproducibility and speed. In most cases, the published instructions for electrolytes tell the user to grind the surface to a 600-grit (P1200) finish and then electropolish for approximately 1 to 2 min. However, the depth of damage after a 600-grit (P1200) finish may be several micrometers, but most electropolishing solutions remove only about 1 7m/min. In this case, the deformation is not completely removed. In general, electropolished surfaces tend to be wavy rather than flat, and focusing may be difficult at high magnifications. Further, electropolishing tends to round edges associated with external surfaces, cracks, or pores. In two-phase alloys, one phase polishes at a different rate than another, leading to excessive relief. In some cases, one phase may be attacked preferentially, and inclusions are usually attacked. Consequently, electrolytic polishing is not recommended for failure analysis or image analysis work, except possibly as a very brief step at the end of a mechanical polishing cycle to remove whatever minor damage may persist.

References cited in this section

24. G.F. Vander Voort, *Metallography: Principles and Practice*, McGraw-Hill, 1984 (reprinted by ASM International, 1999)

25. R.W. Johnson, *Wear*, Vol 16, 1970, p 351–358

Examination of Microstructures

The second main use of light microscopy is to determine the microstructure of the material in question to evaluate its influence on the failure. Such examination is first performed at the origin of the failure to detect any anomalies that may have arisen from inadequate material quality, fabrication or heat treatment deficiencies, or alterations due to service conditions. In failures that do not involve fracture, for example, certain types of wear or corrosion failures, the relationship of the microstructure to the observed damage is assessed at the damage sites (Ref 26, 27). Examples of such work are given in the following to illustrate the value of light microscopy.

Figure 24(a) shows a micrograph of an unusual tensile test fracture that was obtained during the evaluation of a carbon steel (0.25% C, 0.63% Mn, 0.27% Si) casting. Instead of the usual cup-and-cone tensile fracture, the surface was at an angle, approximately 45°, to the tensile axis and was rather rough. Scanning electron microscope examination (Fig. 24b) of the fracture revealed numerous voids typical of shrinkage cavities. Light microscope examination (Fig. 24c) also reveals these cavities. Neither SEM or LM examination by itself was satisfactory for determining why the fracture was unusual, but together they provide a more complete picture. The shrinkage cavities caused failing yield strength, elongation, and reduction of area results.

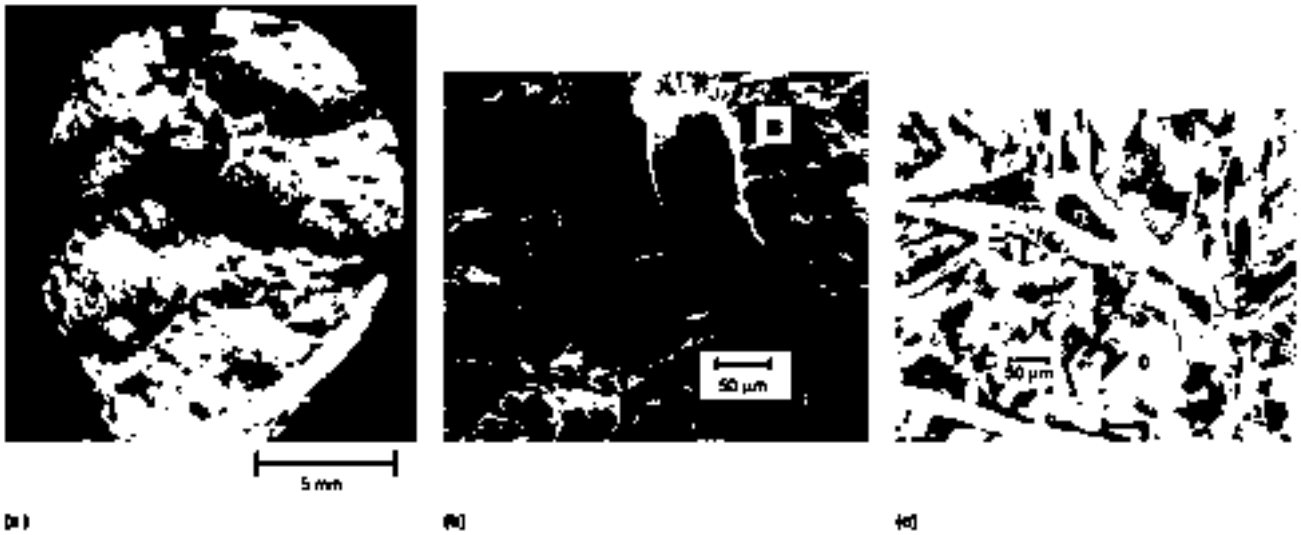


Fig. 24 (a) Macrograph of fracture, (b) SEM fractograph, and (c) light micrograph showing shrinkage cavities in an unusual tensile fracture from a carbon steel casting. The microstructure was revealed using nital.

Another example of a material quality problem is shown in Fig. 25(a) and (b). These micrographs show a seam that was the cause of rejection for a forged pitman arm. In this case, the seam is not perpendicular to the surface because of the metal flow during forging.

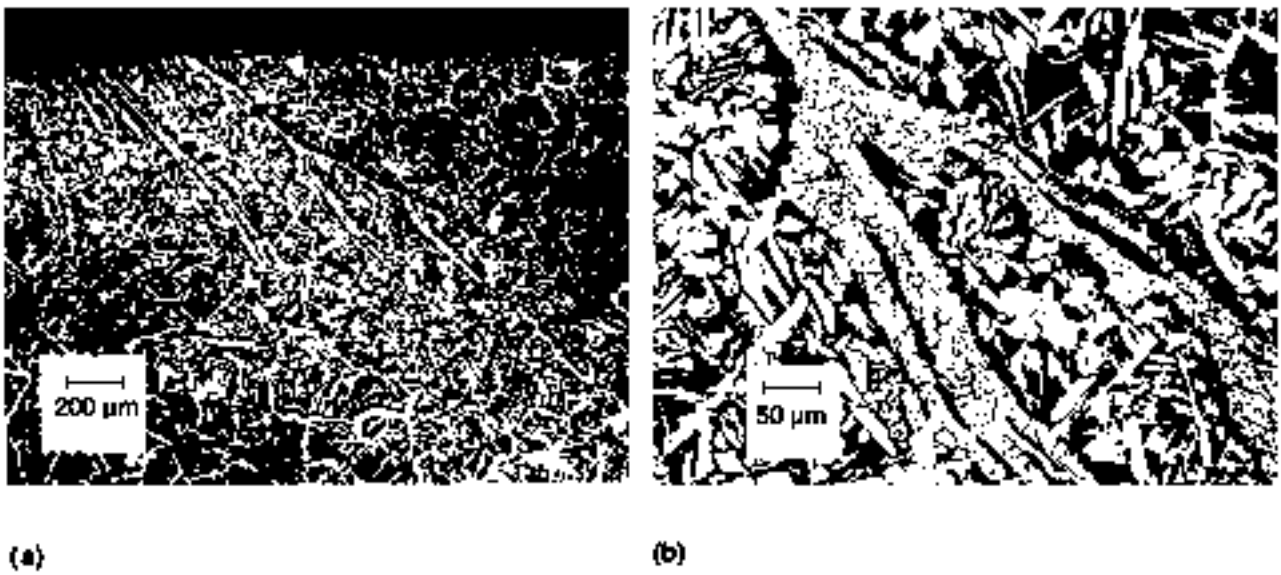


Fig. 25 Light micrographs of two cross-sectional views of a seam found on a closed-die forged pitman arm showing decarburization and internal oxidation. Etched with 2% nital

Another common material problem is the presence of decarburization that may be present on as-rolled stock or may form during heat treatment or, in some cases, during service. Light microscopy is the generally accepted method for detecting decarburization and measuring its extent (Ref 28). Figures 26(a), (b), and (c) show examples of decarburized AISI 5160 spring steel in the as-rolled condition (Fig. 26a) and after heat treatment (Fig. 26b, c). Decarburization of hardened coil springs is not desired, because it reduces the fatigue life of the spring.

Fig. 26 Light micrographs of decarburization observed on cross sections of as-rolled and heat treated AISI 5160H alloy steel spring. (a) Nickel plating on top of scale on an as-rolled specimen. (b) Partial decarburization at the surface of a hardened specimen. (c) Free ferrite and partial decarburization at the surface of a hardened specimen. Etched with 2% nital

A less common example of decarburization is shown in Fig. 27, which shows the surface of a decarburized solution-annealed austenitic manganese steel. At locations where the carbon content is below approximately 0.50%, epsilon martensite is observed (Ref 29, 30). This structure does not possess the remarkable work-hardening capacity typical of such alloys.

Fig. 27 Light micrograph showing epsilon martensite at the surface of a decarburized (less than 0.5% C) austenitic manganese steel specimen. Etched with 2% nital/20% sodium metabisulfite

Figure 28 provides another example of inadequate material quality. This shows a defect observed on the polished inside diameter of an AISI 420 stainless steel mold. Sectioning of the mold at the defect and light microscopy examination revealed a large silicate inclusion that caused the defect.

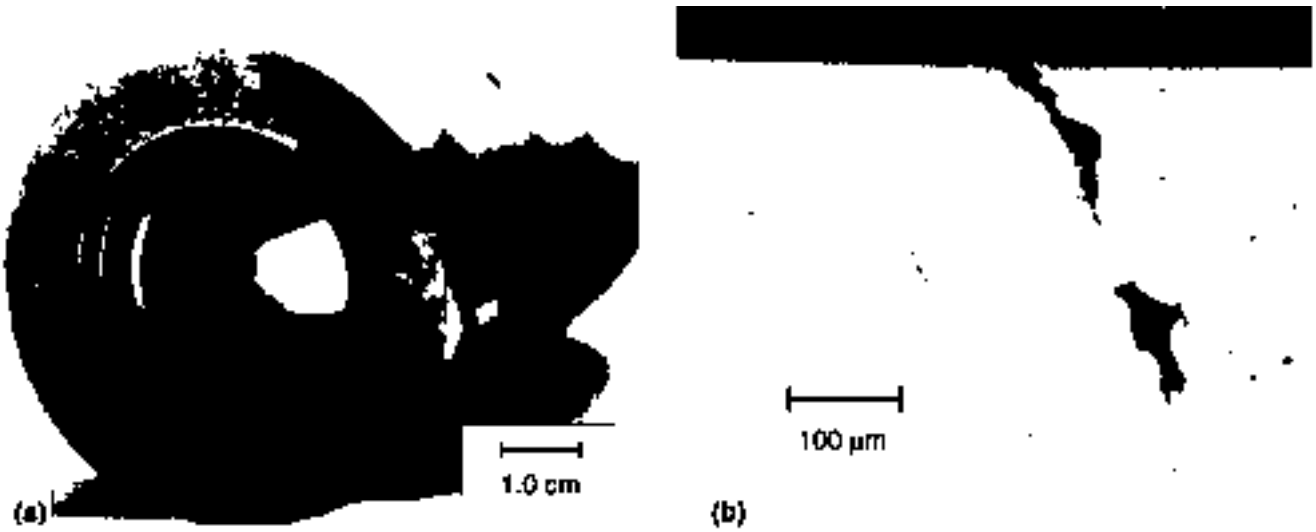


Fig. 28 (a) AISI 420 stainless steel mold containing a defect (arrow) observed after polishing the inside diameter surface. (b) Microscopic examination revealed a large silicate inclusion (unetched).

Numerous failures have been traced to problems occurring during heat treatment. The following examples show how light microscopy was employed to analyze such failures.
Figure 29 shows the operating surface of an AISI L6 tool steel punch that exhibited poor service life. Examination of the microstructure revealed that it was under-austenitized in heat treatment. The hardness was several points HRC below the expected value.

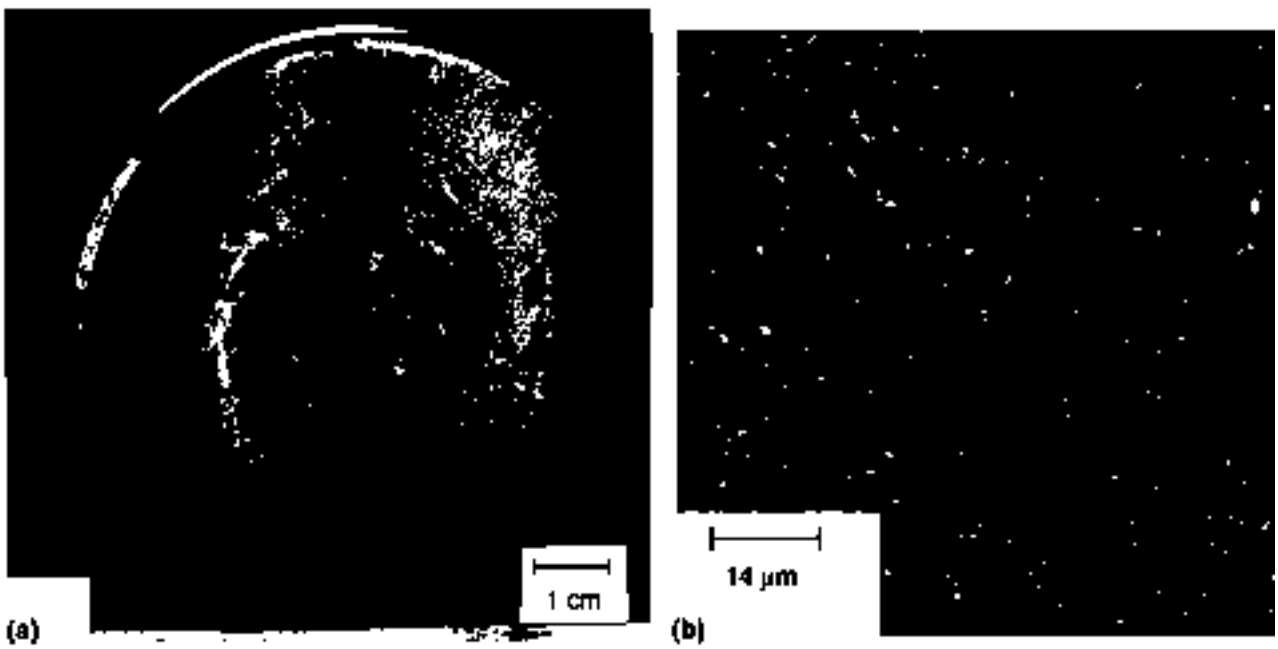


Fig. 29 (a) The working face of an AISI L6 punch that failed after limited service, because (b) the punch was underaustenitized. Specimen etched with nital

Figure 30 shows the microstructure of a roll made from AISI O1 tool steel that cracked during quenching. The quench cracks were located adjacent to deep stamp marks, and the microstructure contained considerable retained austenite. No residual carbide, normally present in this grade when properly heat treated, was present. Consequently, an excessively high austenitizing temperature had been employed. The hardness was 56/57 HRC, which was increased to 62/64 HRC after the specimen was refrigerated in liquid nitrogen.

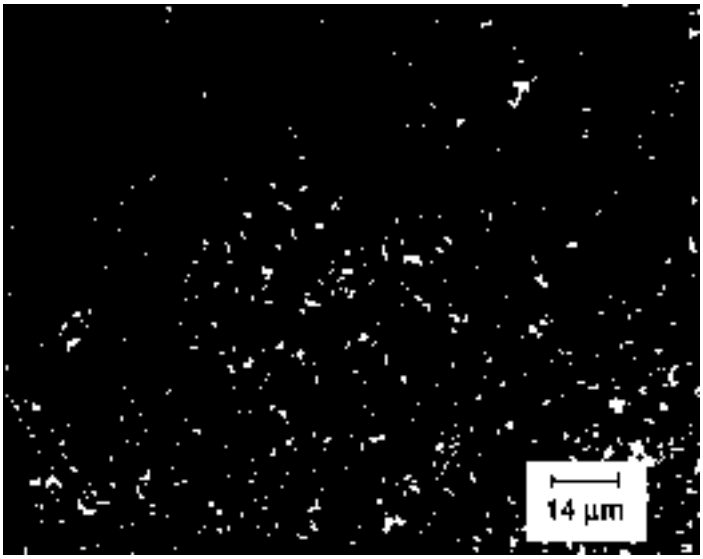


Fig. 30 Light micrograph of overaustenitized AISI O1 tool steel containing coarse plate martensite and substantial unstable retained austenite. Specimen etched with nital

Figure 31 shows two views of the microstructure of a jewelry-striking die made from AISI S7 tool steel that cracked soon after being placed in service. The views show a coarse zone at the surface, consisting of coarse plate martensite and retained austenite. The carbon content at the surface was 0.79%, while the interior was 0.53%. The die had been lightly carburized to a depth of approximately 0.5 mm (0.020 in.) due to improper furnace atmosphere control.

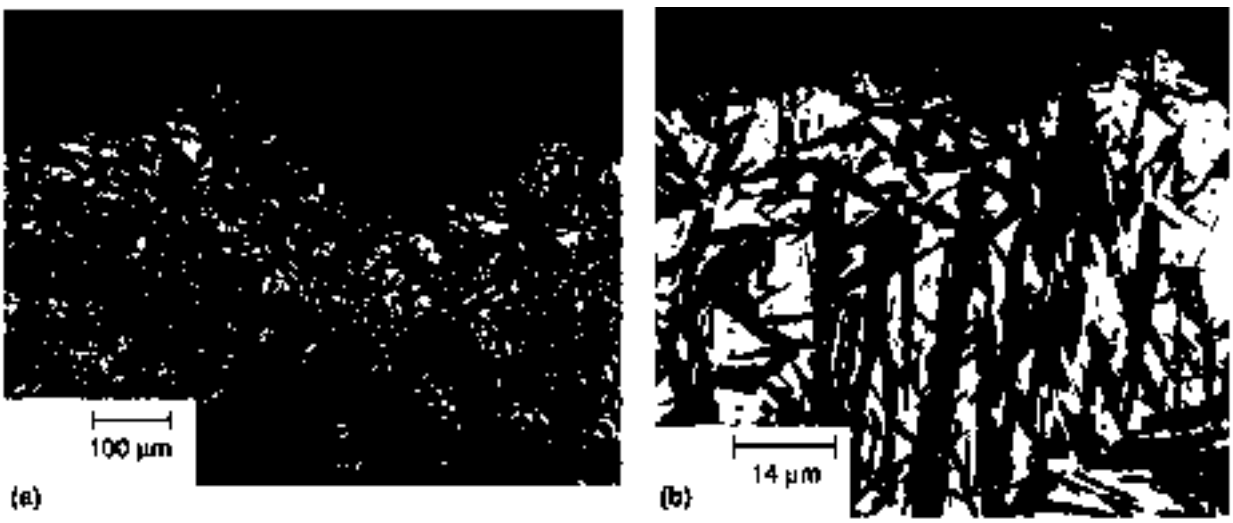


Fig. 31 Light micrographs of an AISI S7 tool steel jewelry-striking die that failed due to the presence of a carbon-enriched surface layer that contained coarse plate martensite and unstable retained austenite. Specimen etched with nital

Figure 32 shows the microstructure of a cracked carburized tread of a track wheel made from AISI 1035 carbon steel. The 60 cm (24 in.) diameter track wheel was carburized at a higher temperature than usual, and the diffusion cycle after carburizing had been omitted. Microstructural examination revealed that cracking followed the grain-boundary carbide network present due to the improper carburizing practice.

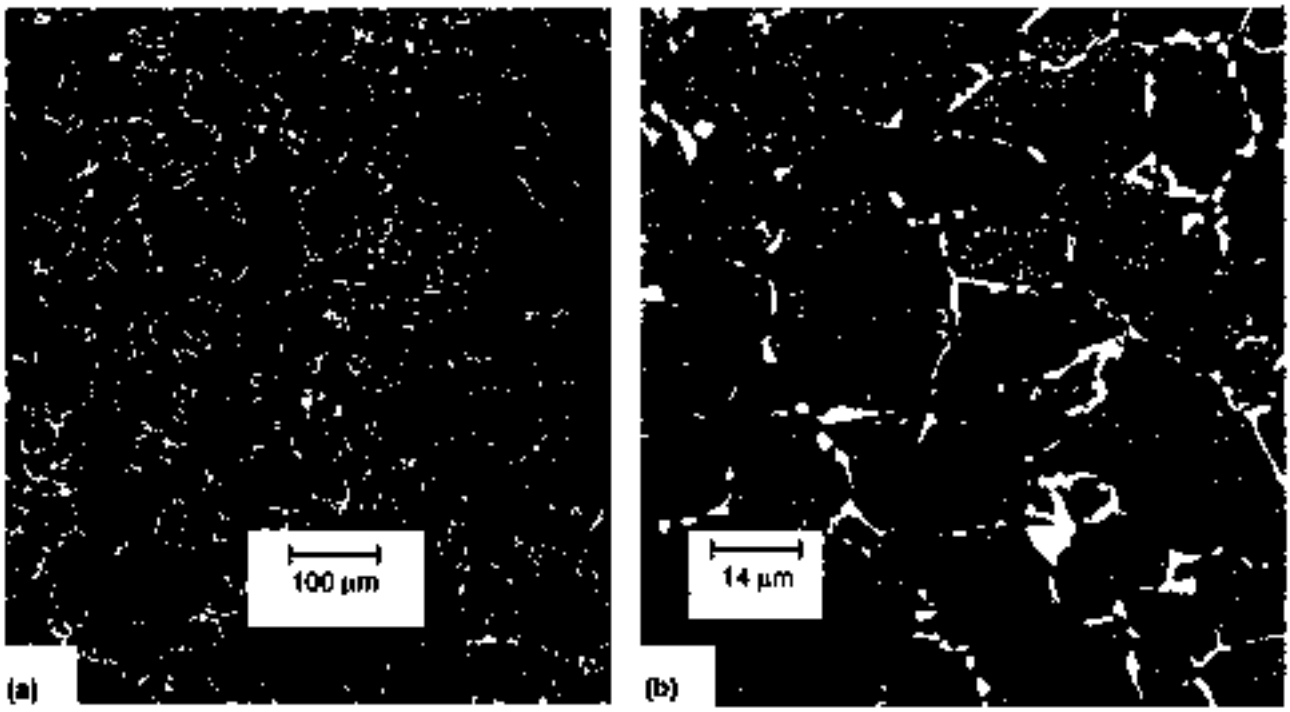


Fig. 32 Light micrographs of a carburized AISI 1035 track wheel that cracked due to the presence of an extensive grain-boundary carbide film. Specimen etched with nital

Figure 33 shows the microstructure of an AISI D2 tool steel powder metallurgy die that was deformed at one end after heat treatment. Light microscopy revealed that the steel was locally melted in this region, apparently due to flame impingement during austenitization.

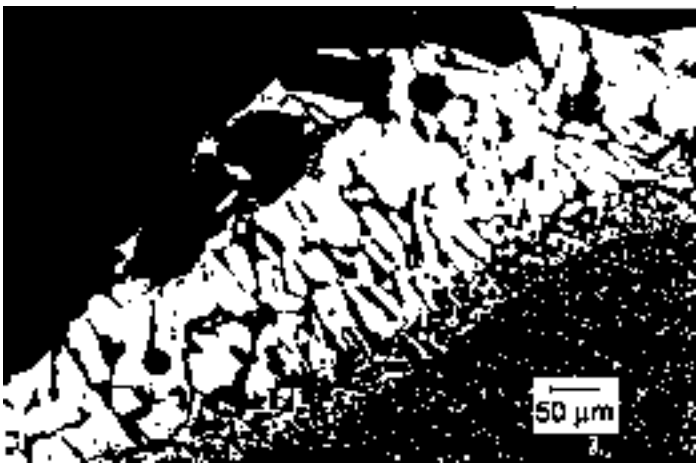


Fig. 33 Light micrograph of a melted region found on an AISI D2 powder metallurgy die after heat treatment. Specimen etched with Marble's reagent

As a final example of a heat-treatment-related failure, Fig. 34 shows the microstructure of an AISI D2 draw die insert that exhibited galling and chipping after limited service. Examination of the microstructure revealed that the austenitizing temperature was well above the recommended 1010 °C (1850 °F) temperature, high enough to cause liquidation at the grain boundaries. Note the nearly complete grain-boundary network of skeletal carbides, similar in appearance to an as-cast condition.

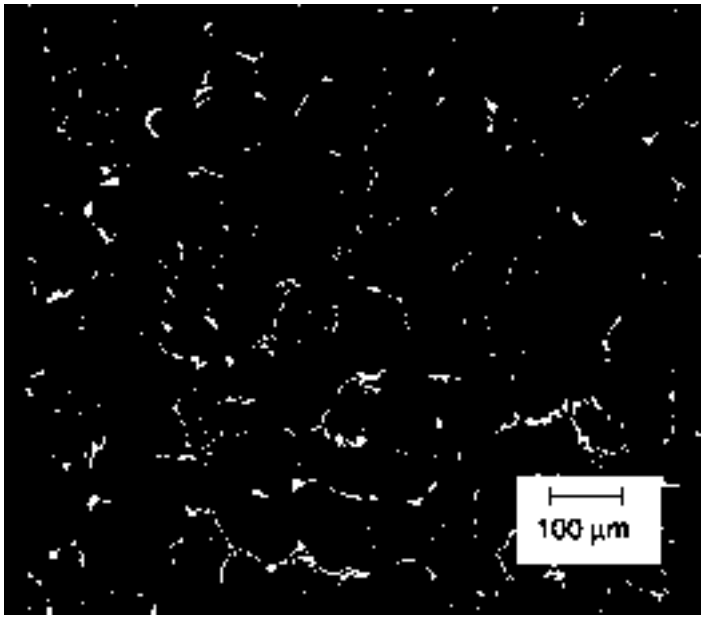


Fig. 34 Light micrograph of a grossly overaustenitized AISI D2 draw die insert. Specimen etched with Marble's reagent

Failures may also arise during fabrication processes. The next several examples illustrate such problems and the use of light microscopy. A wide variety of problems can occur; these examples illustrate only a few of many such problems.
Figure 35 shows planar and through-thickness views of cracks observed in several reaustenitized zones in a forged steel specimen. An electric pencil had been used to identify the component. The cracked regions were present on the face placed against the grounding plate. Apparently, arcing occurred along the edges touching the grounding plate, producing enough heat to locally reaustenitize the steel. On cooling, the hardenability was sufficient to form as-quenched martensite in these regions. The transformation-related expansion caused cracking in these reaustenitized spots.

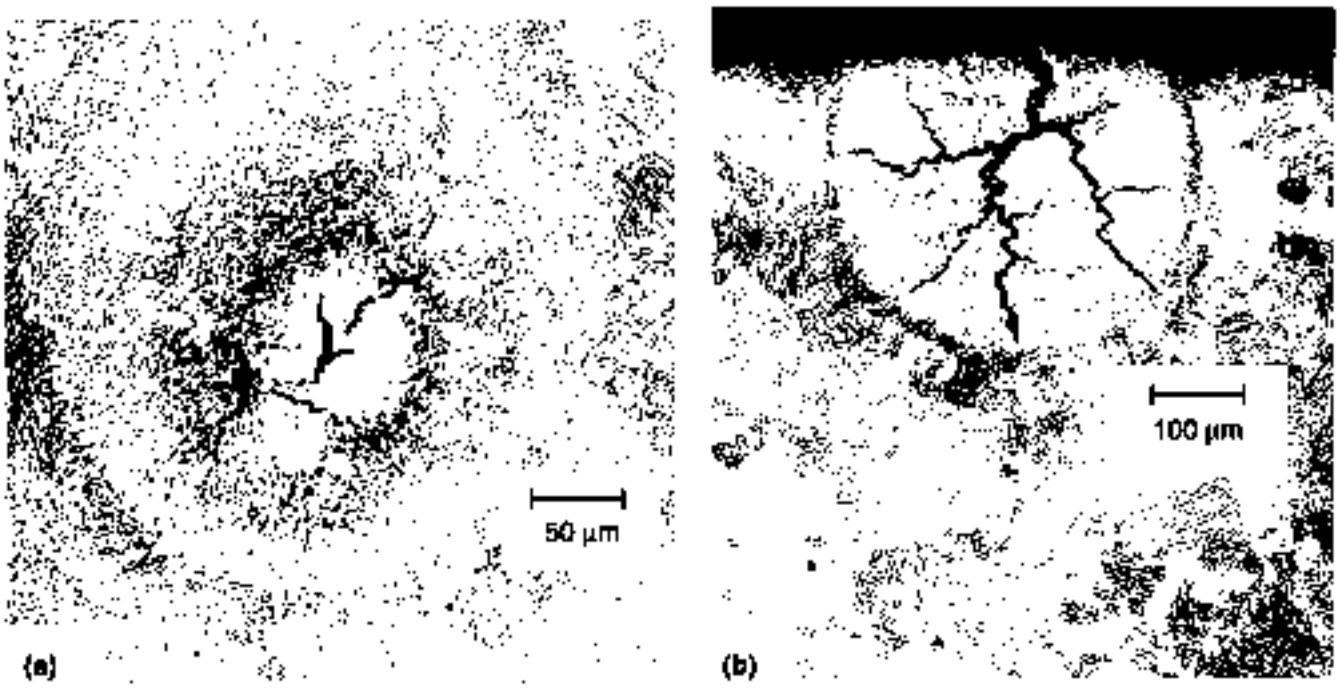


Fig. 35 Light micrographs showing (a) planar and (b) through-thickness views of cracking that occurred at reaustenitized spots due to arcing during electric pencil coding. Specimen etched with nital

Figure 36 shows an example of central bursting (“chevron cracking”) in extruded, as-rolled AISI 4615 alloy steel. Although optimal extrusion parameters can generally minimize such failures, the ductility of the material is also an

important variable. Because this alloy has substantial hardenability, it is difficult to prevent formation of bainite and martensite in small, as-rolled section sizes. In this case, the microstructure consisted of ferrite, bainite, and martensite (the arrows point to microcracks present in martensite patches).

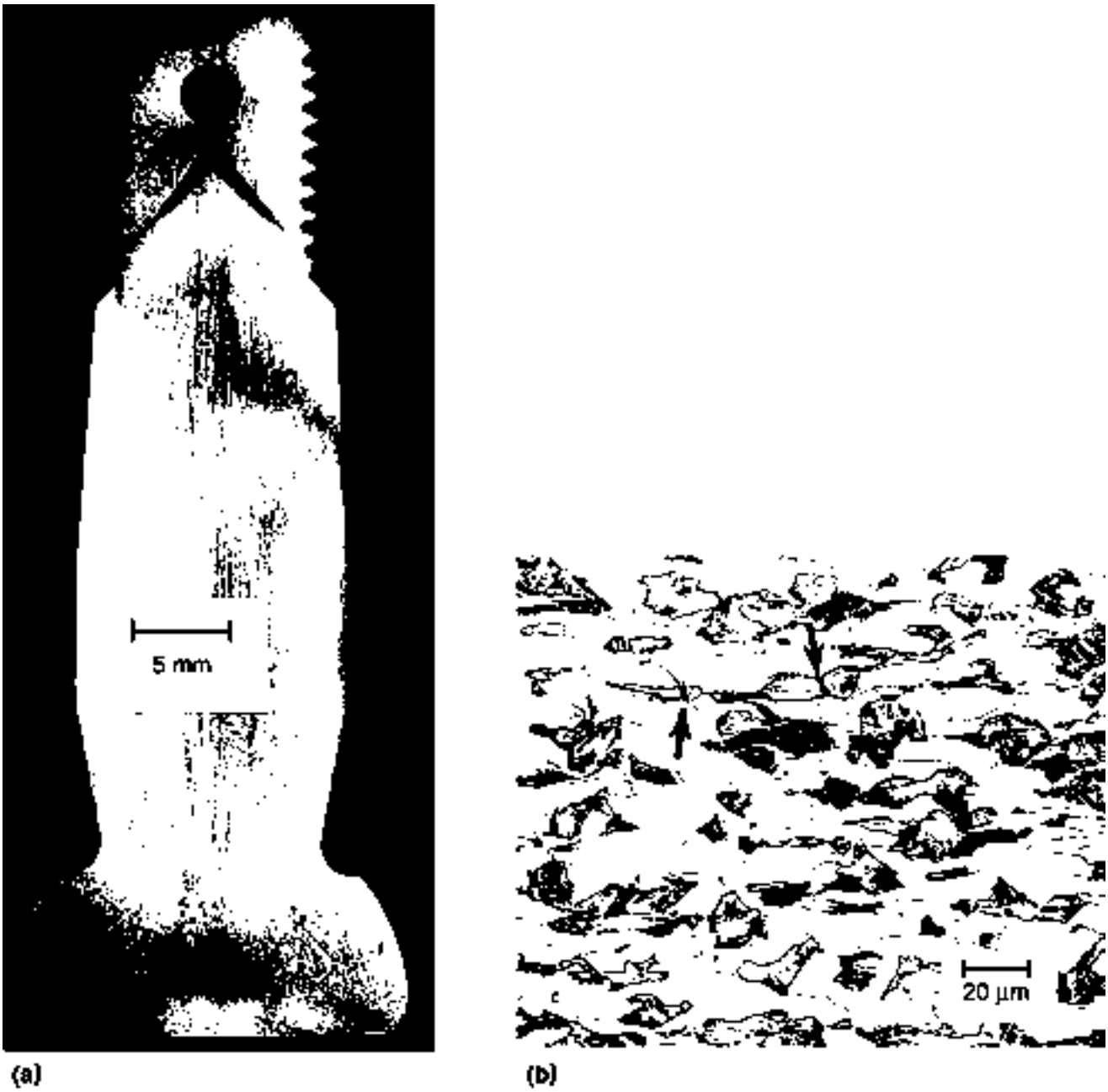


Fig. 36 (a) Central bursting during extruding of AISI 4615 alloy steel specimen was promoted by (b) the presence of bainite and martensite (arrows point to microcracks in the martensite) in the as-rolled stock. Specimen etched with 4% picral, followed by 2% nital

Final grinding of hardened tool steel components is an important processing step that must be carefully controlled. Improper grinding practices can produce a fine network of surface cracking and a characteristic scorch pattern easily revealed by macroetching (Ref 31, 32). Examination of the microstructure at the cracks reveals their shallow nature and a back-tempered condition at the surface, as shown in Fig. 37. In some cases, a shallow, reaustenitized light-etching layer of as-quenched martensite is found at the surface (Ref 31) above the back-tempered zone. In many cases of abusive grinding, microstructural examination reveals that the die was not tempered. As-quenched tool steels are very difficult to grind without causing such cracking. In the example shown, however, the AISI O1 tool steel die was properly tempered, and one must conclude that the grinding operation was at fault.

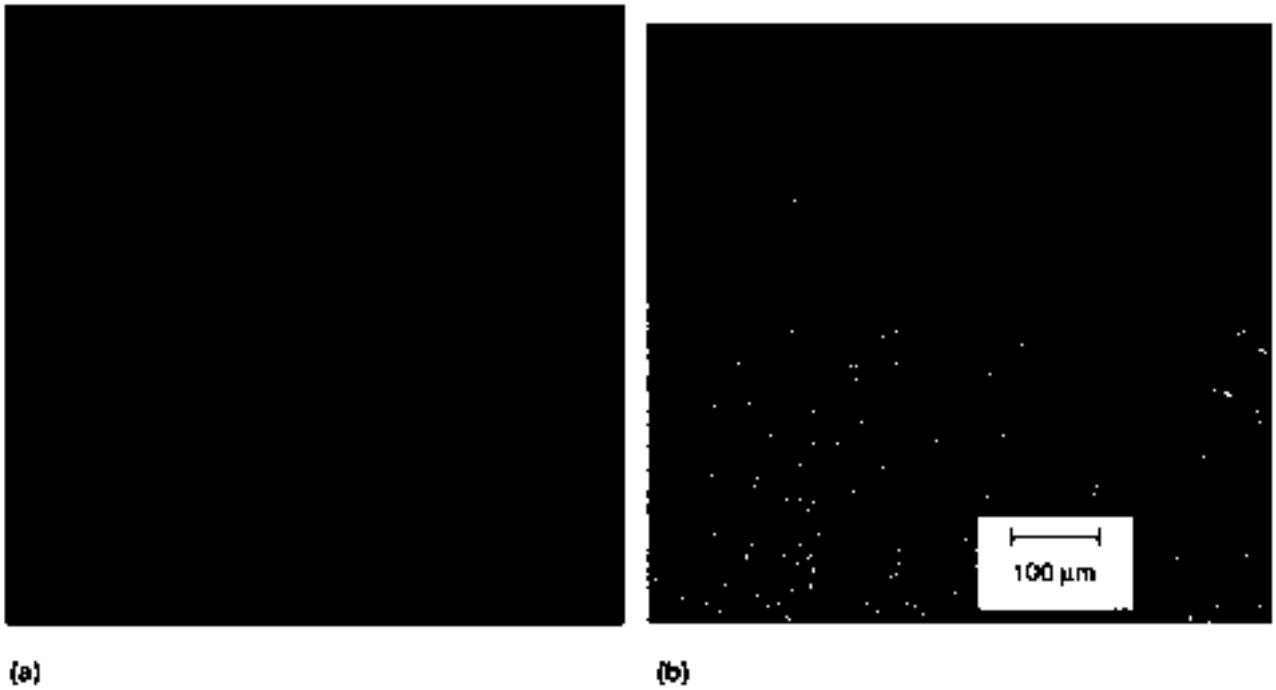


Fig. 37 (a) Abusive grinding caused this 50 mm (2 in.) diameter AISI O1 tool steel die to crack (left, after dye-penetrant inspection). (b) Typical appearance of the cracks (etchant has bled out of the crack, producing a stain around it). Specimen etched with nital

Electrical discharge machining (EDM) is widely used to produce cavities in tool steels. Because it is a spark-erosion process that generates considerable temperature and localized melting, it must be rigorously controlled. After EDM, the cavity surface generally is stoned to remove melted surface layers, and the part is tempered. However, numerous failures have been observed in EDM-processed components (Ref 31, 32). Figure 38 shows a classic example of the microstructure of a failed part that was machined by this technique but not properly posttreated. This was a plastic mold made from AISI S7 tool steel, where pitting was observed on the cavity after polishing. The microstructure at the pit shows a large remelted surface layer (note dendrites). Beneath this layer is a white-etching zone of as-quenched martensite, typical of such failures. Beneath this is a back-tempered zone, where the temperature was below the upper critical temperature for the steel. In many such EDM-related failures, the as-cast layer is not present, but the as-quenched region is always observed.

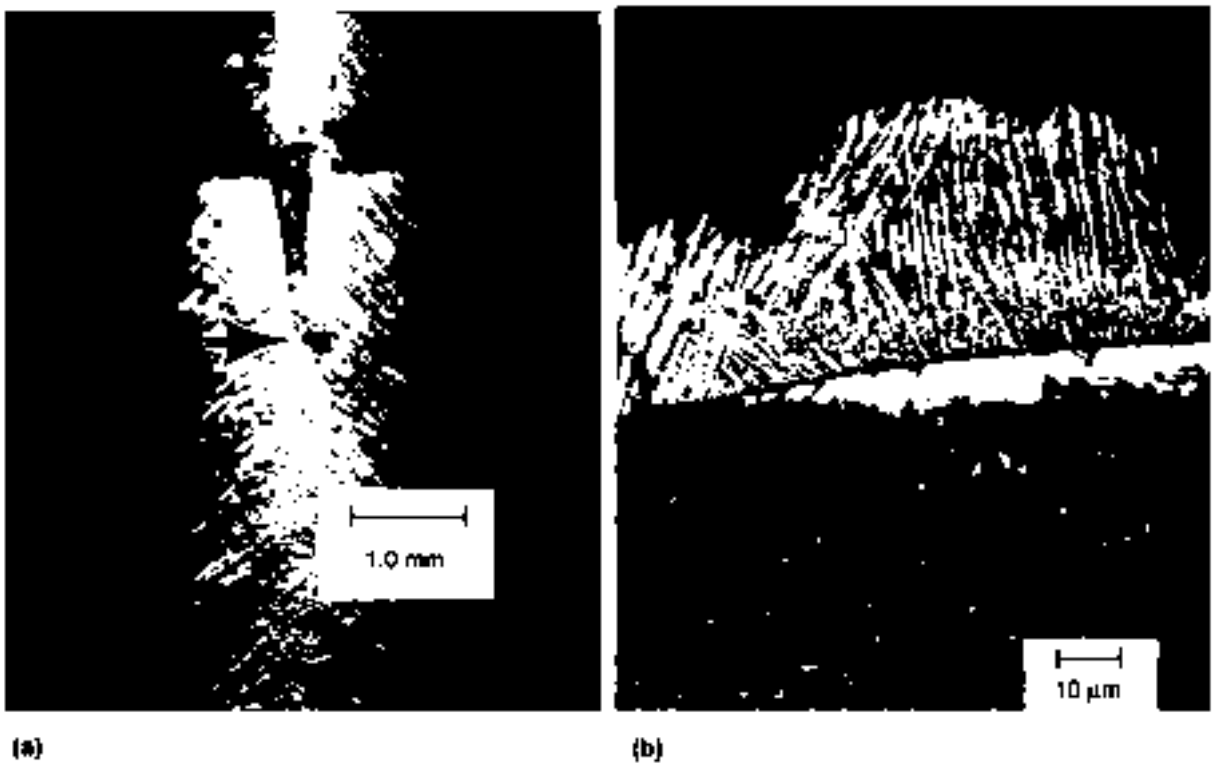


Fig. 38 (a) Pitting on this mold cavity (arrow) was observed after polishing of the AISI S7 plastic mold and was caused by the use of improper post-EDM procedures. (b) The classic appearance of such failures. There is a large, remelted surface layer above a reaustenitized, untempered zone. Specimen etched with nital

Failures also occur due to service conditions, and the next several examples illustrate some of the many problems that can occur. These examples are provided to illustrate the value of light microscopy in such studies.

Figure 39 shows the microstructure of an as-cast 25%Cr-12%Ni heat-resisting alloy used as a hook for holding a basket of parts during austenitizing and water quenching. The alloy contained delta ferrite that transformed to sigma, producing a discontinuous network that caused the hook to crack. Electrolytic etching was used to reveal only the sigma phase.

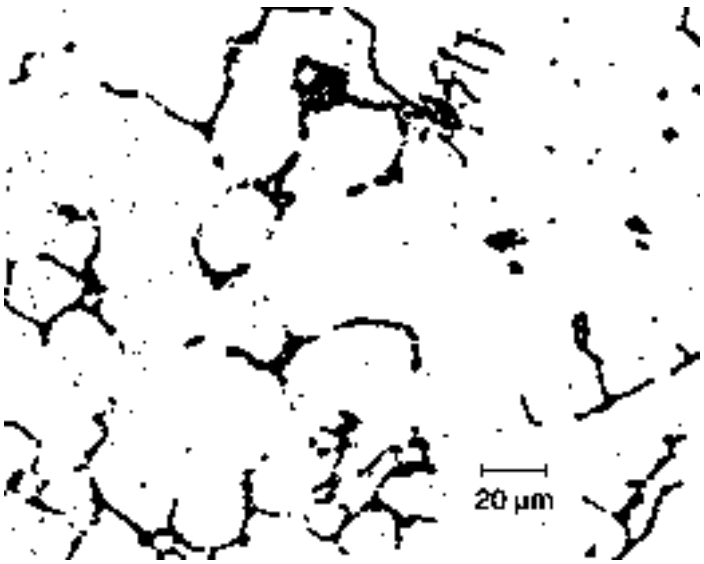


Fig. 39 Light micrograph showing sigma phase revealed by selective etching with 10N KOH (electrolytic). The brittle sigma phase caused extensive cracking in a 25%Cr-12%Ni cast heat treatment basket hook.

Figure 40 shows surface damage due to wear on a 4485 alloy steel medart roll. The specimen was plated with electroless nickel for edge retention. Some of the nickel can be observed in the crack. Note that the surface layer has been reaustenitized due to service-generated heat. Oxidation at the surface and flow in the layer can also be observed.

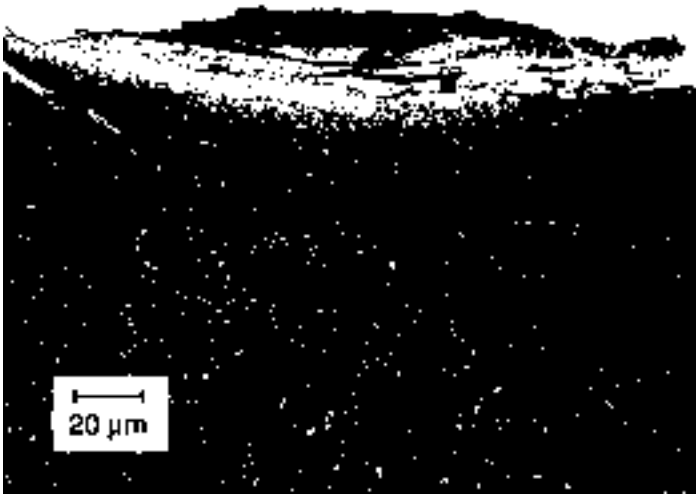


Fig. 40 Light micrograph showing wear damage at the surface of a 4485 alloy steel medart roll. The surface was nickel plated for edge retention and etched with nital.

It is not unusual to observe as-quenched martensitic layers produced on the surface of steels subjected to heavy wear conditions. Figure 41 shows such a layer on the surface of a scrap chopper knife made from a proprietary wear-resistant tool steel. Cracking and spalling may be produced at such layers because of their brittle nature. Such a condition is often observed on steel railroad rails where cracking and spalling has occurred. Figure 42 shows an example of surface cracking near a spall on a 136 lb/yd rail. Microstructural examination revealed three regions at the surface. The outer zone is featureless, typical of heavily deformed as-quenched martensite. The hardness was approximately 60 HRC. Beneath this is a zone containing the same white-etching martensite plus a network of ferrite. Beneath this zone is the normal pearlitic structure of an as-rolled rail steel. Tempering of the sample produced a tempered martensite structure with spheroidized carbides in the outer zones containing the white-etching constituent.

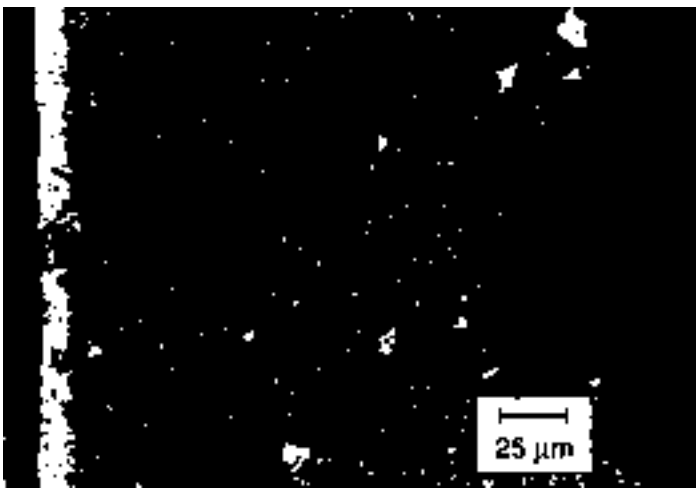


Fig. 41 Light micrograph of the surface of a badly worn steel chopper knife made from an air-hardenable tool steel showing a reaustenitized surface zone and a back-tempered region below it. The large angular particles are nitrides. Specimen etched with nital

Fig. 42 Light micrograph of a white-etching surface layer formed on a rail head due to frictional heat. This specimen was taken adjacent to a spalled area. Specimen etched with picral

References cited in this section

26. W.A. Glaeser, Microscopy and the Study of Wear, *Applied Metallography,* Van Nostrand Reinhold, 1986, p 261–279

27. W.E. White, Microscopy and the Study of Corrosion, *Applied Metallography,* Van Nostrand Reinhold, 1986, p 281–296

28. "Standard Test Methods for Estimating the Depth of Decarburization of Steel Specimens," E 1077, ASTM

29. A.J. Sedriks and T.O. Mulhearn, *JISI,* Vol 202, Nov 1964, p 907–911

30. A.J. Sedriks, *JISI,* Vol 204, Feb 1966, p 142–145

31. G.F. Vander Voort, Macroscopic Examination Procedures for Failure Analysis, *Metallography in Failure Analysis,* Plenum Press, 1978, p 33–63

32. G.F. Vander Voort, Failures of Tools and Dies, *Failure Analysis and Prevention,* Vol 11, *ASM Handbook* American Society for Metals, 1986, p 563–585

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Introduction

THE SCANNING ELECTRON MICROSCOPE (SEM) is one of the most versatile instruments for investigating the microstructure of metallic materials. Compared to the optical microscope, it expands the resolution range by more than one order of magnitude to approximately 10 nm in routine instruments, with ultimate values below 3 nm. Useful magnification thus extends beyond 10,000× up to 100,000×, closing the gap between the optical and the transmission electron microscope. Compared to optical microscopy, the depth of focus, ranging from 1 7m at 10,000× to 2 mm (0.08 in.) at 10×, is larger by more than two orders of magnitude.

Development of SEM Technology

Development of what is now recognized as a scanning electron microscope was begun in 1938 by Ardenne (Ref 1) who hypothesized that an image of a sample surface could be generated by subjecting the sample to a beam of electrons and then monitoring the current that resulted from the sample. The success of these initial experiments led to further work that improved and modified his original design, until the first commercially successful SEM was marketed by Cambridge in the early 1960s. Since that time, the use of SEMs has expanded dramatically until SEMs are commonly found in use today in industry, education, and research applications.

The success of the SEM is directly related to the versatility of the instrument. Typical SEMs have the ability to deliver clear, high-magnification images while also providing the user with the ability to obtain chemical compositions from samples. Recent developments (Ref 2) also have greatly increased the capability of the SEM to obtain crystallographic information from a sample. Thus, the analytical capabilities of an SEM cover a wide range of information, and accessories can be chosen to suit the particular interests of the user.

Scanning electron microscopes have been found particularly useful in failure analysis investigations. This is true for a number of reasons, the most prominent being the ability of the SEM to image large, non-flat samples from low to high magnifications (approximately 10× to greater than 100,000×). The ability to obtain in-focus images of rough samples over a large change in vertical height is termed *depth of field* and it is this trait that gives SEM images their characteristic three-dimensional appearance. The large depth of field is made possible by the relationship between the small size of the electron probe used as related to the size of the imaging pixel determined by the operating magnification. This is explained in more detail in the following section.

While all SEMs exhibit similar characteristics, a number of variations exist, depending on the specific application and/or environment in which they are used. By changing the type of electron source used, the signals to be processed, and the type of vacuum systems, an SEM can be designed to enhance particular capabilities. For example, proper choice of the electron source can be used to enhance imaging, while changing the types of x-ray detectors can provide better chemical analysis. Modifying the vacuum system can allow examination of a wide range of sample types, including nonconductive or wet samples. Thus, a number of types of SEMs now exist with slightly different capabilities.

A conventional SEM typically uses a heated filament to produce electrons. These thermionic sources use tungsten or lanthanum hexaboride (LaB_6) as the filament material, tungsten being less expensive and more robust while LaB_6 is brighter and lasts longer. Scanning electron microscopes that use a high electric field to remove electrons from a tungsten filament are termed cold-field-emission SEMs or high-resolution SEMs and have the highest resolution capabilities of all SEMs. A compromise is reached in the thermal-field-emission (or Schottky gun) SEM, where a combination of heat and electric field are used to produce electrons. These machines have better resolution than conventional SEMs, while being easier to use than the high-resolution machines.

A conventional SEM also requires that the sample be electrically conductive to prevent a charge buildup in the sample that affects the incoming primary and emitted secondary electrons, resulting in a poor, distorted image that is constantly changing in contrast and location. The sample should also be dry when placed in the vacuum chamber and subjected to the beam, since outgassing will similarly prevent a clear image from being obtained. These restrictions on the sample can be overcome by changing the design of the vacuum system and the choice of detection system used to image the sample. Microscopes adapted in this way are termed low- or variable-pressure microscopes. In this type of microscope, the pressure in the sample chamber is raised to a value on the order of 0.1 to 1 torr. Interaction of the electron beam with gas molecules in the region where the beam strikes the sample effectively creates a positively charged “cloud” of ions above the surface of the sample. This positive “cloud” offsets the negative charge buildup that occurs in insulating or poorly conductive samples and allows images of these types of samples to be obtained using back-scattered electrons. Modern microscopes can be purchased that allow the user to operate in either the high-vacuum, conventional imaging mode or in the low- or variable-pressure mode, making them extremely flexible for all types of investigations.

If even greater capability to image nonconductive or wet samples is desired, an environmental SEM can be purchased. This type of SEM operates at vacuum levels of 20 torr and allows both backscattered and secondary images to be obtained from the sample surface. This is made possible by the use of an electron detector that operates on the principle of induced current. In this instance, the interaction of the electron signals from the sample surface with the gas molecules above the sample induces a current in the detector, which consists of two parallel conductive plates, one placed above the sample and one that acts as the sample holder. By proper selection of where the current is measured, a secondary or backscattered image can be obtained.

References cited in this section

1. M. Von Ardenne, *Z. Phys.*, 109, 1938, p 553

2. D. Dingley and Baba-Kishi, Use of Electron Back-Scatter Diffraction Patterns for the Determination of Crystal Symmetry Elements, *Scanning Electron Microsc.*, 2, 1986, p 386

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Operation

All SEMs may be divided into three basic systems of components, namely, the electron optical column, signal detection and display equipment, and the vacuum system.

Electron-Optical Column

A schematic of a typical SEM column is shown in Fig. 1. Electrons are generated by the electron gun and focused and controlled by a series of electromagnetic lenses and coils. The resultant beam is collimated and defined by various apertures placed in the beam path.

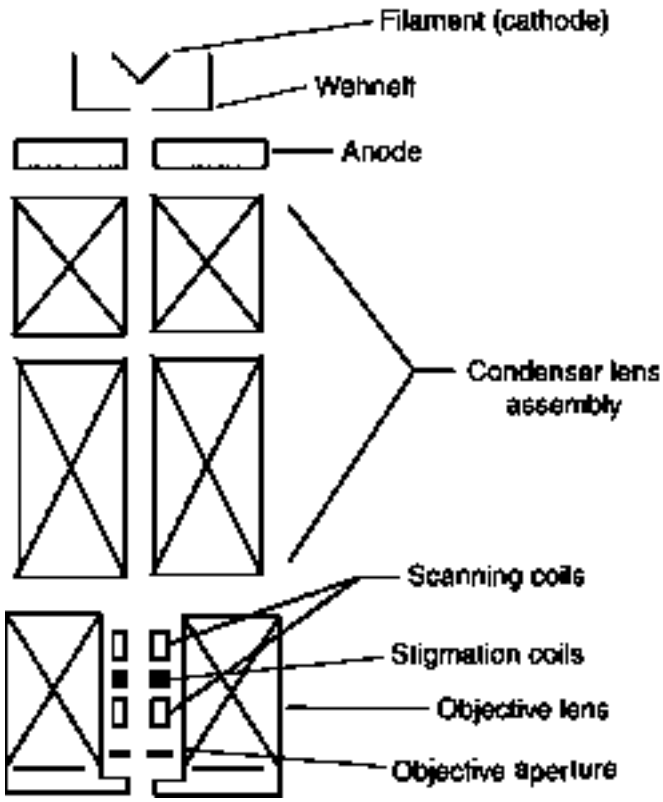


Fig. 1 Schematic showing the general layout of a typical scanning electron microscope

The electron gun consists of a filament, usually of tungsten or LaB_6, placed within a metal cap called the Wehnelt. A current is applied to the filament, causing resistance heating and resulting in emission of light, heat, and electrons. The electrons are attracted down the microscope column by the anode, which is positively biased with respect to the filament. The Wehnelt cylinder placed around the filament carries a slight negative charge with respect to the filament. This charge produces a slight repulsive force on the emitted electrons, causing them to be focused at a point near the exit hole of the filament/Wehnelt gun assembly. The size of this focused crossover point is one factor in determining the ultimate spatial resolution of the instrument. Field emission SEMs have a smaller crossover point, which is the reason for their higher resolving capabilities.

Once the focused electron beam passes the anode, it enters the first set of electromagnetic lenses, collectively called the condenser lens. The condenser lens control is given a variety of labels by manufacturers, such as resolution, probe current, and so forth. Regardless of the name, the condenser lens acts to demagnify the electron beam, producing an even

smaller probe for subsequent imaging purposes. As the strength of the condenser lens increases, the demagnification increases and the probe size becomes progressively smaller. This results in increased resolution. However, the overall signal strength will drop as a finer probe size is only accomplished with a concomitant loss in signal strength for a fixed objective aperture size. Modern digital microscopes can compensate for this signal loss to a certain extent by frame averaging, which allows an image to be acquired over the course of any number of scans. An improper understanding of the trade-off between resolution and signal strength, as determined by the condenser lens, is the most common error associated with SEM operation by a novice. A strong signal, which gives a clear image on the viewing screen, may have poor resolution, whilst a poor, snowy image on the view screen may produce excellent images when a photograph is taken, due to the higher resolution.

After passing through the condenser lens assembly, the electron beam enters the objective lens. The objective lens is responsible for focusing the beam on the surface of the sample, and the control dial for this lens is universally termed "focus" by SEM manufacturers. Unlike the condenser lens, the operation of the objective lens is such that it is always set to produce a focused beam image at the surface of the sample. As the distance from the sample to the objective lens varies, the lens is continually readjusted by the operator to once again bring the probe into focus on the sample surface. Thus, the objective lens current is determined by the distance of the sample to the bottom of the lens. This distance, termed the working distance, is usually given as a readout somewhere on the microscope console. Demagnification of the beam using the objective lens to produce a smaller probe (and therefore higher resolution) is therefore only possible by decreasing the working distance and refocusing the image. However, a decrease in the working distance results in a corresponding loss in depth of field.

Within the objective lens are a series of coils that allow scanning of the beam across the surface of the sample in various ways (scan coils, deflection coils) and correction of any astigmatism associated with the beam (stigmation coils). Also contained within the region of the objective lens is the objective aperture. The objective aperture serves to define the extent of the beam and determines the angular spread. A small aperture will have a small angular spread, which is good for increasing depth of field. However, the smaller the aperture the less the signal that reaches the sample from the condenser lens.

Once the beam strikes the sample, the resolution and depth of field obtained is determined by the relationship between the probe size, the angle of convergence of the beam to produce this probe, and the pixel size on the sample. The pixel size is determined by the magnification and may be thought of as a discreet area on the sample where data is taken and transferred to a corresponding area on a computer screen. This complex relationship is best understood by considering a few simple examples. First consider the schematic shown in Fig. 2. For the sake of convenience, assume an SEM where the beam scans the surface in a square array, transferring information from the sample to the viewing screen 1,000 times per line before dropping to the next line. For a 1,000 line scan, there are, therefore, 1,000,000 pixels of information transferred from the sample to the view screen per scan. For a total scan length on the surface of the sample of 1 cm × 1 cm, each pixel is therefore 0.01 mm square. Magnification is determined simply by dividing the length of the scan of the display screen by the length of scan on the sample. For a 10 cm long scan on the display screen, magnification is just:

$$10\ \text{cm}/1\ \text{cm} = 10\times\ \text{magnification} \qquad \text{(Eq 1)}$$

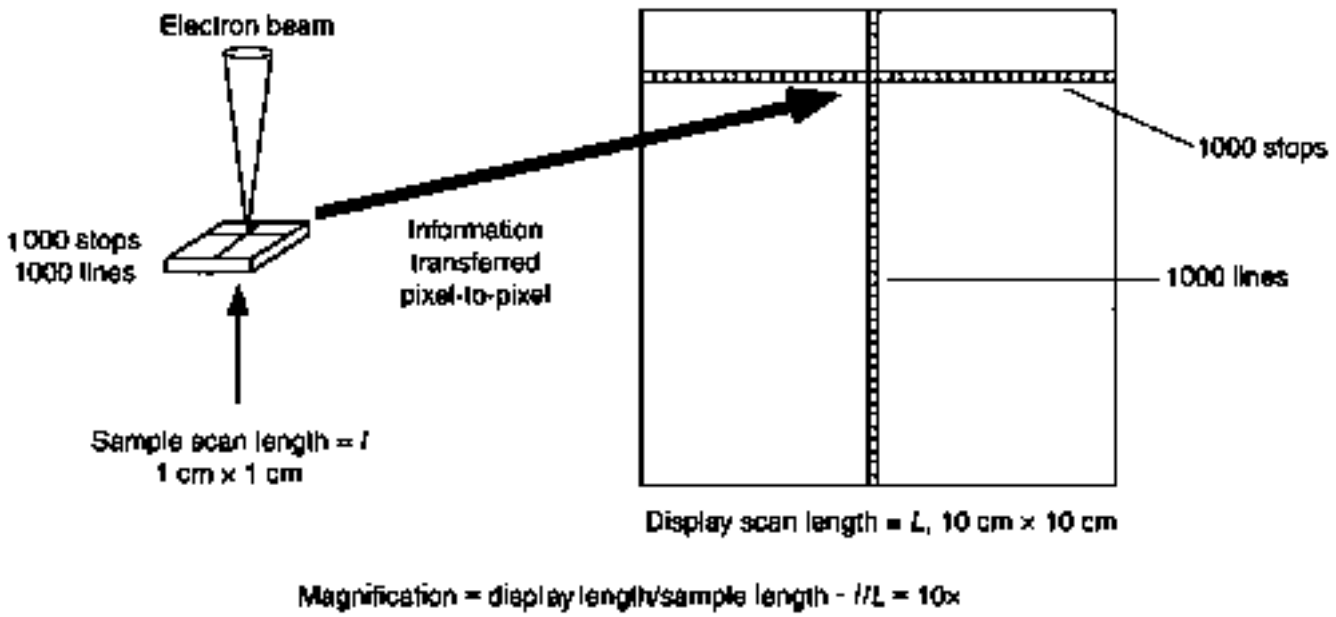


Fig. 2 Illustration of how the information obtained from a square raster on the sample is transferred to the viewing screen to produce magnification

Magnification is increased by decreasing the length of the scan on the sample. The number of pixels scanned remains constant, being 1000 points × 1000 lines. Thus, as magnification increases, the sample pixel size will become progressively smaller. When the electron beam is on the order of the sample pixel size, a loss of resolution occurs and the image is unresolved and blurry, even though the focus (i.e. the setting of the objective lens) is correct. However, as long as the probe size stays less than the pixel size on the specimen, a clear, resolved image will be produced.
If the probe size is significantly smaller than the pixel size, then an in-focus image can be obtained over a considerable vertical distance in a rough sample. This is illustrated in Fig. 3, which schematically shows a view of the electron beam as it strikes a sample. The beam converges to a point and then diverges again, once the image plane is passed. The size of the pixel element on the sample (as determined by the magnification), is shown. It is clear from the illustration that even though the size of the electron beam increases above and below the focus point, that size is still smaller than the pixel element size for a considerable vertical distance. This means that the entire sample surface between the indicated vertical distance will appear in focus. This depth of focus (or depth of field) is what gives SEM images their characteristic three-dimensional appearance. Note that if a larger objective aperture was inserted, the angle of convergence/divergence would be greater, as shown by the dotted lines in Fig. 3. For this greater angle, the probe size becomes greater than the pixel size over a smaller vertical distance, and the depth of field is reduced.

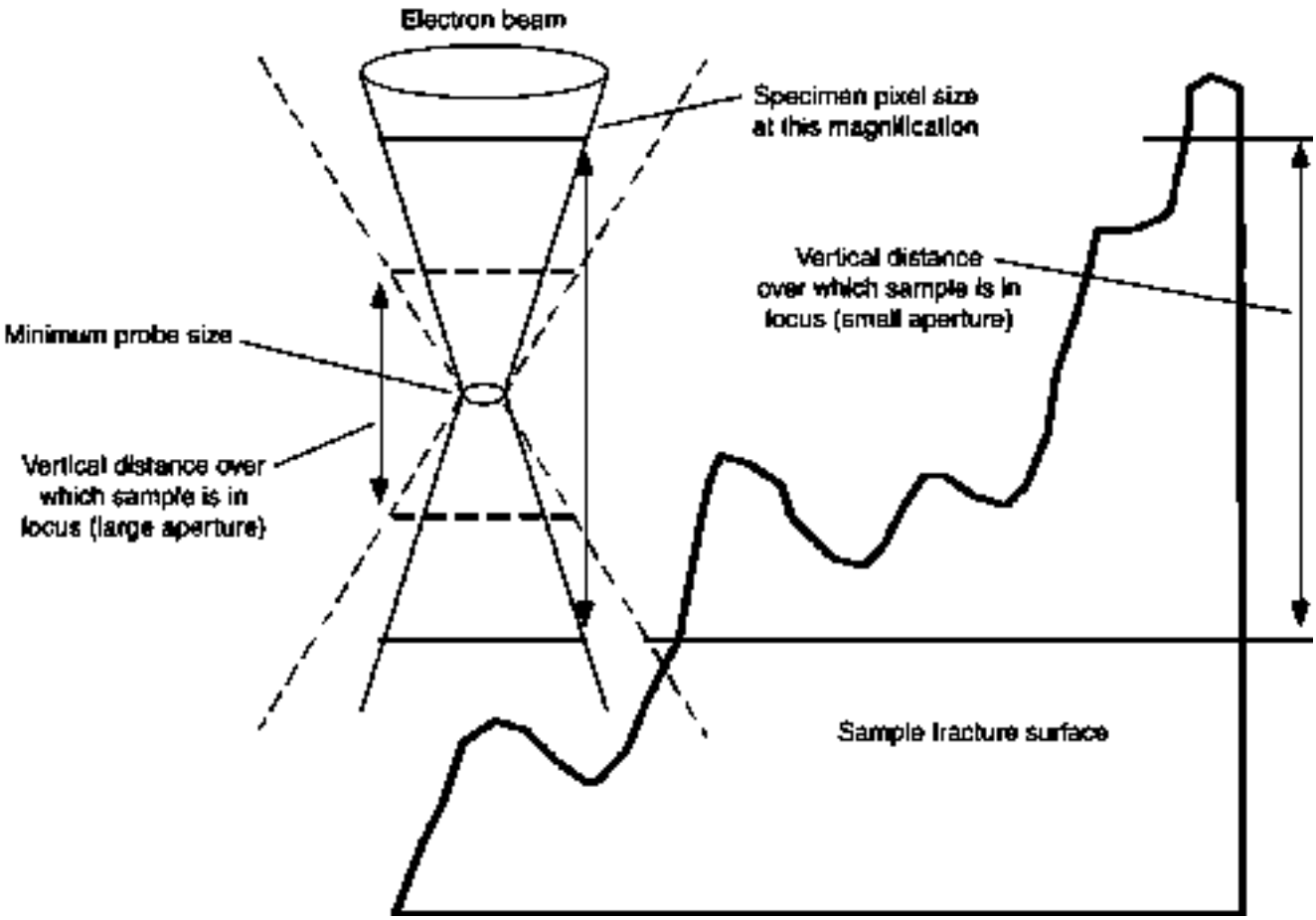


Fig. 3 Illustration showing how depth of focus is dependent on specimen pixel size and aperture size

Successful operation of the SEM, then, depends upon determining which parameter is important for any particular investigation and then optimizing the microscope to obtain the desired information. In a failure analysis investigation, where an extremely rough surface exists, depth of field may be the most important parameter. In this case, depth of field may be enhanced by having a long working distance, low magnification, and a small objective aperture. If one desires to image fatigue striations and high resolution is required, this can be obtained by a strong condenser lens (i.e. high resolution or low probe current) and a short working distance.

Signal Detection and Display

A wide range of signals is available for collection in any SEM. The most common signals used are the secondary and primary backscattered electron signals (SE, BSE) for imaging and the characteristic x-ray signal emitted from the sample for composition identification.

Secondary electrons are low-energy electrons emitted from the sample as a result of coulombic interactions between the incident electrons from the beam and electrons present in the sample. Emission of secondary electrons is such that as the angle between the beam and the surface increases, secondary emission increases (Fig. 4a). Thus, secondary electrons are ideal for imaging rough samples (such as fracture surfaces), where the angle between the beam and the surface is constantly varying. For a smoothly polished sample, where the angle remains constant across the surface of the sample, little or no contrast is seen using secondary electrons.

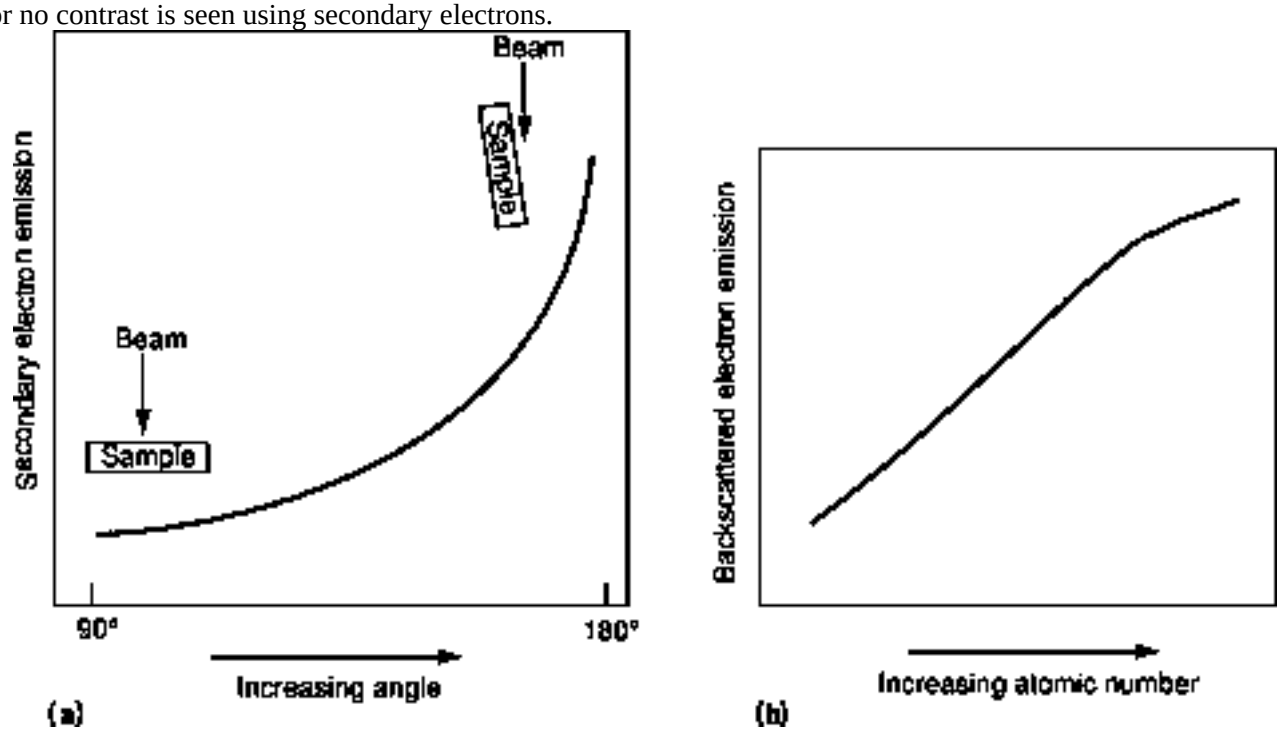


Fig. 4 Emission of (a) secondary electrons as a function of angle of incidence and (b) backscattered electrons as a function of atomic number

Due to their low energy, secondary electrons are typically collected to a photomultiplier/scintillator detector by placing a positive bias on the detector. For each pixel where the beam interacts with the sample, a signal is collected, processed, and displayed as an intensity level on a CRT (technically, a cathode ray tube; as a practical matter, a high resolution black and white or monochrome TV) screen. By coupling the scan of the beam to the scan of the CRT, a point by point correspondence exists between the surface roughness of the sample and the intensity observed on the screen. As the beam from the electron gun scans across a rough surface, the number of emitted secondary electrons continually changes, and the collected signal varies correspondingly. Thus, the intensity displayed on the CRT level changes from white (high-electron-emission) to black (low-electron-emission), and black and white contrast results in an image being produced in a manner exactly analogous to a black and white TV image.

The strength of the secondary electron signal often may be increased by tilting the sample slightly toward the detector. This results in more efficient collection of the secondary signal and can improve the signal-to-noise ratio without decreasing resolution. This may be helpful when trying to get the utmost resolution possible.

Backscattered Electrons. Imaging is somewhat different for primary backscattered electrons (BSE), which are incident electrons that have undergone elastic Rutherford scattering from the nuclei of the atoms that comprise the sample. For this signal, as the size of the nuclei gets larger (i.e., higher atomic number), the BSE emission increases in a nearly linear fashion (Fig. 4b). Thus, for a flat, polished sample containing numerous phases, where little or no contrast may be seen using secondary electrons, an image of high contrast may still be produced using the BSE signal. This is illustrated in Fig. 5, where a layered structure of various metals is imaged using secondary and backscattered electrons. While the layers can be clearly seen in the etched sample, it is only in the BSE image that the large difference in atomic number of the components is discerned. If a polished, homogeneous, single-phase sample were placed in an SEM (e.g., an unetched silicon wafer), little or no contrast would be observed using either technique.

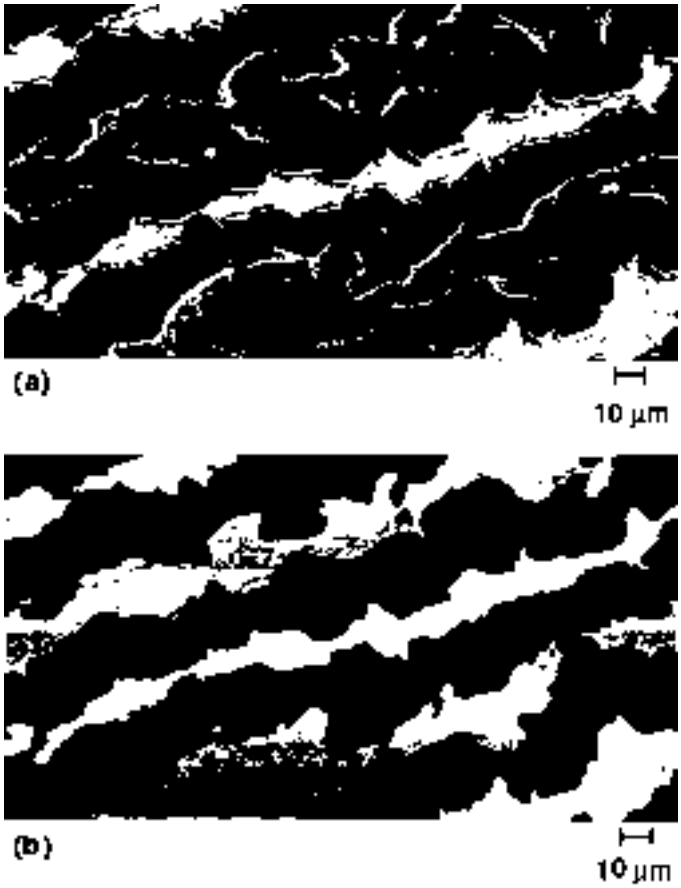


Fig. 5 Example of a layered structure of copper, niobium, and zirconium. (a) Imaged using secondary electrons. (b) Imaged using backscattered electrons. Copper appears dark gray, zirconium is light gray, and niobium is white.

Unlike the secondary electrons that can be attracted to a detector, the high energy BSE signal can only be collected by placing a detector in a position where it is likely to be struck by the emitted electrons. The most advantageous position is directly over the top of the sample. If the detector is designed such that different signals can be collected from different sides of the detector (usually divided into halves or quadrants), enhanced imaging called *topographical imaging* is possible by skillful use of the signals. By manipulating the signals received by different halves or quadrants of the BSE detector, small surface features that are indistinct in a secondary image can be imaged very well with BSE (Fig. 6). Topographical imaging works best on single phase samples that are nearly flat and possess only slight surface roughness, such as pits or ledges.

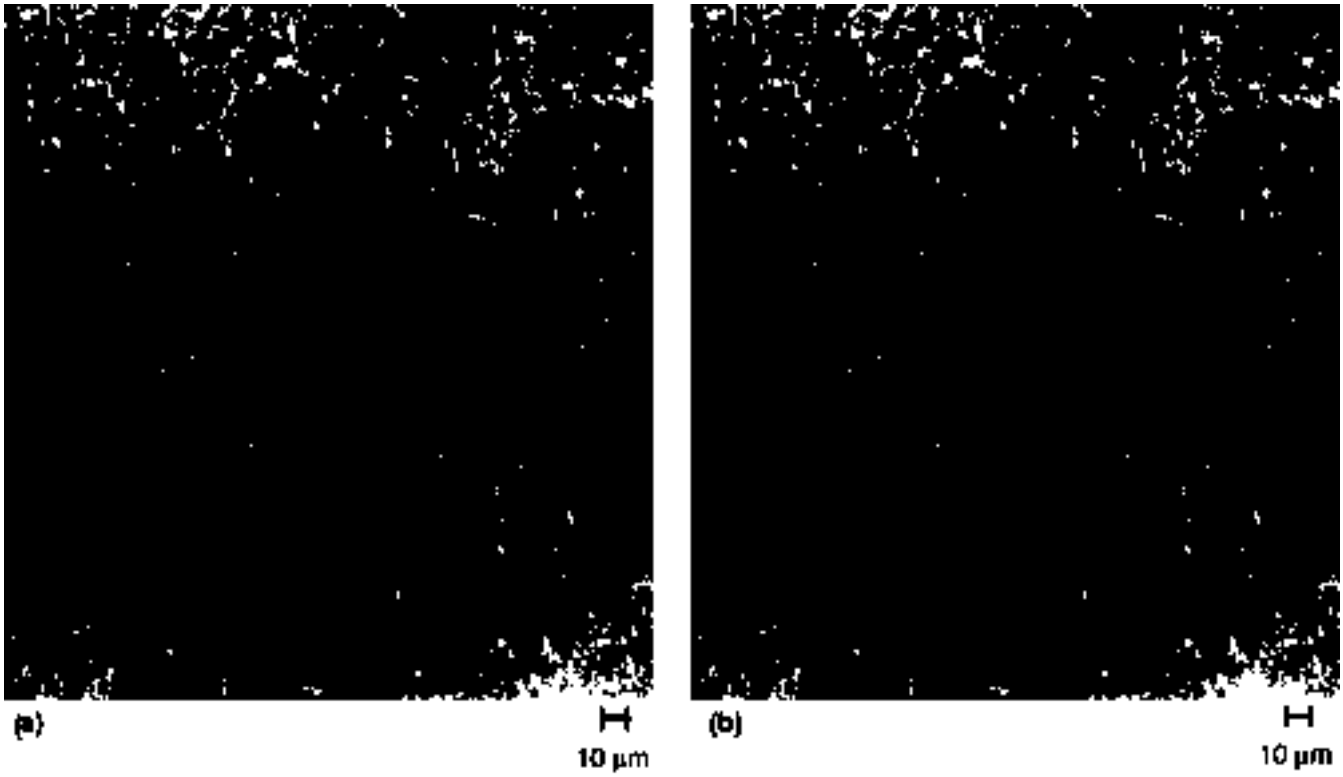


Fig. 6 Illustration of topographical imaging of microhardness marks using a backscattered electron detector. (a) Imaged using secondary electrons. (b) Imaged using backscattered electrons

Topographical imaging is possible due to the highly directional nature of the elastically scattered BSE signal. This characteristic, while useful in many investigations as was just mentioned, also can cause confusion when using the BSE signal to observe rough samples. In this instance, it should be remembered that the image obtained is not directly related to the degree of surface roughness (as is the case for the secondary image) but is a function of how that roughness is angled with respect to the detectors rather than the incident beam. When imaging a rough sample using BSE, the image can be altered radically by changing how the detected BSE signals are manipulated. The signals from all detectors should be added together when viewing BSE images from a rough sample. This should be remembered when using low-pressure or variable-pressure SEMs, where only a BSE signal is available for imaging in the low-pressure mode. This is especially true if the sample is not single phase, since the difference in BSE emission due to varying atomic number adds an additional complication. Having noted these potential problems, BSE imaging of rough surfaces often can still provide a rapid identification of areas that may deserve closer scrutiny. An example of this is shown in Fig. 7, where BSE imaging quickly reveals material of high atomic number embedded in the surface. This particular example is of a failed superalloy shaft. Lead from a nearby failed journal bearing, invisible in the secondary image, is readily apparent in the BSE image.

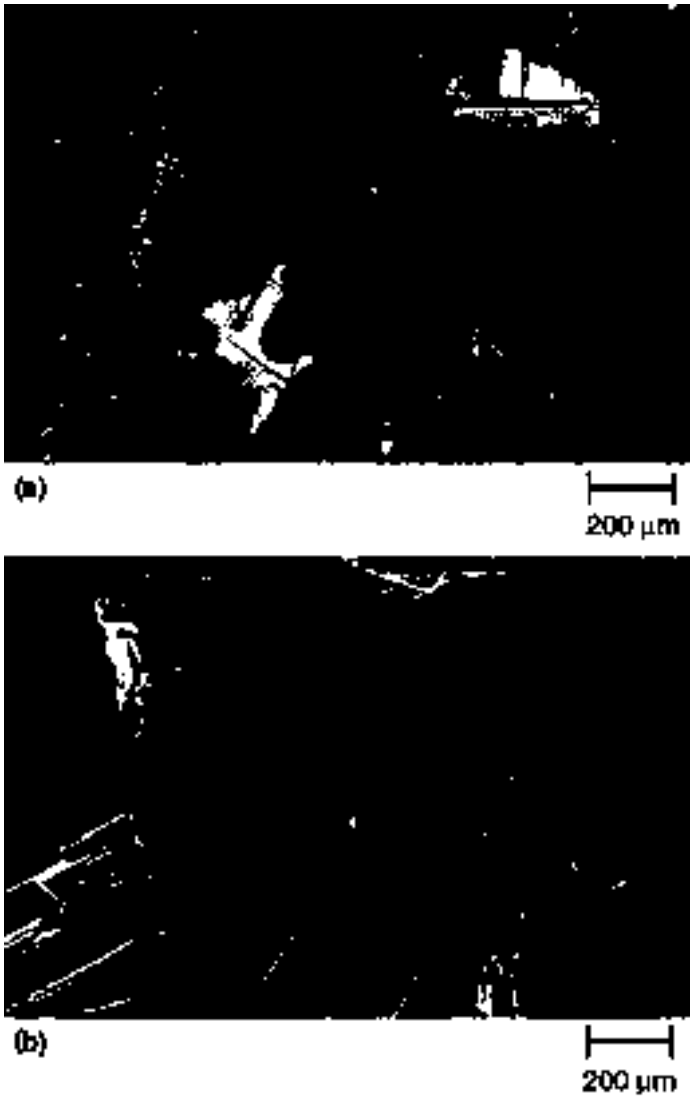


Fig. 7 Electron images of a superalloy shaft fracture. (a) Backscattered electron image. (b) Secondary electron image. Bright areas are lead contamination.

Characteristic x-rays are emitted due to collisions between incoming electrons and bound inner-shell electrons from the atoms that comprise the sample. If the incident electron has enough energy, the collision causes the ejection of the bound electrons. One way the atom decays to the ground state is by replacing the inner shell electron with a higher shell electron. Energy balance is restored by the emission of an x-ray whose energy is equal to the difference in energy levels of the electrons involved. Since the energy levels of all the electrons of all atoms are different, the energy of the emitted x-ray, termed the critical excitation energy, E_c, will be characteristic for that atom. By measuring the energies or the wavelengths of the x-rays emitted from a sample, the composition can be determined. The precision and accuracy of the determination depends on the type of x-ray detector used and how the analysis is carried out. Detectors that determine composition based on the wavelength of the emitted x-ray are termed wavelength dispersive spectrometers or WDS. Using WDS and pure elemental standards for the analysis, typical accuracies of 5% with a precision of 0.001 wt% can be obtained. However, these systems are the most expensive, and obtaining this level requires a fairly skilled operator. The more prevalent detectors determine composition based on x-ray energy. These energy dispersive spectrometers (EDS) are fairly inexpensive, easy to operate, and robust. A careful analysis using elemental standards can produce results of 5% accuracy with a precision of 0.1 wt%.

When conducting an x-ray analysis, it is important to realize that the area sampled is much larger than the size of the electron probe. This is due to spreading of the beam through collisions with the sample atoms, (Fig. 8). The total volume of the sample affected by the beam is termed the *interaction volume.* The shape of the interaction volume depends primarily on the accelerating voltage and the atomic number, Z, of the target material. Depending on these parameters, the shape may vary from being roughly semispherical for low voltages and/or high Z, to bottle-shaped, as shown in Fig. 8. It is clear that for the conditions of Fig. 8, x-rays may be emitted from the sample over an area considerably larger than the size of the electron probe. A reasonable estimate for the smallest area that can be sampled using x-rays is a 1 7m cube.

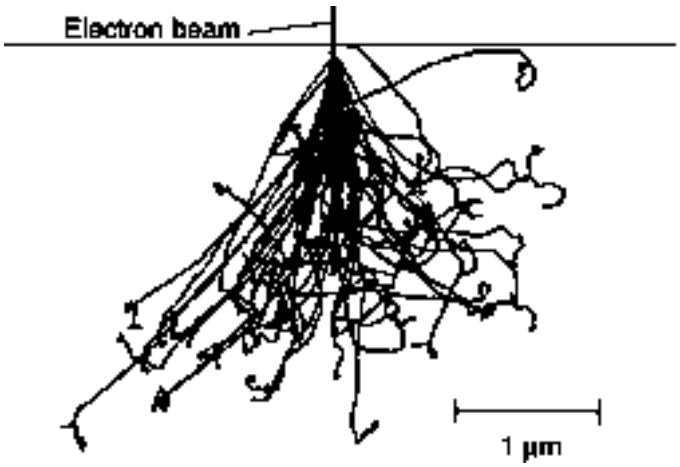


Fig. 8 Simulation of beam spreading of 20 keV electrons in zinc

The condition of the sample surface and the orientation of the surface with respect to the detector greatly affect both the size of the interaction volume and the numbers obtained. Quantitative EDS analyses should only be conducted on flat, polished samples to reduce the possibility of absorption of low energy x-rays by the sample. Analyses from rough surfaces can vary greatly, depending on whether the site sampled is facing toward or away from the EDS detector. Thus, EDS data on rough samples should only be used for qualitative investigations.

X-ray data can be collected for a single point or for a designated area. Point analysis is useful to investigate the compositions of discrete phases, corrosion products, deposits present on a surface, and similar situations. An area analysis averages out point-to-point variations. In either case, it must be remembered that only the composition of a thin surface layer is analyzed, and the results obtained may not be representative of the bulk material.

The working distance at which the signal is collected is another critical factor in quantitative analysis. The EDS detector is usually sighted for best signal collection by the manufacturer at a certain working distance from the objective lens. Acquiring a signal from a sample not at this optimum working distance may result in low and/or inaccurate counts. The manufacturer of the microscope and the EDS system should be consulted to determine the optimum conditions for x-ray collection.

Finally, it should be remembered that a certain amount of overvoltage, approximately $1.5 \times E_c$ (the critical excitation voltage of the x-ray), is required to obtain a desired x-ray. For example, in order to obtain copper k8 x-rays of $E_c = 8.04$ keV, the SEM accelerating voltage must be set greater than or equal to (1.5 × 8.04), or approximately 12.04 keV minimum. If the voltage is too low, these x-rays will not be generated or detected, even though copper may be present.

This leads to a critical point for the failure analyst. During a failure investigation, analyses are often conducted over a number of areas on a number of samples in different laboratories at different times. Before making comparisons regarding similarity or differences among x-ray spectra, one must consider whether the spectra were obtained under the same conditions, especially with respect to accelerating voltage.

Additional signals that can be monitored in an SEM are the specimen current and the electron backscattered pattern (EBSP). Similar in many respects to the specimen current, charge collection techniques are used extensively in the semiconductor industry (Ref 3) whilst the EBSP signal is used to determine crystal structure and orientation (Ref 4). More information concerning the use of these signals is available (Ref 5).

Vacuum System

All SEMS require a vacuum system to allow an electron beam to be produced and directed onto the target material. The constraints placed on the vacuum system depend on the type of SEM. With the advent of low-pressure and environmental SEMs, the design of vacuum systems has changed from a simple vacuum chamber to a multistage chamber with various levels of vacuum produced in different parts of the electron-optical column.

A high vacuum is desired in the gun chamber of all SEMs to minimize filament oxidation and prolong life. This is especially true in high resolution cold field emission SEMs, where oxidation of the tip can seriously reduce the emission of electrons. These SEMs have the most stringent vacuum requirements, and usually a circuit is supplied that allows the operator to heat the filament tip for a short period of time to degas any oxygen that adsorbs on the tip during operation. SEMs with LaB_6, filaments have the next highest requirements while tungsten filaments have the lowest requirements.

Isolation valves that separate the gun chamber from the sample chamber allow a high vacuum to be maintained in the gun area while samples are being inserted or removed. Separate pumps (such as titanium ion-getter pumps) attached to the gun continue to operate while the sample chamber is vented to atmosphere. These features are standard for field emission and LaB_6 guns and add extra complexity (and cost) to these systems. Separate pumps and an isolated gun chamber are often omitted from tungsten thermionic SEMs, which is one reason why they are the least expensive of the SEMs types. A

sample insertion port can often be purchased as an accessory that allows the entire column to remain under vacuum with the electron gun in operation while a sample is inserted through an air-lock into the specimen chamber. Such ports can significantly improve productivity by eliminating the need to turn the microscope off and vent the chamber each time the operator wishes to examine a new sample.

The introduction of low-pressure and environmental SEMs in recent years has resulted in even more complex vacuum systems. Since these microscopes require a given pressure to be maintained in the sample chamber, an even greater sophistication is involved in the vacuum design. The ability to maintain a high vacuum in the gun region and a poor vacuum in the sample chamber is achieved using a series of apertures and vacuum pumps.

References cited in this section

3. D. Newbury, D. Joy, P. Echlin, C. Fiori, and J. Goldstein, *Advanced Scanning Electron Microscopy,* Plenum Press, New York, 1986

4. B.L. Adams, S.I. Wright and K. Kunze, Orientation Imaging: The Emergence of a New Microscopy, *Met Trans A,* April 1993, p 819–831

5. J. Goldstein, D. Newbury, P. Echlin, D. Joy, A. Romig, C. Lyman, C. Fiori, and E. Lifshin, *Scanning Electron Microscopy and X-ray Microanalysis,* Plenum Press, New York, 1991

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Specimen Preparation

Preparation of samples for observation using an SEM has undergone a revolution in recent years with the introduction of low-pressure and environmental SEMs. Prior to their introduction, SEM samples were required to be electrically conductive to enable the electrons from the beam to be carried away, thus preventing charging of the sample and poor imaging. Samples were also required to be dry in order to maintain the vacuum level in the sample chamber and prevent outgassing that would affect the electron beam and again result in poor imaging. These stringent conditions often required samples to be prepared by a variety of methods in order to obtain a useable image. However, nonconductive samples, both dry and moist, can now be imaged successfully depending on the type of microscope used. Thus, sample preparation has become microscope dependent. When purchasing an instrument, consider as an additional cost factor the types of samples to be examined and the necessity for (or elimination of) sample preparation equipment. For example, in failure analysis investigations, parts are often submitted that contain residual oil or grease or show nonconductive corrosion products. For cases such as these, the additional cost of a low-pressure SEM might be offset by the time and money saved in eliminating sample preparation. Furthermore, in some investigations, SEM examination of greasy, dirty, or corroded parts in the as-received condition may be desired to, for example, identify constituents in the corrosion products in situ.

For a conventional SEM, the only constraint placed on examining a dry, conductive sample is that the sample fit in the SEM chamber. The size of the chamber needed should be carefully considered before a purchase is made since a larger, more costly chamber may allow some samples to be examined directly without sectioning. For small samples that need to be secured in some type of resin for polishing and etching, conductive mounting materials are available for purchase. If a nonconductive mount such as normal bakelite is used for polishing, a conductive path from the sample to the SEM specimen holder or stage must be made to prevent sample charging. This is achieved using metal or carbon tape, or by painting a path on the side of the mount using a conductive solution such as Dag (Acheson Colloids Co.). (Dag is available from various manufacturers including Buehler Inc. of Lake Bluff, IL; Struers, Inc. of Westlake, OH, and Structure Probe of West Chester, PA).

If a nonconductive sample is to be examined using a conventional SEM, several methods exist to produce a conductive layer on the surface. The two most common are evaporation and sputter coating. Evaporation coating consists of heating a metal filament to a high temperature in a vacuum chamber to cause evaporation to occur. Atoms that leave the filament travel a line-of-sight trajectory to the surface of the sample where they are adsorbed. A thin coating consisting of a few monolayers of atoms can quickly result, producing a conductive layer across the entire surface of the sample. The heating is usually achieved in a wire filament by passing an electric current through it or, in a small metal sample, by using a levitation induction coil. Carbon evaporation is also possible using a slightly modified process. In this instance, a carbon filament is placed between two electrodes in the vacuum chamber. A current is rapidly introduced through the filament,

causing the carbon fibers to heat and burn out. A large amount of evaporated carbon is produced during the moment of filament burnout.

Evaporation methods are very easy and reliable provided a clean vacuum chamber is maintained, since degradation of the vacuum can cause the quality of the coatings to suffer. Typical metals used for coating are gold and platinum, with finer grained coatings being produced as the evaporation temperature increases. One drawback of metallic films is that the coating produced will be detected in any subsequent x-ray investigations. If chemical analysis of the sample is required, it is best to coat the sample with carbon, which is more difficult to detect using EDS and is less likely to obscure any additional elemental peaks. This is assuming, of course, that one is not wishing to determine the composition of carbon in the sample.

Sputter coating is similar in that the result is a few monolayers of a metal being produced across the surface of the sample. However, in this instance the metal atoms are not removed by evaporation but rather by bombardment from heavy argon ions in a manner analogous to sand blasting. In sputter coating, the sample is placed in a vacuum chamber beneath a metallic target, usually of gold or gold-platinum. After a vacuum is achieved, a partial pressure of argon is introduced into the chamber and a voltage placed across the target material and the sample. The target material is negatively biased with respect to the sample. The presence of the electric field causes an argon plasma to form, producing argon ions. These positive ions bombard the negative target, knocking off atoms of gold (or gold-platinum) that travel by line-of-site to the sample beneath and adhere to the surface. The thickness of the film depends on the time of deposition and the deposition rate that results in forming the plasma.

Many times in failure investigations, it is impractical or impossible to insert a sample into the SEM chamber. In cases such as these, the surface can still be examined at high magnifications using an SEM by making a surface replica of the surface. Production of a replica involves placing a thin sheet of a suitable polymeric material, such as cellulose acetate, on the surface to be examined (Fig. 9). The cellulose acetate is then soaked with acetone to soften the material, and capillary forces between the film and the surface pull the sheet tightly to the surface. Once the cellulose acetate has dried, the material is pulled away from the surface, producing a mirror replication of the surface in the form of the now somewhat stiff acetate. This film is made conductive either by sputter coating or evaporation, and the resultant sample is examined in the SEM. Many times this same technique can be used to remove small particles or inclusions from the surface of a sample. In this instance the particles adhere more tightly to the acetate film than they do to the substrate, and, when the film is removed, the particles can be found clinging to the surface of the acetate. Termed an extraction replica, this technique is useful for removing particles from a rough surface so that more accurate quantitative information can be obtained from the particles using x-ray analysis.

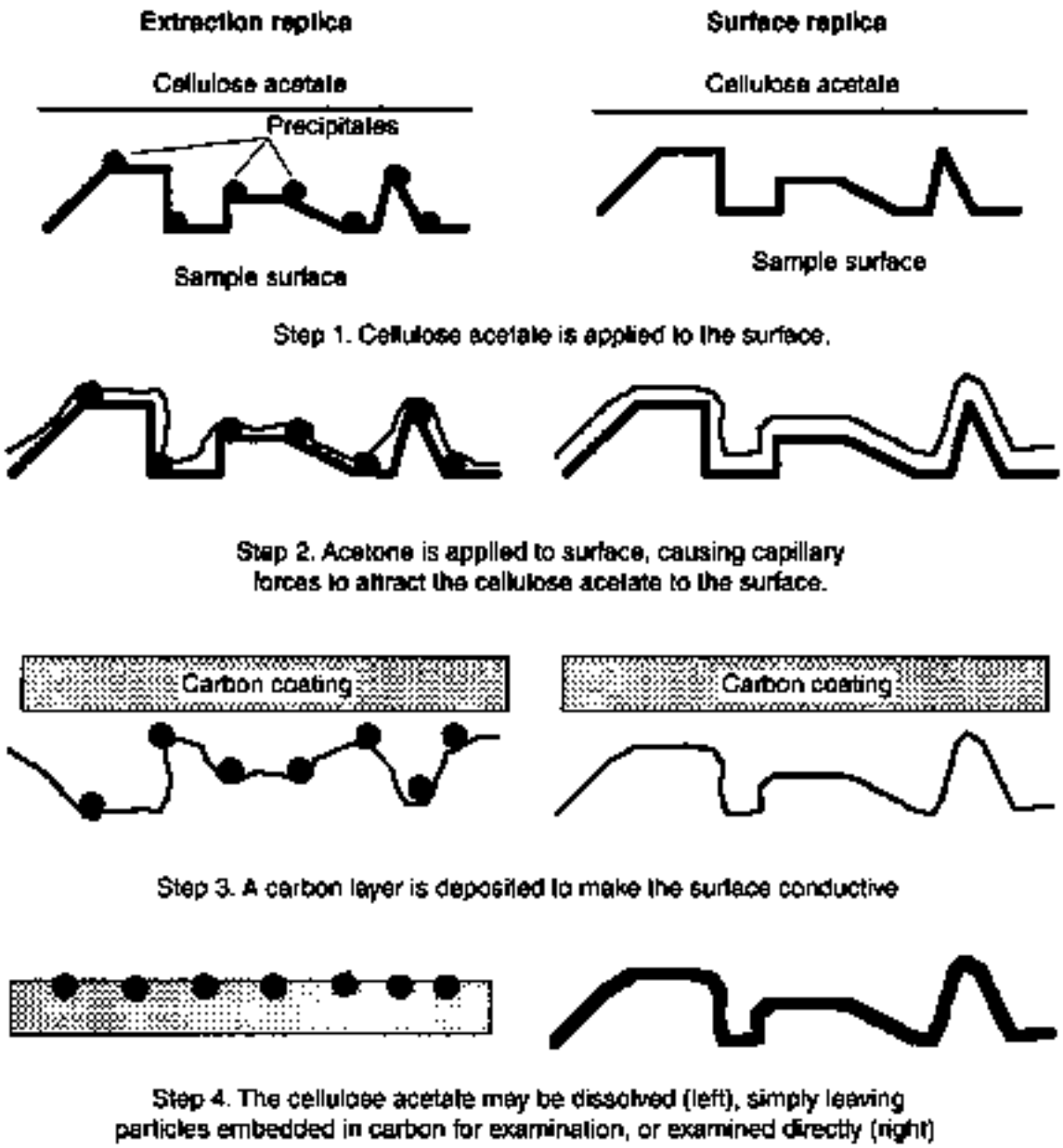


Fig. 9 Schematic showing method for producing extraction replicas and surface replicas for SEM examination

New liquid epoxy resins are also available for making surface replicas for SEM investigations. A small portion of the resin is mixed and applied to the surface. Some manufacturers market kits where the resin is mixed automatically when using an application gun. Different mixtures are available, providing different curing times at various temperatures. Upon drying, the polymeric material is simply peeled from the surface to be examined. Resolution of the cured resin is on the order of 0.1 7m and some mixtures are available that can be examined directly in the SEM.

As discussed in other sections of this Volume, more than one analyst may be interested in examining the specimen(s) involved in a failure investigation. Other interested parties should be given the opportunity to examine the specimen before it is prepared for examination in the SEM (or altered in any other way).

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Application of SEM in Fractography

Fractography is one of the most common applications of SEM. The large depth of focus, the possibility of changing magnification over a wide range, very simple nondestructive specimen preparation with direct inspection, and the three-dimensional appearance of SEM fractographs make the SEM an indispensable tool in failure studies and fracture research. Although the interpretation of microfractographs requires practice and understanding of fracture mechanisms, there are only a small number of basic features that are clearly recognizable and indicative of a particular mode of failure:

- Dimpled fracture, typical of overstress failures of ductile metals and alloys
- Cleavage facets, typical of transgranular brittle fracture of body-centered cubic (bcc) and hexagonal close-packed (hcp) metals and alloys
- Brittle intergranular fracture typical of temper-embrittled steel, in which fracture is due to segregation of an embrittling species to grain boundaries, to intergranular stress-corrosion cracking, or to hydrogen embrittlement
- Stage II striations, typical of fatigue failure

These typical forms of fracture surfaces are briefly described in the following section, “Typical Fracture Surfaces.” More extensive coverage on the interpretation of fracture surfaces is contained in the Section “Fracture” in this Volume. Image dependence on microscope type and operating parameters is discussed in the final section of this article.

Typical Fracture Surfaces

The appearances of fracture surfaces are directly related to the mechanism by which the failure occurs. If the surface remains unaffected and undamaged by events subsequent to the actual failure, it is often a simple matter to determine the failure mode by use of an SEM. In cases where the surface is altered after the initial failure, the case may not be so straightforward. A few typical examples that illustrate these points are shown subsequently. Other sections of this Volume provide extensive further discussion of these fracture mechanisms.

Ductile failures result when a material is plastically deformed, usually in tension, to the point of failure. In the early stages of deformation, small holes called microvoids form, usually around weak points in the material, such as inclusions or second phase particles. As deformation proceeds, the microvoids grow and begin to link (or coalesce), creating larger voids. This process continues until failure occurs. The result of this sequence of events can be observed on the fracture surface of a ductile failure, as shown in Fig. 10. Note the presence of round inclusions at the bottom of several coalesced voids, which appear as small depressions.

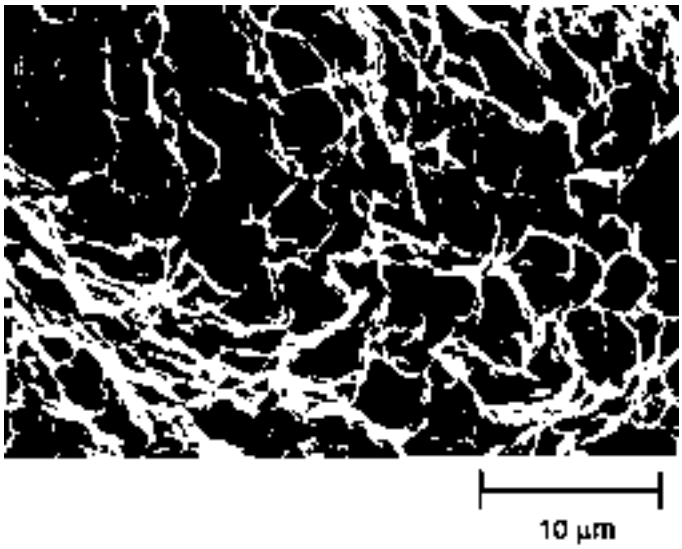


Fig. 10 Typical fracture surface exhibiting ductile failure in a mild carbon steel

Brittle failures typically occur with very little plastic deformation near the formation of the crack prior to rupture. In a brittle fracture, the material may fail by cleavage along specific atom planes or directions, or it may fail due to separation of microstructure at specific sites caused by embrittling impurities or corrosive attack. Regardless of the reason, brittle failures can usually be identified by the sharp, angular appearance of the fracture surface. Figure 11 shows examples of three types of brittle fracture. Figure 11(a) shows the surface of a steel that was fractured below the ductile/brittle transition temperature. The cleavage of the material is apparent, with river markings indicating propagation of the fracture along nearly parallel sets of planes.

Fig. 11 Brittle fractures. (a) Fracture of mild carbon steel below the ductile/brittle transition temperature. Note the appearance of river lines on the faces of the cleavage surfaces. (b) Fracture of a soda-lime glass. Note similarity of river lines to those of (a). (c) Intergranular stress-corrosion cracking in brass. The grain structure is clearly revealed as the material fails along the grain boundaries.

This image should be compared to that of Fig. 11(b), which shows the fracture surface of a common soda-lime glass. Note that both brittle materials exhibit similar river lines marking propagation of the crack along converging lines and pointing back to the origin of the crack. However, the amorphous structure of the glass does not produce the flat cleavage surfaces apparent in the steel. Instead, extremely large, smoothly scalloped regions exist. These are often found in brittle fractures of glasses and polymers.

The brittle failure of the steel and glass is remarkably different from that of Fig. 11(c). Although Fig. 11(c) still exhibits a sharply angular structure common to brittle failures in metals, in this instance the appearance of the surface is due to failure along grain boundaries. The faceted appearance of the grains is quite distinctive and is indicative in this case of intergranular failure due to stress-corrosion cracking.

Mixed Mode Failure. In many instances, the fracture consists of both ductile and brittle regions. An example of such a mixed mode fracture, consisting of ductile tear and brittle cleavage, is shown in Fig. 12. This example is a good illustration of the difference in scale of the surface features associated with ductile and brittle failures.

Fig. 12 Mixed mode fracture in a mild carbon steel cooled just to the ductile/brittle transition

Fatigue Failures. Not all failures will exhibit the characteristic ductile or brittle features displayed in the previous images. This is especially true for fatigue failure where the growth of a crack occurs over an extended period of time as the piece is cycled from one stress state to another while in use. In fatigue, crack propagation occurs with each cycle, and the characteristic features associated with a ductile or brittle failure do not appear. Instead the surface of a classic fatigue failure will be marked by a series of striations, observed at high magnification, which mark the advancement of the crack front with each stress cycle. This is illustrated in Fig. 13 for an aluminum alloy, Fig. 14 for a Type 304 stainless steel, and Fig. 15 for a Grade 250 maraging steel.

Fig. 13 Ductile fatigue striations in aluminum alloy 2024-T3. (a) The uniformity of the crack-propagation process is well illustrated. There was little or no interaction of the fracture process with the inclusion within the rectangle. The long ridges are believed to be high-angle steps at subgrain boundaries. 650×. (b) Higher-magnification view of the region outlined by the rectangle in (a), showing the continuity of the fracture path through and around the inclusion. Fracture "wake lines" are visible and indicate that the direction of crack propagation was toward the lower right corner of the fractograph. These wake lines originate at the inclusion and run normal to the crack front in the direction of crack propagation. 2600×

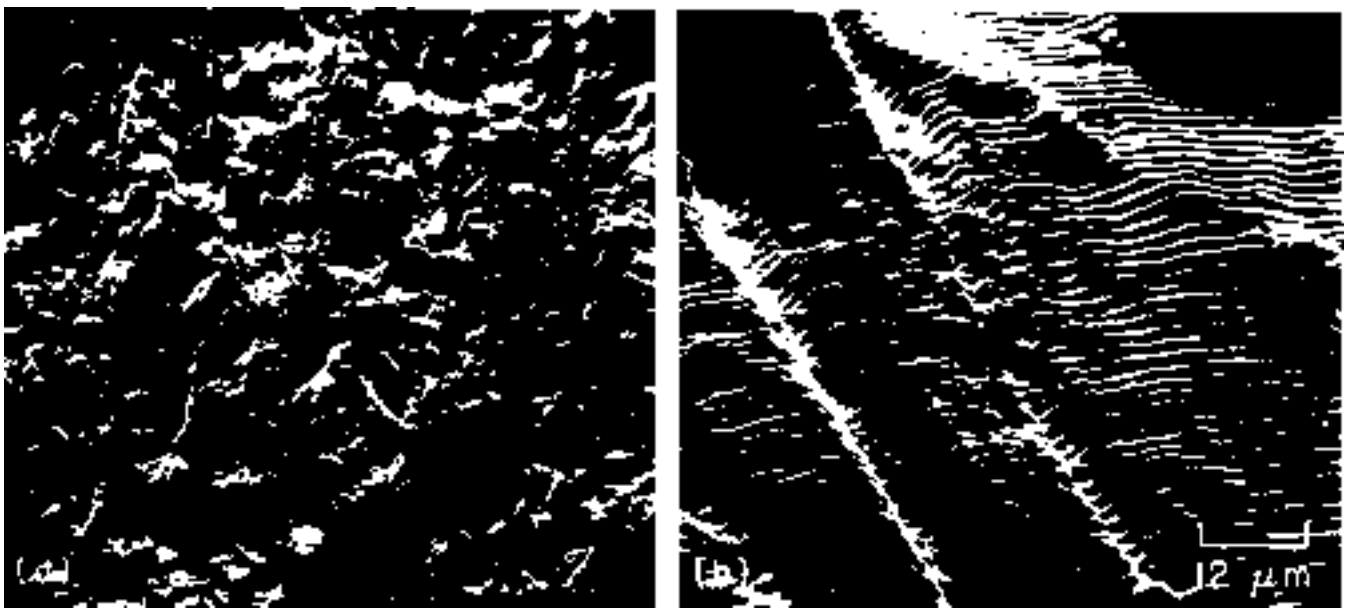


Fig. 14 Fatigue fracture in type 304 stainless steel tested at room temperature. The vertical secondary cracks in (a) are grain-boundary separations. 39×. (b) The well-defined striations resulted from the planar slip characteristic of stainless steels. 810×

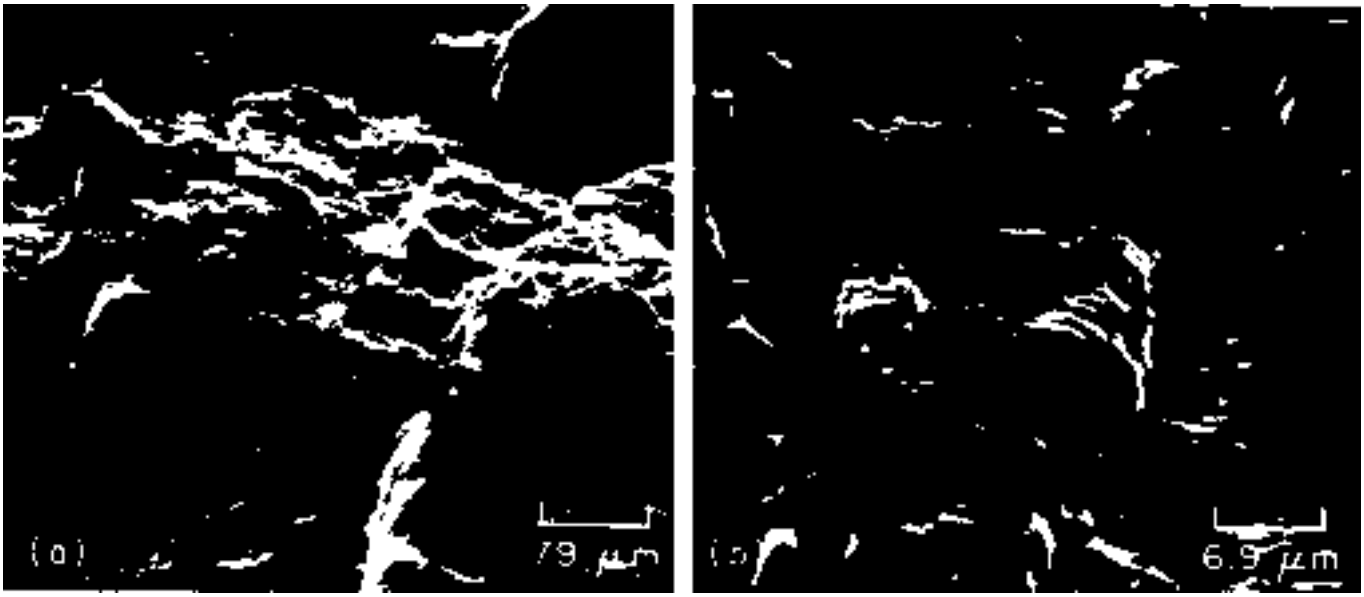


Fig. 15 Fatigue fracture in 18% Ni, grade 250, maraging steel tested at room temperature. (a) Extensive secondary cracking can be seen in a lower-magnification fractograph. 126×. (b) The cracking is clearly evident in a higher 1440× magnification fractograph. Secondary cracks formed at the roots of many fatigue striations. The range of stress intensity (*K*) at the crack tip was 76.9 MPa · $\sqrt{m}$(70 ksi · $\sqrt{\text{in.}}$).

The striations that are formed at very low crack-growth rates (less than 5×10^{-6} in. per cycle) can be difficult to resolve and often cannot be distinguished from the network of slip lines and slip bands associated with plastic deformation at and near the fracture surface. Under these circumstances, measurement of striation spacing is difficult, if not impossible to measure. In steels and other materials, fatigue striations that may form at ordinary crack-growth rates also are not always as well-defined as in aluminum alloys. This is the more common situation, because the fatigue striations are often obscured by subsequent damage to the crack surface as crack propagation proceeds. In this instance the surface of the fracture appears smeared and the fatigue striations may be difficult or impossible to see. The smeared nature of the fracture surface may still permit an identification of the failure as being fatigue related, especially if the suspected fatigue region can be compared to the point of final overload within the sample. An example of this is shown in Fig. 16(a), where the regions of fatigue and final ductile overload are easily identified.

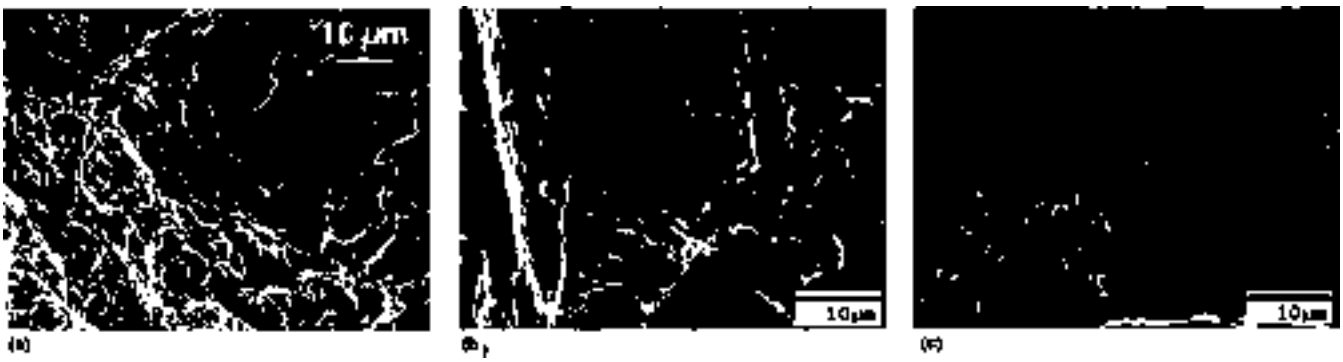


Fig. 16 Fatigue failures. (a) Fatigue (upper right) changing to ductile failure (lower left). (b) Fatigue fracture of class 30 gray iron. (c) Single overload fracture of class 30 gray iron

Comparisons such as that shown in Fig. 16(a) often are the last resort for determining the nature of the failure mode. This is especially true in the fatigue of a brittle material such as cast iron, where the fatigue surface appears almost indistinguishable from the final failure (Fig. 16b and 16c). In cases such as this, identification of the true failure mode may be extremely difficult.

Dependence of Image on Microscope Type and Operating Parameters

A previous section noted that the appearance of any surface imaged with an SEM is influenced to a great extent by the type of SEM employed and the operating parameters used for the investigation. A few specific examples are shown below.

Low-Pressure Microscope. The biggest advantage of low-pressure microscopes is that they allow the imaging of nonconductive surfaces, eliminating the coating that would be required for clear imaging in a conventional SEM. This is illustrated by Fig. 17, which shows a section of PVC tube that failed due to fatigue. The image obtained shows excellent contrast without necessitating that the sample be coated in any way. The chamber pressure for this particular example was 0.3 torr. The importance of selecting a suitable chamber pressure is illustrated in the micrographs of Fig. 18, which present a series of images of nonconductive particulate taken under various conditions using a modern low-pressure SEM. In Fig. 18(a), the microscope is operated as a conventional SEM. Charging of the sample is evident, and an indistinct image is obtained. Figure 18(b) shows the same region now imaged in low-pressure mode. A satisfactory BSE image is obtained showing good contrast and resolution. The quality of the image contrast is degraded, however, if the pressure used for imaging becomes too great, as shown in Fig. 18(c).

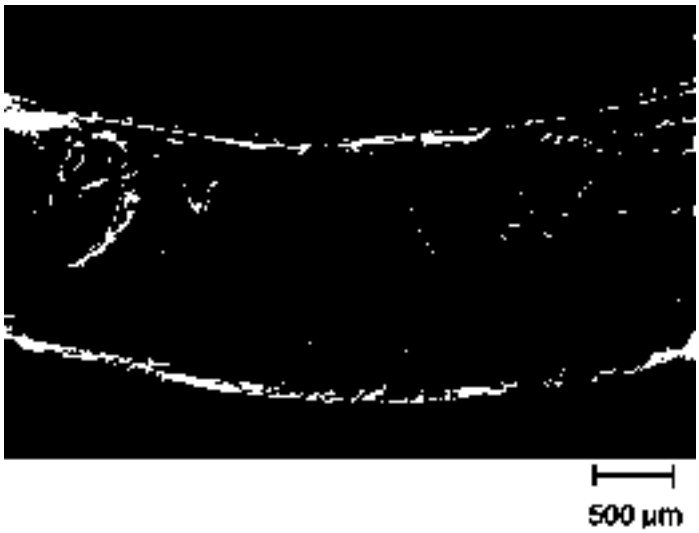


Fig. 17 Fatigue failure of a nonconductive polyvinyl chloride pipe imaged in the uncoated state using a low-pressure microscope

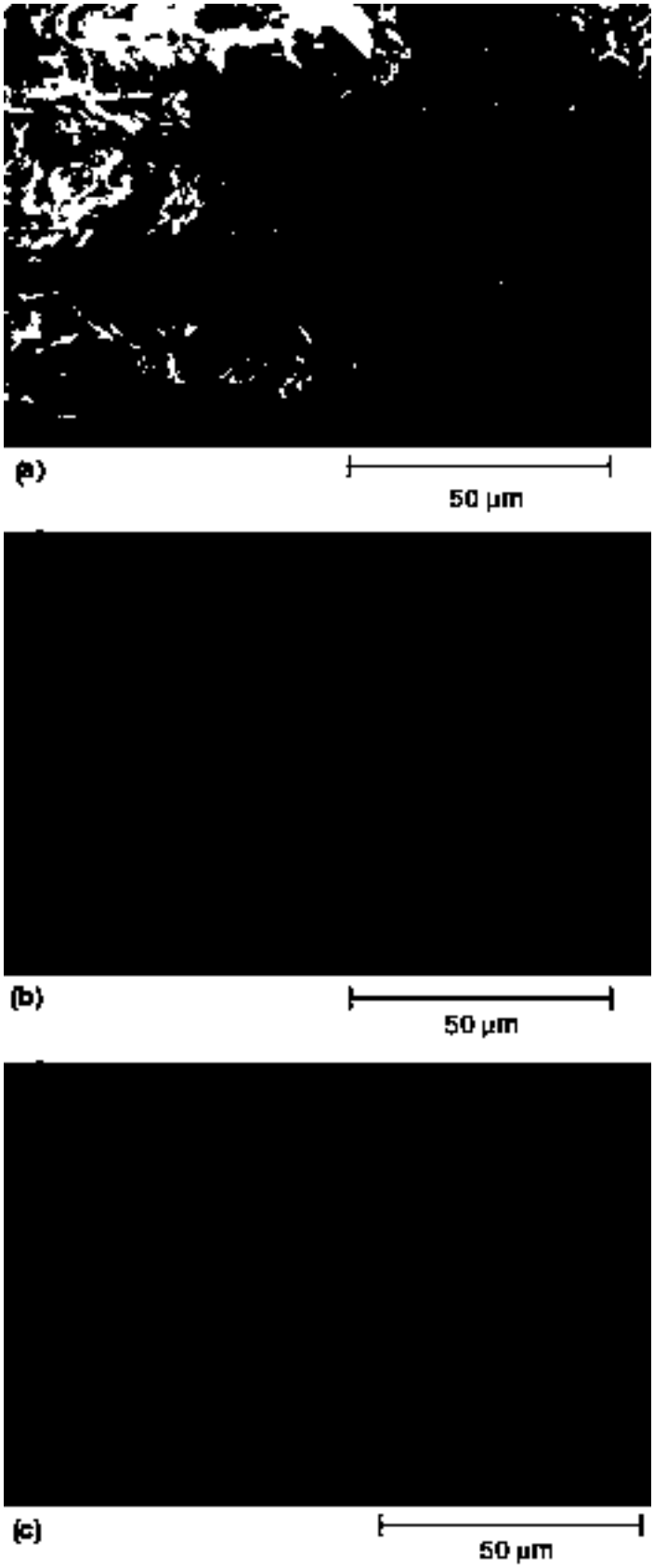


Fig. 18 Nonconductive material imaged using (a) conventional imaging, (b) low-pressure imaging with 20 Pa gas pressure, and (c) using 270 Pa gas pressure

Higher Resolution Field Emission Microscope. Field emission microscopes offer the ability to image samples at higher resolution than a conventional thermionic (LaB_6 or tungsten) microscope. Figure 19 shows a comparison of pearlite lamellae imaged using an inexpensive tungsten-filament microscope and a field emission instrument under identical operating conditions (15 kV; 16 mm working distance; moderate condenser lens strength). The field emission machine clearly produces a better image. However, this comes at an increase in instrument cost. The question often arises when

purchasing a new microscope: at what point is the increased cost justified? This question cannot be definitively answered for all samples and situations. Typical sample condition, operator expertise, and the information to be obtained all must be considered in purchasing to determine what the normal operating mode will be. For the example given in Fig. 19, the inexpensive tungsten-filament instrument still provides much of the same information as the (much more costly) field emission machine. If most investigations will be conducted at magnifications of 10,000× or below, the wisest choice may be to purchase the less expensive and easier to maintain thermionic machine.

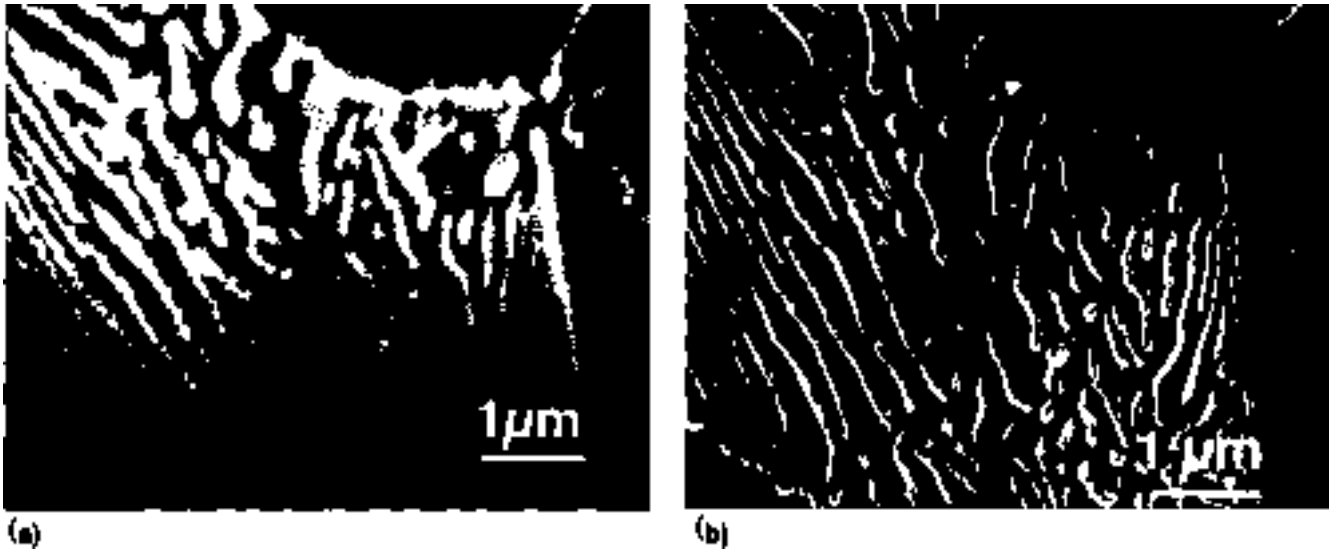


Fig. 19 Comparison of (a) conventional tungsten filament SEM to (b) field emission instrument. Samples imaged at 10,000×

High Accelerating Voltage versus Low Voltage. Accelerating voltage also plays a large role in the appearance of the observed secondary electron image. As accelerating voltage increases, the interaction volume shape changes from semispherical to bottle shaped. Most secondary electrons are emitted from within 100 Å of the surface, so as the neck gets smaller and smaller with increasing voltage the relative number of secondaries coming from near the surface becomes less. Conversely, as accelerating voltage decreases, the secondary signal from the near-surface region increases. This results in an image with greater surface information than is possible at the higher voltage. This effect is illustrated in Fig. 20. The images were obtained using a convention tungsten-filament SEM operated under identical conditions except for accelerating voltage.

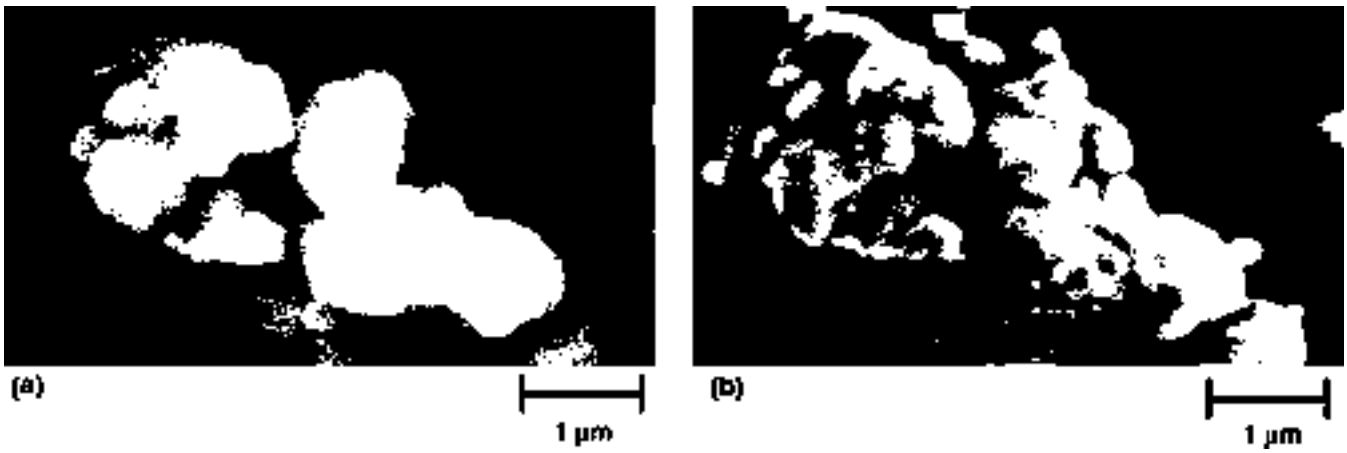


Fig. 20 Affect of accelerating voltage on imaging. Sample is gold-coated ZnS powders imaged using (a) 25 kV and (b) 5 kV. Note enhanced surface information at 5 kV. Both 20,000×

Surface films are also best evaluated at lower accelerating voltage. Figure 21 shows a defect area in an anodized coating on a titanium component for a medical device. Using light microscopy, a small white spot was visible on the blue

anodized coating. With the SEM at 20 keV, the spot was barely detectable. The defect area was easily characterized as a void in the anodized coating by inspection using a 2 keV accelerating voltage.

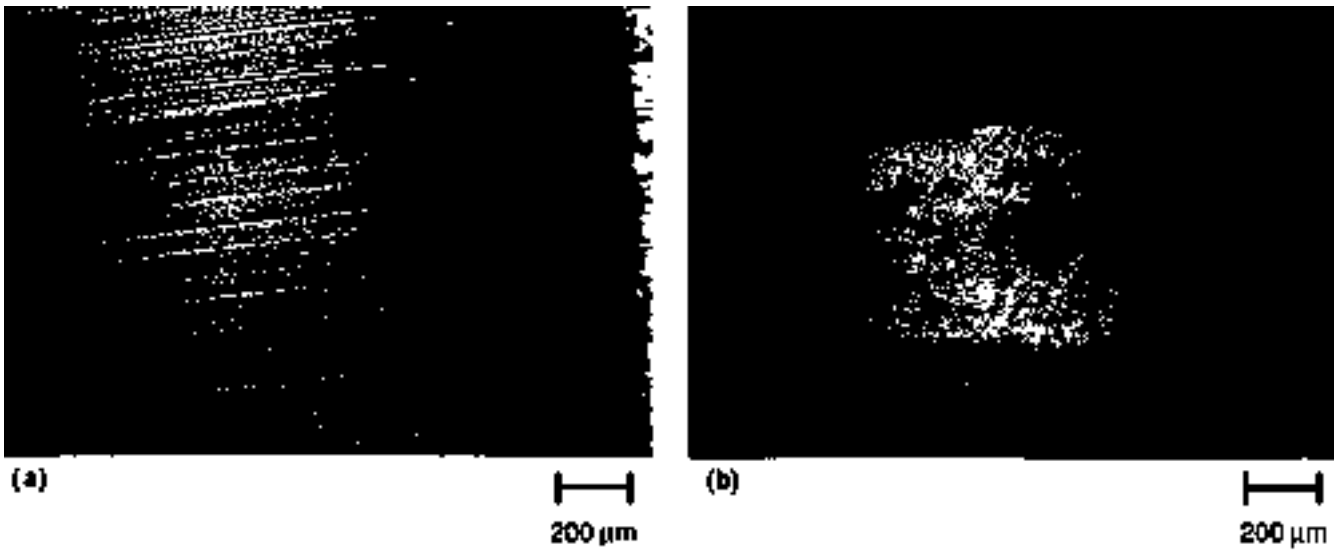


Fig. 21 Examination of a surface film at (a) 20 kV and (b) 2 kV

Figures 20 and 21 clearly show the advantages in imaging that can be obtained by going to lower voltages. However, it must be remembered that if one wants to simultaneously obtain chemical compositions using x-ray analysis, the accelerating voltage must be sufficient to excite these x-rays. Thus, it may not be possible to obtain the best imaging conditions while also obtaining the best x-ray signal. It should also be noted that in some instances, using higher accelerating voltages may be useful in obtaining images of features that are invisible using other techniques. Figure 22 is a good illustration of this. Optical observation of this painted surface revealed only a small surface bump, as did low voltage imaging using the SEM. Imaging using 30 keV, however, revealed microscopic particles under the surface at the bump that could be identified as iron. Thus, the higher voltage allowed identification of these near-surface contaminant particles when other methods failed.

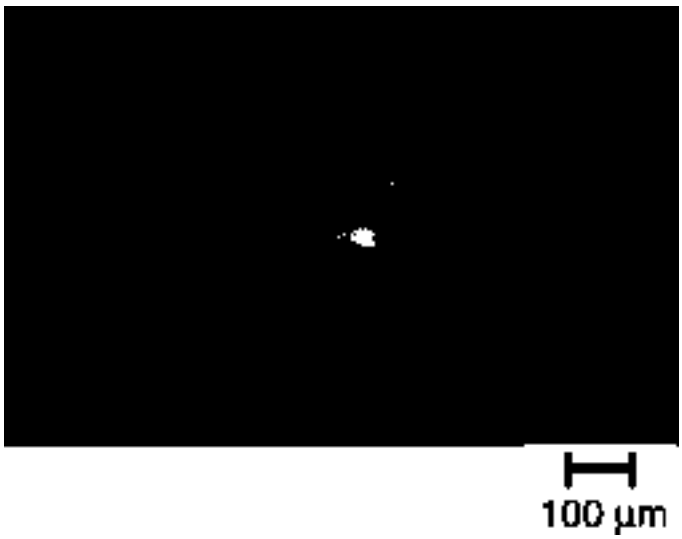


Fig. 22 Near-surface contaminant particles visible using 30 kV accelerating voltage

Comparison and Recommendations

The preceding discussion is meant to convey that a large number of factors need to be considered when deciding which instrument is best suited for any particular application. In many failure applications, the need for SEM work is limited to observations below 10,000× magnification. Essentially, any SEM should be able to obtain good images at these magnifications. Purchase decisions, therefore, are usually based on other factors than mere magnification considerations. Types of samples examined, the size of the samples, information required, and so on, all play a role in choosing an

instrument, with cost usually being the overriding consideration. If 95% of the specimens examined are large, metallic samples where only low magnifications are desired, an inexpensive W-filament machine equipped with an EDS detector provides extremely good value for the amount of funds expended. As sample variety increases to include insulating polymer and ceramic materials, a low-pressure SEM may be more advantageous. If higher resolutions are required for a considerable amount of work, the field emission machines may need to be purchased.

The wide range of capabilities now available in SEMs makes it possible to purchase machines especially designed for specific applications. Only the operator can determine which SEM features are necessary and which are optional. A skilled operator with a thorough knowledge of the operation of the instrument may be able to obtain images beyond the capability of a novice. Thus, ease of operation may be the final factor in determining the best instrument for any given application.

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Acknowledgments

Paul Stevenson of Perkins Engines, Ltd. is gratefully acknowledged for his assistance in acquiring several images.

References

1. M. Von Ardenne, *Z. Phys.*, 109, 1938, p 553
2. D. Dingley and Baba-Kishi, Use of Electron Back-Scatter Diffraction Patterns for the Determination of Crystal Symmetry Elements, *Scanning Electron Microsc.*, 2, 1986, p 386
3. D. Newbury, D. Joy, P. Echlin, C. Fiori, and J. Goldstein, *Advanced Scanning Electron Microscopy*, Plenum Press, New York, 1986
4. B.L. Adams, S.I. Wright and K. Kunze, Orientation Imaging: The Emergence of a New Microscopy, *Met Trans A*, April 1993, p 819–831
5. J. Goldstein, D. Newbury, P. Echlin, D. Joy, A. Romig, C. Lyman, C. Fiori, and E. Lifshin, *Scanning Electron Microscopy and X-ray Microanalysis*, Plenum Press, New York, 1991

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

Introduction

MANY ANALYTICAL TECHNIQUES are available for the study and characterization of surfaces. These techniques provide data about the physical topography, physical properties, chemical composition, and chemical structure of the surfaces under study. Most of these techniques are based on bombarding the surface with photons, x-rays, ions, neutrons, or electrons and analyzing the radiation emitted and/or reflected from the surface. Other techniques use other interactions, such as physical probing of the surface.

A few key techniques have become standardized and are available to most failure analysts, either through an in-house laboratory or from an outside service laboratory. The most common analytical methods for chemical characterization of surfaces are shown in Table 1. The techniques to be applied to a particular failure depend on the type and size of the

sample, the depth of analysis, the type of information sought, the ease of performing the analysis, the allowable destruction of the sample in either preparation or analysis, and the cost/time required. The information required about the surfaces in a failure analysis varies from failure to failure. No one technique can fully characterize a surface, but a full characterization is seldom required to solve a particular problem.

Table 1 Evaluation techniques for chemical characterization of surfaces

Technique	Information	Analysis depth	Analysis area	Detection limit	Ease of use
EDS	Elemental	<5 7m	<1 7m	<1 at.%	Easy
WDS	Elemental	<5 7m	>1 7m	<0.1 at.%	Easy
AES	Elemental	<5 nm	>100 nm	<0.5 at.%	Moderate
XPS	Elemental, chemical structure	<5 nm	>10 7m	<0.1 at.%	Moderate
TOF-SIMS	Elemental, molecular structure	<2 nm	<5 7m	<1 ppm	Moderate
FTIR	Chemical structure	<5 7m	>10 7m	<100 ppm	Moderate
Raman	Chemical structure	>1 7m	>1 7m	<0.1 at.%	Difficult

Note: EDS, energy-dispersive spectroscopy; WDS, wavelength-dispersive spectroscopy; AES, Auger electron spectroscopy; XPS, x-ray photoelectron spectroscopy; TOF-SIMS, time-of-flight secondary ion mass spectrometry; FTIR, Fourier transform infrared (spectroscopy)

Understanding the various analytical techniques allows an analyst to select the most appropriate method(s) to obtain the data needed for each failure. In many cases, a combination of analytical techniques may be required to evaluate the physical and chemical nature of the surface under study. This article covers the chemical characterization of surfaces by Auger electron spectroscopy (AES), x-ray photoelectron spectroscopy (XPS), and time-of-flight secondary ion mass spectrometry (TOF-SIMS). Chemical characterization of surfaces by energy-dispersive spectroscopy (EDS) instrumentation, which is commonly a module integrated with modern scanning electron microscopes, is discussed in many other articles in this Volume; an overview of EDS and wavelength-dispersive spectroscopy is provided in the article "Chemical Analysis of Metals in Failure Analysis" in this Volume. Fourier transform infrared spectroscopy is used more extensively in the analysis of polymers (see the article "Characterization of Plastics in Failure Analysis" in this Volume).

Table 2 is a summary chart of techniques discussed in this article. In this brief review, it is not possible to develop all the capabilities of each technique, but some of the more important attributes are highlighted for a preliminary insight into the strength and usefulness of these techniques for chemical characterization of surfaces.

Table 2 Summary chart of techniques discussed in this article

Feature	Technique		
	AES	XPS	TOF-SIMS
Probe beam	Electrons	X-ray photons	Ions
Analyzed beam	Electrons	Electrons	Ions
Average sampling depth	5 nm	5 nm	2 nm
Detection limits	10^{-3}	10^{-4}	10^{-6}
Spatial resolution	10 nm	5–10 7m	150 nm
Information	Mostly elemental, SEM photos	Elemental, chemical	Elemental, molecular
Strengths	Ultimate small area analysis, imaging	Ease of use, quantification	Chemical and molecular analysis, imaging
Limitations	Semiconductive	Very few	Quantification difficult
Major applications	Semiconductors, electronics	All industries	Polymers, contamination, trace metal analysis

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

Overview of Surface Analysis

What Is Surface Analysis? The “surface” of a sample can mean different things to different people. To those performing analytical tests at the air/specimen interface of samples, the word surface can be anywhere from the top monolayer to as deep as several micrometers into the sample, depending on the depth of analysis of the technique that is being applied. Although there are scores of different techniques currently being used to study the surfaces of materials, only a fraction of these can be classified as true surface analysis tools that derive the majority of their analytical signal from the top few atomic layers. Three such techniques include AES, XPS—also known as electron spectroscopy for chemical analysis (ESCA)—and TOF-SIMS. The electron spectroscopy techniques of AES and XPS have depths of analysis, on average, of approximately 5 nm, while TOF-SIMS is even more surface sensitive, deriving the majority of its signal from the top 2 nm. Because these techniques are so surface sensitive, they are often employed in failure analysis or general surface characterization. Typical applications include identification of thin layers of contaminants on surfaces or at interfaces, evaluation of cleaning processes, and the identification of stains and discolorations.

Successful failure analysis can often be a matter of piecing together the different bits of information that different techniques can provide. Each of these three surface analysis techniques provides a different view of the sample surface and thus, a different piece of the puzzle. This article describes the basic theory behind each of the different techniques, the types of data produced from each, and some typical applications. Table 2 summarizes the different features of these techniques to allow for at-a-glance comparisons. Also discussed are the different types of samples that can be analyzed and the special sample-handling procedures that must be implemented when preparing to do failure analysis using these surface-sensitive techniques.

Analytical Considerations. Proper design of experiment and the choice of samples for analysis are important factors in the likelihood of success when determining the cause of a failure. Failures can occur in the manufacturing of a device or some time after the device has been in use. In the ideal situation, good versus bad samples can be compared. This allows for differences in elemental concentrations and chemistries to be observed and related to their effects on device failure. Attempts are then made to link these differences to a known step in the manufacturing process or suspected contaminant in the end use of the product. The cause of failure may be obvious from the contaminants observed, or it may be extremely subtle. If a contaminant is found, but its root source is not known, then the next step, if possible, is to analyze samples and materials extracted at various steps in the manufacturing cycle or end use of the product.

In other instances, a direct comparison of a failed material to a good sample is difficult or even impossible, such as in interface delamination problems, where the bonding failure does not occur on good samples. In these cases, data from the failed area should be compared to reference materials to infer what should or should not be present. Reference materials can include starting substrate materials (before and after cleaning processes); pure compounds and coatings used in the manufacturing of the device; dried solvents, oils, greases, and cleaning solution residues used in processing; and any material that may have come in contact with the device.

In all cases, the surface analyst must rely on the knowledge of the process engineer to provide as much detail as possible regarding the nature of the failure and the history of the sample of interest. Doing analysis on samples where no processing and handling history are provided can result in the incorrect interpretation of the data and results that are very misleading.

Sample-Handling Issues. In no other class of analytical techniques is sample handling as critical an issue. Because AES, XPS, and TOF-SIMS provide information from the top few atomic layers, extreme care must be taken to keep the surface of interest from being covered up with extraneous contamination. Physically touching a sample with ungloved and even gloved hands is a common source of handling contamination. With ungloved hands, finger oils, salts, hand lotions, and other contaminants are easily transferred to touched surfaces. Many hand lotions contain siloxane, commonly referred to as silicone, as one of their main ingredients. Siloxanes are known to easily spread across surfaces. Thus, touching any part of a sample with siloxane-contaminated hands can lead to large areas of the sample becoming contaminated. Many assume that if sample handling is performed with gloved hands, their samples are safe from contamination. Unfortunately, the surfaces of many gloves are also laden with oils, salts and/or lubricants that can be transferred to the sample during handling (Ref 1). Even so-called clean-room gloves are not necessarily clean of such contaminants, because they are manufactured more to be particle-free than contaminant-free. Common surface contaminants on powder-free gloves include Si, Cl, Na, and Zn. Touching the area to be investigated with unclean tools can also introduce impurities onto the area of interest. In general, nothing should ever come in physical contact with the area of interest. Even the air surrounding the sample should be kept free of smoke and particles. Caution should also be taken to make sure that samples do not come in contact with common plastic bags during storage or shipping, because additives in the plastic can transfer to the sample and be detected with surface-sensitive tools. Clean, plastic petri dishes (polystyrene or polypropylene are common), glass containers, and even plain white typing paper are all acceptable ways of storing or shipping samples that will later be analyzed by one or more surface analysis methods. If tape must be used to secure samples during shipping, avoid using tapes that are laden with silicones, such as many conducting carbon tapes. Many clear “cellophane” tapes are fairly clean of silicones, but make sure to avoid placing the tape in the vicinity of the area of interest on the sample.

Sample Types. Most solid materials, if they are vacuum compatible and are of the proper geometric size to be accepted into the surface analysis instrument, can be analyzed by one or more of these surface analysis tools. This includes, but is not limited to, wafers, sheets, films, coatings, foils, chunks, powders, wires, tubes, printed circuit boards, computer chips, and other whole or partial pieces of devices. While surface analysis tools are used to analyze solid surfaces, dried residues from liquids and even low-vapor-pressure oils and greases can also be looked at. A minute drop or thin layer of the liquid or oil of interest can be placed on a clean substrate (silicon wafer, aluminum foil, etc.) and allowed to dry in air prior to its analysis.

One consideration, however, is the electrical property of the sample—is it conducting or insulating? In general, AES is better adapted to running conducting and semiconducting materials, while both conducting and insulating materials are readily analyzed by XPS and TOF-SIMS.

Reference cited in this section

1. "XPS Analysis of Disposable Gloves," XPS application brief, Charles Evans & Associates, Sunnyvale, CA, 1999

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

---

Auger Electron Spectroscopy

Auger electron spectroscopy is a surface analysis technique used to determine the elemental composition of the top few atomic layers of a surface or exposed interface in a solid material. Auger electron spectroscopy can detect all elements except hydrogen and helium and can provide semi-quantitative information, with detection limits of 0.1 to 1 at.% for most elements. The technique is named after Pierre Auger, who first described the process in 1925. The most attractive attribute of AES is its ability to analyze extremely small features. Modern AES instruments with field-emission electron sources can produce secondary electron micrographs with spatial resolutions as small as 10 nm and can characterize sample features as small as 25 nm.

Auger electron spectroscopy is performed under ultrahigh vacuum conditions, using an electron beam typically in the 3 to 25 keV range. The AES process begins with electron bombardment of the sample material causing an atom to eject an inner-shell electron, thus forming a vacancy. A second electron from a higher shell fills this inner-shell vacancy in an energy gain process. This energy can then cause the ejection of a third electron, referred to as an Auger electron. The energy of the escaping Auger electron is analyzed by an electron spectrometer. Because each element has a unique set of electron energies surrounding the nucleus, it also has a unique set of Auger peaks, and the resulting spectrum provides a fingerprint of the probed surface. The kinetic energy of the Auger electrons is typically between 40 and 2500 eV. In this energy regime, electrons can travel only a short distance before interacting with other atoms and losing energy. This short distance is referred to as the escape depth of the electron. Escape depths range from 0.5 to 10 nm, depending on the kinetic energy of the emitted electron and the material being analyzed. It is this small range of escape depths that gives AES its surface sensitivity. Those electrons that are close enough to the surface to escape without loss of energy are detected as Auger electron peaks. Those electrons that lose energy before leaving the sample surface add to the background of the spectrum.

Scanning Auger microscopy is accomplished by scanning an electron beam across the surface of a sample while measuring resultant electron signals. This scanning process generates secondary electron microscopy (SEM) images, backscattered electron images, and Auger maps. Secondary electron microscopy images, which provide a topographic view of the sample by detecting low-energy electrons emitted from the surface, are used to locate specific areas for more detailed study. Backscattered electron images, involving higher-energy electrons that have undergone scattering processes before escaping from the sample, reveal atomic number contrast and crystallographic information. Auger maps, obtained by measuring the emitted Auger electron intensity while scanning the electron beam, reveal the lateral distribution of elements across the sample surface.

Because of the uncompensated loss of secondary electrons during electron bombardment, AES typically has a difficult time with insulating materials and, therefore, is primarily used for analyzing conducting and semiconducting solids. However, AES analysis of some inorganic insulating materials is also possible.

A competing process to Auger electron generation is that of x-ray emission, the analytical signal for the technique known as energy dispersive x-ray spectroscopy (EDS). In general, x-ray emission is more pronounced for heavier elements,

while the probability of Auger electron emission is greater from lighter elements. While AES can detect elements above He in the periodic table, EDS is usually limited to elements heavier than beryllium. Auger electron spectroscopy is also more appropriate for submicrometer particles, because the analyzed volume of material with AES is much less than that for EDS. In EDS, the x-ray emission is generated from regions a micrometer or more in width and depth, even when much-finer electron probes are used.

The nomenclature used to identify the various Auger peaks consists of listing the three electron shells involved in the formation of the Auger electron. For example, for low-atomic-number elements, the most probable transitions occur when a K-level electron is ejected by the primary electron beam, an L-level electron drops into the vacancy, and another L-level electron is ejected. This is referred to as a KLL Auger electron.

Because of its excellent analytical spot size, AES is highly used in semiconductor, electronics, computer disk drive, and metallurgical applications where extremely small area analyses (spectra, depth profiling, or mapping) are required. The basic types of AES data, which are briefly described subsequently, include:

- Scanning electron micrographs
- Survey spectra
- Depth profiles
- Maps and line scans

For further information, see the Selected References at the end of this article for recommended reading on AES.

Scanning Electron Micrographs. Because the electron beam used in modern-day AES systems is similar to those used in scanning electron microscopes, AES systems can provide very high-spatial-resolution secondary electron micrographs. These images are used for the location and documentation of the areas of interest (Fig. 1).

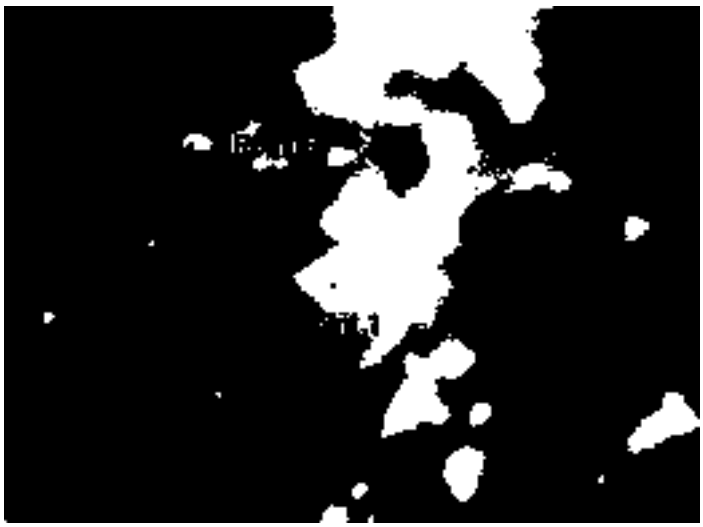


Fig. 1 Scanning electron microscopy photo of the surface of a 300-series stainless steel sample obtained from AES instrument. Field of view, 1 μm

Survey Spectra. Typically, the first piece of analytical data obtained in an AES experiment is a survey spectrum. A survey spectrum is a plot of AES signal intensity versus the kinetic energy of the detected electrons. Survey spectra show which elements are present and, after applying sensitivity factors (Ref 2), can be used to quantify the amount of each element detected. Survey spectra are typically displayed on a differentiated scale of signal intensity to better observe the small Auger peaks present on a large, sloping background (Fig. 2). In some instances, when excessive atmospheric or handling contamination obscures much of the underlying surface of interest, inert gas sputtering can be employed to remove at least some of the obstructing contamination layer. Care must be taken to prevent sputter removing too much of the sample, or the surface of interest will also be removed. Typically, only 0.5 to 3 nm is presputtered (relative to the sputter rate of SiO_2) for this type of “subsurface” analysis.

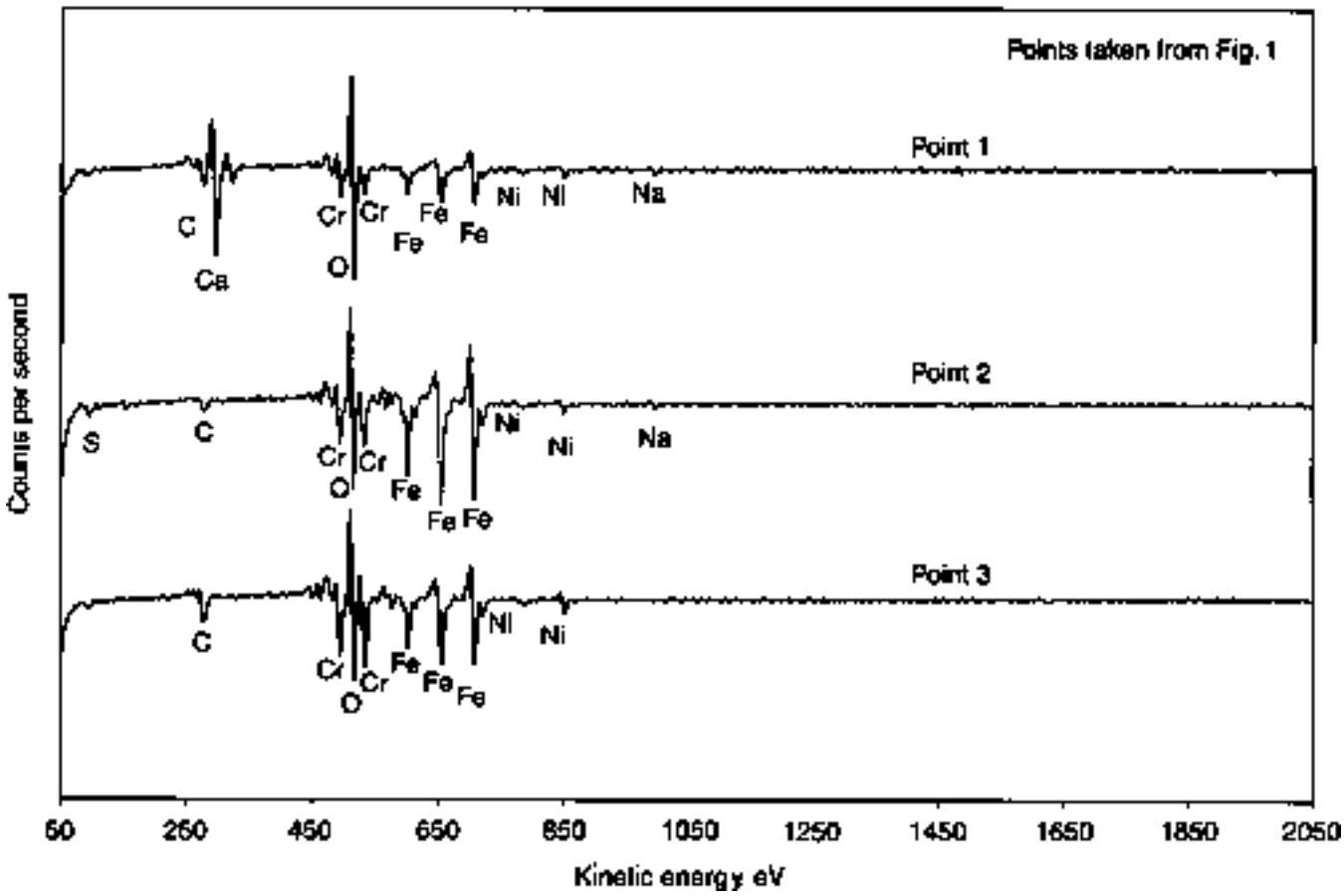


Fig. 2 Auger electron spectroscopy survey spectra obtained from features observed in secondary electron microscopy photo of stainless steel (Fig. 1). Point 1, large particle feature near center of Fig. 3. Point 2, small particle feature. Point 3, control point off of defects

Depth Profiles. Compositional depth profiling is possible with AES with the aid of an inert gas sputter ion beam. The ion beam is used to slowly remove material in a controlled manner, thus exposing a new surface to be analyzed. By sequentially sputtering and taking AES data, a compositional depth profile may be generated that shows the AES signal for selected elements as a function of sputter time (depth). Exact sputter depths are often difficult to obtain, because exact sputter rates are seldom known, and, on rough surfaces, the sputter crater depths are difficult to measure. Therefore, sputter rates are often specified relative to a known sputter rate on a reference material, such as thermally grown SiO_2 on single-crystal silicon. Practical maximum profile depths are on the order of a couple micrometers or less. Compositional depth profiling is particularly effective for thin-film analysis and for solving materials problems that require identifying the thickness of a thin surface layer, the composition of a thin film deposit, the presence of interdiffusion between thin films, or the presence of contamination at an interface between two layers.

Maps and Line Scans. Because of its excellent spatial resolution, elemental mapping is one of the most common modes of operation with AES. A map is a two-dimensional display of Auger electron signal intensity for a select element from a given area on the sample surface. High intensities (bright areas) on the map indicate that more of that particular element is present at that point than at lower-intensity (darker) points (Fig. 3). Mapping can be an excellent way of detecting localized (unknown) contamination. After first obtaining survey spectra from the area of interest, Auger maps are obtained for the elements of interest. The maps are then studied to identify localized areas that have low signal for those elements. Additional spectra can then be obtained in these low-intensity areas to determine if other localized elements are also present. Line scans are similar to maps, except that the data is obtained in one direction only.

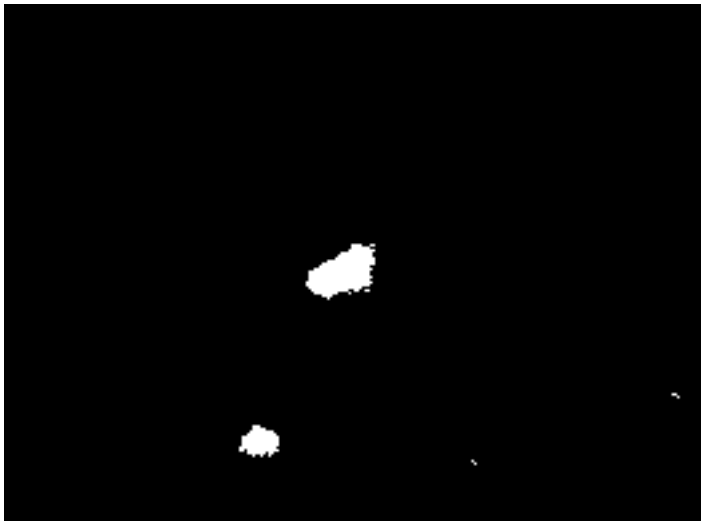


Fig. 3 Auger electron spectroscopy map of calcium contamination on stainless steel surface. Field of view, 1 %m

Reference cited in this section

2. K. Childs et al., *Handbook of Auger Electron Spectroscopy,* 3rd ed., C.L. Hedberg, Ed., Physical Electronics, Inc., 1995

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

---

X-Ray Photoelectron Spectroscopy

X-ray photoelectron spectroscopy, also known as ESCA, is a surface analysis technique that can be used to determine both the elemental and chemical composition of the outermost atomic layers of a solid material. With the exception of H and He, all elements can be detected. Detection limits are, on average, approximately 0.1 at.%; however many of the heavier elements can be detected down to about 0.01 at.% (100 ppm). By monitoring subtle shifts in the atomic binding energies of the emitted photoelectrons, chemical speciation can be obtained from both organic and inorganic materials. Similar to AES, the average depth of analysis is approximately 5 nm.

X-ray photoelectron spectroscopy is accomplished by flooding the sample with x-rays of a known energy (typically Mg K8 at 1253.6 eV or monochromated Al K8 at 1486.7 eV). Absorption of these x-rays by the sample atoms causes photoelectrons to be emitted. The kinetic energy of the emitted photoelectrons is measured with an electron spectrometer. To a first approximation, the kinetic energy is determined from the following equation:

$$KE = h(- BE$$

where KE is the measured kinetic energy of the emitted photoelectron, h(is the energy of the x-rays being used, and BE is the atomic binding energy associated with the emitted photoelectron. To maintain consistency in the spectral energy scale when using different x-ray energies, the data are displayed on a binding-energy, rather than kinetic-energy, scale. The binding energy associated with a peak is then used to establish its elemental identity and chemical state. The incoming x-rays penetrate microns into the surface of the sample. However, the escape depths for these low-kinetic-energy (less than 1500 eV) photoelectrons are less than 10 nm, thus making XPS an extremely surface-sensitive technique.

Quantification is possible with the use of elemental sensitivity factors (Ref 3) that have been determined empirically and found to be in agreement with the current theoretical models for quantification of XPS data.

If the sample is insulating, x-ray bombardment can create a positive-charge buildup on the surface of the sample, causing all of the spectral peaks to shift to apparently higher binding energies. To minimize this charging phenomenon, low-energy electrons or a combination of electron and ion floods (Ref 4) may be added to the sample surface. All peaks are then referenced to a peak of known energy, such as C because hydrocarbons are present on most surfaces. This charge correction method allows one to obtain useful chemical-state information from insulating samples.

Common materials analyzed by XPS include metals, lubricants, semiconductors, metal oxides, glasses, ceramics, catalysts, plastics/polymers, coatings/thin films, and paper.
Because the technique is so flexible in its sample-handling capabilities, the industries served by XPS are quite varied. Application areas include bond pads, paint and other thin-film adhesion problems, cleanliness concerns, corrosion analysis, identification of debris and discolorations, characterization of polymer surface functionalization, lubricant thickness measurements, optical and other thin-film profiling, and determination of oxidation state and oxide thickness of alloys. Types of XPS data, as briefly described subsequently, include:

- Survey spectra
- High-resolution spectra
- Depth profiles
- Angle-dependent analysis
- Maps and line scans

See the Selected References at the end of this article for recommended reading on XPS.
Survey Spectra. Similar to AES, the first piece of analytical data obtained in an XPS experiment is typically a survey spectrum (Fig. 4). Survey spectra are acquired to determine which elements are present and, after applying sensitivity factors (Ref 3), can be used to quantify the amount of each element detected. Survey spectra are obtained using high-sensitivity analyzer conditions to maximize the possibility of detecting low-level species.

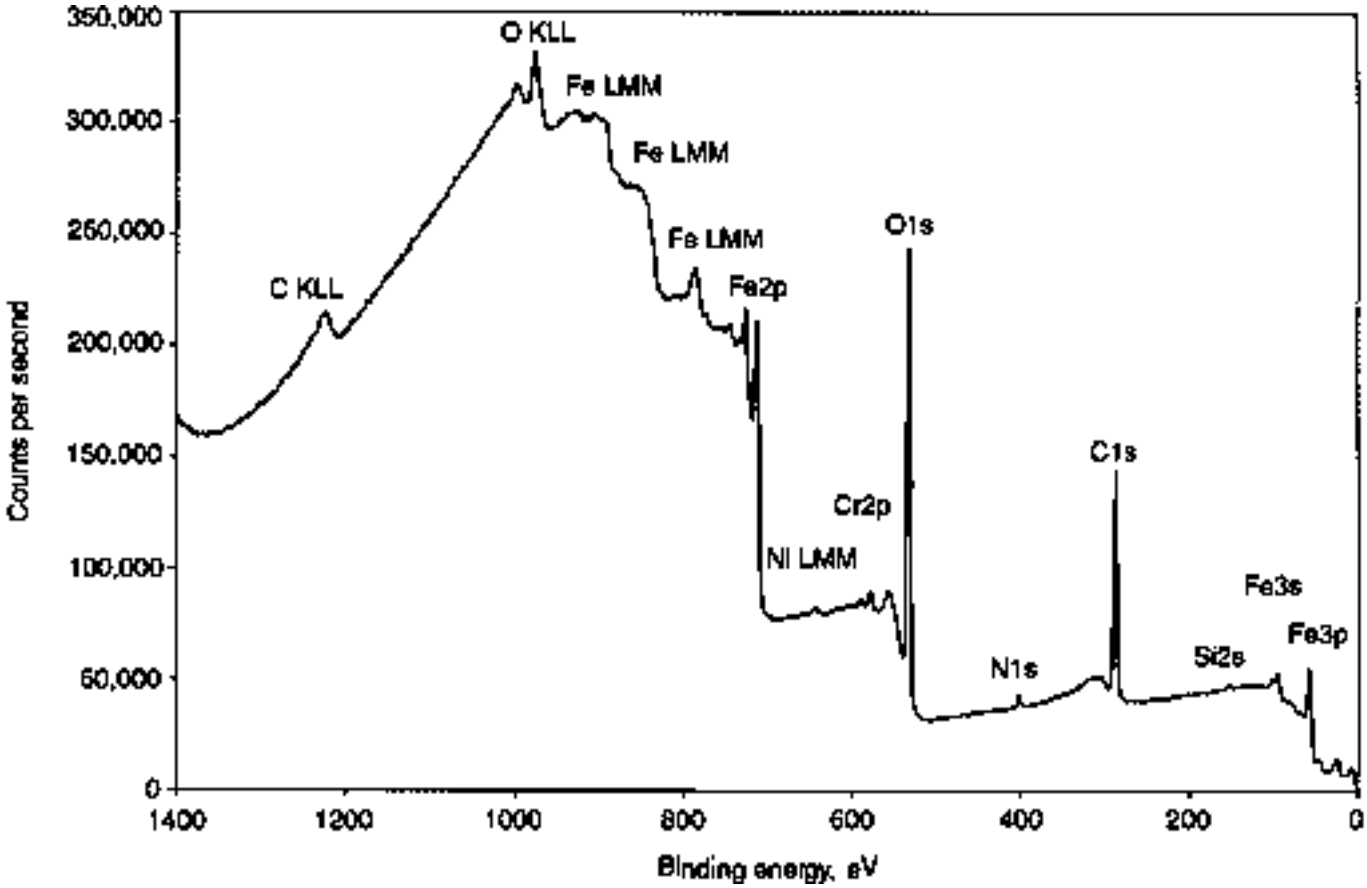


Fig. 4 X-ray photoelectron spectroscopy survey spectrum of stainless steel surface

High-Resolution Spectra. Once it is determined what species are present on a surface, high-resolution scans are typically acquired on specific elements of interest to monitor binding-energy shifts that can provide chemical information. These scans are obtained using much higher energy-resolution conditions on the spectrometer compared to the survey spectra. Because the peak shapes obtained in this mode are more accurate, quantification is also better. It is on these high-resolution scans where curve-fits and other mathematical routines are performed to extract chemical-state information and quantification. An example of the type of localized chemical bonding information that can be obtained with XPS is shown in the high-resolution C spectrum of polyethylene terephthalate (PET) (Fig. 5). Three relatively large peaks and one smaller peak are observed. The largest peak in the spectrum, located at a binding energy of 284.8 eV, is assigned to hydrocarbon-type carbon (C-C or C-H) where C is bound only to other C or H. In PET, this peak is associated with the nonoxidized carbons in the aromatic ring. At about 286.3 eV, a C-O peak is observed that is characteristic of the ethylene carbons in PET that are singly bound to oxygen. Located at 288.7 eV is a peak associated with the carboxylate carbons

(ester O=C-O) that are bound to two oxygens. A small satellite peak is also observed at about 291.5 eV. This peak is indicative of the aromaticity within the PET molecule.

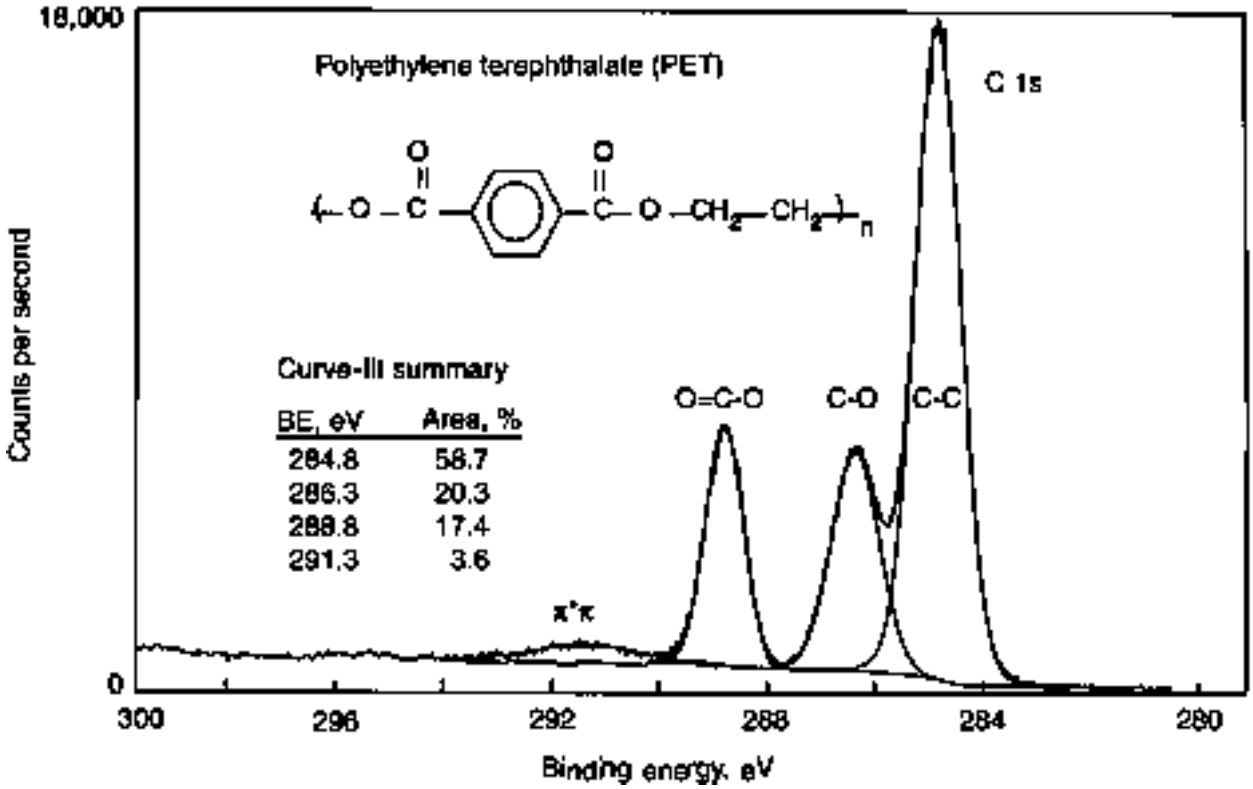


Fig. 5 X-ray photoelectron spectroscopy high-resolution spectrum of polyethylene terephthalate (PET)

Depth Profiles. Elemental and, in some cases, chemical depth profiling is possible with the use of an inert gas sputter ion gun. The ion gun is used to slowly remove material, thus exposing a new surface to be analyzed. By sequentially sputtering and taking XPS data, a compositional depth profile can be generated (Fig. 6). If the chemical matrix of the sample is not severely damaged by the sputter ion beam, chemical information may also be obtained as a function of depth. However, many oxides and especially organics are very susceptible to chemical modification during ion bombardment. Similar to AES, exact sputter rates are often difficult to determine. Therefore, sputter rates are often specified relative to a known sputter rate on a reference material, such as thermally grown SiO_2 on single-crystal silicon.

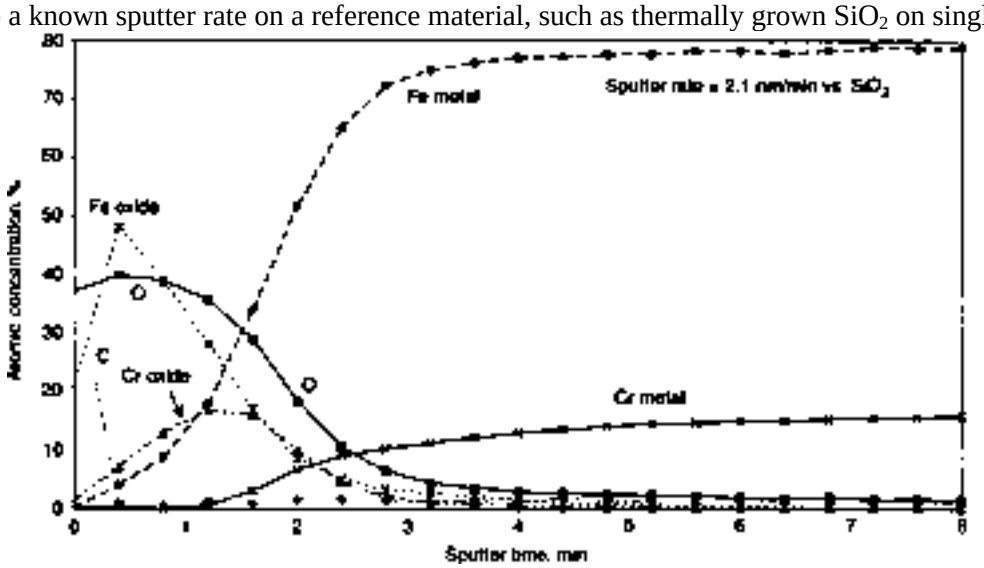


Fig. 6 X-ray photoelectron spectroscopy compositional depth profile of stainless steel

Angle-Dependent Analysis. Angle-dependent XPS is a method for nondestructively analyzing the in-depth chemical gradients in the top surface layers (<10 nm) of a smooth sample. By changing the sample tilt with respect to the energy analyzer (takeoff angle), the effective depth of analysis can easily be changed. This technique can be useful in cases where ion bombardment is known to modify the chemical makeup of the material being probed. It has been greatly used

for studying such things as the depth of modification of plasma- and corona-treated polymers, the oxidation of metals and semiconductors, polymer-metal adhesion, additive surface migration, and the lubrication of various materials.

Maps and Line Scans. Many newer-generation XPS instruments are capable of obtaining spatially resolved information from a sample surface by acquiring photoelectron maps and line scans. A photoelectron map is a two-dimensional display of photoelectron intensity for a specific element from a given area on the sample surface. High intensities (bright areas) on the map indicate that more of that particular element or chemistry is present at that point than at lower-intensity (darker) points. Line scans are similar to maps, except that the data is obtained in one direction only. Spatial resolutions for state-of-the-art instruments are on the order of 5 to 10 7m.

References cited in this section

3. J. Moulder et al., *Handbook of X-Ray Photoelectron Spectroscopy*, J. Chastain and R.C. King, Jr., Ed., Physical Electronics, Inc., 1995

4. P.E. Larson and M.A. Kelly, Surface Charge Neutralization of Insulating Samples in X-Ray Photoemission Spectroscopy, *J. Vac. Sci. Technol. A,* Vol 16 (No. 6), 1998, p 3483

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

---

Time-of-Flight Secondary Ion Mass Spectrometry

Time-of-flight secondary ion mass spectrometry is an analytical technique that uses a high-energy, primary ion beam to probe the surface of a solid material. The instrument is typically operated in the “static” mode for obtaining elemental and molecular-chemical information from both organics and inorganics. In this mode of operation, the sample integrity and chemistry are preserved by applying extremely low primary ion doses (less than 1×10^{12} ions/cm^2) during the entire experiment. This ensures that roughly less than 0.1% of the surface atoms or molecules are ever struck and damaged by the primary ion beam. Time-of-flight secondary ion mass spectrometry is the most surface sensitive of the surface analytical techniques, with a depth of analysis of only about 2 nm. The technique has extremely good detection sensitivities, with detection limits for most elements in the parts-per-thousand to parts-per-million range. By using a finely focused ion beam (typically Ga or In) it is also possible to record the lateral distribution of chemical species across a surface with micron to submicron spatial resolution. Because secondary ion intensities vary dramatically from element to element and are highly dependent on the matrix from which they are sputtered, quantification with TOF-SIMS can be extremely difficult. Thus, in most cases, the technique is used more for qualitative purposes than for quantitative analyses. However, in select cases where appropriate standards are available (e.g., metals on silicon), accurate quantification can be performed (Ref 5).

When elemental depth profiles are required, most TOF-SIMS systems can also be operated as a “dynamic” SIMS instrument. In the dynamic mode, very high primary ion doses are used in order to obtain in-depth information. However, in this mode of operation, very little chemical specificity can be gleaned, due to the destructive nature of the primary ion beam. Therefore, the following discussion only focuses on the static mode of operation.

In a TOF-SIMS experiment, the desorption/ejection of “secondary” ions from the surface of a material is initiated by a short pulse (~1 ns) of primary ions that impinges on the surface at high angles of incidence. The momentum transfer from the primary beam to the solid initiates a “collision cascade” within the solid, much like a microscopic billiard game. A portion of this momentum is redirected back toward the surface, resulting in the ejection of atomic and molecular ions. Greater than 90% of these secondary ions originate from the outermost one to two layers of the solid, thus defining TOF-SIMS as an extremely surface-sensitive technique. By applying a potential between the sample surface and the mass analyzer, the desorbed secondary ions are extracted into a time-of-flight mass spectrometer where their masses are separated in flight time, with very high accuracy, based on their mass-to-charge ratio (m/z). The resulting mass spectrum, typically in the 1 to 2000 dalton range, allows for the unambiguous identification of chemical moieties that can often be correlated to the original surface structure. An example of the type of chemical/molecular information that can be obtained from secondary ion mass spectra is shown in the positive ion spectrum of PET (Fig. 7). The spectrum shows a variety of large molecular cluster ions that are identified as fragments of the PET polymer chain. This provides information on the long-range molecular makeup of the sample as well as providing a unique fingerprint for this material.

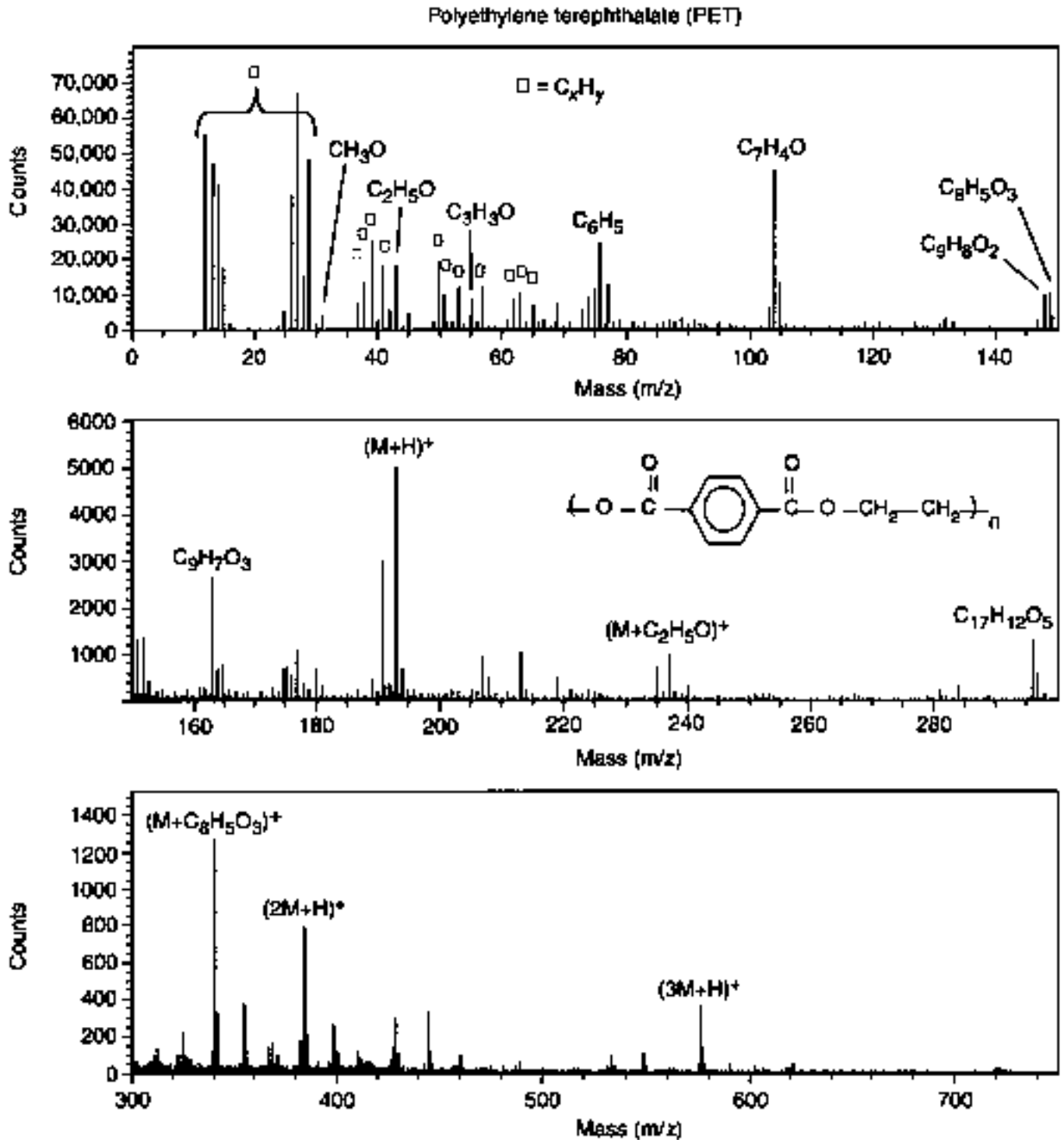


Fig. 7 Time-of-flight secondary ion mass spectroscopy mass spectrum of polyethylene terephthalate (PET)

On insulating materials, the loss of secondary electrons during ion bombardment can lead to a positive charge buildup and loss of signal from the sample surface. Therefore, auxiliary electron sources are often used to supply electrons to the sample surface to help neutralize the excess charge.

Because TOF-SIMS can easily handle insulating as well as conducting materials, the types of samples analyzed by TOF are often similar to those analyzed with XPS. However, because of its much smaller analytical probe size, TOF-SIMS systems can analyze features as small as a few micrometers in width and are commonly used for mapping distributions of elements and molecules across surfaces.

Application areas for TOF-SIMS include organic and inorganic contaminant identification, cleaning studies, surface segregation and modifications, corrosion, discolorations, drug distribution, diffusion studies, and chemical characterization. Two types of secondary ion data are simultaneously obtained in a TOF-SIMS experiment:

- A total-area mass spectrum

- A total secondary ion image

As the focused primary ion beam is digitally rastered across the sample surface, complete mass spectra of ion intensities versus mass-to-charge ratio are obtained at every pixel (256 × 256) within the raster. The summation of these spectra produces a total area mass spectrum (Fig. 8). From the same acquisition, a total ion image is also obtained that shows the lateral distribution of secondary ion signals from across the area of analysis (Fig. 9). High intensities (bright areas) on the image indicate that more secondary ion signal is present at those points than at lower-intensity (darker) points. The variation in secondary ion intensity can be due to topographical effects or from differences in chemical composition. The total ion image is used to selectively analyze the mass spectra (chemistries) from areas that show differing amounts of brightness. Conversely, the total ion mass spectrum can be used to select specific elements or molecules for display in secondary ion maps that show the relative localized abundance of these species. Because both positive and negative secondary ions are created during ion bombardment, two separate experiments must be performed for complete characterization of a given sample.

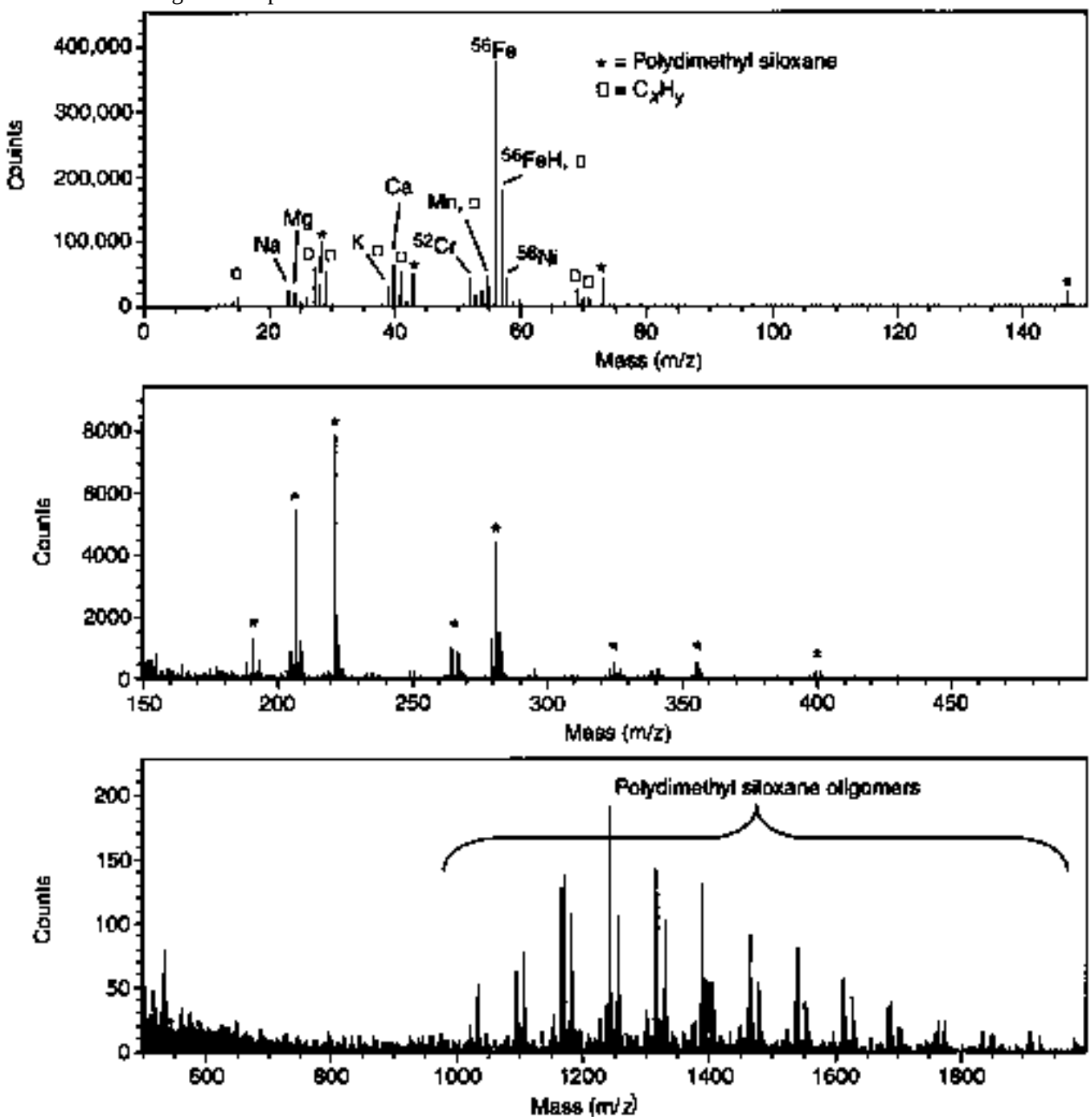


Fig. 8 Time-of-flight secondary ion mass spectrometry positive ion spectrum of stainless steel surface

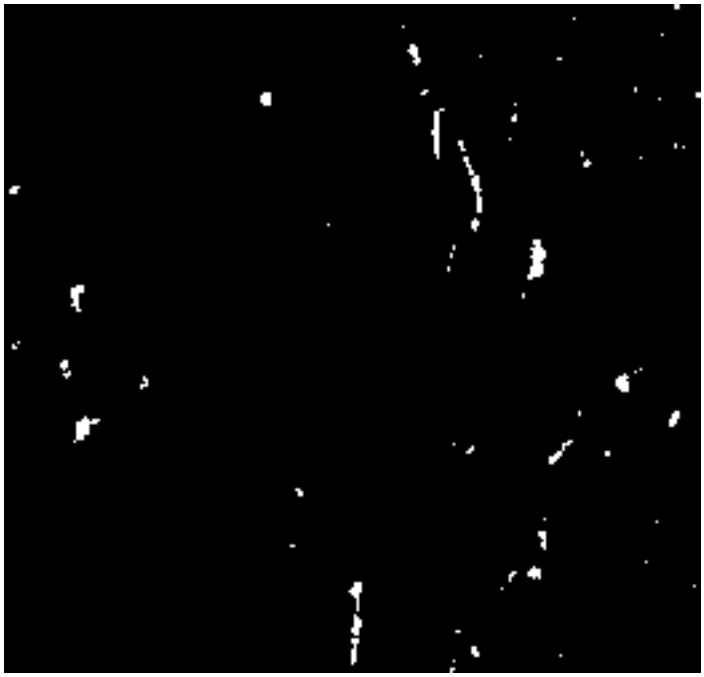


Fig. 9 Total positive ion image of stainless steel surface (50 by 50 %m) revealed by TOF-SIMS

Reference cited in this section

5. M.A. Douglas and P.J. Chen, Quantitative Trace Metal Analysis of Silicon Surfaces by TOF-SIMS, *Surf. Interface Anal.*, Vol 26, 1998, p 984–994

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

---

Example: Stainless Steel Analysis

To exemplify the strengths of AES, XPS, and TOF-SIMS, data obtained from the surface of a slightly corroded stainless steel sheet are presented for each technique. Each method provides a different view of the surface and a different piece of the sample characterization puzzle. Note that in these examples not every type of analysis possible is shown for each technique (e.g., survey spectra and Auger maps but no depth profile are shown for AES).

Example 1: AES Analysis of Stainless Steel. Figure 1 shows a high-magnification secondary electron micrograph (100,000×) obtained from an AES instrument on the as-received surface of a piece of 300-series stainless steel sheet. The micrograph shows a series of what appear to be extremely small particles or defects, most of which are only about 25 nm in diameter. This SEM image was used to define points on which AES was to be performed. In this case, one large and one small particle (points 1 and 2, respectively) were analyzed along with a control region (point 3) off of the defects. Prior to analysis, a light sputter etch (approximately 1.5 nm removed) was performed to remove some of the atmospheric contamination to better observe the underlying stainless components. All three of the survey spectra (Fig. 2) show the presence of the expected C, O, Cr, Fe and Ni from the air-exposed stainless steel surface. In addition, the spectrum from point 1 (larger defect) shows the presence of a large amount of Ca along with a smaller Na peak. The AES spectrum from the small defect (point 2) shows the presence of S and Na contaminants but also shows a much different Fe to Cr peak ratio than the spectrum from the control area (point 3). Quantification of the elements detected from these three points is shown in Table 3. Note that for each point, the concentrations are normalized to 100% in an assumption that all elements were detected. Auger maps were then acquired for several of the elements of interest, including Ca, Cr, Fe, and O. The map for Ca is shown in Fig. 3, which reveals a lateral distribution of Ca within a 1 7m field of view. Although scores of particle defects were observed in the SEM image, only a few of them appear to have significant Ca, as shown in the Ca

Auger map. Most of the smaller particles correspond to larger concentrations of Fe, compared to the background area (point 3, Fig. 2). This indicates that localized corrosion has occurred within these 25 nm-sized defects, with an obvious potential cause being the presence of Na.

Table 3 Auger electron spectroscopy atomic concentrations (at.%). 1.5 nm sputter removed

Sample	C	O	Na	S	Ca	Cr	Fe	Ni
Point 1, large defect	13.1	30.2	3.5	nd	38.0	3.4	10.0	1.9
Point 2, small defect	8.6	35.2	5.6	0.5	nd	9.7	37.9	2.6
Point 3, control area	21.4	33.4	nd	nd	nd	16.7	24.1	4.5

nd, not detected

Example 2: XPS Analysis of Stainless Steel. Figure 4 shows a survey spectrum obtained from a millimeter-sized area of the as-received 300-series stainless steel. Several elements are observed, including C, N, O, Si, Cr, Fe, and Ni. Integrating the area under the peaks and applying empirically derived sensitivity factors yields the atomic concentration values shown in Table 4.

Table 4 X-ray photoelectron spectroscopy atomic concentrations (at.%)

Sample	C	N	O	Si	Cr	Fe	Ni
Stainless steel	44.6	1.6	40.2	0.1	0.9	12.5	0.2

Although a certain amount of organic contamination is expected due to atmospheric exposure, the amount of C found on the as-received surface is relatively high. Some of this organic may have been introduced from surface or vapor-phase transfer of plastic additives or contaminants from the plastic bag in which the steel was stored. The C high-resolution spectrum (Fig. 10) shows a C peak that can be mathematically fit into at least three types of carbon, with energies (but not peak ratios) similar to that of the spectrum of PET (Fig. 5). The largest peak, occurring at 284.8 eV, is from hydrocarbon-type carbon. A second peak, located at 286.3 eV, is attributed to either C-O (ether or hydroxyl) or possibly C-N (organic-based N), because a small amount of N is observed in the survey spectrum. At 288.4 eV, a peak typically associated with carboxylate chemistry (O=C-O) is also observed. Thus, the XPS reveals which organic functional groups are present on the stainless steel and suggests possible contaminants. However, further characterization with a technique such as TOF-SIMS would be needed before definite conclusions should be drawn regarding the identity of the organic.

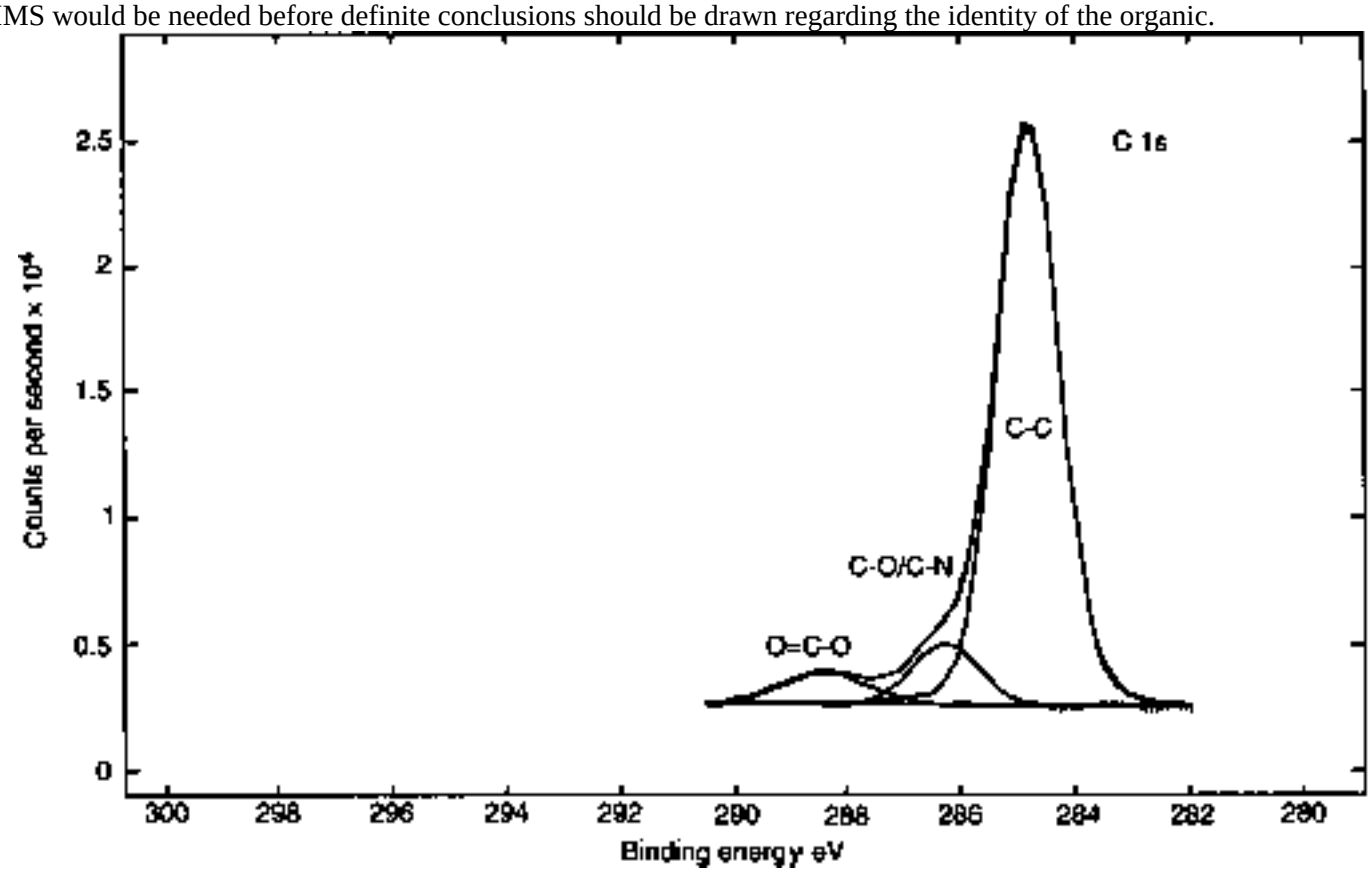


Fig. 10 X-ray photoelectron spectroscopy high-resolution carbon spectrum of stainless steel surface

The ratio of Cr to Fe is often used as an indication of the "goodness" of passivation (Ref 6), with ratios of 1 to 2 being optimal (Ref 7) for well-passivated stainless steel. Clearly, this stainless steel sample has a very poor Cr to Fe ratio. Well-passivated stainless steel contains a thin Cr oxide layer as the outermost surface, protecting the underlying metallic Cr, Fe, and Ni from becoming oxidized. If the Cr oxide surface layer is performing properly, the XPS high-resolution spectral analysis should be able to see through the passivation layer and detect the presence of subsurface metallic Fe that has been protected from oxidation. On this particular stainless steel sample, little, if any, metallic Fe is observed (Fig. 11). Figure 12 shows a high-resolution Fe spectrum obtained from a well-passivated stainless steel. Note that the Fe metal component on the well-passivated steel is much more prominent compared to this "corroded" steel surface.

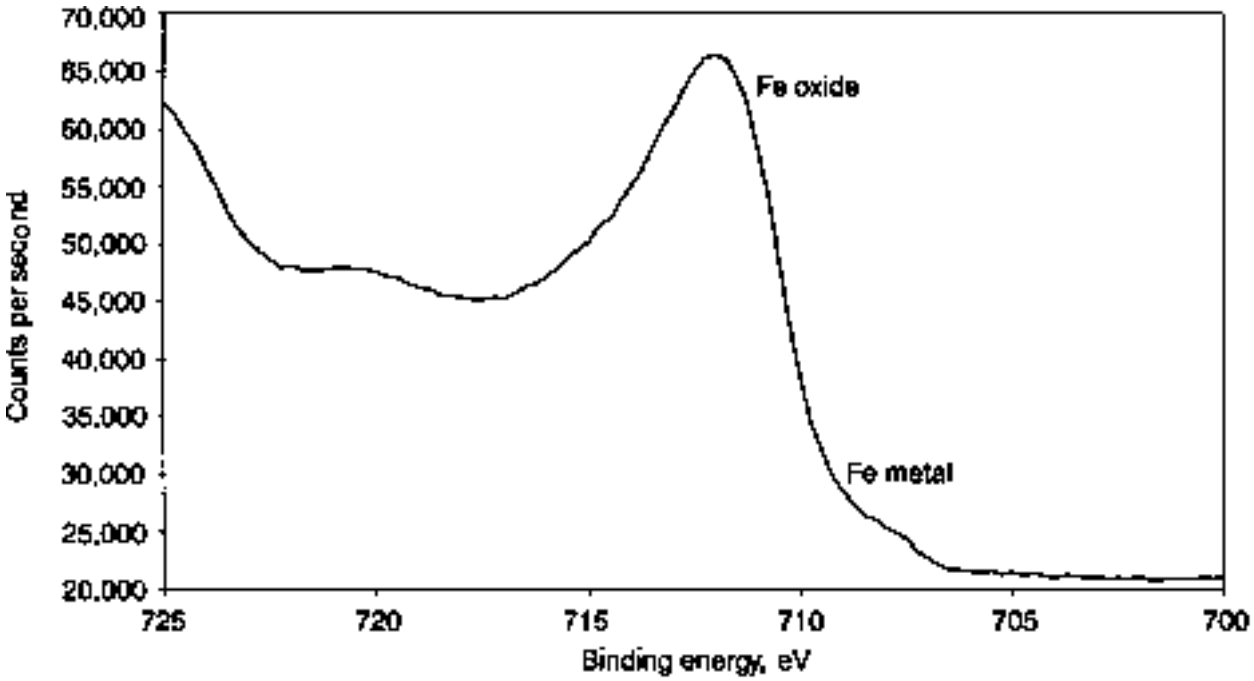


Fig. 11 X-ray photoelectron spectroscopy high-resolution iron spectrum of stainless steel surface

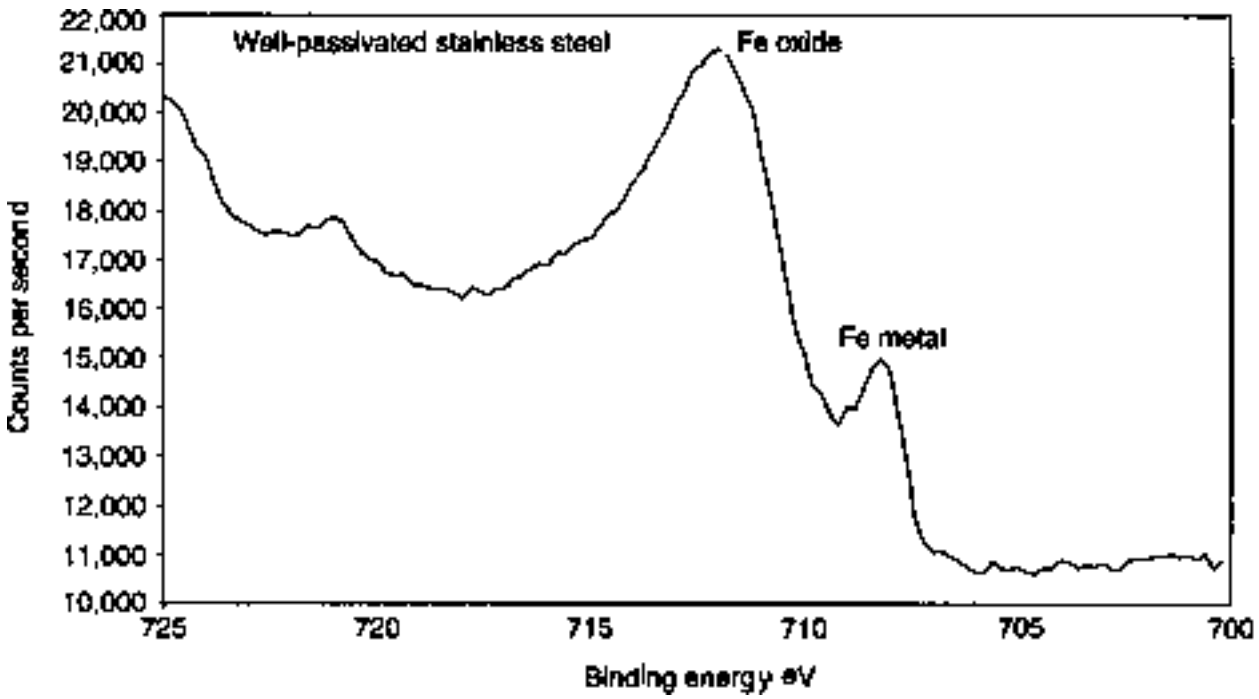


Fig. 12 X-ray photoelectron spectroscopy high-resolution iron spectrum obtained from well-passivated stainless steel surface

Figure 6 shows the compositional depth profile of the 300-series stainless steel sample. The sputter rate for the profile, relative to the sputter rate of SiO_2, is 2.1 nm/min, with data points collected every 0.8 nm. The Fe and Cr energy windows for the first several cycles of the profile are shown in Fig. 13 and 14, respectively. Both graphs show primarily oxide (higher binding-energy peaks) for the first few sputter cycles, after which the metallic form (lower binding-energy peaks) becomes dominant. These binding-energy shifts allow for easy mathematical separation of the oxidized and metallic forms within the profile (Ref 8, 9). The rapid decrease in surface C in the first half-minute of sputtering indicates that this contaminant is present as a thin layer on the surface of the stainless steel rather than in thick, uneven patches. Just under the C layer, Fe (oxide) and O become the primary components, with relatively low amounts of Cr being observed. Below the Fe oxide, a thin Cr oxide-enriched layer is detected, under which the bulk constituents of the stainless steel are found. On a well-passivated steel surface, a Cr oxide layer, rather than Fe oxide, would dominate the outermost surface region. The XPS results on this sample are consistent with those of the AES that found localized regions of corrosion occurring.

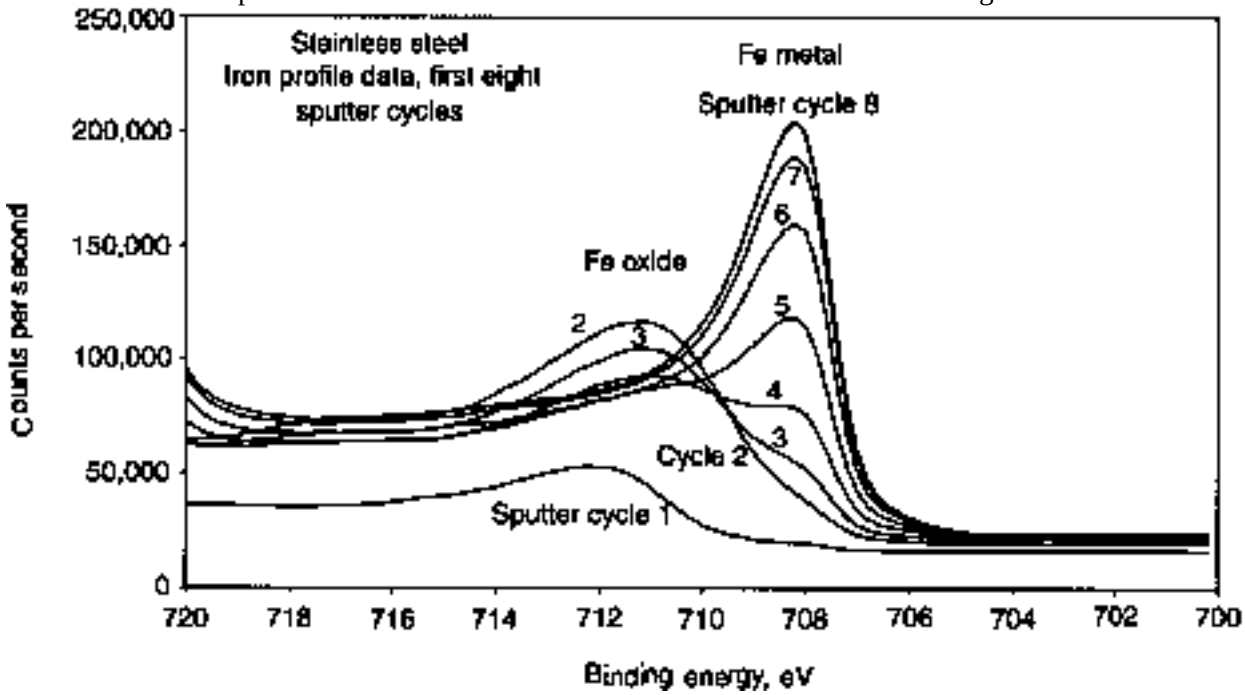


Fig. 13 X-ray photoelectron spectroscopy montage display of iron in the first eight sputter cycles of the depth profile (Fig. 6)

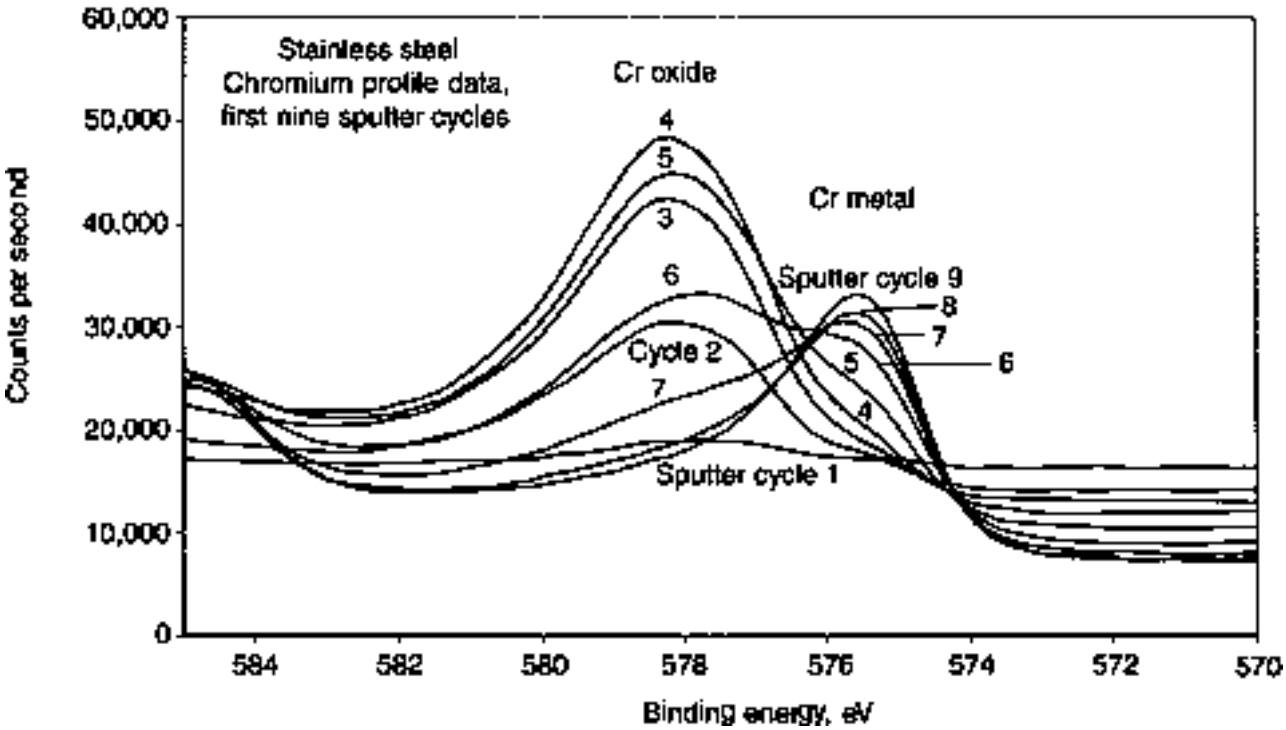


Fig. 14 X-ray photoelectron spectroscopy montage display of Cr in the first nine sputter cycles of the depth profile (Fig. 6)

Example 3: TOF-SIMS Analysis of Stainless Steel. The positive and negative mass spectra obtained from a 50 7m area on the surface of the as-received stainless steel sample are shown in Fig. 8 and 15. The positive ion spectrum shows the presence of the expected stainless components (Fe, Cr, Ni, and Mn) along with a variety of other species. Inorganic impurities include Na, Mg, K, and Ca while the organic contaminants observed include generic hydrocarbon fragments and polydimethyl siloxane (PDMS), often referred to as silicone. (Note that XPS and AES found little, if any, Si on the stainless steel surface.) A characteristic PDMS fragmentation pattern is seen with peaks at positive ion masses 73, 133, 147, 207, 221, 281, 295, 325, and 355. Between 1000 and 2000 daltons, an oligomer distribution for the PDMS is observed that provides an indication of the molecular weight of this contaminant. When looking at TOF-SIMS spectra, two things should be kept in mind. First, this technique is only sampling the outer few atomic layers of the sample and thus highlights any impurities present on the surface. Second, secondary ion yields play a huge role in how intense a peak for a particular specie appears. For example, on this steel sample, relatively large peaks for alkali and alkali earth impurities are observed. However, these elements have some of the highest positive secondary ion yields, because they all have relatively low electron affinities. Thus, they give strong peaks, even though their concentrations may be very low.

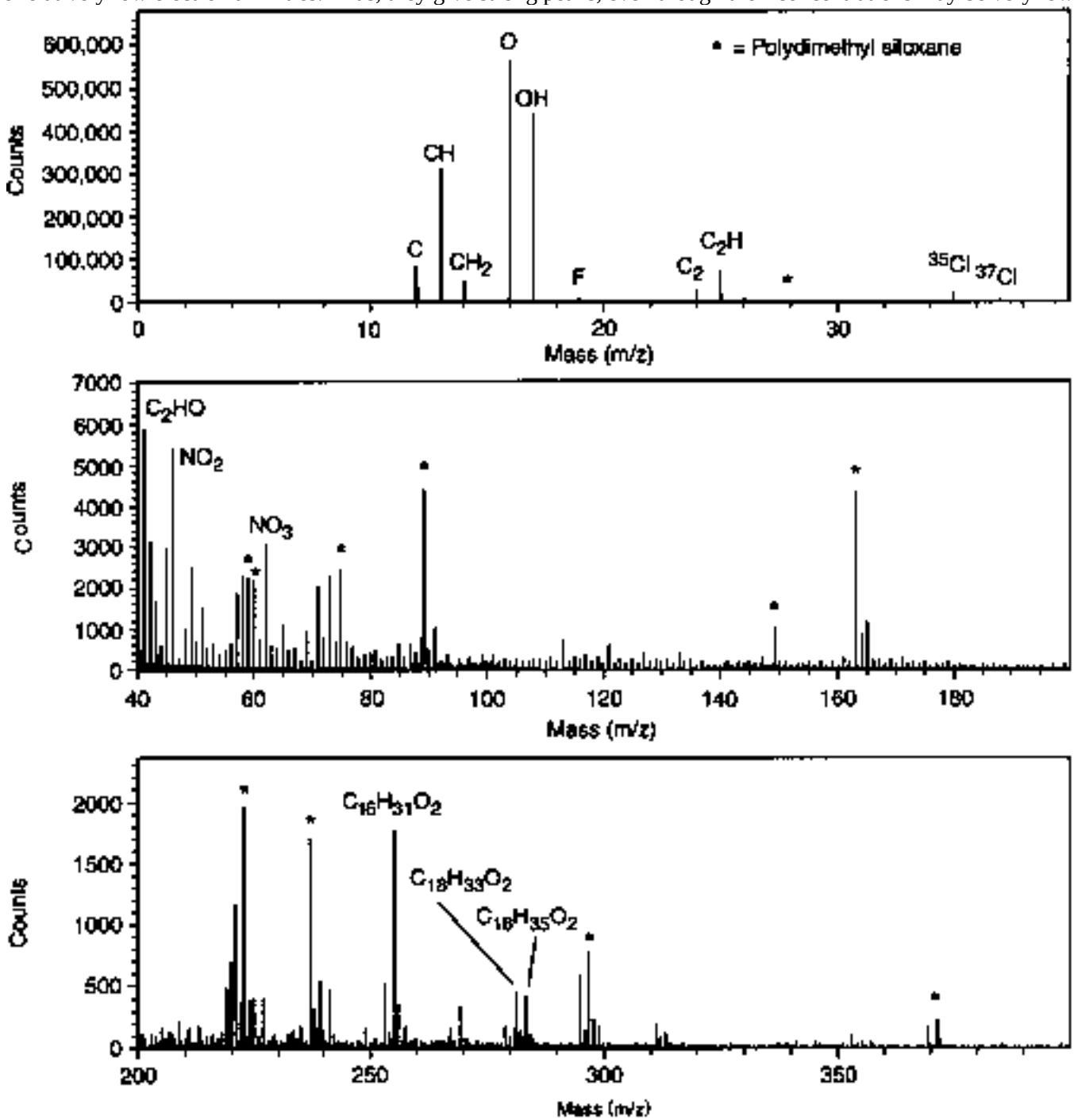


Fig. 15 Time-of-flight secondary ion mass spectrometry negative ion spectrum of stainless steel surface. Postive ion spectrum is in Fig. 8.

The negative ion spectrum provides complimentary information, showing the presence of several species not observed in the positive spectrum. Inorganic impurities seen on the stainless steel surface include O, OH, F, Cl, and NO_x (nitrate). Note that the negative ion spectrum shows little evidence for the stainless steel matrix material, because the transition metals have relatively poor negative secondary ion yields. Organic contaminants include general low-mass hydrocarbon fragments; PDMS at masses 59, 60, 75, 119, 149, and 223; and fatty carboxylates (masses 255, 281, and 283). Possible sources of the fatty carboxylate contaminant include finger oils, lubricants, and polymer additives. The finding of a carboxylate on the steel is consistent with the XPS carbon high-resolution spectrum. However, it is clear from the TOF-SIMS data that there are multiple sources of organic contamination on the surface, not just one.

Another important feature of TOF-SIMS instrumentation is its ability to perform high-mass resolution analyses. In many cases, peaks with the same nominal mass must be separated and their masses accurately measured for proper peak identification. For example, Fig. 16 shows an expanded scale of the positive ion mass spectrum centered around mass 63. Two peaks are observed that are separated by less than 0.1 mass unit. From their exact masses, the peaks can be positively identified as the 63 isotope of Cu and a C_5H_3 fragment. This capability is especially important for separating and identifying inorganic masses from organic contaminants that are typically present on a surface.

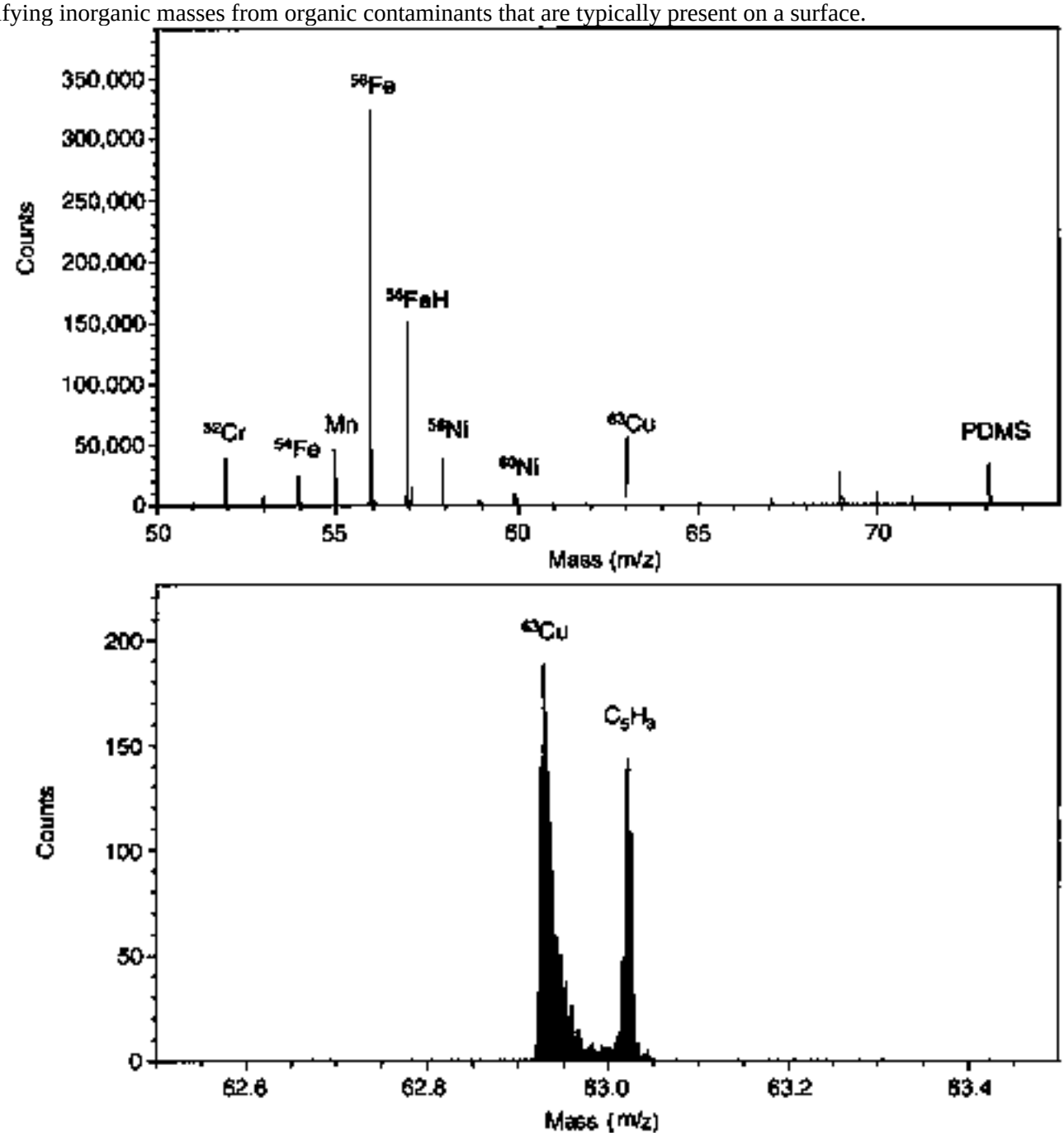
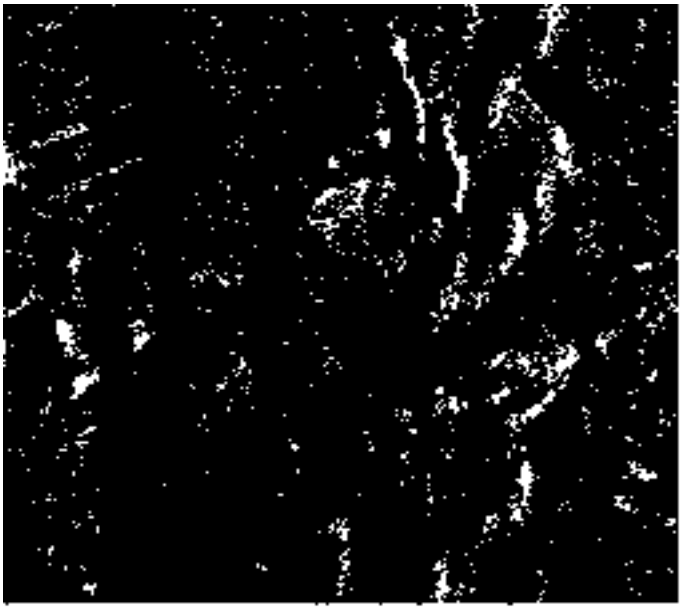
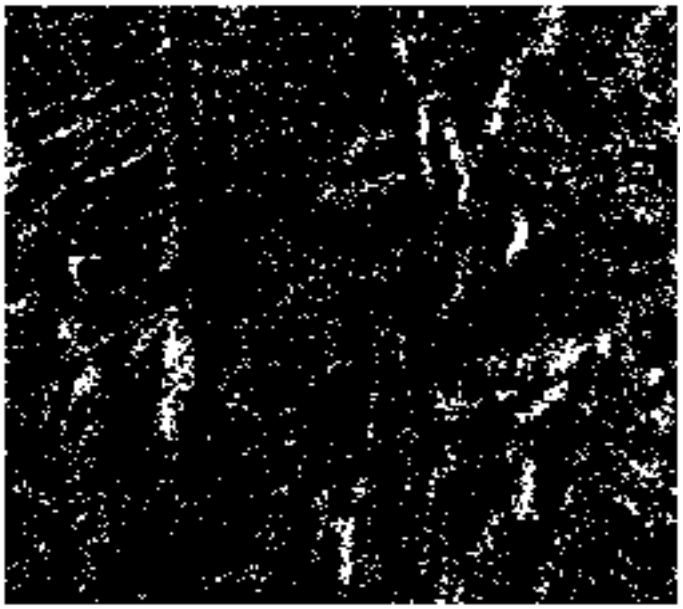


Fig. 16 Time-of-flight secondary ion mass spectrometry spectrum showing mass separation of Cu and C_5H_3 peaks, both at a nominal mass of 63

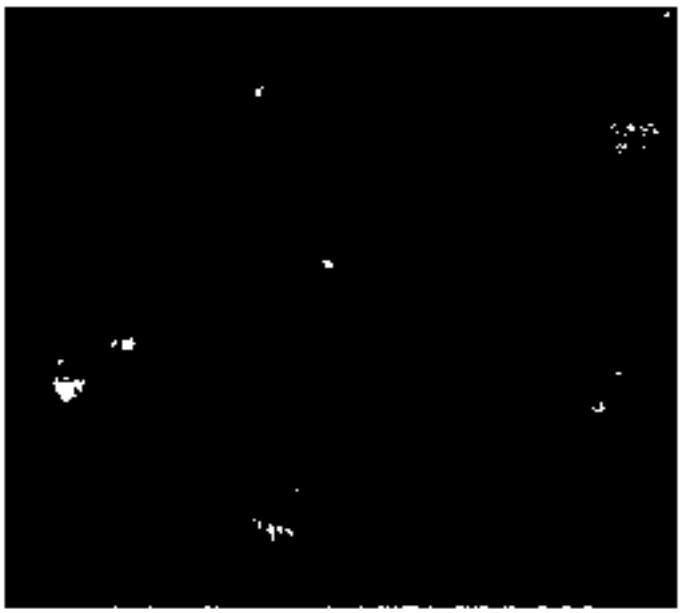
Figure 9 is the total positive ion image from the steel surface and shows the variation in secondary ion signal from the 50 7m square area where the spectra were collected. The point-to-point variation in signal can be caused by differences in chemistry or can be due simply to topographical effects, because sputter and secondary ion yields can be altered by surface morphology and changes in the effective angle of incidence. The fact that topography effects can be observed in the total ion image can be very useful in locating small surface features of interest. Secondary ion images for select elements, such as Cr, Fe, and Na are shown in Fig. 17(a–c). The Fe and especially the Cr maps are very similar to the total ion image. However, note that the vertical dark line or “scratch” seen in the left side of the Cr image is nicely filled in with a strong Fe signal. This indicates that the protective Cr passivation layer has been lost in this region and thus is an active site for corrosion to occur. The Na image (Fig. 17c) reveals both small and larger regions of Na contaminating the surface. The presence of ionic species on steel can be especially detrimental, because localized areas of electrochemical activity can be created that are known to accelerate corrosion (Ref 10).



(a)



(b)



(c)

Fig. 17 Time-of-flight secondary ion mass spectrometry images of 50 by 50 μm stainless steel surface area. (a) Map of Cr (b) Map of Fe, (c) Map of Na

Summary of Example of Stainless Steel Analysis with AES, XPS, and TOF-SIMS. The previously mentioned results show three separate analytical views of a stainless steel surface. Auger electron spectroscopy provided the best spatially resolved data and showed the presence of localized corrosion. X-ray photoelectron spectroscopy provided a view of the average surface chemistry and substantiated the AES results by showing that the Cr to Fe ratio was relatively low and that the expected Cr oxide passivation layer was lacking on this surface. The XPS also showed a lack of metallic Fe indicating that the surface was more fully oxidized than expected. Finally, the TOF-SIMS showed what organic contaminants, as well as other low-level impurities, were present in the top few atomic layers.

References cited in this section

6. "Test Method for AES Analysis of Surface and Oxide," 91060573B-STD, and "Test Method for XPS Analysis of Surface," 90120403B-STD, International Sematech SemaSpecs; see also www.sematech.org/public/resources/stds/semaspec.htm

7. S.J. Kerber and J. Tverberg, Stainless Steel Surface Analysis, *Adv. Mater. Process.*, Nov 2000, p 33–36

8. W.F. Stickle and D.G. Watson, *J. Vac. Sci. Technol. A*, Vol 10, 1992, p 2806

9. D.G. Watson, *Surf. Interface Anal.*, Vol 15, 1990, p 516–524

10. J.R. Davis, Ed., *Metals Handbook Desk Edition*, 2nd ed., ASM International, 1998, p 304, 379–386

Acknowledgments

The author would like to thank the rest of the Evans PHI and Physical Electronics Laboratory staff for their help in the preparation of this work.

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

References

1. "XPS Analysis of Disposable Gloves," XPS application brief, Charles Evans & Associates, Sunnyvale, CA, 1999

2. K. Childs et al., *Handbook of Auger Electron Spectroscopy*, 3rd ed., C.L. Hedberg, Ed., Physical Electronics, Inc., 1995

3. J. Moulder et al., *Handbook of X-Ray Photoelectron Spectroscopy*, J. Chastain and R.C. King, Jr., Ed., Physical Electronics, Inc., 1995

4. P.E. Larson and M.A. Kelly, Surface Charge Neutralization of Insulating Samples in X-Ray Photoemission Spectroscopy, *J. Vac. Sci. Technol. A*, Vol 16 (No. 6), 1998, p 3483

5. M.A. Douglas and P.J. Chen, Quantitative Trace Metal Analysis of Silicon Surfaces by TOF-SIMS, *Surf. Interface Anal.*, Vol 26, 1998, p 984–994

6. "Test Method for AES Analysis of Surface and Oxide," 91060573B-STD, and "Test Method for XPS Analysis of Surface," 90120403B-STD, International Sematech SemaSpecs; see also www.sematech.org/public/resources/stds/semaspec.htm

7. S.J. Kerber and J. Tverberg, Stainless Steel Surface Analysis, *Adv. Mater. Process.*, Nov 2000, p 33–36

8. W.F. Stickle and D.G. Watson, *J. Vac. Sci. Technol. A*, Vol 10, 1992, p 2806

9. D.G. Watson, *Surf. Interface Anal.*, Vol 15, 1990, p 516–524

10. J.R. Davis, Ed., *Metals Handbook Desk Edition*, 2nd ed., ASM International, 1998, p 304, 379–386

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

Selected References

General

- C.R. Brundle et al., *Encyclopedia of Materials Characterization*, L.E. Fitzpatrick, Ed., Butterworth-Heinemann, 1992, p 282–299, 310–323, 549–558
- E.A. Leone and A.J. Signorelli, Surface Analysis, *A Guide to Materials Characterization and Chemical Analysis*, 2nd ed., J.P. Sibilia, Ed., VCH Publishers, Inc., 1996, p 221–259
- F. Settle, Ed., *Handbook of Instrumental Techniques for Analytical Chemistry*, Prentice Hall PTR, 1997, p 793–808, 811–827, 831–847

AES, XPS, and TOF-SIMS

- D. Briggs and M.P. Seah, Ed., *Practical Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*, John Wiley & Sons, 1983
- J.C. Vickerman and D. Briggs, *The Wiley Static SIMS Library*, John Wiley & Sons, 1996
- J. Moulder et al., *Handbook of X-Ray Photoelectron Spectroscopy*, J. Chastain and R.C. King, Jr., Ed., Physical Electronics, Inc., 1995
- K. Childs et al., *Handbook of Auger Electron Spectroscopy*, 3rd ed., C.L. Hedberg, Ed., Physical Electronics, Inc., 1995

Quantitative Fractography

Arun M. Gokhale, Georgia Institute of Technology

Introduction

A FRACTURE SURFACE is generated by the deformation and fracture processes operative at numerous length scales ranging from nanometers to millimeters. The morphology and topography of the fracture surface are generally dictated by the material microstructure and active failure mechanisms. Therefore, quantitative characterization of the fracture surface

geometry, that is, quantitative fractography, can provide useful information regarding the microstructural features and failure mechanisms that govern the material fracture. The objective of quantitative fractography is to describe the geometric attributes of the topographic and microstructural features present in the fracture surface in quantitative terms, such as number, size, length, area, spacing, and orientation.

In the 1940s, fracture surfaces were studied using light microscopy (Ref 1). Such observations were useful for qualitative description of fracture surface topography. However, low magnifications and depth of field limited the utility of the technique for quantitative characterization. Development of the scanning electron microscope (SEM), and more recently, of powerful digital image analysis equipment, has led to significant advances in quantitative fractography. Numerous quantitative correlations between the material properties such as strength, ductility, toughness, and fatigue life, and the quantitative fractographic parameters have been reported in the literature during the past decade or so (Ref 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16). Modern quantitative fractographic techniques can be divided into the following categories:

- Profilometry-based quantitative fractographic techniques
- SEM-based quantitative fractographic methods
- Three-dimensional fracture surface reconstruction methods
- Internal fractographic techniques

Fracture profilometry concerns characterization of fracture *profiles* generated by intersections of the fracture surface with metallographic planes (Ref 17, 18, 19, 20, 21, 22). A fracture profile can be observed by using a standard metallurgical light microscope or an SEM. An important advantage of fracture profilometry is that the underlying microstructure can be observed simultaneously and, therefore, it can be correlated to the fracture path. Nonetheless, a fracture profile is essentially a lower-dimensional section through a nonplanar fracture surface present in the three-dimensional (3D) space. Consequently, the geometric characteristics of the features observed in a fracture profile do not necessarily represent the true values in the corresponding fracture surface. Therefore, it is essential to use unbiased statistical *stereological relationships* (Ref 23, 24) for estimation of the geometric attributes of the features in the fracture surface from the measurements performed on the fracture profiles. Significant progress has been made in the development of practical, unbiased, and efficient stereological techniques during the last decade or so (Ref 25, 26, 27, 28, 29). Several geometric parameters that quantify various aspects of fracture surface geometry can be estimated from the profilometric measurements by using general and unbiased stereological relationships. Characteristics such as fracture surface roughness (Ref 21, 22), fracture surface anisotropy (Ref 4, 30), extent of overlaps (Ref 4), average asperity height (Ref 6, 31), fraction of fracture path through different constituents or phases (Ref 32), and fracture surface fractal dimension (Ref 33, 34) can be estimated from profilometric data.

A fracture surface can be conveniently observed using an SEM. In most cases, no special specimen preparation is required for SEM fractography. Fractography by SEM is particularly useful for quantitative characterization of specific features (for example, dimples, striations, pores, facets, inclusions, etc.) present in the fracture surface (Ref 35, 36). The features of interest are characterized via estimation of their number density, average size, shape, spacing, orientation, and the distributions of these parameters, as required. Fractography by SEM is also useful for estimation of the fractions of fracture surface area generated due to different failure mechanisms, such as cleavage fracture, ductile fracture, and intergranular fracture (Ref 2, 37, 38). It is, however, important to recognize that a SEM fractograph is essentially a projected image of the nonplanar fracture surface present in the 3D space. Clearly, the geometric characteristics measured from such a projected image do not necessarily represent the corresponding true values in the fracture surface. Thus, it is essential to use general stereological relationships to estimate the true feature-specific fractographic attributes from the measurements performed on the SEM fractographs.

It is possible to characterize the true 3D geometry of a fracture surface by using stereo-pair based imaging (Ref 39, 40, 41, 42, 43, 44, 45, 46, 47). Modern image processing techniques permit reconstruction of the 3D fracture surface and its quantitative characterization using stereo-pair SEM data in an efficient manner (Ref 41, 42, 43). The basic image analysis techniques are described in detail in the article “Image Analysis” in *Materials Characterization*, Volume 10 of *ASM Handbook*. The electron backscatter diffraction (EBSD) technique of crystallographic orientation imaging and its combination with stereo-pair based SEM fractography is useful for characterization of crystallographic orientations of the features present in the fracture surface (Ref 46). Direct simultaneous observations of a nonplanar fracture surface in 3D space and the 3D microstructure just beneath it can be extremely useful in understanding and quantifying the failure modes and the relationships between the microstructure and the fracture path. Such observations are possible using a combination of SEM fractography, serial sectioning, and image analysis (Ref 48, 49).

Internal quantitative fractography concerns quantitative characterization of the microstructural damage features such as microcracks, particle cracks and debonds, microvoids, delaminations, and so forth and their evolution in the 3D microstructure as a function of stress, strain, stress state, number of stress cycles, and so on (Ref 50, 51, 52, 53, 54, 55, 56, 57). Quantitative characterization of the microstructural damage evolution is useful for understanding the local fracture processes at lower length scales and their contributions to the global fracture. Such data are required for

development of internal state variables-based predictive fracture models that are applicable under multiaxial loading conditions (Ref 58, 59). However, the internal quantitative fractographic parameters can be estimated by using the same stereological methods that are available for characterization of 3D microstructures from metallographic sections and projections. Therefore, the internal quantitative fractographic techniques are covered in the article on "Quantitative Metallography" in *Metallography and Microstructures,* Volume 9 of *ASM Handbook.*

Different categories of quantitative fractographic techniques are described in the following sections in this article. In each case, sufficient methodological details, governing equations, and practical examples are given. The next section is devoted to the fractographic techniques that are based on fracture profilometry. It is followed by a section describing methods based on SEM fractography. The final section addresses procedures for 3D fracture surface reconstruction.

References cited in this section

1. C.A. Zapffe and G.A. Moore, A Micrographic Study of the Cleavage of Hydrogenised Ferrite, *Trans. AIME,* Vol 154, 1943, p 335–359

2. N.U. Deshpande, A.M. Gokhale, D.K. Denzer, and John Liu, Relationship Between Fracture Toughness, Fracture Path, and Microstructure of 7050 Aluminum Alloy—I: Quantitative Characterization, *Metall. Mater. Trans.,* Vol 29A, 1998, p 1191–1201

3. A.M. Gokhale and G.R. Patel, Origins of Variability in the Mechanical Properties of AM60 Magnesium Alloy Castings, *Proceedings of Symposium Magnesium Techology-2001,* John N. Hryn, ed., TMS, 2001, p 195–200

4. A.M. Gokhale, W.J. Drury, and S. Mishra, Recent Developments in Quantitative Fractography, *Fractography of Modern Engineering Materials,* STP 1203, J.E. Masters and L.N. Gilbertson, Ed., ASTM, 1993, p 3–22

5. A. Shyam, S.A. Padula II, D.L. Davidson, and W.W. Miligan, Fatigue Crack Propagation Thresholds in Nickel base Super Alloys at High Frequencies and Temperatures, *Metall. Mater. Trans.,* Vol 32A, 2001

6. W.J. Drury, A.M. Gokhale, and S.D. Antolovich, Effect of Crack Surface Geometry on Fatigue Crack Closure, *Metall. Mater. Trans.,* Vol 26A, 1995, p 2651–2663

7. H. Kaneko, H. Ishikawa, T. Konishi, and M. Yamada, Study on Estimation Methods of Applied Stress Using Fractography Analysis, *Mitsubishi Tech. Rev.,* Vol 38 (No. 3), 2001, p 156–160

8. B.B. Mandelbrot, D.E. Passoja, and A.J. Paullay, Fractal Character of Fracture Surfaces of Metals, *Nature,* Vol 308, 1984, p 721–722

9. J.E. Dove and J.D. Frost, A Method for Measuring Geomembrane Surface Roughness, *Geosynthetics International,* Vol 3, 1996, p 369–392

10. E.E. Underwood and K. Banerji, Fractals in Fractography, *Mater. Sci. Eng.,* Vol 80A, 1986, p 1–14

11. X. W. Li, J. F. Tian, Y. Kang, and Z. G. Wang, Quantitative Analysis of Fracture Surface Roughness, *Scr. Metall. Mater.,* Vol 33, 1995, p 803–809

12. R. G. Baggerly and T. Archbold, Quantitative Fractography of Fatigue Fracture in Nodular Iron, *Proceedings of Third International Conference on Low Cycle Fatigue and Elasto-Plastic Behavior of Materials,* K.T. Rie, Ed., Elsevier Science, Berlin, 1992, p 118–123

13. N. Ranganathan, N. Gerard, A. Tougui, R. Leroy, M. Benguediab, M. Mazari, Y. Nadot, and J. Petit, Analysis of Fatigue Crack Propagation by Quantitative Fractography, STP 1402, ASTM, 2001, p 52–84

14. W.J. Drury, Involvement of Processing Defects in Failure of FP/Al-Li, *Metall. Trans.,* Vol 20A, 1989, p 2175–2179

15. A. Krasowsky, Advanced SEM Methods in Fatigue and Fracture Research, *Met. Mater.* Vol 36, 1998, p 193–200

16. A.M. Brandt and G. Prokopski, On the Fractal Dimension of Fracture Surfaces of Concrete Elements, *J. Mater. Sci.,* Vol 28, 1993, p 4762–4766

17. J.R. Pickens and J. Gurland, Metallographic Characterization of Fracture Profiles on Sectioning Planes, *Proceedings of 4th International Conference on Stereology,* E.E. Underwood, R. De Wit, and G.A. Moore, Ed., NBS Publication 431, National Bureau of Standards, 1976, p 269–272

18. E.E. Underwood and S.B. Chakraborty, Quantitative Fractography of a Fatigued Ti-28V Alloy, *Fractography and Materials Science,* L.N. Gilbertson and R.D. Zipp, Ed., STP 733, ASTM, 1981, p 337–354

19. S.M. El-Soudani, Profilometric Analysis of Fracture, *Metallography,* Vol 11, 1978, p 247–336

20. E.E. Underwood, Stereological Analysis of Fracture Roughness Parameters, *Acta Stereol.,* Vol 6 (Supplement II), 1987, p 169–178

21. A.M. Gokhale and E.E. Underwood, A General Method for Estimation of Fracture Surface Roughness, Part-I: Theoretical Aspects, *Metall. Trans.-A,* Vol 21A, 1990, p 1193–1199

22. A.M. Gokhale and W.J. Drury, A General Method for Estimation of Fracture Surface Roughness, Part-II: Practical Considerations, *Metall. Trans.,* Vol 21A, 1990, p 1201–1207

23. E.E. Underwood, *Quantitative Stereology,* Addison Wesley Publishing Co., 1970

24. R.T. DeHoff and F.N. Rhines, *Quantitative Microscopy,* McGraw Hill Publishing Co., New York, 1968

25. A.J. Baddeley, H.J.G. Gundersen, and L.M. Cruz Orive, Estimation of Surface Area Using Vertical Sections, *J. Microsc.,* Vol 142, 1986, p 259–276

26. A.M. Gokhale and W.J. Drury, Efficient Estimation of Microstructural Surface Area Using Trisector, *Metall. Trans.,* Vol 25A, 1994, p 919–928

27. A.M. Gokhale, "Unbiased Estimation of Curve Length in 3D using Vertical Slices," *J. Microsc.,* Vol 159, 1990, p 133–141

28. D.C. Sterio, The Unbiased Estimation of Number and Size of Arbitrary Particles Using Disector, *J. Microsc.,* Vol 134, 1984, p 127–136

29. A.M. Gokhale and A. Tewari, Efficient Estimation of Number Density in Opaque Material Microstructures Using Large Area Disector (LAD), Part 3, *J. Microsc.,* Vol 200, 2000, p 277–283

30. S.D. Antolovich, A.M. Gokhale, and C. Bathias, Application of Quantitative Fractography and Computed Tomography to the Fracture Processes in Materials, *Quantitative Methods in Fractography,* STP 1085, B.M. Strauss and S.K. Putatunda, Ed., ASTM, 1990, p 3–25

31. W.J. Drury and A.M. Gokhale, Digital Profilometry of Fracture Surfaces, *Quantitative Microscopy and Image Analysis,* D.J. Diaz, Ed., ASM International, 1994, p 83–87

32. M.D. Dighe and A.M. Gokhale, Relationship Between Microstructural Extremum and Fracture Path in a Cast Al-Si-Mg Alloy, *Scr. Mater.,* Vol 37, 1997, p 1435–1440

33. W.J. Drury and A.M. Gokhale, Measurement and Interpretation of Fracture Surface Fractal Dimension, *Metallography: Past, Present, and Future,* STP 1165, G.F. Vander Voort, Ed., ASTM, 1993, p 295–309

34. E. Charkaluk, M. Bigerelle, and A. Lost, Fractals and Fracture, *Eng. Fract. Mech.,* Vol 61, 1998, p 119–139

35. A.S. Argon and J. Im, Separation of Second Phase Particles in Spheroidized 1045 Steel, Cu-0.6% Cr Alloy, and Maraging Steel in Plastic Straining, *Metall. Trans.,* Vol 6, 1975, p 839–851

36. A.M. Gokhale and G.R. Patel, Effect of Casting Defects on Mechanical Properties of Magnesium Alloy Castings, *Proceedings of TMS Symposium on Automotive Alloys 2002,* S.K. Das, Ed., TMS, in press

37. N.U. Deshpande and A.M. Gokhale, Relationship Between Fracture Path and Microstructure of Aluminum Alloy 7050, *Proceedings of 4th International Conference on Aluminum Alloys,* T.H. Sanders, Ed., Georgia Institute of Technology, Atlanta, Vol 1, 1994, p 685–692

38. T. Kawabata and O. Izumi, The Relationship Between Fracture Toughness and Transgranular Fracture in Al-6% Zn-2.5% Mg Alloy, *Acta Metall.,* Vol 24, 1977, p 817–825

39. A. Boyde, Quantitative Photogrammetric Analysis and Quantitative Stereoscopic Analysis of SEM Images, *J. Microsc.,* Vol 98, 1973, p 452–471

40. A. Tatschl, C.J. Gilbert, V. Schroeder, R. Pippan, and R.O. Ritche, Stereophotogrammetric Investigation of Overload and Cyclic Fatigue Fracture Surface Morphologies in a Zr-Ti-Ni-Cu-Be Bulk Metallic Glass, *J. Mater. Res.,* Vol 15, 2000, p 898–903

41. J.J. Ammann, L.R. Hein, and A.M. Nazar, Three-Dimensional Reconstruction of Fracture Surfaces, *Mater. Charact.,* Vol 36, 1996, p 379–386

42. I. Rogerio de Oliveira Hein, F. de Azevedo Silva, A. Maria, M. Nazar, and J.J. Ammann, Three-Dimensional Reconstruction of Fracture Surfaces: Area Matching Algorithms for Parallax Measurements, *Scanning,* Vol 21, 1999, p 253–263

43. L.R.O. Hein, Quantitative Fractography by Digital Image Processing: NIH Image Macro Tools for Stereo Pair Analysis and 3-D Reconstruction, Part 1, *J. Microsc.,* Vol 204, 2001, p 17–28

44. C. Po-Shou and R.C. Wilcox, Stereographic Technique for Quantitative Analysis of Cleavage Plane Orientation, *Mater. Charact.,* Vol 26, 1991, p 9–15

45. G. Koenig, W. Nickel, J. Storl, D. Meyer, and J. Stange, Digital Stereophotogrammetry for Processing SEM Data, *Scanning,* Vol 9, 1987, p 185–193

46. P.A. Davis and V. Randle, Combined Application of Electron Backscatter Diffraction and Stereo-Photogrammetry in Fractography Studies, Part 1, *J. Microsc.,* Vol 204, 2001, p 29–38

47. H.E. Exner and M. Fripan, Quantitative Assessment of Three-Dimensional Roughness, Anisotropy, and Angular Distributions of Fracture Surfaces by Stereometry, Part 2, *J. Microsc.,* Vol 139, 1985, p 161–178

48. M.D. Dighe, "Development of Digital Image Processing Based Methodology to Quantify and Correlate the Microstructure and Three-Dimensional Fracture Surface Morphology of Aluminum Alloy 7050," Ph.D. dissertation, Georgia Institute of Technology, 2000

49. M.D. Dighe and A.M. Gokhale, Correlation Between the Microstructure and Three-Dimensional Fracture Surface Morphology in 7050 Alloy, *Proceedings of James T. Staley Honorary Symposium on Aluminum Alloys-2001,* TMS, in press

50. A.M. Gokhale, Estimation of Bivariate Size and Orientation Distribution of Microcracks, *Acta Metall. Mater.,* Vol 44, 1996, p 475–485

51. V. Benes, A.M. Gokhale, and M. Slamova, Unfolding the Bivariate Size-Orientation Distribution Function, *Acta Stereol.,* Vol 15, 1996, p 9–14

52. W.T. Whited, A.M. Gokhale, and N.U. Deshpande, Characterization of Thermally Induced Microcracks in a Metal Matrix Composite, *Microstruct. Sci.,* Vol 21, 1994, p 107–120

53. M.D. Dighe, A.M. Gokhale, and M.F. Horstemeyer, Effect of Loading Condition and Stress State on Damage Evolution of Silicon Particles in an Al-Si-Mg Base Cast Alloy, *Metall. Mater. Trans.,* Vol 33A, 2002, p 555–565

54. M.D. Dighe, A.M. Gokhale, and M.F. Horstemeyer, Effect of Strain Rate on Damage Evolution in a Cast Al-Si-Mg Alloy, *Metall. Mater. Trans.,* Vol 31A, 2000, p 1725–1731

55. J.-W. Yeh and W.-P. Liu, The Cracking Mechanism of Si Particles in A357 Alloy, *Metall. Trans.,* Vol 27A, 1996, p 3358–3368

56. A.M. Gokhale, Utility of Stereological Counting Measurements in the Study of Creep Cavitation Kinetics, *MiCon-90: Advances in Video Technology for Microstructural Control,* STP 1094, G.F. Vander Voort, Ed., ASTM, 1991, p 332–339

57. D. Kwon and R.J. Asaro, A Study of Void Nucleation, Growth, and Coalescence in Spheroidized 1518 Steel, *Metall. Trans.,* Vol 21A, 1990, p 117–134

58. Z.H. Huang, J.F. Tian, and Z.G. Wang, A Study of the Slit-Island Analysis as a Method for Measuring Fractal Dimension of Fractured Surface, *Scr. Metall.,* Vol 24, 1990, p 967–972

59. Z.H. Huang, J.F. Tian, and Z.G. Wang, *Mater. Sci. Eng.,* Vol A118, 1989, p 19–24

Quantitative Fractography

Arun M. Gokhale, Georgia Institute of Technology

Profilometry-Based Quantitative Fractography

A fracture profile is a tortuous irregular line generated by the intersection of a fracture surface with a sectioning plane. Several experimental methods are available for generation of fracture profiles. These experimental methods yield the (x, y) coordinates of the set of closely spaced points in the fracture profile. From these basic data, important geometric attributes of the fracture profile can be computed. Stereological equations can be then used for unbiased statistical estimation of the geometric attributes of the corresponding fracture surface from the computed fracture profile parameters. The experimental methods for profile generation, computation of fracture profile parameters, and the stereological procedures for estimation of fracture surface attributes from the profilometric data are described in the following subsections.

Experimental Methods for Profile Generation

Physical metallographic sectioning of the specimen, sectioning of fracture surface replicas, and nondestructive profile generation by using stereo pairs and other methods are the important experimental techniques used for generation of fracture profiles. These methods are described as follows.

Metallographic Technique for Fracture Profile Generation. Depending on the angular orientation of the sectioning plane with respect to the fracture surface, it is possible to generate horizontal (Ref 8, 58, 59), slanted (Ref 60), conical (Ref 61), and vertical section (Ref 17, 18, 21, 22) fracture profiles. The vertical section fracture profiles are most widely used as they are easy to generate using well-known metallographic procedures, and general stereological equations are available for estimation of fracture surface characteristics from the measurements performed on vertical section fracture profiles (Ref 4, 21, 22). Figure 1 shows a schematic fracture surface, a sectioning plane, and a schematic representation of the fracture profile generated by intersection of the fracture surface with the sectioning plane. Figure 2(a) shows a fracture profile generated from the fracture surface of a tensile test specimen of a metal-matrix composite (MMC) containing unidirectionally aligned alumina fibers in the matrix of an aluminum alloy. Figures 2(b) and 2(c) show a fracture profile from the tensile fracture surface of a low-alloy steel specimen. These fracture surfaces were electroplated with copper. As illustrated by Fig. 2, an important advantage of profilometry is that the fracture profile and the underlying microstructure can be observed simultaneously, and therefore, the fracture path can be correlated to the microstructure.

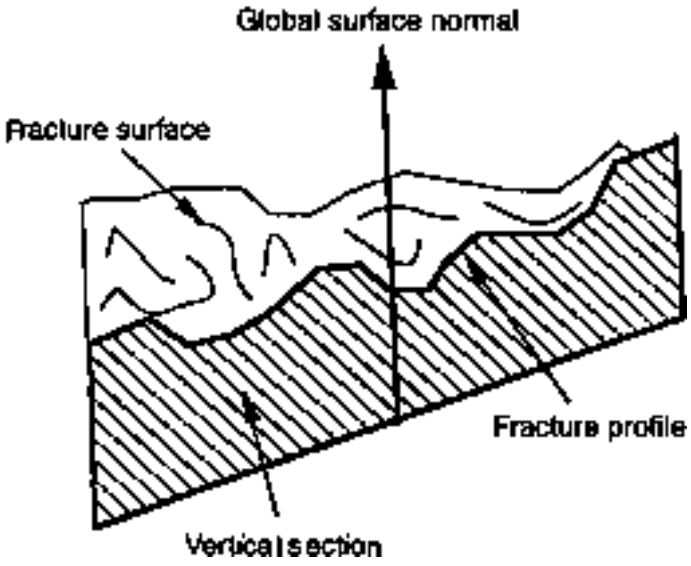
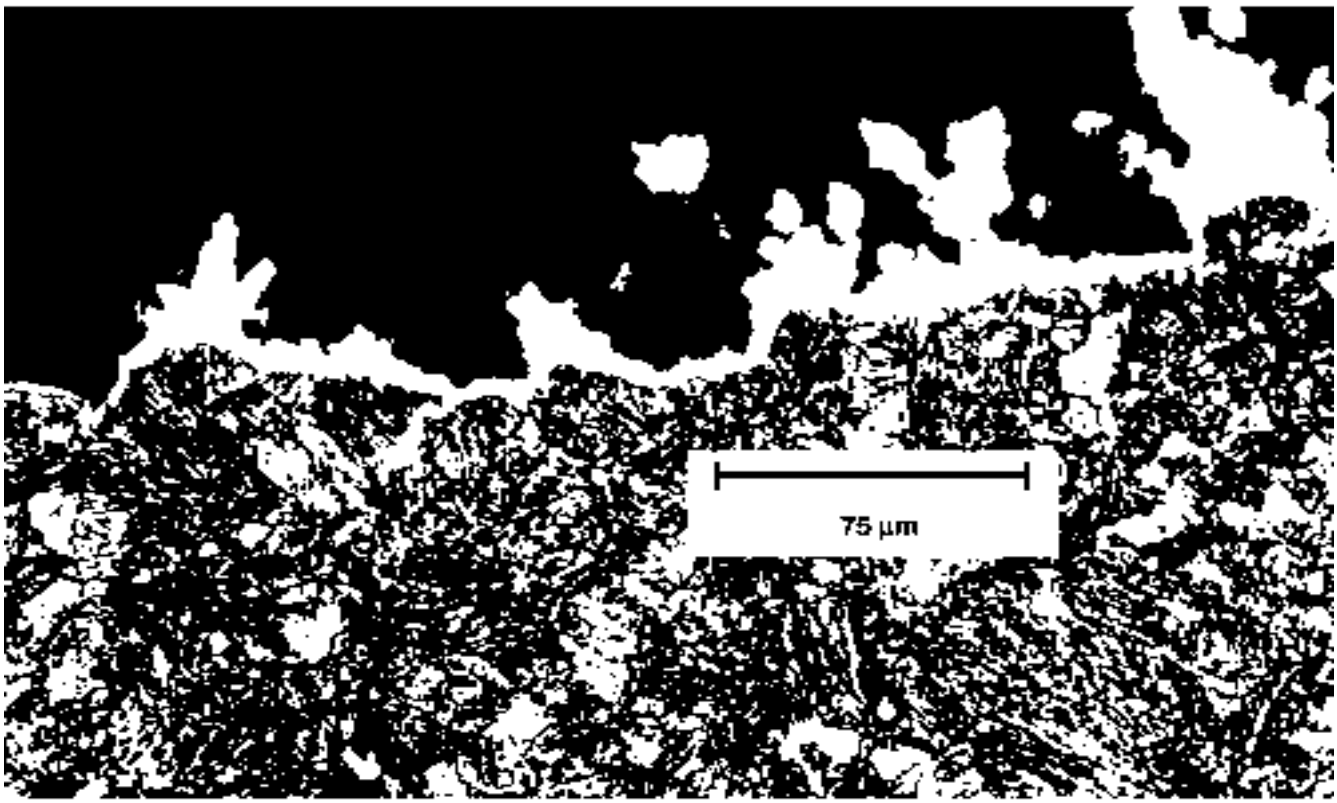
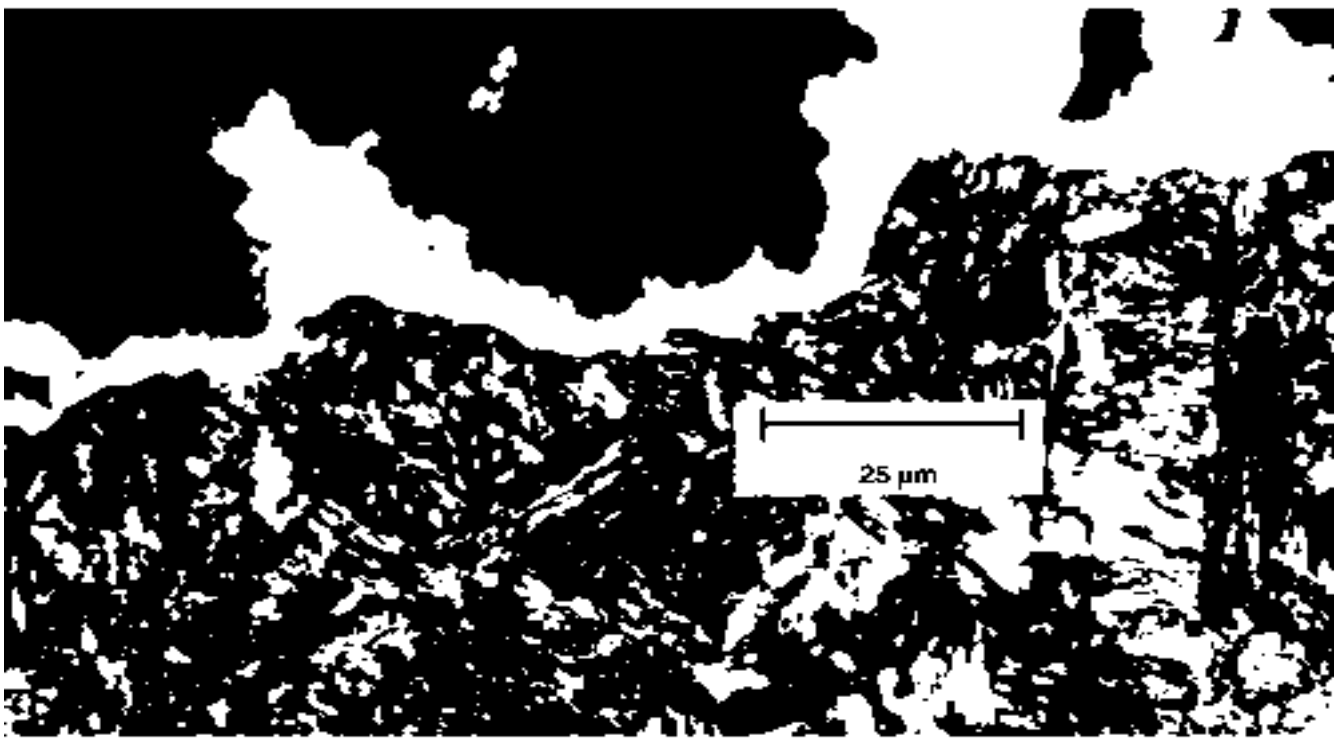


Fig. 1 A schematic fracture surface, a sectioning plane, and the fracture profile generated by intersection of the sectioning plane with the fracture surface. Source: Ref 4

(a)



(b)



(c)

Fig. 2 Fracture profiles. (a) Fracture profile generated from the fracture surface of a metal-matrix composite containing alumina fibers distributed in the matrix of an aluminum alloy. (b) Fracture profile generated from the fracture surface of a tensile test specimen of low-alloy steel. (c) A high magnification field of view of the fracture profile in (b)

Nevertheless, metallographic fracture profile generation is destructive of the fracture surface and sample. Consequently, additional detailed scrutiny of special areas in the fracture surface is not possible once it is coated and cut. However, the fracture surface can be carefully inspected by SEM in advance, and the images of areas of interest (or the image montage of the whole fracture surface) can be grabbed and stored in the memory of an image analysis computer before coating and cutting take place for fracture profilometry (Ref 48). The experimental procedures for fracture profile generation are straightforward. When the specimen is fractured, two nonplanar fracture surfaces are produced. After the completion of SEM observations and image storage, one or both surfaces can be electrolytically coated (see Fig. 2) to preserve the edge on subsequent sectioning (Ref 62). The coated specimen is then mounted metallographically and prepared according to conventional metallographic procedures (Ref 62).

The complete information regarding the fracture profile topography is contained in the set of (x, y) coordinates of the points in the profile. These data can be obtained by digitizing the fracture profile via either interactive or automatic digital image analysis (Ref 22). Some popular image manipulation software packages also have these capabilities. Typically, the process involves tracing the fracture profile using an electronic mouse for interactive measurements, or appropriate image segmentation for automatic measurements. The image analyzer records the coordinates of closely spaced points in the fracture profile at preselected fixed intervals. The profile is thus approximated by a series of straight-line segments as illustrated in Fig. 3. As explained in the subsequent section, numerous geometric characteristics of the fracture profile can be computed from these data.

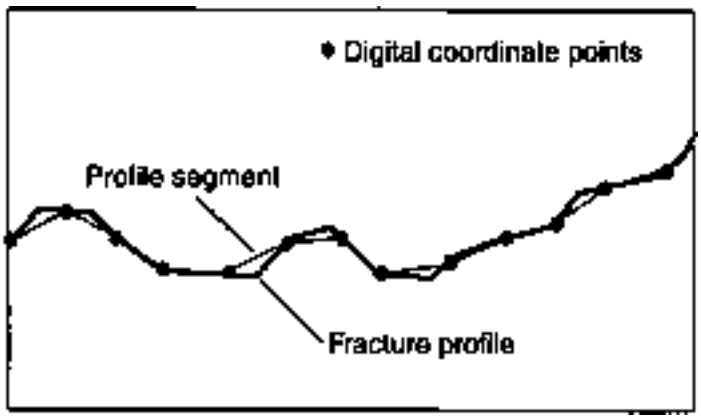


Fig. 3 Digitization of the fracture profile yields (x, y) coordinates of a set of closely spaced points on the profile. The profile is then approximated by a series of straight-line segments joining the adjacent points.

Replica Technique For Fracture Profile Generation. Planar fracture profiles can be also generated by sectioning replicas of the fracture surface (Ref 63). In this manner, the fracture surface itself can be preserved for subsequent studies. The suitable foil materials and procedures for coating and striping replicas are well established. The replicas can be sectioned by using a microtome, which is also a useful instrument for serial sectioning of the replicas for 3D fracture surface reconstruction. However, care must be taken to minimize the distortion of the replica during the sectioning process. Dimples in fractured 2024 aluminum alloy have been studied using microtomed slices from replicas of the fracture surface (Ref 64). Even though the replica technique is quite attractive, its applications are limited to relatively smooth surfaces in order to avoid tearing the replica.

Nondestructive Profile Generation Methods. The techniques for nondestructive fracture profile generation include laser profilometry (Ref 65), atomic force microscopy (Ref 66), confocal scanning microscopy (Ref 67), light profile microscopy (Ref 68), interferometry (Ref 69), stylus profilometry (Ref 70), and SEM stereo-pair based methods (Ref 71). For laser profilometry (Ref 65), the fracture surface is illuminated using a thin slit of laser at an angle of 45 degrees to the surface, and the surface is observed at an angle of 90 degrees. The projected slit of laser light appears as a straight line if the surface is flat and as a progressively more undulating rough line as the roughness of the surface increases. The output image of the laser slit is grabbed using an image analyzer and digitized to obtain the (x, y) coordinates of closely spaced points along the profile, and the profile characteristics are computed from these data.

In confocal laser scanning microscopy (CLSM), the fracture surface is scanned with a fine focused laser beam (Ref 67). As a result of point illumination and point detection, only light reflected from features within the focal plane enters the detector, which then generates an optical slice containing sharp details. Such a slice is essentially a horizontal profile through the fracture surface. By changing the focal plane (*z*-shift of the specimen or the microscope objective), a series of optical slices is received from which the (*x*, *y*, and *z*) coordinates of the points on different slices can be obtained. The data can be used for 3D fracture surface reconstruction and also to compute the geometric characteristics of the profiles and/or the surface.

The methods based on interferometry (Ref 69) utilize the interference fringes produced when a laser is reflected off a rough fracture surface and a flat reference surface. The fringes contain the topographic information regarding the geometry of the profile along the scan line. The stylus profilometers are widely used to characterize the topography of machined surfaces, but for fractography, they are useful only for very rough surfaces (Ref 70). In these profilometers, the vertical deflection of the stylus is recorded as a function of its position, which essentially yields the (*x*, *y*) coordinates of the points in the profile along the chosen direction.

In light profile microscopy (Ref 68), a narrow illuminated line is projected on the fracture surface that reveals the profile of the surface in a specified direction. The line can be observed through a light microscope, and its image can be stored for subsequent analysis. The resolution obtained depends on the objective lens optics. Roughness of fatigue fracture profiles from 1080 steel was characterized using the light-profile microscopy (Ref 61). In these profiles, the mean peak-to-trough distance had values in the range of 50 ± 5 7m.

An SEM-based method for nondestructive profile generation involves the use of the SEM line scan mode (Ref 71). In this method, a narrow contamination line is deposited on the specimen fracture surface using the line scan mode. The specimen is then tilted, preferably around an axis parallel to the line-scan direction, by an angle 8. The contamination line then appears as an oblique projection of the profile that can be evaluated (point by point) by the following relationship:

$$\Delta z = \left[\frac{1}{M\cos(90^\circ - \alpha)}\right]\Delta k \qquad \text{(Eq 1)}$$

In Eq 1, " *k* is the displacement of the contamination line at the point of interest, *M* is the magnification, and " *z* is the corresponding height (elevation) of the profile at that point. The tilt angle should be as large as possible. The major source of error comes from broadened contamination lines with diffuse contours. Fortunately, the lines appear considerably sharper in the tilted position. Overall, both the contamination line and light-beam profile techniques are useful for quantitative characterization of the fracture profiles.

Estimation of Geometric Attributes of Fracture Profiles

All experimental profilometric methods yield (*x*, *y*) coordinates of a set of closely spaced points in the fracture profile under investigation. From these basic data, numerous geometric attributes of the fracture profile can be calculated in a fairly straightforward manner as to be described.

Fracture Profile Roughness Parameters. An important topographic attribute of a fracture profile is its *roughness.* The motivation for the estimation of roughness is its likely correlation with fracture toughness, the extent of crack deflection, and fatigue-crack growth behavior. Several parameters have been proposed to quantify fracture profile roughness. The most important ones are those that can be defined uniquely and can be used to estimate the corresponding fracture *surface* roughness attributes in an unbiased manner.

The lineal profile roughness parameter, R_L, is perhaps the most important profile roughness parameter for quantitative fractographic analysis. It is defined as the ratio of the true total fracture profile length, 2, and its *apparent* projected length, *L* (Ref 17, 18, 19, 20, 21):

$$R_L = 2/L \qquad \text{(Eq 2)}$$

Figure 4(a) shows a schematic fracture surface of area *S* enclosed in a cylindrical reference volume. The axis of the cylinder (*z*-axis) is along the direction perpendicular to the average topographic plane of the fracture surface. In the present context, this direction is called vertical axis (Ref 21), and any sectioning plane containing (or parallel to) this vertical axis is called a vertical sectioning plane (see Fig. 4b). A fracture profile generated by a vertical sectioning plane is called a vertical section fracture profile (see Fig. 4c). Note that the vertical axis can be identified in the vertical sections. To calculate the roughness parameter R_L, it is necessary to measure the total length of the fracture profile 2. Once the fracture profile is digitized, the total profile length 2 is simply given by the sum of the lengths of the straight line segments joining the adjacent digitized points in the profile (see Fig. 3), and therefore, the parameter R_L can be calculated. In general, R_L can have any value from one to infinity. For a randomly oriented fracture profile that does not have any reentrant regions or folds (Ref 20), R_L is equal to +/2. Experimental values of R_L reported in the literature are in the range of 1.05 to 3.75. For fatigue crack growth studies, the value of the parameter R_L in the crack growth direction can be used to calculate the geodesic crack length from the experimentally measured conventional projected crack length (Ref 72).

For ductile fracture of ultrahigh-strength steels where the fracture micromechanisms and microstructure lead to smooth dimple interfaces, the lineal profile roughness parameter exhibits quantitative correlation with the critical value of the crack opening displacement, $*_{Ic}$ (Ref 73).

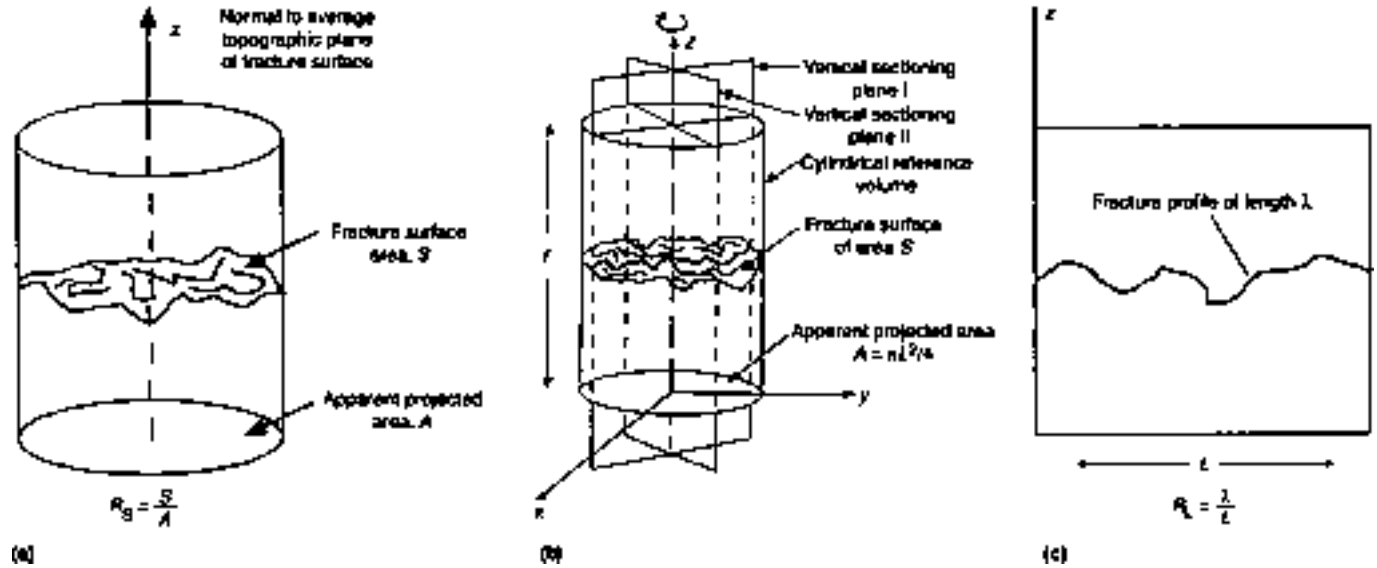


Fig. 4 Vertical section planes. *L*, cylinder diameter. (a) A schematic fracture surface of area *S* enclosed in the cylindrical reference volume. (b) Fracture surface in (a) sectioned by vertical sectioning planes. The vertical planes are the sectioning planes that contain (or are parallel to) the vertical axis. (c) A vertical section fracture profile generated by a vertical plane. The vertical axis can be identified in the vertical sections. *L*, apparent projected length of fracture profile

It is important to recognize that the geometry of a fracture profile depends on the geometry and topography of the fracture surface as well as on the orientation of the sectioning plane with respect to the fracture surface. Therefore, the value of R_L (and of other profile parameters) can systematically vary with the orientation of the vertical sectioning plane (Ref 20, 21). As illustrated in Fig. 5, if the fracture surface anisotropy is not symmetric with respect to the vertical axis, then different vertical plane orientations yield significantly different values of R_L, which has been observed experimentally (Ref 22). It follows that the value of R_L is not unique for a given fracture surface, and contrary to El-Soudani's earlier theorem (Ref 19), a higher value of R_L (or of any other profile roughness parameter) on a single sectioning plane does not necessarily imply a rougher fracture surface.

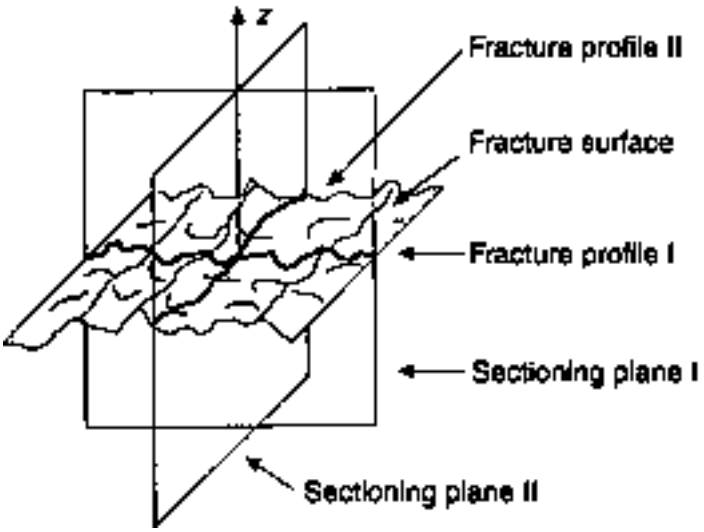


Fig. 5 Schematic illustrating why vertical planes of different orientations may yield fracture profiles having different R_L values.

Profile configuration parameter, R_P, is another descriptor of fracture profile roughness. R_P is equal to the average height of the peaks on the fracture profile divided by the average spacing of the peaks (Ref 74), and it also can be calculated from the digitized profile data. Vertical roughness parameter, R_V, is defined as the total projected length of the fracture profile

on the *y*-axis (i.e., vertical axis for the vertical section fracture profiles) divided by the apparent projected length of the fracture profile on the *x*-axis (i.e., mean line through the fracture profile). According to Underwood and Banerji (Ref 75), for the fracture profiles that do not have overlaps (reentrant segments), R_V is equal to $2R_P$.

It is important to recognize that the roughness parameters R_L, R_P, and R_V are dimensionless, and therefore they do not contain information about the length scale of the topographic details on the fracture surface. These dimensionless parameters have the same values for geometrically similar fracture profiles that differ only in scale. A profile roughness parameter that is length-scale dependent is the average value of peak-to-trough height, H_a, in the fracture profile. The parameter H_a can also be calculated from the digitized profile data. Another important length scale dependent profile roughness parameter is the average asperity height • *h*• in the fracture profile. The asperity heights have to be measured with respect to some reference base line, which is not always naturally defined in the complex tortuous fracture profiles. A least-square straight line through the fracture profile segment may be used as a base line, but the position of such a base line is dependent on the length of the profile segment. Therefore the calculated average asperity height may vary with the increase in the length of the profile segment under analysis. However, analysis of a number of fracture profiles generated by fatigue crack propagation has revealed that although initially the calculated average asperity height increases with the length of the base line, it quickly approaches a saturation value (Ref 6, 31) for baseline lengths larger than about 500 7m (see Fig. 6). For the nickel-base alloy 718, the saturation value of the average asperity height • *h*• for fatigue crack propagation fracture profiles correlates with the roughness-induced crack closure stress intensity K_{cl} (see Fig. 7). Recently, a quantitative correlation between the fatigue crack propagation threshold and the standard deviation of the fracture surface asperity height distribution has been reported for some nickel-base superalloys (Ref 5). Through the extrapolation, the linear correlation enables calculation of an "intrinsic" threshold corresponding to zero roughness, which is in agreement with the one predicted by theoretical models.

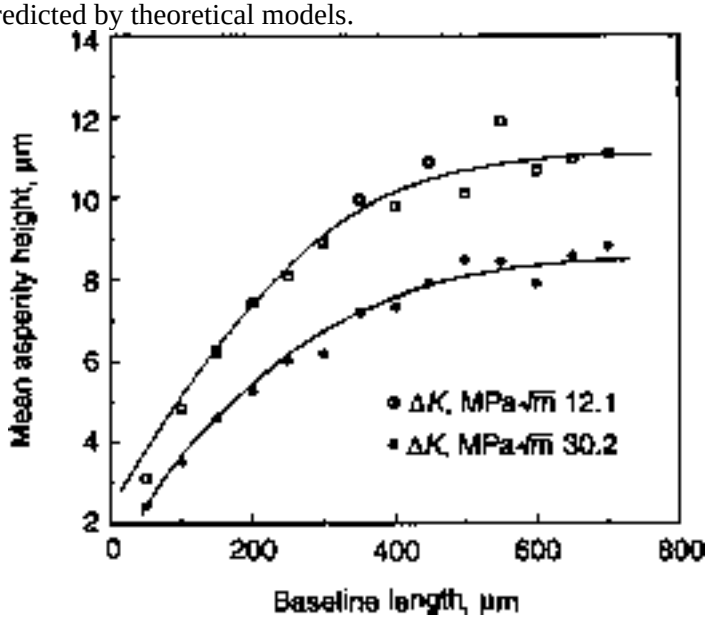


Fig. 6 Variation of the calculated average asperity height with the baseline length for the fatigue crack propagation fracture profile of a IN-718 nickel-base alloy having a coarse grain structure (grain size, 23 %m). Observe that average asperity height reaches a saturation value for baseline lengths larger than 500 %m. Source: Ref 6, 31

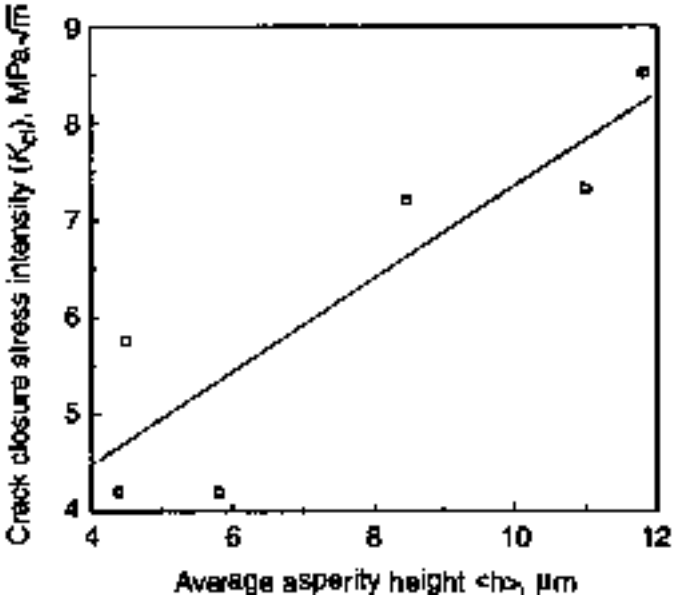
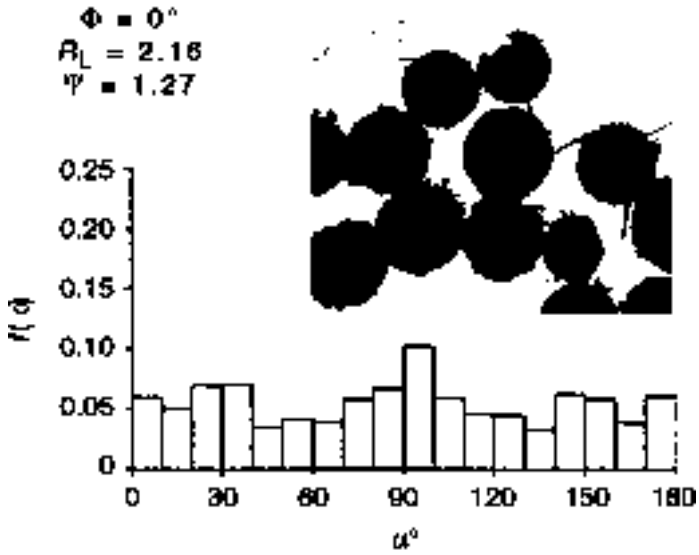
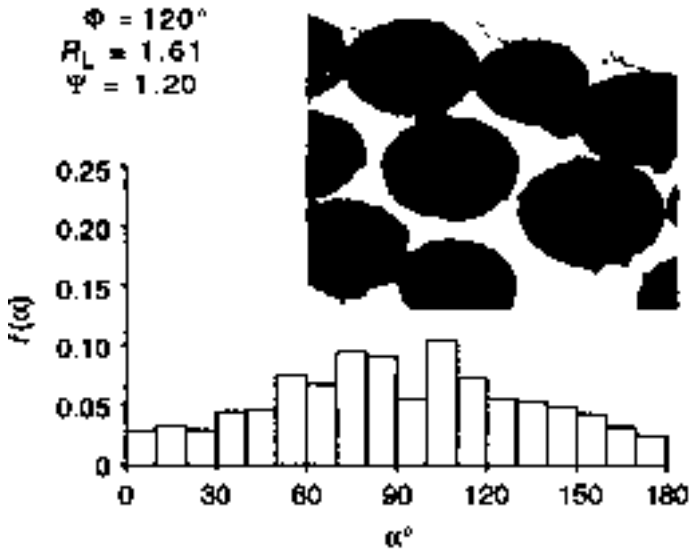


Fig. 7 Plot of crack-closure stress intensity versus average asperity height for fatigue crack propagation in IN-718 nickel-base alloy. The data include specimens of two different grain sizes tested at three different *K* levels. Source: Ref 6

Angular Orientation Distribution of the Fracture Profile. The angular orientation of a line element on a vertical-section fracture profile is specified by angle 8 between the line element and the vertical axis. The angle 8 can be calculated for each straight-line segment connecting the consecutive points of the digitized profile (see Fig. 3). The orientation distribution function of the fracture profile, *f*(8), is defined such that *f*(8)*d*8 is the fraction of profile length in the orientation range 8 to (8 + *d*8), where 0 9 8 9 +. The orientation distribution function, *f*(8) quantifies the anisotropy of the fracture profile. For a randomly oriented fracture profile, *f*(8) is equal to [1/+], as all angular orientations are equally likely in such a profile. The orientation distribution function can be computed in histogram form from the experimentally measured values of the orientation angles of the line elements in the fracture profile. The profile orientation distribution function is required for estimation of the fracture surface roughness as well as for estimation of the orientation distribution of surface elements in the fracture surface from the profilometric data. In general, the profile angular orientation distribution may vary with the orientation of the vertical section. Figure 8 shows experimentally measured profile orientation distribution functions, lineal profile roughness parameters, and profile structure factors for fracture profiles observed in three vertical sections of tensile fracture surface of a metal-matrix composite (MMC) generated by uniaxial load perpendicular to the fibers. The three vertical sections are mutually at an angle of 120° to one another, and as they are vertical sections, all of them contain the vertical axis. The MMC contains aligned alumina fibers in an aluminum-alloy matrix (Ref 22).


Φ = 0°
R_L = 2.16
Ψ = 1.27
0.25
0.20
0.15
0.10
0.05
0
f(α)
0
30
60
90
120
150
180
α°


Φ = 120°
R_L = 1.61
Ψ = 1.20
0.25
0.20
0.15
0.10
0.05
0
f(α)
0
30
60
90
120
150
180
α°

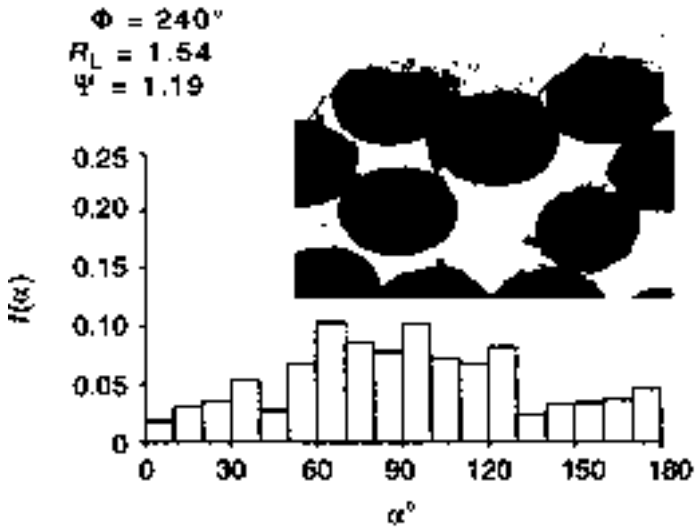

Φ = 240°
R_L = 1.54
Ψ = 1.19
0.25
0.20
0.15
0.10
0.05
0
f(α)
0
30
60
90
120
150
180
α°

Fig. 8 Profile orientation distribution functions, R_L, and profile structure factors & for fracture profiles observed in three vertical sections of the tensile fracture surface of a metal-matrix composite (MMC) generated by uniaxial loading perpendicular to the fibers. The three vertical sections are mutually at an angle of 120° to one another, and all three sections contain the vertical axis. The MMC contains aligned alumina fibers in an aluminum alloy matrix. Source: Ref 22

Extent of Overlaps in the Fracture Profile. Fracture surfaces, and therefore, fracture profiles may contain reentrant segments or overlaps. Figure 9 depicts a schematic fracture profile containing overlapped (or reentrant) segments. Figure 10(a) shows the fracture profile from the fracture surface of copper that failed in creep, whereas Fig. 10(b) shows a typical room-temperature tensile fracture profile from the same material (Ref 4, 76). Observe that the creep fracture profile contains significant overlapped/reentrant segments, whereas the tensile fracture profile contains no overlaps. The overlaps are obviously not observed in SEM images; profilometry is an important experimental technique to observe and quantify the fracture surface overlaps. The extent of overlaps in the fracture profile is quantified by the profile overlap parameter O_P, which is simply equal to the fraction of the profile length containing overlapped or reentrant segments. The profile overlap parameter can be computed via interactive image analysis, where the operator identifies the overlapped segments using an electronic mouse. The lengths of the overlapped segments are then measured by using the digitized coordinate data. The angular orientation distribution of the overlapped segments and their roughness can be also computed from the same data.

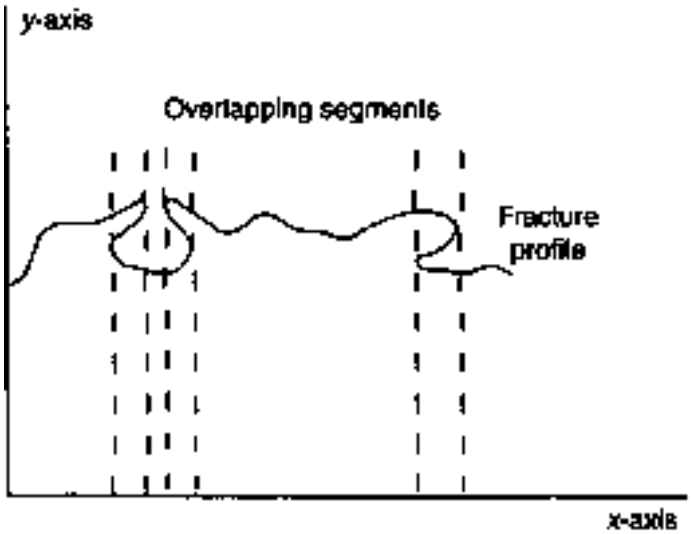


Fig. 9 A schematic fracture profile containing overlapped (or reentrant) segments

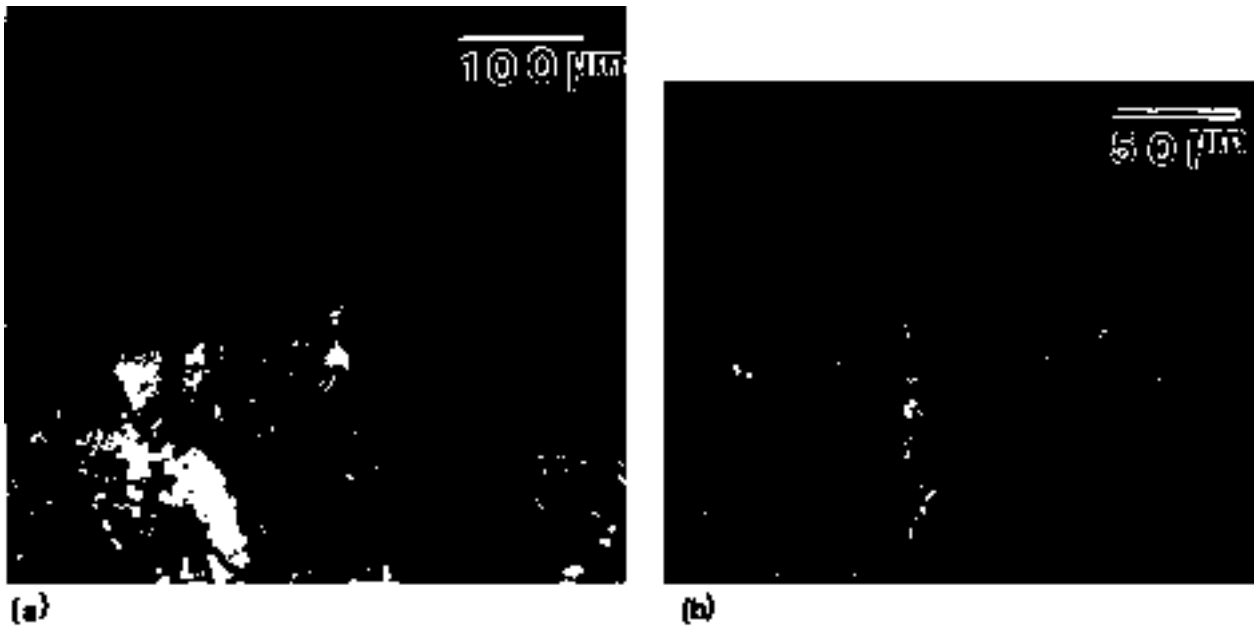


Fig. 10 Fracture surface overlaps. (a) Overlapped/reentrant segments in the fracture profile from the fracture surface of a copper specimen that failed in creep. Source: Ref 4, 76 (b) A typical room temperature tensile fracture profile of copper does not show overlaps.

Correlation between Fracture Profile Path and Microstructure. An important advantage of fracture profilometry using metallographic sections is that the fracture profile and underlying microstructure can be observed simultaneously, which is useful for characterizing correlations between the microstructure and the fracture path. A simple parameter that quantifies such a correlation is the fraction of fracture profile length through a phase or constituent of interest, P_f. In one study (Ref 32), the fraction of fracture profile length through Si particles was observed to be 0.16 in the vertical section fracture profile of a tensile fracture surface of chill cast A356 aluminum alloy having a volume fraction of Si particle equal to 0.074. In this case, as P_f is significantly higher than the silicon particle volume fraction, the fracture profile appears to preferentially traverse through the silicon particles present in the interdendritic eutectic regions of the cast alloy. Another profile parameter, $P(k)$, represents the probability of a microstructural constituent i being associated with a fracture mode k. The probability parameter is defined as (Ref 77):

$$P_i(k) = \frac{\sum (\lambda_i)_k}{\sum (\lambda_i)_{profile}} \qquad \text{(Eq 3)}$$

where the denominator denotes the total profile length through constituent i, and the numerator represents only that portion of the profile length through constituent i that is associated with fracture mode k. Three-dimensional analogs to the preceding linear parameters are also proposed. The index k for fracture modes is assigned according to $k = 0$, cleavage fracture; $k = 1$, quasi-cleavage fracture; $k = 2$, quasi-dimple fracture; and $k = 3$, dimple fracture. A linear fracture path preference index, Q_i, has also been proposed for each microstructural constituent i, according to (Ref 77):

$$Q_i = \frac{\sum (\lambda_i)_{profile}}{\sum (L_i)_{||}} \qquad \text{(Eq 4)}$$

where $\%(2_i)_{profile}$ is the total length of the fracture profile that runs through the ith constituent, and $\%(L_i)_{||}$ is the total intercept length through the microstructure across the ith constituent and along a straight test line parallel to the effective fracture direction. This parameter can be recognized as a form of the lineal profile roughness parameter R_L for the corresponding phase or constituent.

Fractal Characteristics of Fracture Profiles. It is well known that the lengths of many naturally occurring irregular lines such as coastlines, mountain ranges, fracture profiles, and so forth depend on the size of the measuring unit or the yardstick used for the length measurements. This concept is illustrated in Fig. 11, which shows that the measured length $2(0)$ of the irregular curve increases as the size of the measuring unit 0 decreases. Mandelbrot (Ref 79) has given the following relationship between the measured length $2(0)$ and the yardstick length 0:

$$2(0) = 2_0[0]^{(1-D)} \qquad \text{(Eq 5)}$$

$$R^2 = \left(\frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{xy}^2 \qquad \text{(Eq 17)}$$

The result is that stress transformations can be performed by using the geometric principles of a circle. For example, if stresses are known at a point, they are plotted on a figure that has horizontal axis σ and vertical axis τ so that (σ_x, τ_{xy}) is the coordinate point at horizontal position σ_x and vertical position τ_{xy} and so that (σ_y, τ_{xy}) is the coordinate that corresponds to σ_y and τ_{xy}. Since the center of the circle is on the σ axis, this can be easily found by Eq 16. The radius of the circle is given by Eq 17. From this, the entire circle can be drawn. This is illustrated by the schematic Mohr's circle in Fig. 5.

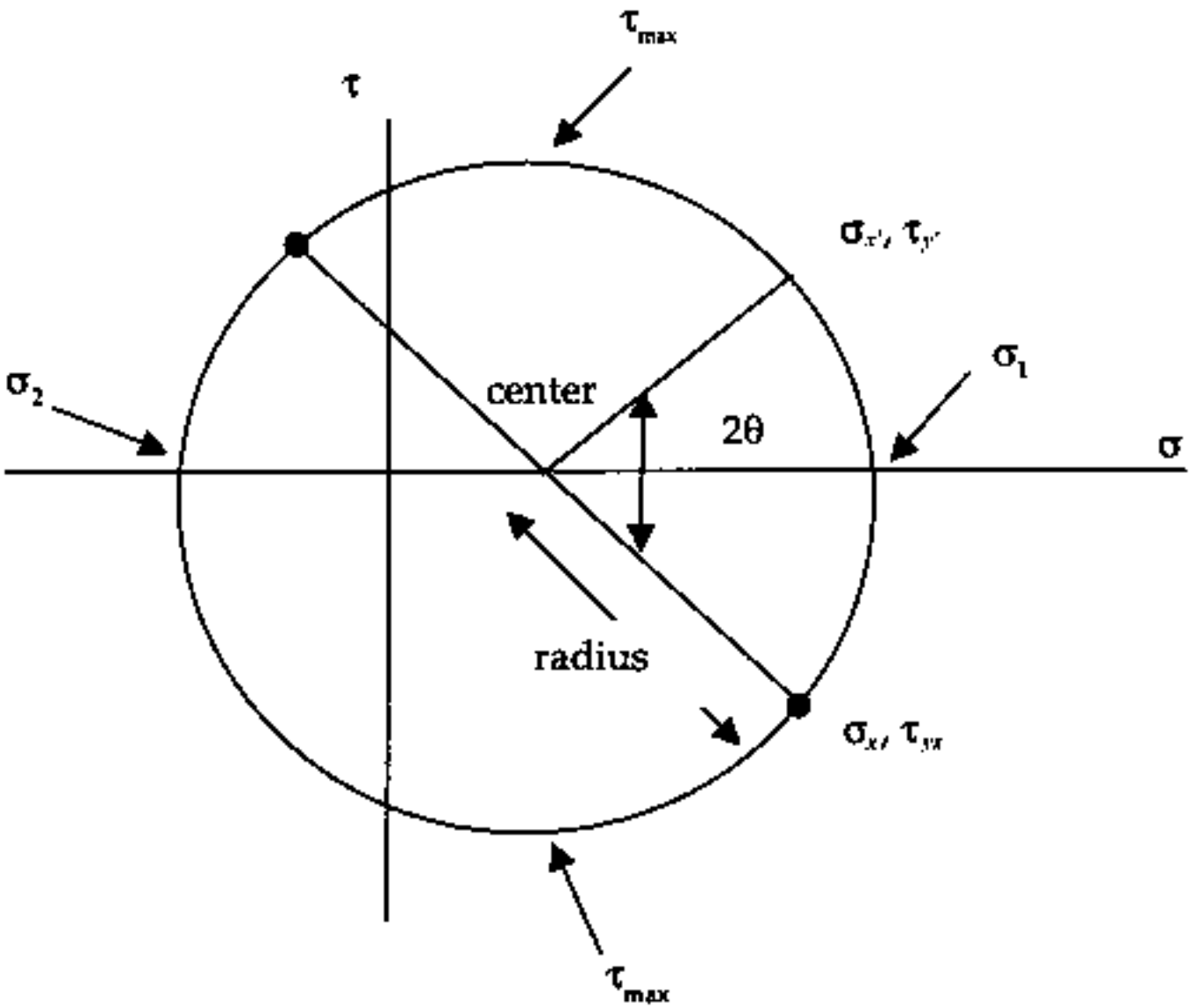


Fig. 5 Mohr's circle for two-dimensional stress transformation

Then for a rotation of axes by θ in the stress element, the position on the Mohr's circle must go through a rotation of 2θ in the same direction but around the circumference of the circle. Every point on the circle corresponds to a possible stress pair. Note that a rotation of 180° of the circle corresponds to a rotation of the stress element by 90° so that σ_x is transformed to σ_y as shown in Fig. 5.

Principal Stresses. The purpose of a stress transformation is mainly to find the stresses that can be used in a failure criterion. These would be either the largest magnitude stresses in any direction or the magnitude of stresses on a weak plane. For the former, the extreme stresses can be found by taking Eq 11 and applying the calculus principle $d\sigma_{X'}/d\theta = 0$. That is, principal stresses are normal (perpendicular) stresses on planes for which the shear stresses are zero. The values of principal stresses are given by:

Consider radial loading of a cylindrical pin in a hole. First, look at the case in which significant clearance exists between the pin and the hole. In that event, a near-Hertzian stress state may exist beneath the pin, given sufficient material between the hole and an edge. However, if the hole is located near an edge and the loading is perpendicular (or nearly so) to that edge, then a large tensile stress exists at that edge beneath the hole. The magnitude of this tensile stress is strongly influenced by the stress state below the contact. Now examine the other extreme: little or no clearance between the pin and hole. In this event, Hertzian theory cannot be employed to develop the near-field subsurface stress state. Due to better load distribution, the edge tensile stress now will be significantly lower, as will the maximum subsurface shear stress. These two cases define the envelope in which pin/hole contact occurs. Differences between the two edge tensile stresses could approach 100%. Again this serves to highlight the need for proper stress analysis.
For additional information on complex stresses in components, see the article "Overview of Mechanical Properties and Testing for Design" in *Mechanical Testing and Evaluation,* Volume 8 of *ASM Handbook.*

References cited in this section

1. J.E. Shigley and C.R. Mischke, *Mechanical Engineering Design,* 5th ed., McGraw-Hill, 1989
2. W.K. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, Inc., 1997
3. W.C. Young, *Roark's Formulas for Stress and Strain,* 6th ed., McGraw-Hill, 1989
4. A. Blake, *Practical Stress Analysis in Engineering Design,* 2nd ed., Marcel Dekker, 1990

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Stress Analysis Of Common Geometries

The stress analysis of common geometries can be made with either a mechanics of material approach or a theory of elasticity approach. The mechanics of materials approach follows four steps to arrive at a solution; these steps are taken in a prescribed sequence. The mechanics of materials always proceeds as follows:

1. Define the geometry of deformation.
2. Define strain from the geometry of deformation.
3. Define stress from this strain through the stress-strain relationships.
4. Relate the stresses to the boundary loadings through equilibrium.

Following these four steps always leads to a solution. In the procedure, the first step is the most difficult, defining the geometry of deformation. This has been done in the past for most common geometries by applied mechanics theoreticians who studied deformation models. In developing the mechanics of materials solutions here, the author uses these classical deformation patterns invents rather than new ones.
The theory of elasticity method requires that a number of conditions be satisfied simultaneously, namely, the body must satisfy internally a set of equilibrium equations and a set of compatibility equations. The equilibrium equations ensure that all parts of the body are in equilibrium under the stress field, and the compatibility equations ensure that the displacements are single-valued and possible. The compatibility equations derive from strain-displacement equations. The theory of elasticity always gives an exact set of equations for the solution to a problem. However, the equations are often a partial differential equation that cannot be easily solved and/or which must be solved numerically.
Therefore, the common elements that are presented in the following sections for stress analysis follow the mechanics of materials method for solution. Examples presented include the tensile geometry, torsion geometry, bending and beam shear, pin shear, thin-walled pressure vessels, curved beams, and springs.

Simple Shaft or Bar

Fig. 14 Pin and clevis loading. (a) Pin going through loaded clevis and plate. (b) Shear planes in a clevis and pin connected to a threaded eye bolt

The pin often is in double shear as it is in this case. That means that it is stressed on both sides of the center. Then a load P is applied to the middle of the pin, and the internal shear force $V = P/2$ is resisted on either side of the central load. Given a pin diameter, d, and the pin shear stress, which is usually taken as the average stress given by:

$$\tau = \frac{V}{A} = \frac{P/2}{(\pi/4)d^2} = \frac{2P}{\pi d^2} \qquad \text{(Eq 62)}$$

The pin could be analyzed as a beam to get a maximum shear stress at the neutral axis but that is not commonly done. The clevis or plate with a pinhole can also be analyzed. The stress from the pinhole loading is not large enough to cause a problem unless the pinhole is too close to an edge, in which case it could be critical. The pinhole resists a shear load over the area that goes from the middle of the pinhole to the plate edge along a distance d (Fig. 15). That distance could go from the middle of the pinhole to the surface of the clevis as shown on the right; however, it is often advised to take d from the top of the pinhole to the edge of the clevis as shown on the left, here labeled d_{min}. That provides a conservative calculation of the stress. If the d_{min} dimension is about the same as the pin diameter, the design is usually okay. The stress then is:

$$\sigma = \frac{P}{2td_{min}} \qquad \text{(Eq 63)}$$

where t is the thickness of the plate. The stress concentration associated with the hole is not taken to influence the area defined by d_{min}. Here P is the total load applied to the one side of the clevis, as shown in Fig. 15, and not the load applied to the pin, as shown in Fig. 14.

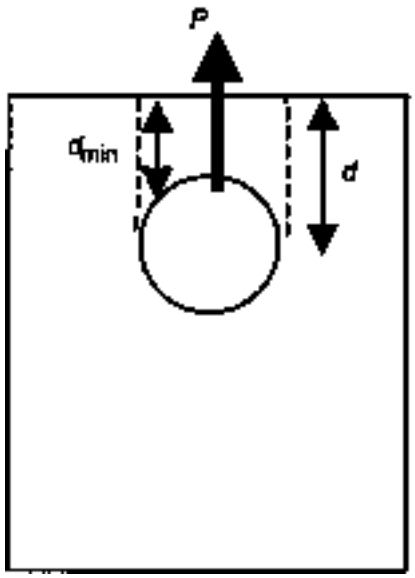


Fig. 15 Distance *d* for stress evaluation of a pin and clevis

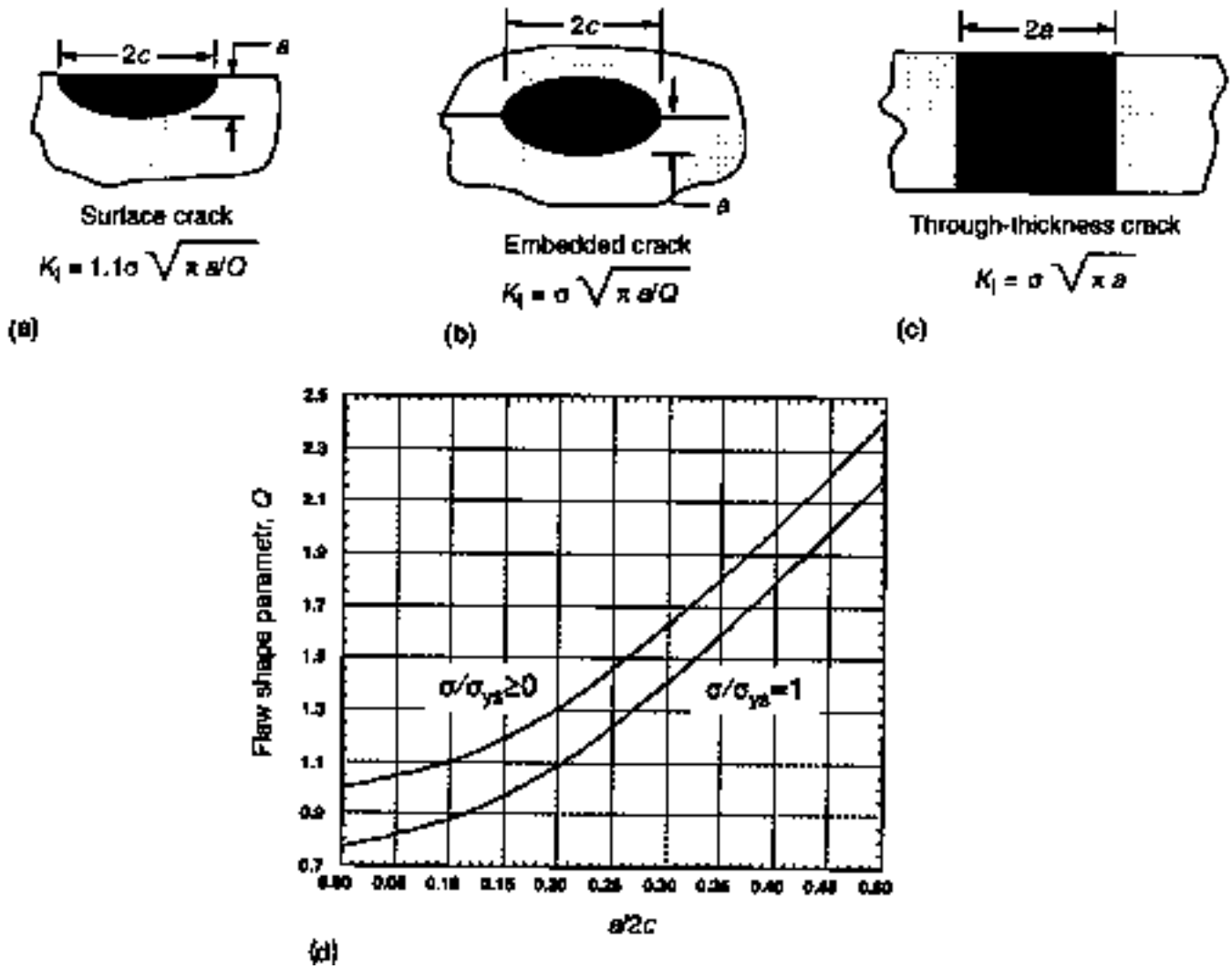
Spring Stresses. Spring stress analysis can be done for two types of springs, the linear helical spring and the helical torsion spring. The linear helical spring has a coil of wire and is wound in a round helical shape. It is loaded along the axis of the spring (Fig. 16). In this case the spring wire is round as well. In Fig. 16, D is the diameter of the coil taken at midwire, d is the wire diameter for a round wire spring, and C is the spring index (D/d).

The constant, C, depends on various factors including thickness and deformation characteristics of the material. It usually has a value between $1/\pi$ and $\frac{1}{3}\pi$. This equation is only valid for linear elastic bodies; that is, when R_P is much smaller than the crack size (i.e., R_p < 10% of the crack size) and other physical dimensions of the body (Ref 5). As K increases (e.g., by increasing the load) R_p also increases, and the material in the plastic enclave becomes more severely strained. This plastic enclave is very important in metals and polymers since it is the region in which energy is absorbed and is responsible for the relatively high fracture toughness of metals compared to all other materials.

As a result of the direct relationship between K and the strains at the crack tip, a critical value of K can be defined for when fracture takes place. When the region of plastic deformation around a crack is small compared to the size of the crack (as is often true for large structures and high-strength materials), the magnitude of the stress field (in tension) around the crack is thus related to the stress-intensity factor, K, as follows:

$$K = Y\sigma\sqrt{\pi a}$$

where Y is a geometric factor, σ is the gross stress across the fracture plane, and a is the crack length. In this way, linear elastic analysis of small-scale yielding can be used to define a unique factor, K, that is proportional to the local crack-tip stress field outside the small crack-tip plastic zone. The stress-intensity factor can have a simple relation to applied stress and crack length, or the relation can involve complex geometry factors for complex loading, various configurations of real structural components, or variations in crack shapes. For example, the geometric factor, Y, is typically on the order of 1 for a tension-loaded plate with a through-thickness crack (Fig. 21). Geometric factors (expressed as a flaw shape parameter, Q) for surface cracks and embedded cracks are also shown in Fig. 21 (Ref 6). Specific formulas for K depend on the load/crack geometry, and numerous solutions for K can be found in Ref 7 and in the article "Summary of Stress-Intensity Factors" in *Fatigue and Fracture*, Volume 19 of *ASM Handbook* (Ref 8).



9. A. Liu, Chapter 7, Fracture Mechanics for Mixed Crack Tip Displacement Modes, *Structural Life Assessment Methods*, ASM International, 1998

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Subcritical Fracture Mechanics

Fracture mechanics analysis dealing with cracking (as opposed to fracture) is sometimes referred to as subcritical fracture mechanics. The final fracture is by overload. In subcritical cracking, the amount of plastic deformation during cracking is very small, and so LEFM concepts are typically valid for crack-growth analysis. An exception is creep cracking, for which EPFM concepts must be used. Creep crack growth is discussed in Ref 10.

Fatigue. In the case of cyclic loading, subcritical crack growth from fatigue loading, da/dN, is characterized by the stress-intensity range, $\Delta K = K_{max} - K_{min}$. During a load cycle, K varies from a minimum and maximum:

$$K_{min} = Y\sigma_{min}\sqrt{\pi a}$$

$$K_{max} = Y\sigma_{max}\sqrt{\pi a}$$

where the stress ratio is $R = K_{min}/K_{max}$. The crack growth rate, da/dN, depends on ΔK, and R and must be provided by a test of the material. Crack growth rates versus ΔK typically have three distinct regions of behavior, as discussed in more detail in the article "Fatigue Failures" in this Volume. Region 1 is the stage of early crack growth with a threshold value (ΔK_{th}) below which cracks rates become diminishingly small. If values of ΔK are above ΔK_{th}, the crack size grows and thus causes the stress intensity to increase at the crack tip. The crack growth rates accelerate during stage 1 until they reach stage 2, where rates are described by the Paris equation:

$$da/dN = C(\Delta K)^n$$

where C and n are empirical constants, which depend on the stress ratio. In practice, fracture mechanics testing is typically done at a small positive R value (n, $R \sim 0.1$), while service conditions range from -1 to +1 for R. Because the Paris equation covers only one R value, an alternate form is:

$$\frac{da}{dN} = \frac{C}{(1-R)^m}\Delta K^n = C\frac{\Delta K^m}{(1-R)^m}\Delta K^{n-m}$$

This is known as the Walker equation. Many other equations are used, but it should be pointed out that the equations are phenomenological curve-fitting equations. Conditions, especially environments, can alter the shape of the curves. For example, Fig. 24 shows three cases of environmental effects:

- *True corrosion fatigue* (TCF): Fig. 24(a). Material-environmental combination where the threshold value is lowered and the chemical reaction is enhanced by the plastic behavior at the crack tip under fatigue loading. The chemical reaction is not influenced by the static stress at the crack tip.
- *Stress corrosion fatigue* (SCF): Fig. 24(b). Material-environmental combination where the chemical reaction is strongly influenced by the static stress at the crack tip (i.e., stress corrosion is observed) but is not influenced by the plastic behavior at the crack tip as a result of fatigue loading.
- *TCF combined with SCF:* Fig. 24(c)

The latter is the most common.

$$
\mathbf{T}' = \begin{bmatrix} 0.866 & 0 & -0.500 \\ 0 & 1 & 0 \\ 0.500 & 0 & 0.866 \end{bmatrix} \begin{bmatrix} 10 & 4 & 8 \\ 4 & 20 & 12 \\ 8 & 12 & -5 \end{bmatrix}
$$

$$
\times \begin{bmatrix} 0.866 & 0 & 0.500 \\ 0 & 1 & 0 \\ -0.500 & 0 & 0.866 \end{bmatrix}
$$

$$
= \begin{bmatrix} -0.678 & -2.536 & 10.495 \\ -2.536 & 20 & 12.392 \\ 10.495 & 12.392 & 5.678 \end{bmatrix}
$$

If another coordinate transformation (into the double primed system) is necessary, then a second matrix of direction cosines, **B**, is required. If **B** represents the orientation between the double primed system and the primed system, then the transformation equation becomes:

$$
\mathbf{T}'' = \mathbf{BATA}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}
$$

The transformations **A** and **B** can be combined by defining **C** = **BA**. In that case

$$
\mathbf{T}'' = \mathbf{CTC}^{\mathrm{T}}
$$

which is in the same form as the single transformation.

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

References

1. J.E. Shigley and C.R. Mischke, *Mechanical Engineering Design,* 5th ed., McGraw-Hill, 1989
2. W.K. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, Inc., 1997
3. W.C. Young, *Roark's Formulas for Stress and Strain,* 6th ed., McGraw-Hill, 1989
4. A. Blake, *Practical Stress Analysis in Engineering Design,* 2nd ed., Marcel Dekker, 1990
5. S. Antolovich and B. Antolovich, An Introduction to Fracture Mechanics, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 371–380
6. H. Nelson, Fractography in the Failure Analysis of Engineering Structures, seminar presented at ASM Annual Meeting, 10 Sept 1997
7. H. Tada, P.C. Paris, and G. Irwin, *The Stress Analysis of Cracks Handbook,* 3rd ed., American Society of Mechanical Engineers, 2000
8. A. Liu, Summary of Stress-Intensity Factors, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 980–1000
9. A. Liu, Chapter 7, Fracture Mechanics for Mixed Crack Tip Displacement Modes, *Structural Life Assessment Methods,* ASM International, 1998

beyond the scope of this article, XRD techniques can also be used for the measurement of percent retained austenite (Ref 19), and other phases, pole figures (Ref 7), and other analyses too numerous to mention.

References cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

2. V. Hauk, *Structural and Residual Stress Analysis by Nondestructive Methods,* Elsevier, 1997

3. M.E. Hilley et al., "Residual Stress Measurement by X-Ray Diffraction" J784a, Society of Automotive Engineers, 1971

4. P.S. Prevey, X-Ray Diffraction Residual Stress Techniques, *Materials Characterization,* Vol 10, *ASM Handbook,* American Society for Metals, 1986, p 380–392

5. J. Lu, et al., *Handbook of Measurement of Residual Stress,* Fairmont Press, 1996

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* 2nd ed., Wiley-Interscience, 1974

8. C.O. Ruud, Residual Stress Measurements, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000

9. H.H. Lester, and R.H. Aborn, *Army Ordnance,* Vol 6, 1925–1926, p 129, 200, 283 and 364

10. C.O. Ruud, P.S. DiMascio, and D.J. Snoha, A Miniature Instrument for Residual Stress Measurement, *Advances in X-Ray Analysis,* Vol 27, Plenum, 1984

11. H. Berger, Sensing Stress, *Research/Penn State,* Vol 10 (No. 2), 1989

12. M.E. Brauss, G.V. Gorveatte, and J.F. Porter, Development of a Miniature X-Ray Diffraction Based Stress Analysis System Suitable for Use on Marine Structures, *Nondestructive Evaluation of Materials and Composites—SPIE,* Vol 2944, 1996

13. J.A. Pineault, and M.E. Brauss, Stress Mapping—A New Way of Tackling the Characterization of Residual Stress, *Exp. Tech.,* March 1995

14. J.A. Pineault, M.E. Brauss, and J.F. Porter, Characterization of Plastic Strain in HY-80 Using X-Ray Diffraction Techniques, *Proceedings of the Conference on Naval Applications of Material Technology,* (Halifax), D.R.E.A., 1993

15. M.G. Carfagno et al., *X-Ray Diffraction Measurement of Stresses in Post-Tensioning Tendons: Extending the Lifespan of Structures,* IABSE, Zurich, Switzerland, 1995

16. M.E. Brauss and J.A. Pineault, Residential Strain Measurements of Steel Structures, *NDE for the Energy Industry,* NDE-Vol 13, American Society of Mechanical Engineers, 1995

17. M.E. Brauss et al., "Deadload Stress Measurements on Brooklyn Bridge Wrought Iron Eye Bars and Truss Sections Using X-Ray Diffraction Techniques: Proceedings of the 14th Annual Meeting of the International Bridge Conference," Engineers' Society of Western PA, 1997

The examples shown in Fig. 3 demonstrate that an elliptical fit gives good results (Fig. 3a), whereas a linear fit using either the $\psi > 0$ or the $\psi < 0$ branch leads to incorrect results (Fig. 3b). In this case, the results from the two opening branches of the curve differ by approximately 54 ksi.

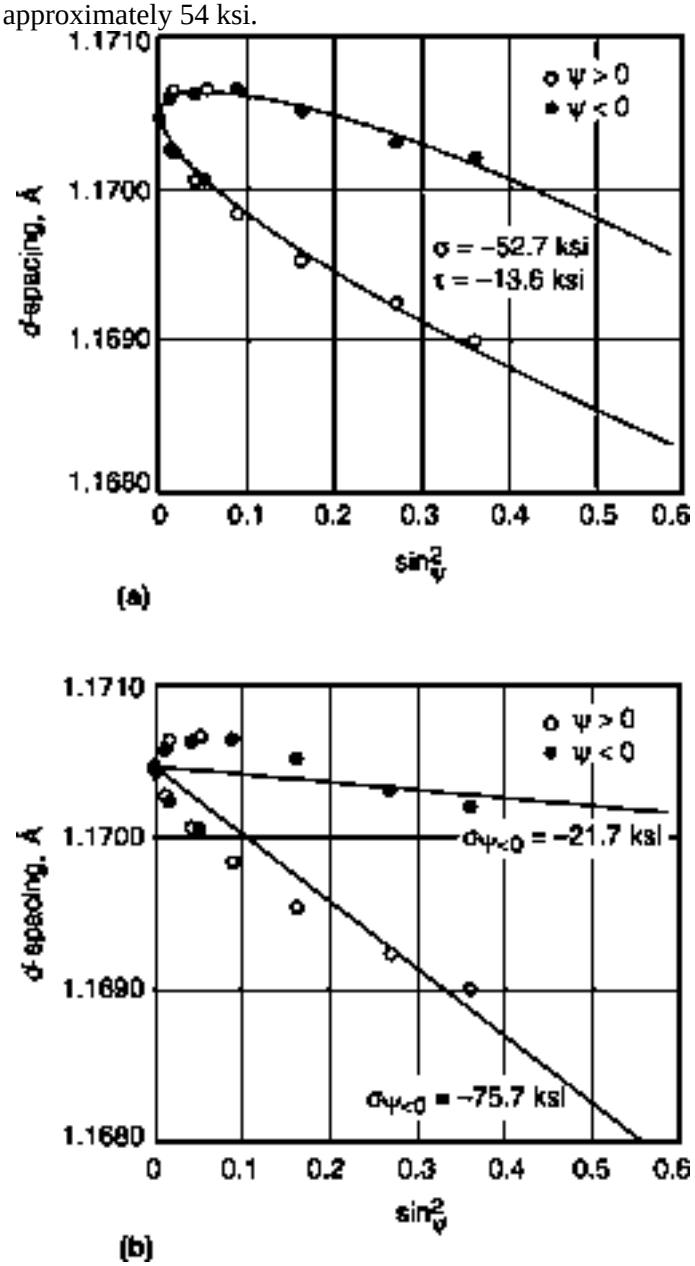


Fig. 3 ψ-splitting on steel using (a) elliptical fit and (b) linear fit for ψ > 0 and ψ < 0

With recent advances in equipment speed, the full XRD peak profile can be obtained quickly and easily. When using instrumentation having single-channel detectors, the full diffraction profile should always be collected. Equipment that allows the use of only two ψ tilts (typically, $\psi = 0°$ and $\psi = 45°$) has limited applications. The use of only a few ψ tilts can lead to erroneous results; thus, the double-exposure method with $\psi > 0$ or $\psi < 0$ is only viable in special cases and when it is certain that the *d*-spacing versus $\sin^2\psi$ plot is linear. This means that, to ensure accurate results on a wide variety of samples and measurement locations, many ψ tilts using both $\psi > 0$ and $\psi < 0$ and elliptical fitting methods should be applied. The use of true ψ or χ (chi) geometry, sometimes referred to as Ω (omega) and ψ, respectively, is recommended. The use of χ goniometer geometry has advantages, primarily a slightly increased tilt range where the ψ tilt range is limited due to sample geometry constraints. However, when using the modified χ (or side-inclination) geometry, errors may be introduced due to the ψ angle offset and therefore, should only be used when the specimen geometry eliminates the preferred methods (i.e., true ψ or χ). True-χ and modified-χ geometry are more sensitive to goniometer alignment, particularly with small goniometer radii; thus, care must be taken in sample and goniometer alignment.

References cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

A summary in Fig. 7 shows simplified, 2D free-body diagrams illustrating the orientations of the normal-stress and shear-stress systems at any internal point in a shaft loaded in pure tension, torsion, and compression. The single-overload fracture behavior of both ductile and brittle materials also is illustrated for each type of load.

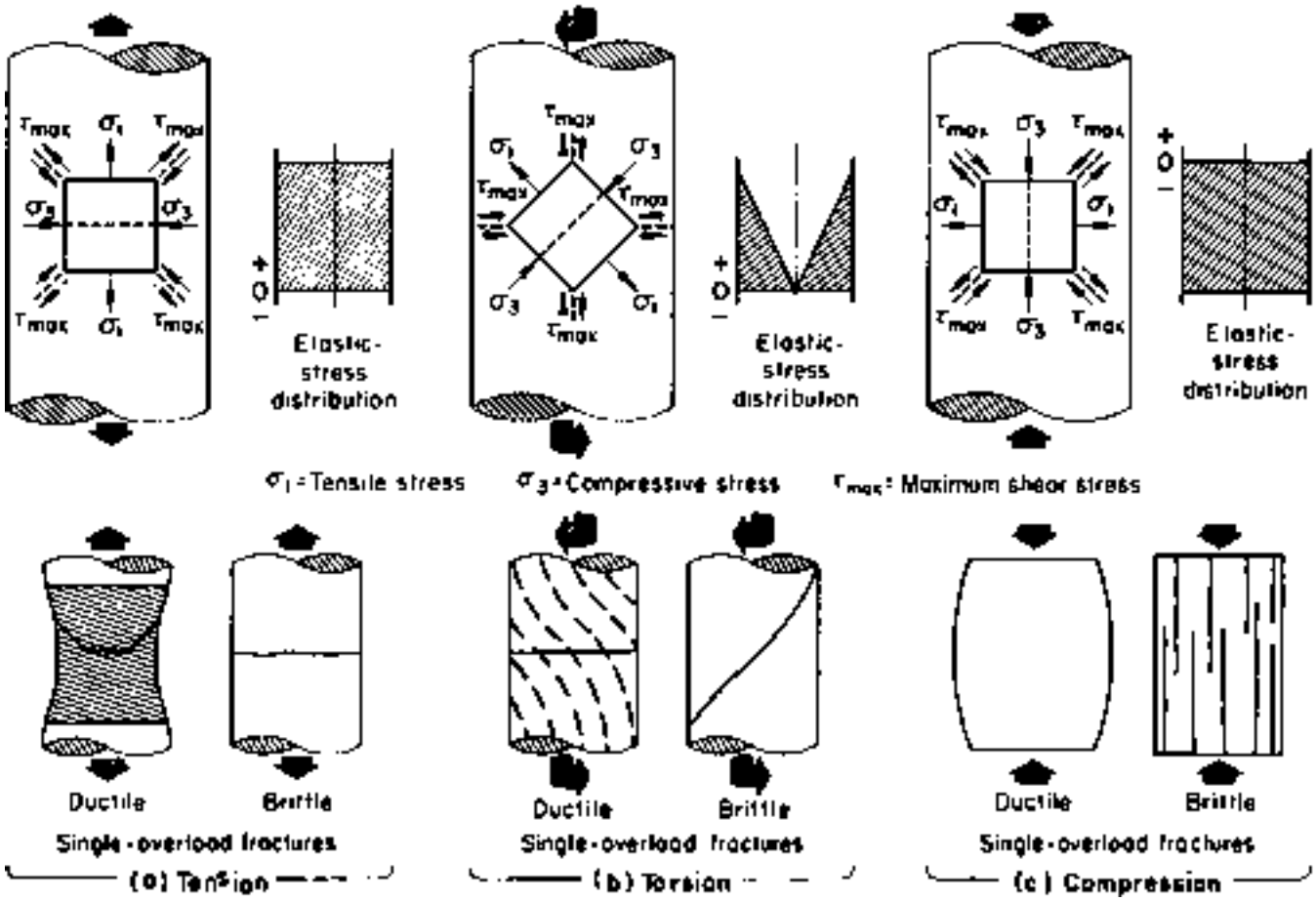


Fig. 7 Free-body diagrams showing orientation of normal stresses and shear stresses in a shaft and the single-overload fracture behavior of ductile and brittle materials. (a) Under simple tension. (b) Under torsion. (c) Under compression loading

Tensile Bars. The tensile bar has an axis, *x*, that goes through the centroid of a cross-sectional area. The area is often constant along the *x* axis but need not be. Tensile bars also are often round but need not be. Any shape of a tensile bar is permissible, but it must have a cross-sectional area with a centroid that lies on the load line. Assume that the load, *P*, is directed along the axis, that is, through the centroid of the area. Off-center loads, those not along a centroid, result in more complex loading and would not be analyzed as simple axial loading.

Step 1 in the mechanics of materials assumes that the deformation mode of the bar is uniform elongation, that is, the strain is averaged over the entire length (*L*) of the body. Step 2 gives the strain as the stretch, δ, divided by the total length of the bar, *L*. The only nonzero stress is σ_x, which lies along the *x* axis so that step 3 gives the stress from the stress-strain equation as $\sigma_x = E\varepsilon_x$. The equilibrium condition from step 4 gives the relationship between the stress, σ, and the external loading, *P*, as load equals the stress times area (*A*). These steps result in following equations:

$$\sigma = \frac{P}{A} \qquad \sigma = E\varepsilon \qquad \varepsilon = \frac{\delta}{L}$$
$$\delta = \int \frac{P dx}{A E} = \frac{PL}{A E} \qquad \text{(Eq 42)}$$

The strain in the above equation should be indicated as the strain in the direction of load (e.g., ε_y)

The second result for the deflection, δ, is correct when the load, *P*; area, *A*; and modulus, *E* do not vary along the *x*-axis. For simple geometries like the tension bar, there is only one stress component. As discussed in preceding sections, there can be transverse strains even though no stress is applied in these directions.

Under tension loading, the tensile stresses (σ_1 in Fig. 7) are longitudinal, whereas the compressive stress components (σ_3 in Fig. 7) are transverse. The stresses created by tensile loading of cylindrical or approximately square specimens cause considerable deformation (elongation and necking) before fracture, which originates near the center of the shaft within a

18. M.E. Brauss, J.A. Pineault, M. Belassel, and S.I. Teodoropol, Nondestructive, Quantitative Stress Characterization of Wire Rope and Steel Cables, *Proceedings of SPIE, Structural Materials Technology,* Vol 3400, Society of Photo-Optical Instrumentation Engineers, 1998

19. C.F. Jatczak et al., "Retained Austenite and Its Measurement by X-Ray Diffraction," SP-453, Society of Automotive Engineers, 1980

Residual Stress in Failures and XRD Analysis

The residual stresses present in a component can arise in almost every step of processing (Ref 1) and, in many cases, can play a key role in determining the effective service life of a component. When the failure analyst suspects residual stress as a contributing factor to a premature failure, such suspicions may be validated by experiment and by measurement. X-ray diffraction residual stress measurement analysis should be specified when damaged components:

- Experience static in service loads that are a significant fraction of the maximum allowable equivalent stress (stress concentrations, overloaded)
- Distort or form cracks without applied loads
- Are placed in corrosive environments (stress corrosion)
- Experience cyclic loading in service (mechanical or thermal fatigue)
- May have been subject to improper processing on manufacture (shot peening, grinding, milling, etc.)
- May have been subject to inappropriate heat treatment (stress relief, induction hardening, service temperature, thermal strains, etc.)

Reference cited in this section

1. I.C. Noyan and J.B. Cohen, *Residual Stress: Measurement by Diffraction and Interpretation,* Springer-Verlag, 1987

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

X-Ray Diffraction Theory and Residual Stress Measurement

The XRD technique uses the distance between crystallographic planes, that is, *d*-spacing, as a strain gage. This method can only be applied to crystalline, polycrystalline, and semicrystalline materials (Ref 1, 20), where thousands of grains are sampled in a typical measurement. When the material is in tension, the *d*-spacing increases in the direction of stress, and when the material is in compression, the *d*-spacing decreases. The presence of residual stresses in the material produces a shift in the XRD peak angular position that is directly measured by the detector (Ref 6). For a known x-ray wavelength, λ, and n equal to unity, the diffraction angle, 2θ, is measured experimentally, and the *d*-spacing is then calculated using Bragg's law:

Dislocations are typically introduced into materials by cold working. The formation of small crystallites (coherent domains) and the introduction of elastic microstrains both result in broadening of the XRD line profile (Ref 7). Line broadening is also a function of instrumental broadening, that is, the XRD instrument optics.

Despite trends in macrostress (Fig. 28), a trend of decreased hardness with increased heat treatment temperature is often observed in iron alloys sample sets (Fig. 29a). An associated decrease in the XRD peak breadth is also often observed (Fig. 29b). In this example, the range of heat treatment temperatures remained below the annealing temperature for each alloy; thus the instrumental and particle size contributions to the broadening remained more or less constant. If the heat treatment temperature exceeds the annealing temperature, the observed XRD peak breadth may be reduced either by recovery, recrystallization, or grain growth (Ref 6). Heat treatment with slow cooling has the effect of relieving or reducing microstresses as well as decreasing the dislocation density, thus decreasing the hardness.

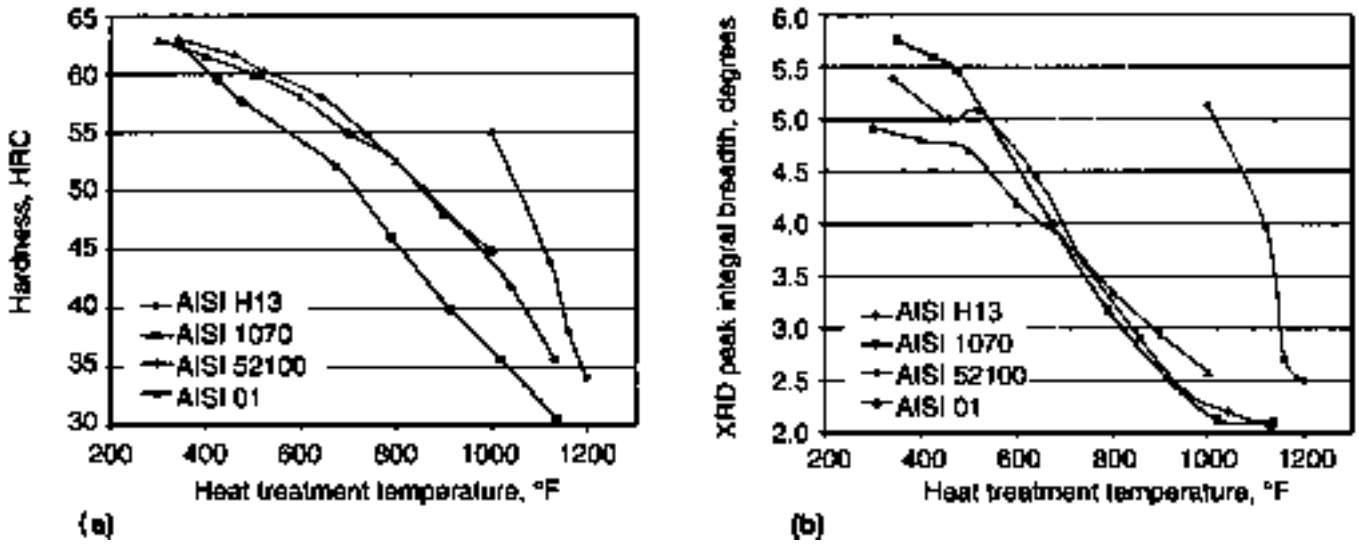


Fig. 29 Effect of heat treatment temperature on (a) hardness (HRC) and (b) XRD peak integral breadth.

The effects of these mechanisms can thus be observed in the XRD peak breadth. In general, this is used as a trending tool, and the data are most easily used in empirical form to assess the relative hardness, cold working, or dislocation density in a given material.

More detailed analyses of the source of XRD line broadening can be performed using Fourier space methods, such as the Warren-Averbach method (Ref 54), or real space methods, such as Voigt deconvolution (Ref 55). The XRD peak integral breadth and/or full width at half maximum is generally characterized as a function of depth with the use of electropolishing techniques in parallel with the measurement of residual macrostresses.

References cited in this section

6. B.D. Cullity, *Elements of X-Ray Diffraction,* 2nd ed., Addison-Wesley, 1978

7. H.P. Klug and L.E. Alexander, *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* 2nd ed., Wiley-Interscience, 1974

54. B.E. Warren, *X-Ray Diffraction,* Addison-Wesley, 1969

55. J.I. Langford, A Rapid Method for Analyzing the Breadths of Diffraction and Spectral Lines Using the Voigt Function, *J. Appl. Crystallogr.,* Vol 11, 1978

important variable. Because this alloy has substantial hardenability, it is difficult to prevent formation of bainite and martensite in small, as-rolled section sizes. In this case, the microstructure consisted of ferrite, bainite, and martensite (the arrows point to microcracks present in martensite patches).

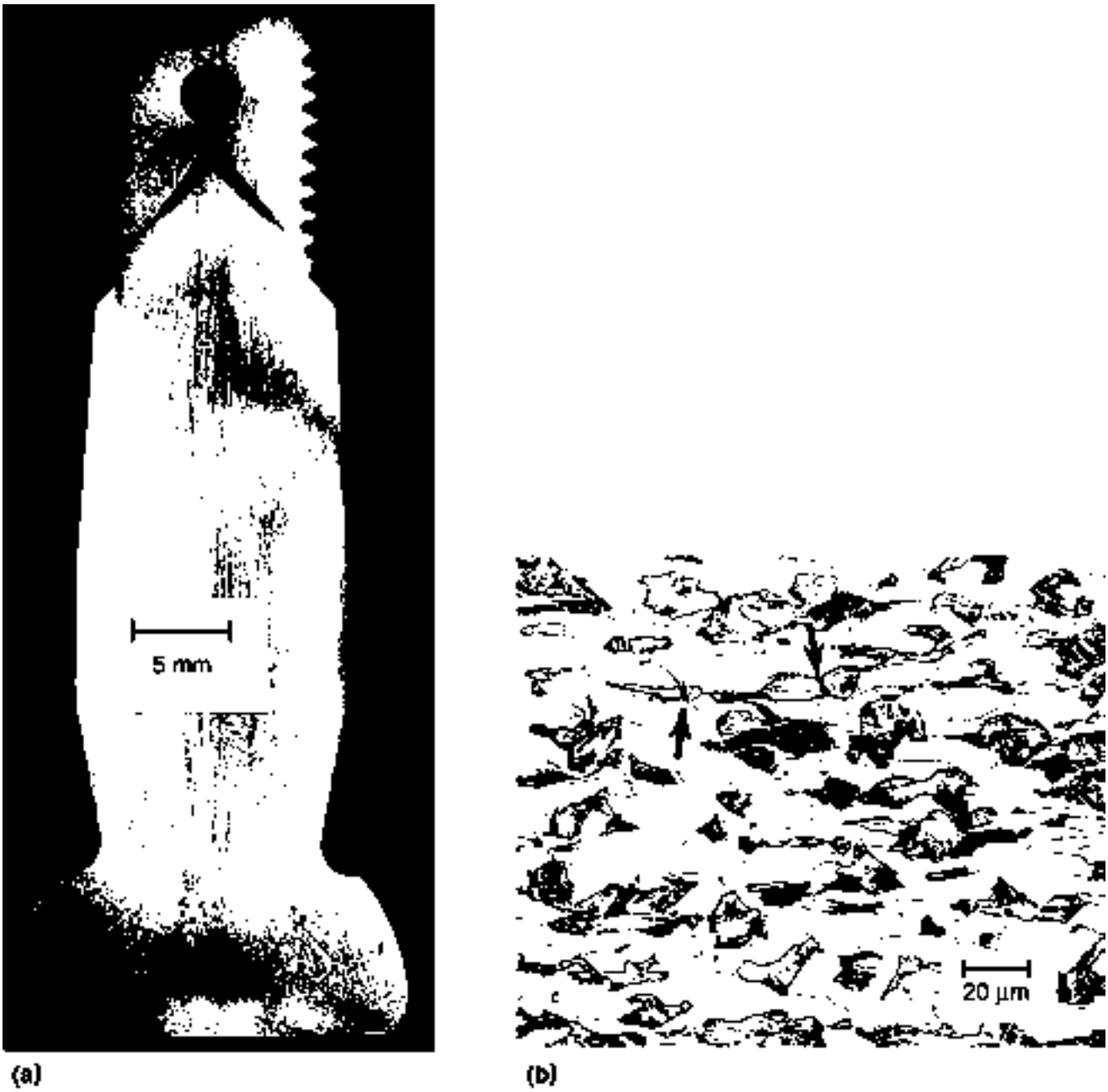


Fig. 36 (a) Central bursting during extruding of AISI 4615 alloy steel specimen was promoted by (b) the presence of bainite and martensite (arrows point to microcracks in the martensite) in the as-rolled stock. Specimen etched with 4% picral, followed by 2% nital

Final grinding of hardened tool steel components is an important processing step that must be carefully controlled. Improper grinding practices can produce a fine network of surface cracking and a characteristic scorch pattern easily revealed by macroetching (Ref 31, 32). Examination of the microstructure at the cracks reveals their shallow nature and a back-tempered condition at the surface, as shown in Fig. 37. In some cases, a shallow, reaustenitized light-etching layer of as-quenched martensite is found at the surface (Ref 31) above the back-tempered zone. In many cases of abusive grinding, microstructural examination reveals that the die was not tempered. As-quenched tool steels are very difficult to grind without causing such cracking. In the example shown, however, the AISI O1 tool steel die was properly tempered, and one must conclude that the grinding operation was at fault.

Fig. 6 Variation of the calculated average asperity height with the baseline length for the fatigue crack propagation fracture profile of a IN-718 nickel-base alloy having a coarse grain structure (grain size, 23 µm). Observe that average asperity height reaches a saturation value for baseline lengths larger than 500 µm. Source: Ref 6, 31

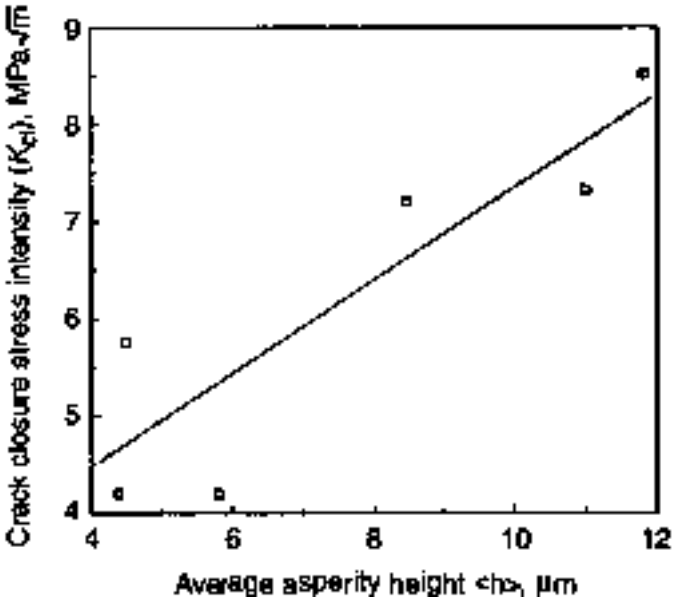


Fig. 7 Plot of crack-closure stress intensity versus average asperity height for fatigue crack propagation in IN-718 nickel-base alloy. The data include specimens of two different grain sizes tested at three different Δ*K* levels. Source: Ref 6

Angular Orientation Distribution of the Fracture Profile. The angular orientation of a line element on a vertical-section fracture profile is specified by angle α between the line element and the vertical axis. The angle α can be calculated for each straight-line segment connecting the consecutive points of the digitized profile (see Fig. 3). The orientation distribution function of the fracture profile, $f(\alpha)$, is defined such that $f(\alpha)d\alpha$ is the fraction of profile length in the orientation range α to $(\alpha + d\alpha)$, where $0 \leq \alpha \leq \pi$. The orientation distribution function, $f(\alpha)$ quantifies the anisotropy of the fracture profile. For a randomly oriented fracture profile, $f(\alpha)$ is equal to $[1/\pi]$, as all angular orientations are equally likely in such a profile. The orientation distribution function can be computed in histogram form from the experimentally measured values of the orientation angles of the line elements in the fracture profile. The profile orientation distribution function is required for estimation of the fracture surface roughness as well as for estimation of the orientation distribution of surface elements in the fracture surface from the profilometric data. In general, the profile angular orientation distribution may vary with the orientation of the vertical section. Figure 8 shows experimentally measured profile orientation distribution functions, lineal profile roughness parameters, and profile structure factors for fracture profiles observed in three vertical sections of tensile fracture surface of a metal-matrix composite (MMC) generated by uniaxial load perpendicular to the fibers. The three vertical sections are mutually at an angle of 120° to one another, and as they are vertical sections, all of them contain the vertical axis. The MMC contains aligned alumina fibers in an aluminum-alloy matrix (Ref 22).

Other dimple/feature characteristics of interest are the depth (or height, *H*) of the dimple/feature and the height-to-width ratio, ξ. The height-to-width ratio of the dimples (or other features of interest) can be directly measured by using stereophotogrammetry (Ref 97). Measurements on fracture profiles have also been used for the estimation of the height-to-width ratios of dimples (Ref 64, 72, 77, and 75), and asperities (Ref 6, 31, and 90) on fracture surfaces. The dimple/asperity heights and widths can be extracted directly from the set of (*x*, *y*) coordinates of the points on the digitized fracture profiles, from which the height-to-width ratios can be calculated (Ref 6, 31, 72, 75, and 90).

Estimation of Striation Spacing in the Fatigue Fracture Surfaces. Fatigue striations are observed in the fatigue fracture surfaces of numerous materials. The close correlation between the striation spacing, macroscopic crack growth rate (*da*/*dN*), and effective stress intensity range, Δ*K*, is well known. Consequently, measurement of striation spacing is an important aspect of the quantitative characterization of fatigue fracture surfaces. Figure 21 shows fatigue striations in the fracture surface of a fatigue test of a permanent-mold cast modified A356 aluminum alloy tested at 0.4% strain amplitude. The fatigue striations are observed in the fracture regions through the aluminum-rich dendrites but not in the interdendritic regions. Further, the local striation normals for all colonies are not along the same direction.

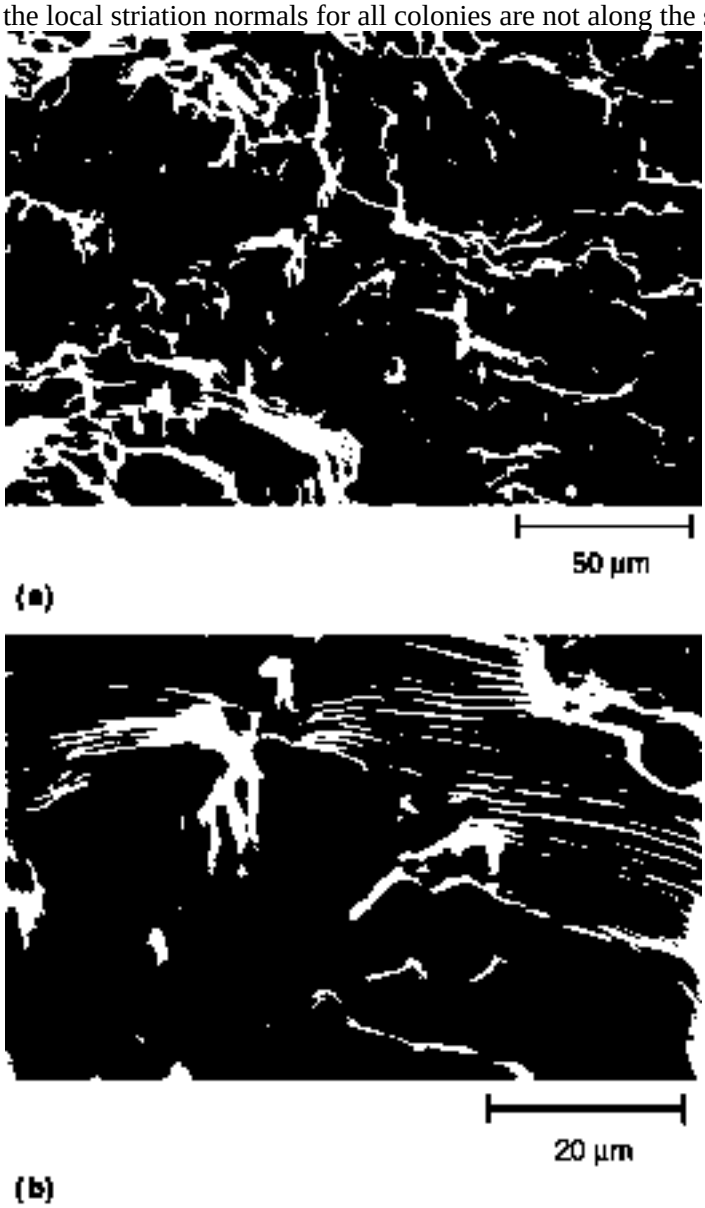


Fig. 21 Fatigue striations in a cast A356 aluminum alloy. (a) 500×. (b) 1500×

Figure 22 illustrates the relationship of the true striation spacing on the fatigue fracture surface to the projections of the striations observed in the SEM fractographs. The problem of estimation of true striation spacing from the measurements performed on the projected SEM images has been analyzed (Ref 99). To estimate the average projected striation spacing normal to the striations, • λ_{Pr}• , superimpose a straight test line on a striation colony and count the number of striations intersecting the test line. The test line should be oriented normal to the striations. Repeat these measurements on numerous randomly sampled striation colonies in the region of interest, each time placing the test line normal to the striations. From these measurements, calculate the average number of striations intersected per unit test line length, • N_L• . The average projected striation spacing normal to the striations • λ_{Pr}• is equal to [1/• N_L•]:

Table 1 and 2. Examination of the information in these tables shows that the fracture features provide information about:

- The crack initiation site and crack propagation direction
- The mechanism of cracking and the path of fracture
- The load conditions (monotonic or cyclic)
- The environment
- Geometric constraints that influenced crack initiation and/or crack propagation
- Fabrication imperfections that influenced crack initiation and/or crack propagation

In the latter case, it is very important to make the distinction between a *manufacturing imperfection* and a *manufacturing flaw* (or *defect*). Manufactured components typically contain geometric and material imperfections, but whether an imperfection caused a failure is critical in the determination of root cause. Manufacturing imperfections are not necessarily defects, and in many (if not all) situations, quantitative analysis should be considered to determine whether an imperfection is actually a root-cause flaw responsible for failure. Fabrication imperfections are discussed in more detail in the article "Mechanisms and Appearances of Ductile and Brittle Fracture in Metals" in this Volume.

It should also be clear from examination of these tables that not all features created by a given cause for failure are necessarily present on a given fracture surface. For example, beach marks (at low magnification) and striations (at higher magnification) are well-known features of fatigue cracks but are not always present or visible. In addition, not all fracture mechanisms have unique appearances. For example, intergranular fracture can be caused by a number of mechanisms, as described in more detail in the article "Intergranular Fracture" in this Volume.

It is also important to understand that the fracture surface only provides evidence of the crack propagation process; it does not reveal evidence of events prior to nucleation and growth. Examination beyond the fracture surface also provides information. For example, visual inspection of a fractured component may indicate events prior to fracture initiation, such as a shape change indicating prior deformation. Metallographic examination of material removed far from the fracture surface also can provide information regarding the penultimate microstructure, including the presence of cold work (bent annealing twins, deformation bands, and/or grain shape change), evidence of rapid loading and/or low-temperature service (deformation twins), and so forth. This also is so very necessary to the failure investigation.

Macroscopic Features. Macroscopic features typically help identify the fracture-initiation site and crack-propagation direction. The orientation of the fracture surface, the location of crack initiation site(s), and the crack-propagation direction should correlate with the internal state of stress created by the external loads and component geometry. When the failed component is in multiple pieces, and chevrons are visible on the fracture surface, analysis of crack branching (*crack bifurcation*) (Fig. 1) (Ref 25) can be used to locate the crack initiation site. Fracture initiates in the region where local *stress* (as determined by the external loading conditions, part geometry, and/or macroscopic and microscopic regions of stress concentration) exceeds the local *strength* of the material. Thus, variations in material strength and microscale discontinuities (such as an inclusion or forging seam) must be considered in conjunction with variations in localized stress that is determined by applied loads and macroscopic stress concentrations (such as a geometric notch or other change in cross section).

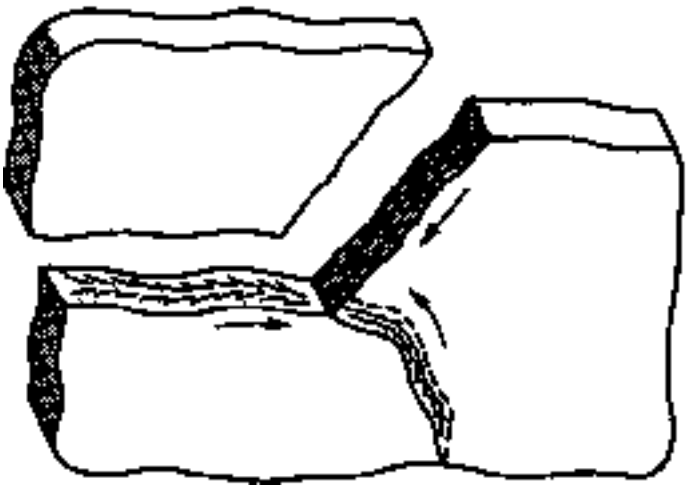


Fig. 1 Schematic view of a component that has fractured in multiple pieces. If chevrons are visible on the fracture surface, the sequence of crack formation can be used to obtain the crack formation sequence and the location of the initiation site. Source: Ref 25

The fracture surface orientation relative to the component geometry may also exclude some loading conditions (axial, bending, torsion, monotonic versus cyclic) as causative factors. For example, crack initiation is not expected along the centerline of a component loaded in bending or torsion, even if a significant material imperfection is present at that location because no normal stress acts at the centerline. (There is a shear stress at this location in bending, but in a homogeneous material, it is too small to initiate fracture. That might not be the case for a laminated structure loaded in bending.) Alternatively, brittle torsion failure is readily identified at the macroscale in cylindrical sections because of the unique helical nature of the fracture surface (Fig. 2) (Ref 26).

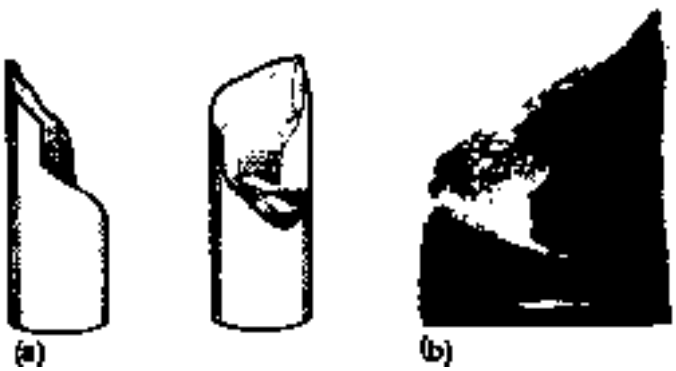


Fig. 2 Wolf's ear helical fracture due to torsion loading. (a) Schematic of brittle torsion fracture of chalk. (b) Helical tensile fracture of oxygen-free high-conductivity copper bar prestrained in torsion to a shear strain of 4. 3×. Source (b): Ref 26

Surface roughness and optical reflectivity also provide qualitative clues to events associated with crack propagation. For example, a dull/matte surface indicates microscale ductile fracture, while a shiny, highly reflective surface indicates brittle cracking by cleavage or intergranular fracture. In addition, when intergranular fracture occurs in coarse-grained materials, individual equiaxed grains have a distinctive rock-candy appearance that may be visible with a hand lens.

Surface roughness provides clues as to whether the material is high strength (smoother) or low strength (rougher) and whether fracture occurred as a result of cyclic loading. The surfaces from fatigue crack growth are typically smoother than monotonic overload fracture areas. The monotonic overload fracture of a high-strength quenched and tempered steel is significantly smoother overall than is the overload fracture of a pearlitic steel or annealed copper. Also, fracture surface roughness increases as a crack propagates so the roughest area on the fracture surface is usually the last to fail (Fig. 3) (Ref 27). Fracture surface roughness and the likelihood of crack bifurcation also increase with magnitude of the applied load and depend on the toughness of the material (Fig. 4) (Ref 28).

Fig. 3 Change in surface roughness due to crack propagation. Fracture surface roughness increases with distance of propagation, crack propagation rate, and decreased strength level. This component failed in fatigue. Crack initiation was on a longitudinal plane visible at the top in a surface hardened region. The crack then propagated on a helical plane in torsion. Note the change in surface roughness as the crack propagates from the surface hardened region at the top into the core and finally into the hardened case at the bottom of the photograph. The roughest region on the fracture surface is the final overload region at the bottom. Source: Ref 27

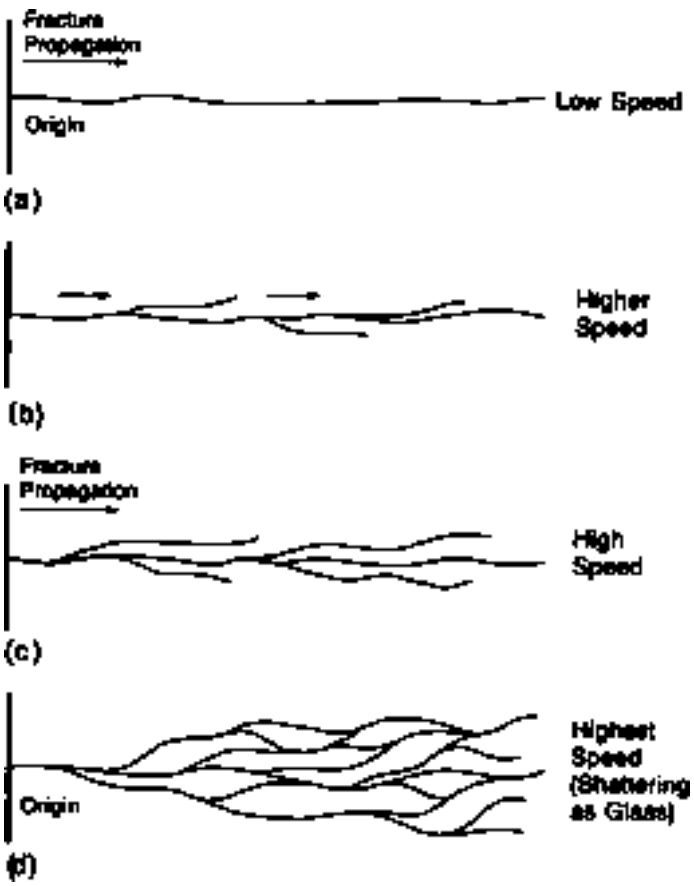


Fig. 4 Change in crack bifurcation with magnitude of the load and speed of crack propagation. (a) Low speed (low load) and high toughness. (b) Higher speed (higher load) and high toughness. (c) High speed (high load) and low toughness. (d) Higher speed (higher load) and low toughness. Source: Ref 28

Radial marks (Fig. 5) (Ref 29) and chevrons (Fig. 6), if present, are macroscopic surface features that indicate the region of crack initiation and propagation direction. They are common and dominant macroscopic features of the fracture of wrought metallic materials but are often absent or poorly defined in castings. The “v” of a chevron points back to the initiation site, and a sequence of “v”s across the fracture surface indicates the crack propagation direction. Radial marks typically fan out from the initiation site, although Fig. 7 shows an exception to this.

Fig. 5 Macroscale radial marks. The fan array of the radial marks points back to the crack initiation site and is created under conditions of “rapid” crack propagation. Source: Ref 29

Fig. 6 Chevrons on the fracture surface of an induction-hardened axle fabricated from 1541 steel. The V-shaped chevrons point back to an initiation site marked by the arrow at the top of the figure. Component shows fatigue crack growth initiating at the arrow creating the circular-shaped region at the top. Overload then occurred, and fast fracture created chevrons in the hardened case but not in the tougher core. Note also the increase in surface roughness in the pearlitic core as the overload crack grows. Source: Ref 29

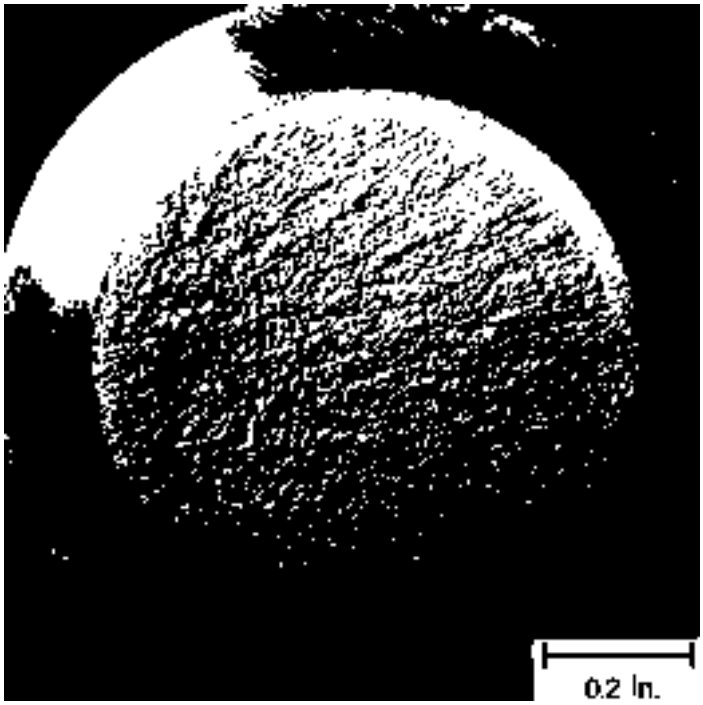


Fig. 7 Macroscale brittle fracture in tensile loading. A light ring is visible around the outside circumference. A faint radial pattern is visible from approximately 11 to 4 o'clock and running towards a dark spot near 9:00 The roughest area on the fracture surface is near the dark spot (see text for discussion). Source: Ref 30

Figure 7 (Ref 30) illustrates the importance of utilizing fracture surface roughness in conjunction with radial marks to identify the initiating location for fracture. There is a light colored region around the perimeter of the specimen, and there is a small dark region slightly in from the surface at the 9 o'clock position. Light-colored surface regions are often associated with surface hardening. If no information is available except the photograph, it might be (incorrectly) concluded from the radial markings that the dark region was the location of some material imperfection that initiated fracture. Typically, it is common for fracture initiation to occur within a relatively small region, where the center of the radial-fan markings provides a strong indication of the crack initiation region.

The center of the radial-fan markings are usually at or near the fracture initiation site, but this interpretation would be incorrect in the case of Fig. 7. The specimen shown in Fig. 7 was in fact pre-fatigue-cracked and then loaded in monotonic tension to failure. Cracking initiated within a wide arc at the surface (at positions from approximately 11 to 4 o'clock), and the radial lines grew together rather than fanned out as the crack propagated. The dark region at 9 o'clock is not an imperfection but rather contained within a rougher region with surface steps, indicative of the last region of fracture. Surface roughness increases as the crack propagates, and the rough area surrounds the dark spot. The spot appears dark because of the angle of the illumination. The deciding issue is the surface roughness, as it indicates the later material to fail was near the dark region and not at the surface.

Typically, it is common for fracture initiation to occur within a relatively small region (in which case the center of the radial markings would fan out from the small region of crack initiation). However, as demonstrated in Fig. 7, this is not the case when crack initiation occurs over a wider region.

Microscopic Examination and Appearances. Microscopic examination can help identity imperfections that initiate cracking, the path of fracture (intergranular or transgranular), and the microscale mechanisms of cracking (i.e., microvoid coalescence, cleavage, or fatigue). Localized directions of crack growth also can be determined from the river lines of cleavage fracture surfaces. In the case of microvoid coalescence, the shape of the dimples on the fracture surface can be correlated with loading conditions (i.e., equiaxed dimples from tensile loading; elliptical dimples from shear or torsional loading).

Microscale fractographic features help identify the microscopic mechanism(s) causing fracture. Figure 8 (Ref 31) is a summary of the possible microstructural features associated with the basic types of external load conditions (overload, fatigue, and environmentally assisted sustained-load cracking). As indicated in the figure,

a dimpled fracture surface is uniquely associated with the microscale mechanism of microvoid coalescence, which typically is associated with macroscopic ductile fractures. However, macroscale brittle fractures can also occur when plastic deformation is limited to a small volume of material and not macroscale visible, while the fracture process is still microvoid coalescence. This is the case when the ductile fracture mechanism of microvoid coalescence is constrained to a plane-strain strain fracture mode (referred to as plane-strain microvoid coalescence) or occurs preferentially in the limited region adjacent to the grain boundary (resulting in a dimpled intergranular fracture surface). These types of ductile and brittle fractures are discussed in more detail later in this article.

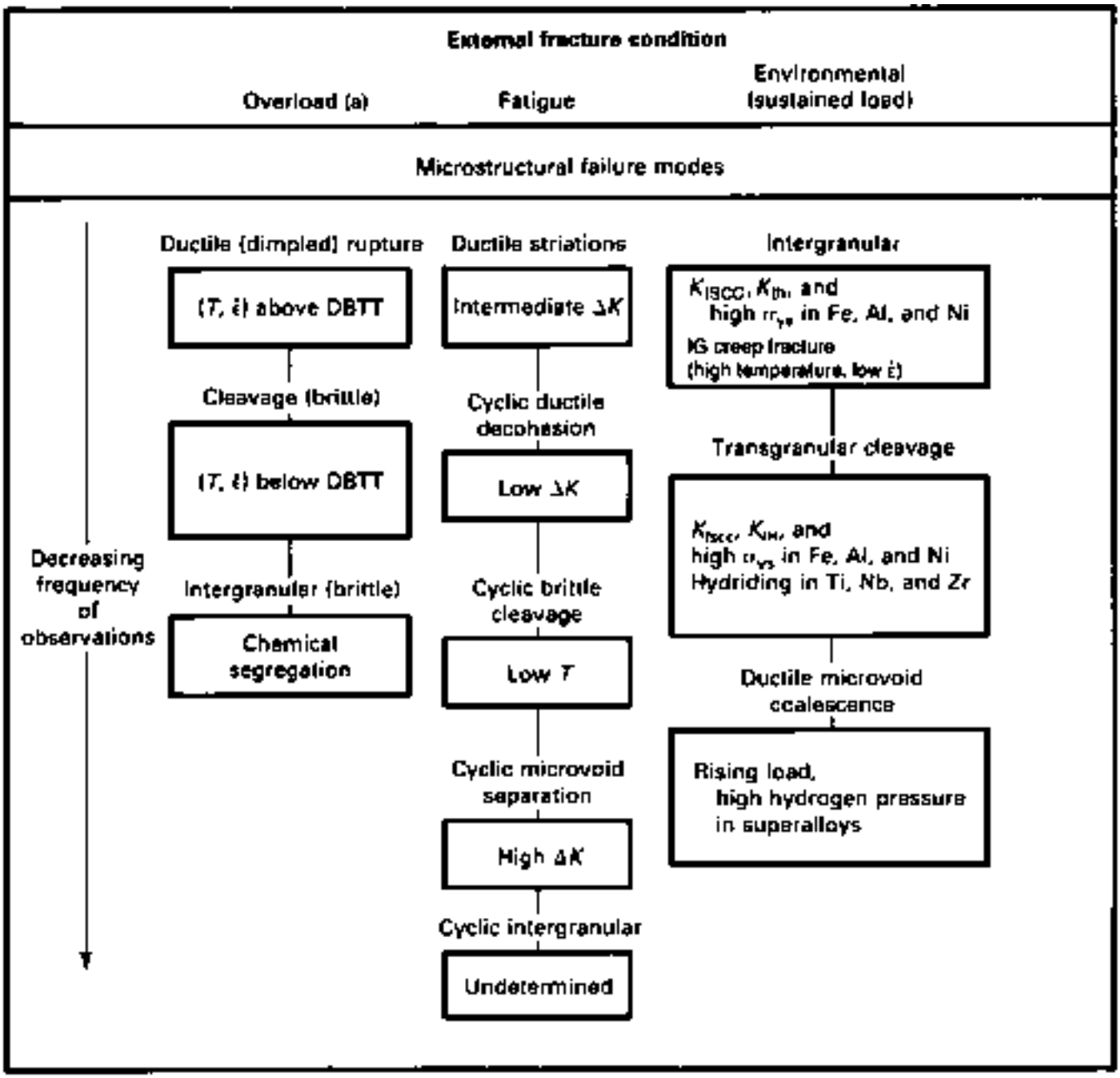


Fig. 8 Observed microscopic fracture mechanisms for different loading conditions and environments. DBTT is the ductile brittle transition temperature, and K_{ISCC} is the stress corrosion threshold. K_{IHE} is the hydrogen embrittlement threshold. Note 8(a): See Fig. 13 and discussions for conditions of macroscale ductile and brittle fracture. Adapted from Ref 31

In a more general sense, the microscale features of fractures in crystalline materials can be described as either transgranular (TG) or intergranular (IG). Transgranular crack propagation is caused by competing mechanisms of ductile crack nucleation, growth by slip deformation mechanism, and brittle cracking by cleavage. (As described later in this article, twinning is a transgranular mechanism of plastic deformation. Deformation

twinning provides a limited amount of ductile deformation but also provides an alternative for initiation of cleavage cracks. What can result from twinning deformation is cleavage crack nucleation at the intersection of mechanical twins, for example, as discussed further in the next article, “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals.”)

These two mechanisms of TG cracking have distinct appearances on the microscale characterized by dimpled fracture surface for ductile TG fractures (Fig. 9) and the distinctive river lines of cleavage for brittle cracking (Fig. 10). The occurrence and appearances of these TG cracking mechanisms are influenced by crystal structure, microstructure, loading rate, and temperature, as briefly discussed later in this article and in more detail in the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals,” in this Section of this Volume.

Fig. 9 Dimpled rupture created by microvoid coalescence in a quenched and tempered steel. Note the presence of carbide particles in the bottom of several dimples. Palladium shadowed two-stage carbon replica. Because the image is a replica of the fracture surface, there is a reversal of fractographic features. Source: Ref 7

Fig. 10 River lines on a cleavage fracture surface. Direction of growth is parallel to the direction of crack coalescence as indicated by the arrow. Cracks must reinitiate at a boundary containing a twist (mode III) deformation component.

Intergranular fracture is clearly distinguished from TG fracture, but unlike TG fracture, the microscopic appearances of IG fractures are not uniquely associated with a specific microscale mechanism. (It is not uncommon to see a fracture surface containing both IG and TG fracture.) Preferential cracking in (or near) the grain boundaries may be related to various types of IG mechanisms (which may be diffusion-related processes such as creep void formation, hydrogen void formation, or impurity segregation in the grain boundaries). Various microstructural and mechanical conditions also may influence crack path. For example, Table 3 (Ref 32) focuses on the influence of microstructure on fracture toughness and the fracture path in some high-strength alloys. Sometimes the facets of IG-fracture are almost featureless, perhaps containing only the presence of a second phase when they occur truly in the grain boundary. However, it not uncommon to see very small dimples on grain boundary facets. Dimpled intergranular fracture can occur from void formation in the grain boundaries or from the mechanism of microvoid coalescence in the region adjacent to the grain boundary.

Table 3 The influence of microstructure on fracture path and fracture toughness for several high strength alloys

Microstructural feature	**Specific aspects**	**Effects on fracture path**	Effects on fracture mechanics properties
Phase morphology in α-β titanium alloys	Change from α and β grains to α in β	Crack branching and irregularity greatly increased	• **Higher K_{Ic} and K_{Issc}** • **Increased resistance to fatigue crack growth**
Large particles in steels and aluminium alloys	Randomly distributed	Large microvoids (mainly transgranular)	• **Lower K_{Ic}** • **Acceleration of high stress intensity fatigue**
	Intergranular carbides in steels	Nucleation of cleavage	
Dispersoids in steels and aluminium alloys	Randomly distributed	Sheets of small transgranular microvoids	• **Lower k_{Ic}** • **Acceleration of high**

			stress intensity fatigue
	Increased spacing between dispersoids in aluminium alloys	Faceted crack growth maintained to higher ΔK	**Transition from region I to region II fatigue crack growth postponed to higher ΔK**
Precipitates and particles in Al-Zn-Mg-Cu alloys	Larger matrix precipitates and grain boundary particles due to overaging	Increase in intergranular microvoid coalescence during unstable fracture	**Lower K_{Ic} and K_c**
		None for stress corrosion	**Higher K_{Iscc} and lower region II crack growth velocity**
Larger grain size	Overaged Al-Zn-Mg-Cu alloys	Increase in intergranular microvoid coalescence	**Lower K_{Ic} and K_c**
	Steels	Faceted crack growth maintained to higher ΔK	**Transition from region I to region II fatigue crack growth postponed to higher ΔK: higher ΔK_{th}**
Conventional α-β titanium alloys	Crack branching and irregularity maintained to higher ΔK	Increased resistance to region II fatigue crack growth	
Fibering in aluminium alloys	Alignment of grains and particles	Branching and irregularity much less when crack plane normal to short transverse direction	• **Lower K_{Ic} and K_{Issc}** • **Decreased resistance to fatigue crack growth**
Crystallographic texture in α-β titanium alloys	**Preferred orientation of the hexagonal α phase in microstructures with individual α and β grains**	**Crack blunting, branching and irregularity depend on texture**	**Complex and strong influences on K_{Ic}, K_{Issc}, ΔK_{th}, and the resistance to fatigue crack growth**

Source: Ref 32

The possible influence of the microstructure or the environment sometimes can be determined more easily by microstructural examination, especially with specimens taken perpendicular to the fracture surface and containing the fracture surface in edge view. Specific examples include identification of IG fracture and the possible role of inclusions and/or second phases in influencing the direction of crack propagation. In summary, microscopic examination of the fracture surface can but may not always or uniquely provide information regarding:

- Whether crack propagation (i.e., the fracture progression mechanism) is ductile or brittle
- Whether loading was axial, bending, or torsion and monotonic or cyclic
- Whether fracture may have occurred at a high fraction of the melting point (high homologous temperature, T_H)
- Whether the environment played a role in the fracture
- Whether the thermal processing history of the material was improper

Mud Cracks. The presence of oxidation products may indicate elevated temperature service, and the presence of surface discoloration may indicate a corrosive service environment. Thicker surface deposits from liquid on the surface often crack in a distinctive manner as they dry and result in *mud-cracks* (Fig. 11 and 12). These mud cracks in the surface deposit may indicate the possibility of an environmentally induced fracture; that is, stress-corrosion cracking (SCC). Unfortunately, mud cracks can also be created by caustic cleaning residue from attempts to clean the fracture before providing the specimen to the analyst.

Fig. 11 Mud cracks on the surface of an intergranular fracture in 7079-T651 aluminum that failed under SCC conditions in a 3.5% chloride solution. TEM replica

Fig. 12 Mud cracks on the fracture surface of a quenched and tempered 4340 steel exposed to a marine environment. TEM replica

References cited in this section

7. *Electron Fractography,* STP 436, ASTM, 1968

25. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials*, 4th ed., John Wiley, 1996

52. W.T. Whited, A.M. Gokhale, and N.U. Deshpande, Characterization of Thermally Induced Microcracks in a Metal Matrix Composite, *Microstruct. Sci.*, Vol 21, 1994, p 107–120

53. M.D. Dighe, A.M. Gokhale, and M.F. Horstemeyer, Effect of Loading Condition and Stress State on Damage Evolution of Silicon Particles in an Al-Si-Mg Base Cast Alloy, *Metall. Mater. Trans.*, Vol 33A, 2002, p 555–565

54. M.D. Dighe, A.M. Gokhale, and M.F. Horstemeyer, Effect of Strain Rate on Damage Evolution in a Cast Al-Si-Mg Alloy, *Metall. Mater. Trans.*, Vol 31A, 2000, p 1725–1731

55. J.-W. Yeh and W.-P. Liu, The Cracking Mechanism of Si Particles in A357 Alloy, *Metall. Trans.*, Vol 27A, 1996, p 3358–3368

56. A.M. Gokhale, Utility of Stereological Counting Measurements in the Study of Creep Cavitation Kinetics, *MiCon-90: Advances in Video Technology for Microstructural Control*, STP 1094, G.F. Vander Voort, Ed., ASTM, 1991, p 332–339

57. D. Kwon and R.J. Asaro, A Study of Void Nucleation, Growth, and Coalescence in Spheroidized 1518 Steel, *Metall. Trans.*, Vol 21A, 1990, p 117–134

58. Z.H. Huang, J.F. Tian, and Z.G. Wang, A Study of the Slit-Island Analysis as a Method for Measuring Fractal Dimension of Fractured Surface, *Scr. Metall.*, Vol 24, 1990, p 967–972

59. Z.H. Huang, J.F. Tian, and Z.G. Wang, *Mater. Sci. Eng.*, Vol A118, 1989, p 19–24

Quantitative Fractography

Arun M. Gokhale, Georgia Institute of Technology

Profilometry-Based Quantitative Fractography

A fracture profile is a tortuous irregular line generated by the intersection of a fracture surface with a sectioning plane. Several experimental methods are available for generation of fracture profiles. These experimental methods yield the (x, y) coordinates of the set of closely spaced points in the fracture profile. From these basic data, important geometric attributes of the fracture profile can be computed. Stereological equations can be then used for unbiased statistical estimation of the geometric attributes of the corresponding fracture surface from the computed fracture profile parameters. The experimental methods for profile generation, computation of fracture profile parameters, and the stereological procedures for estimation of fracture surface attributes from the profilometric data are described in the following subsections.

Experimental Methods for Profile Generation

Physical metallographic sectioning of the specimen, sectioning of fracture surface replicas, and nondestructive profile generation by using stereo pairs and other methods are the important experimental techniques used for generation of fracture profiles. These methods are described as follows.

Metallographic Technique for Fracture Profile Generation. Depending on the angular orientation of the sectioning plane with respect to the fracture surface, it is possible to generate horizontal (Ref 8, 58, 59), slanted (Ref 60), conical (Ref 61), and vertical section (Ref 17, 18, 21, 22) fracture profiles. The vertical section fracture profiles are most widely used as they are easy to generate using well-known metallographic procedures, and general stereological equations are available for estimation of fracture surface characteristics from the measurements performed on vertical section fracture profiles (Ref 4, 21, 22). Figure 1 shows a schematic fracture surface, a sectioning plane, and a schematic representation of the fracture profile generated by intersection of the fracture surface with the sectioning plane. Figure 2(a) shows a fracture profile generated from the fracture surface of a tensile test specimen of a metal-matrix composite (MMC) containing unidirectionally aligned alumina fibers in the matrix of an aluminum alloy. Figures 2(b) and 2(c) show a fracture profile

amorphous and crystalline, by cleavage. Additionally, in polycrystalline polyphase metallic materials, fracture may occur at or adjacent to grain boundaries, and cleavage fracture may occur in large second phases present in the microstructure, which then provide crack nuclei for ductile crack propagation in the matrix.

Cleavage fracture at a macroscale (low magnification) is characterized by high light reflectivity and a relatively flat surface. At the microscale, it is a series of flat regions or ledges that are often faceted. Higher magnification shows the ledges to be connected by ligaments described as *river lines* (Fig. 30) or *fans* (Fig. 31, Ref 45 and 46). The river lines coalesce as they propagate and have the appearance of water flowing downstream (Fig. 10). However, river lines cannot always propagate across grain boundaries in crystalline material, and at the microscale, crack initiation may occur in more than one location. Thus the river lines only indicate the *local* direction of crack propagation, which can be opposite that of the crack propagation at the macroscale.

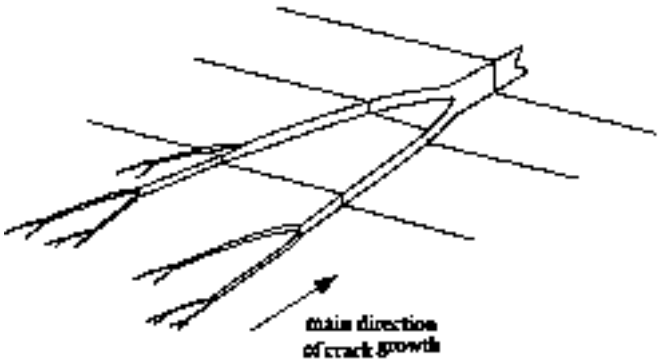


Fig. 30 Schematic of a river pattern. Crack growth is in the direction of crack coalescence. River patterns may be visible at the macroscale in organic glasses and brittle polymers but are visible only at the microscale in metallic materials. Source: Ref 11

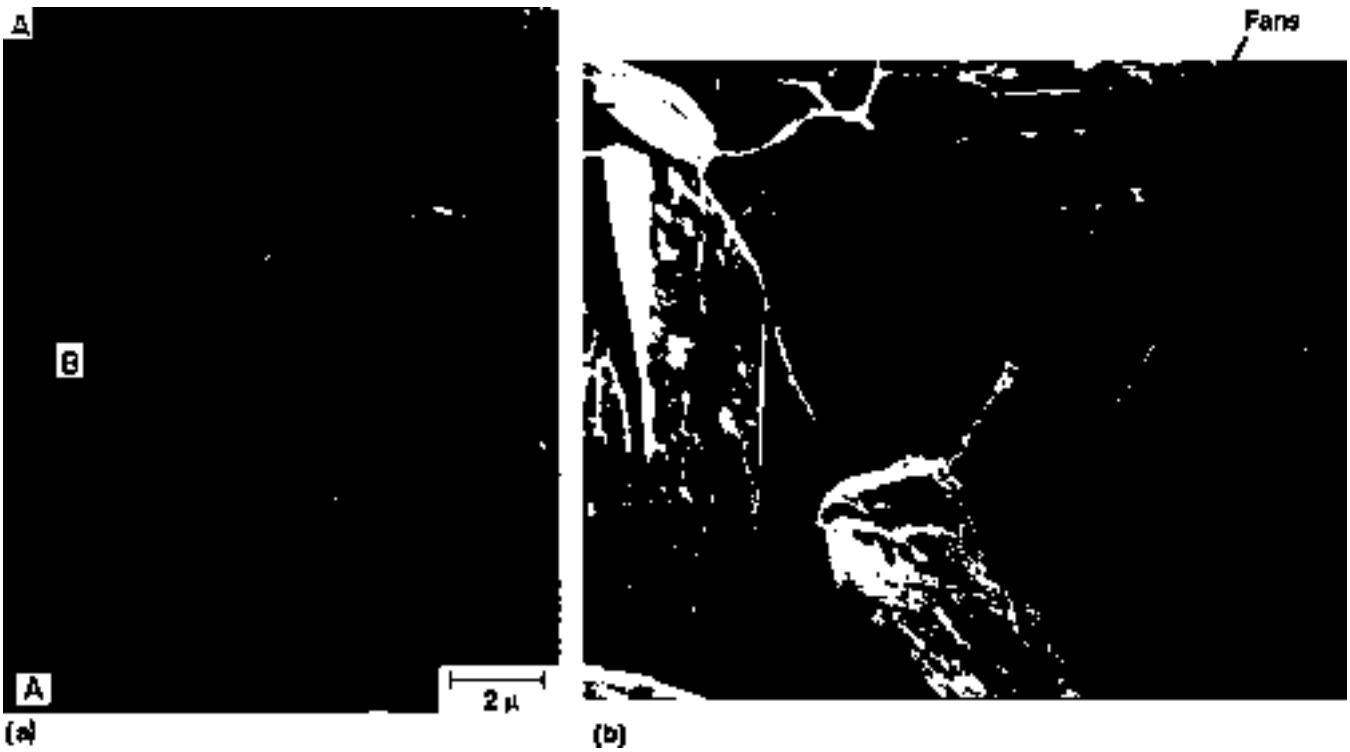


Fig. 31 Fans. (a) Examples of fans in a two-stage TEM replica of a cleavage fracture surface of iron. The river lines point back to the crack initiation site. (b) Fans on SEM image. Source: Ref 44, 46

The normal stress acting on a plane is of importance in propagating cleavage cracks. Consider the cracked plate shown in Fig. 32 (Ref 11). When crack propagation from the preexisting imperfection (assumed through-thickness crack) occurs, the crack develops “wings” as it propagates and curves so as to reorient to propagate on the macroscale plane of maximum normal stress. At the microscale in crystalline material, a more

complicated process is required to cause this macroscale appearance to be curved (see the section “Tilt and Twist Boundaries and Loading Mode,” following in this article.)

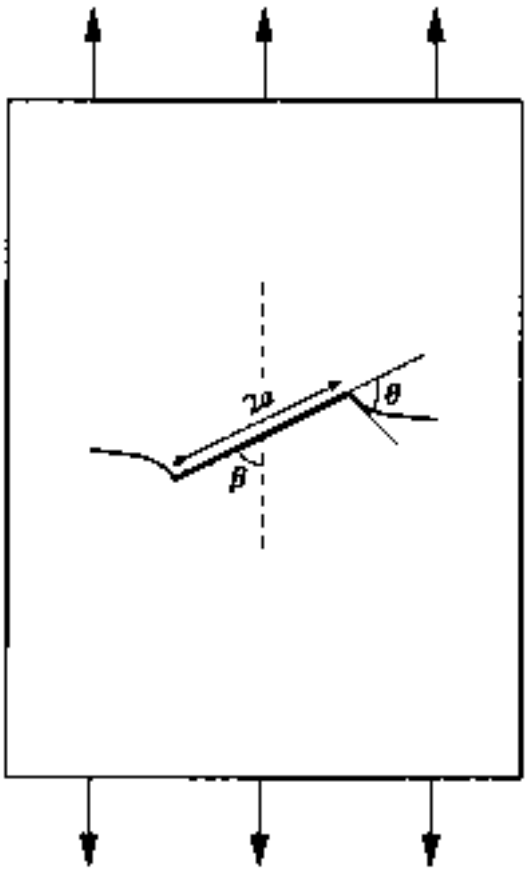


Fig. 32 Macroscale brittle crack propagation due to combined mode I and mode II loading. As cracks grow from the preexisting cracklike imperfection, crack curvature develops because of growth on a plane of maximum normal stress. Source: Ref 11

Cleavage fracture can also occur in some polymeric materials and in ceramics and inorganic glasses creating a mirror region visible at the macroscale (Fig. 33) (Ref 11). Cleavage fracture in polymeric materials and inorganic glasses also results in flat featureless regions, sometimes quite large. River lines are again present at the microscale but may also often be visible at the macroscale. The flat, high-reflectivity region where fracture initiates (*mirror region*) is surrounded by *mist* and *hackle marks* (Fig. 33). The features may be visible at the macroscale or with the aid of a small hand lens.

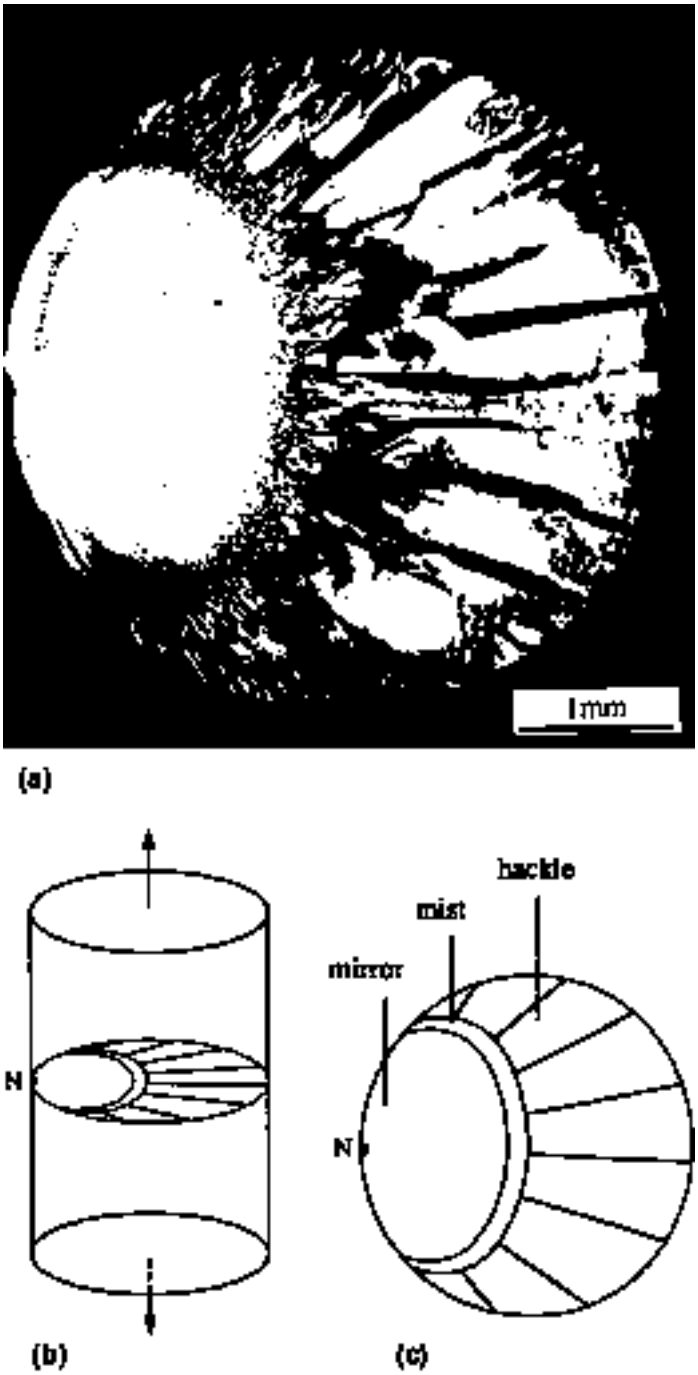


Fig. 33 Cleavage fracture in a soda lime glass. Crack progresses from left to right. (a) Fracture surface shows the initiation region (featureless mirror region), mist surrounding the mirror region and hackle. (b) Geometry of tensile test showing position of fracture surface normal to tensile axis. (c) Arrangement of mirror, mist, and hackle regions on fracture surface. See text for discussion. Source: Ref 11

Totally brittle fracture in metals at the microscopic level ("ideal cleavage" or "pure cleavage") occurs only under certain well-defined conditions (primarily when the component is in single crystal form and has a limited number of slip systems) and is correctly described as "cleavage fracture." More commonly in metals, the fracture surface contains varying fractions of transgranular cleavage and evidence of plastic deformation by slip. When both fracture processes operate intimately together, and especially in the fracture of quenched and tempered steels, the fracture process is termed *quasi-cleavage.* The dividing line between the terms "cleavage" and "quasi-cleavage" is somewhat arbitrary.

Metallic material fracture surfaces showing large fractions of cleavage may show cracking on more than one crystallographic plane within a given grain, leading to the most common feature associated with brittle faceted fracture—*river lines,* as previously noted. These lines may form by a ductile process, but the slip deformation that created them is not resolvable and such fracture is not described as quasi-cleavage. Alternatively, with less

constraint, the connecting ligaments may become sufficiently large that microvoid coalescence is observed in thin bands weaving through the general cleavage surface (Fig. 34).

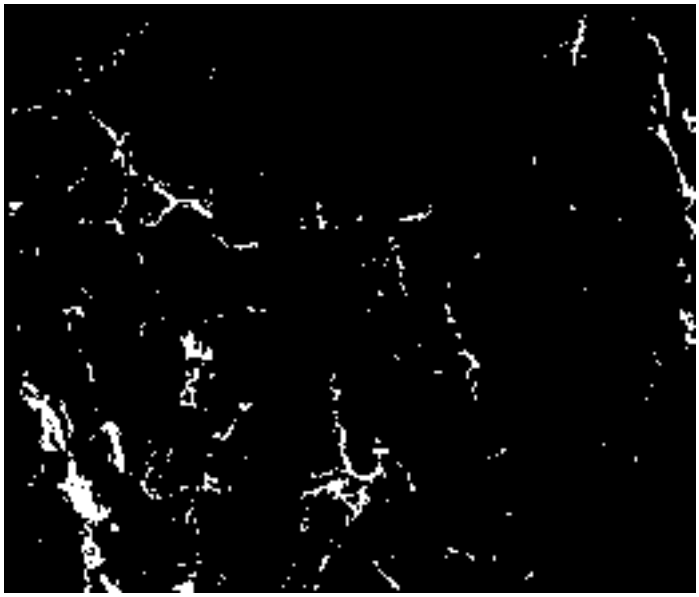


Fig. 34 Microscale quasi-cleavage fracture in an O1 tool steel tested at room temperature. Predominantly cleavage cracking with patches and ribbons of microvoid coalescence. Source: Ref 35

Considerable evidence exists that cleavage fracture is facilitated by the presence of a normal stress on the cleavage plane. However, in crystalline materials there is convincing evidence that cleavage fracture occurs by prior dislocation motion to create a cleavage crack nucleus and therefore a shear stress is also required on one or more slip planes, depending on the model (see the article "Mechanisms and Appearances of Ductile and Brittle Fractures in Metals" in this Volume. If the material is amorphous and isotropic, the cleavage plane is simply the macroscale plane having the largest principal stress. However, if interatomic bond energy is anisotropic, the resolved normal stress on the weakest bonded plane apparently controls cleavage fracture; for example, in a layered structure.

If the material is crystalline, bonding energies across planes of atoms are typically anisotropic so that cleavage fracture occurs on specific crystallographic-plane families, typically of low index, in individual grains. Cleavage is in general observed in less symmetrical crystal structures (hcp, orthorhombic, etc.) and in the bcc materials, but cleavage is unlikely in the fcc lattice with the possible exceptions discussed in the article in this Section "Mechanisms and Appearances of Ductile and Brittle Fractures in Metals."

The specific cleavage plane family and the cleavage plane multiplicity (number of nonparallel planes) vary with crystal structure, and there may be more than one cleavage plane family in a particular crystal structure. For example, bcc materials typically cleave on the family of basal planes (three planes, mutually perpendicular), but cleavage has also been reported on the {1,1,0} family of planes (Ref 47). Cleavage fracture has not been observed in the bcc alkali metals (Ref 48). The hcp materials are routinely considered to cleave on the basal plane (only one plane in the family), but this is not necessarily true. There may be multiple cleavage plane families. For example, zinc cleaves readily on the basal plane, but magnesium cleaves only with difficulty on the basal plane. Beryllium cleaves on both the basal plane and on the {1, 1, -2,0} family of planes (3 planes, not mutually perpendicular but perpendicular to the basal planes). More details of cleavage fracture and fractographic appearance in crystalline metallic materials are discussed in the article "Mechanisms and Appearances of Ductile and Brittle Fractures in Metals."

Tilt and Twist Boundaries and Loading Mode. Assume that a crack exists inside a small region in a body on the *y*-plane and that its crack front is parallel to the *z*-direction (Fig. 35a) (Ref 11). With reference to the coordinate system shown, stresses σ_{yy}, or σ_{xy} (and σ_{yx}) will propagate the crack in the *x*-direction on the *y*-plane. In fracture mechanics terminology, the stress σ_{yy} creates opening mode (mode I) loading, and the shear stress σ_{xy} creates in-plane shear loading (mode II), both of which propagate the crack in the *x*-direction. If both stresses act, the plane of maximum normal stress is some new plane designated as y' in Fig. 35(b). The crack can move onto this new plane y' while maintaining a contiguous crack front. The conclusion is that combined mode I-mode II

loading will propagate the crack in the x' direction. Furthermore, if a boundary exists in the body that lies parallel to the z-direction, the crack can continue to propagate across the boundary. This boundary is identified as a tilt boundary. Now consider the case in Fig. 35(b). Again assume a preexisting crack lying in the y-plane. Subject this crack to pure mode III (out-of-plane shear) loading creating stresses σ_{zz} and/or σ_{yz} or in combination with mode I (σ_{yy}) loading. The plane of maximum normal stress (y') no longer contains the crack front (parallel to z). That is, under combined mode I-mode III loading, the z-direction becomes z', rotating as shown, and the plane of maximum normal stress is now the y'-plane. It can be concluded that combined mode I-mode III loading tries to twist the crack onto the y'-plane, and it cannot remain contiguous if it does this. If a (grain) boundary exists in the material such that the cleavage plane in grain A has a normal y and the cleavage plane normal in grain B is y', the grain boundary is a twist boundary and the cleavage crack cannot propagate across this boundary. To cross the boundary at z, the crack must re-nucleate new cleavage plane defined by its normal y'.

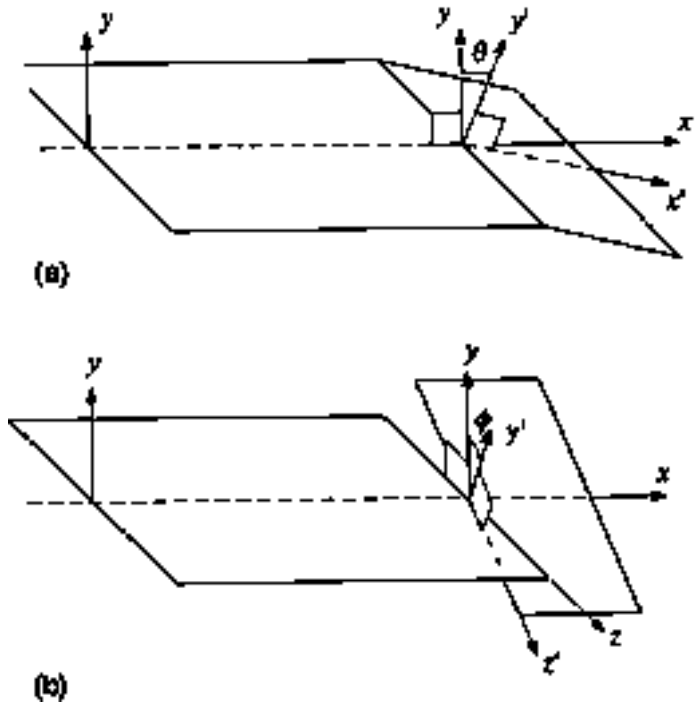


Fig. 35 Propagation of cracks in (a) Combined mode 1—mode II loading and (b) Combined mode I—mode III loading. The boundary between the planes y and y' in (a) is a tilt boundary. The boundary between the planes y and y' in (b) is a twist boundary. Cracks can propagate across the boundary in (a) but not in (b). See text for discussion. Source: Ref 11

When a component contains multiple cracks in close proximity to each other or a crack approaches a second discontinuity, the stress fields associated with the two cracks overlap, causing a complex state of stress in the ligament between the cracks; crack curvature may result, as shown in Fig. 36 (Ref 11). Similarly, when a propagating crack approaches a free surface, constraint is relaxed and curvature may result.

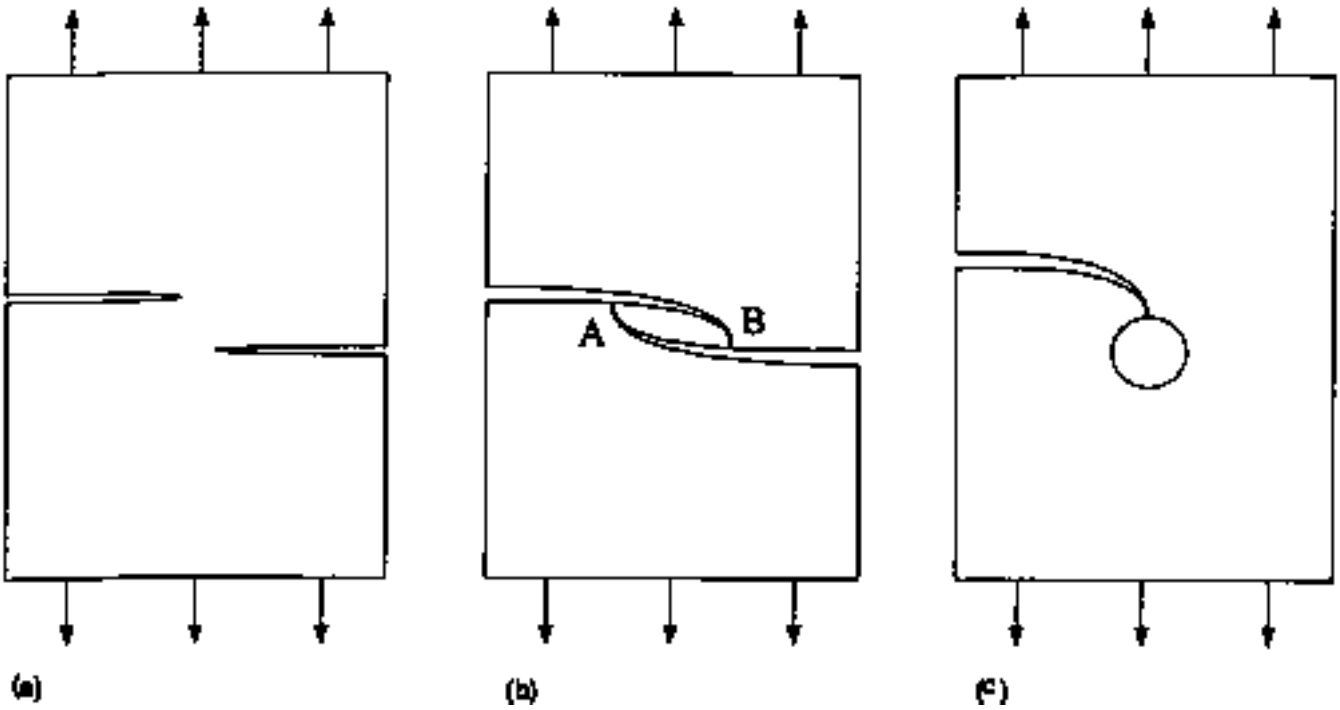


Fig. 36 For loading that is a combination of mode I and mode II loading, crack curvature occurs when the stress field of two cracks interact with each other or when the stress field of a growing crack interacts with a free surface, for example that created by a hole. (a) Non-coplanar parallel cracks. (b) Interaction between non-coplanar parallel cracks. (c) Interaction between a crack and a hole. Source: Ref 11

References cited in this section

11. D. Hull, *Fractography*, Cambridge University Press, 1999

35. W.S. Lin, The Ductile-Brittle Fracture Transformation: A Comparison of Macro and Microscopic Observation on Compact Tension Specimens, master's thesis, University of Tennessee, 1974

44. G. Bodine, “The Effect of Strain Rate Upon the Morphology of High Purity 26% Chromium, 1% Molybdenum Ferritic Stainless Steel”, master's thesis, University of Tennessee 1974

45. C.R. Brooks, and A. Choudhury, *Failure Analysis of Engineering Materials*, McGraw-Hill, New York, 2001, p 136

46. C.D. Beachem, “Microscopic Fracture Processes” *Fracture*, Vol 1, Academic Press, 1968

47. C.S. Barret and R. Bakish, Trans. AIME, Vol 212, 1958, p 122

48. R.E. Reed-Hill, *Physical Metallurgy Principles*, Van Nostrand, 1973, p 755

W.T. Becker, University of Tennessee, Emeritus; S. Lampman, ASM International

Fatigue Fracture

Fatigue fracture is another example of a macroscale brittle fracture associated with microscale ductile fracture processes. On a macroscopic scale, however, long-life fatigue failures are brittle; that is, there is little or no net-section distortion associated with the fatigue region of the fracture surface (as the nominal stress is less than the net-section yield strength). Thus, a damaged component containing a tightly closed crack *may* indicate loading below the net-section yield strength and potentially could be an indication of a long-life fatigue failure.

A fatigue fracture surface generally consists of three regions associated with crack initiation, crack growth, and final overload (e.g., Fig. 40). Fatigue cracks are typically transgranular, but intergranular fatigue does occur in some circumstances such as high strain (low-cycle) fatigue or at low ΔK when environmental factors become more dominant or when microstructural variations influence the crack path. This section briefly describes the general features of fatigue fractures in terms of crack initiation and fatigue crack propagation. More detailed information also is contained in the articles “Fatigue Fracture Appearances” and “Fatigue Failures” in this Volume.

Fig. 40 A fracture surface created by fatigue. A fatigue fracture surface consists of 3 zones. The initiation zone is often black in steels due to oxidation and may be associated with ratchet marks. The fatigue region often shows beach marks. The overload region is rougher than the fatigue region and may show chevrons or radial marks. In this example rubbing has obliterated the early stages of fatigue cracking, but ratchet marks are present to indicate locations of crack initiation. Material is 1046 steel with a hardness of approximately HRC 30. Source: Ref 29

Fatigue Crack Initiation. Macroscale fatigue-crack initiation sites in steel are often black in color. In aluminum alloys, they may be discerned at high magnification due to a rounding of the fracture surface as a result of oxidation. In castings the distinction between the region of fatigue crack growth and the overload region is typically less distinct and in some cases not identifiable.

Fatigue cracks typically originate at the surface of a component, often in a region associated with a macro- or microscale stress riser. Exceptions occur in the case of contact loads where maximum shear stresses develop in a subsurface region (see the section “Contact Fatigue” in the article “Fatigue Failures” in this Volume) and in some surface treated components. Fracture “at the surface” includes fracture slightly below the surface if subsurface imperfections or discontinuities are severe enough stress risers. In some cases, multiple fatigue cracks are initiated on the surfaces where stresses are highest. For example, Fig. 40 shows a fracture surface resulting from cyclic reversed bending. Cracks initiated at several locations on two sides approximately near the three o'clock and nine o'clock positions, where maximum tensile stresses developed during cyclic reversed

bending. Rubbing has removed evidence on the fracture surface of crack initiation, but the initiation locations can still be identified by the presence of ratchet marks (see below). In this instance, the overload region (visibly rougher along the north-south diameter of the section) also indicates completely reversed loading (S_{max}/S_{min} = -1), because the regions of fatigue crack growth are both of the same size, causing the overload region to be centered on the original neutral axis.

Ratchet Marks. Another important indicator of fatigue is the presence of "ratchet marks," where fracture initiates at the surface of a component on two closely spaced parallel planes. As the cracks propagate, coalescence occurs and the cracks eventually join to form a steplike feature or ratchet mark (Fig. 41) (Ref 50) on the fracture surface in the vicinity of the multiple initiation sites. Ratchet marks may also be formed in monotonic loading but are significantly smaller in number and less frequently seen.

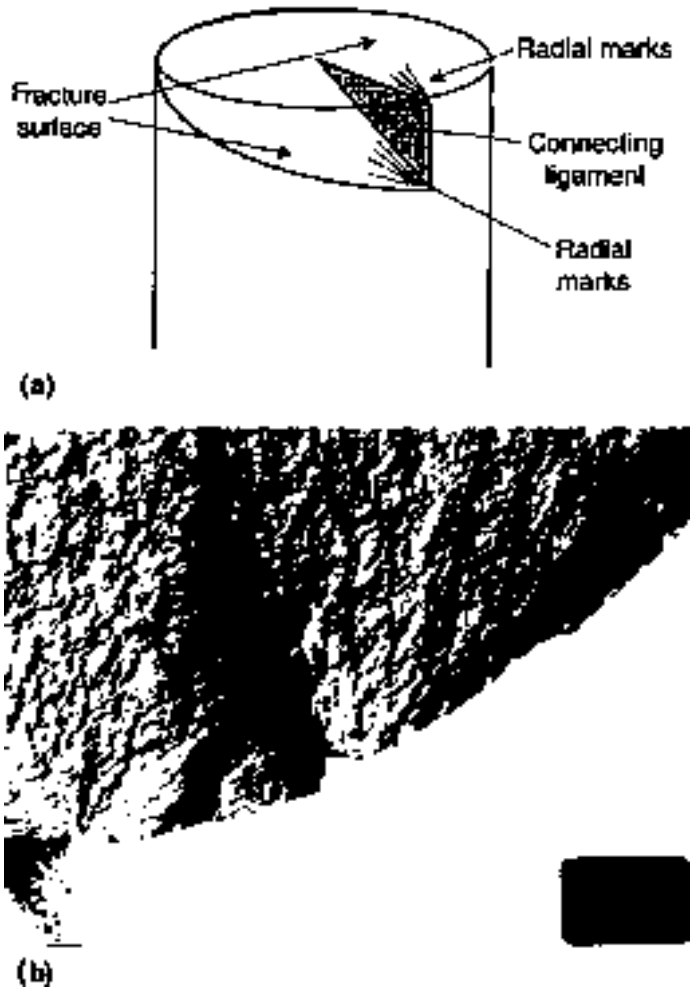


Fig. 41 Ratchet marks. (a) Schematic of two cracks that initiate in close proximity to each other. The two cracks then coalesce after some growth. (b) A ratchet mark on a cylindrical section of a 1042 steel loaded in rotating bending. Radial marks indicate the two initiation sites. Source 41(b): Ref 50

Along with initiation sites, fatigue ratchet marks can then often be used to identify the type of cyclic loading. For example, reversed bending produces crack initiation on opposite sides (Fig. 40) where maximum stresses alternately occur on opposite sides during each cycle of reversed bending. Likewise, if the cyclic loading is not fully reversed, ratchet marks would be expected on only one side.

Surface rubbing due to crack closure during the compressive part of the cycle may destroy a radial pattern and/or beach marks, but ratchet marks are still usually present to indicate the location of crack initiation. Ratchet marks are distributed in the areas of maximum stress. For example, every point on the circumference of a cylindrical section at a constant distance from a support sees the same nominal maximum stress in rotating bending loading. In the absence of surface imperfections, the nominal stress is constant at every point of the circumference. Therefore, crack initiation at any point on the circumference at that distance from the support is equally likely. Many ratchet marks are then distributed around the circumference. That is a common situation for moderate levels of nominal stress (but not for very high or very low stress levels, which both produce a

smaller number of ratchet marks). Alternatively in reversed or partially reversed bending, a much smaller circumferential arc sees a high stress so the distribution of ratchet marks along the circumference is more limited.

Fatigue Crack Propagation. The relative portions of the fracture surface from fatigue crack propagation and the final overload region depend on the level of applied stress relative to the nominal strength of the material and the surface geometry of the part. The overload region is usually visibly rougher (Fig. 40) than the fatigue region, and the fatigue region is dull and matte. Fatigue crack growth behavior is generally characterized by fracture mechanics in terms of crack growth rates (da/dN) and the applied stress intensity range (ΔK) (Fig. 42). The underlying microscopic processes of fatigue crack growth include (Fig. 8):

- Ductile striations
- "Crystallographic" fatigue—(i.e., fatigue crack growth on specific crystallographic planes)
- Cyclic brittle cleavage
- Cyclic microvoid processes
- Cyclic intergranular processes

Ductile striations are the most commonly observed fatigue feature with intermediate ΔK values (i.e., the power-law regime of the Paris equation where fatigue striations can be analytically correlated to a normalized stress-intensity range, ΔK, as discussed in succeeding paragraphs). However, there is a tendency for some alternative fracture mechanisms to occur at either very low or very high ΔK values. For example, the effect of ΔK values on the occurrence of different fracture appearances is illustrated in Fig. 43 for three classes of materials. The effect of ΔK values on fatigue fracture appearances of a nickel-base alloy (Incoloy X750) is also shown in Fig. 846 of the "Nickel-Base Superalloys: Atlas of Fractographs" in *Fractography*, Volume 12 of *ASM Handbook.*

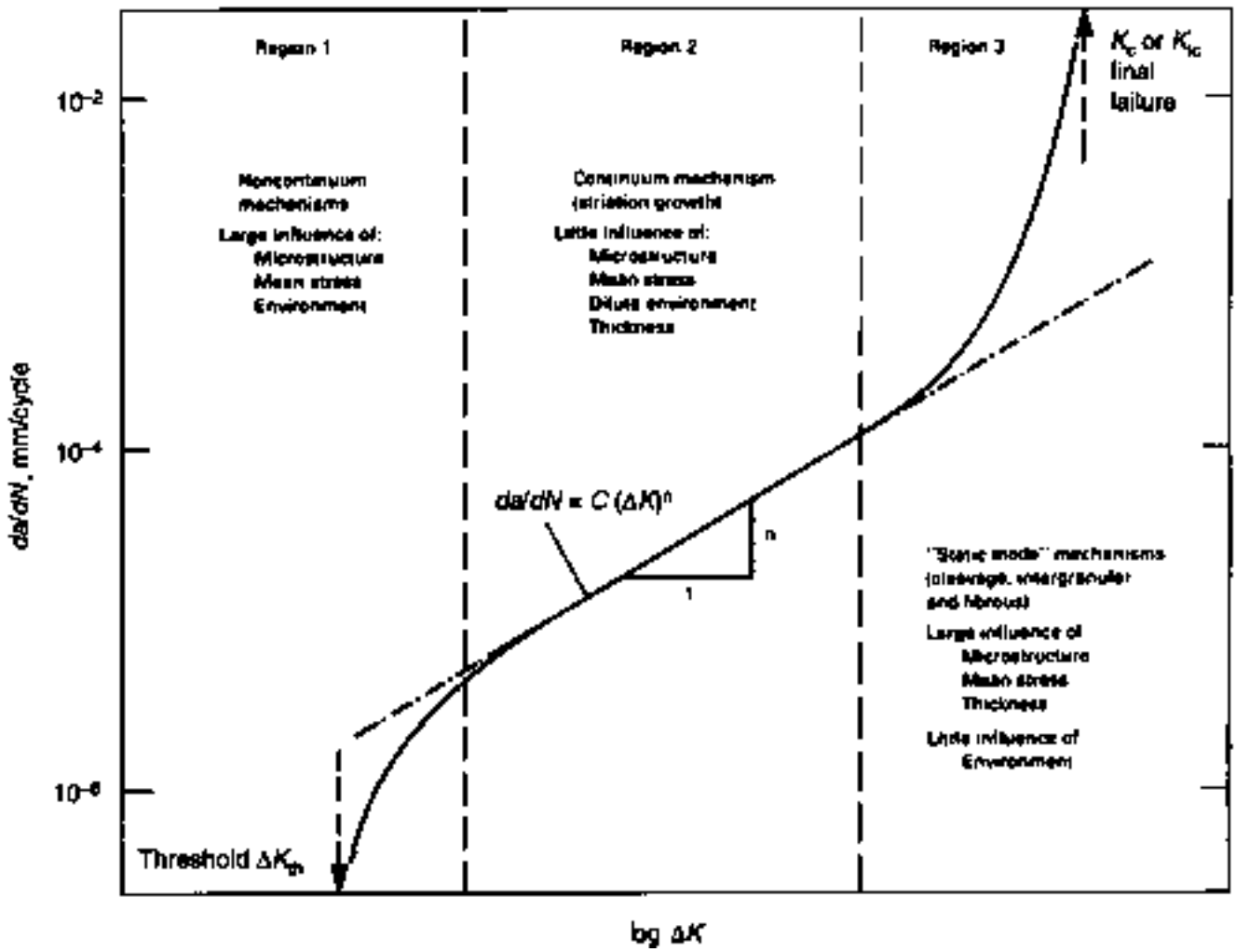


Fig. 42 Fatigue crack propagation rate versus stress intensity factor range. Fatigue striations may be present on the fracture surface for loading in the linear portion of the curve (Paris Law region), and

permit analytical estimations of life to fracture. Just as fracture toughness varies with crack propagation direction in monotonic loading, the Paris Law exponent also depends on crack propagation direction.

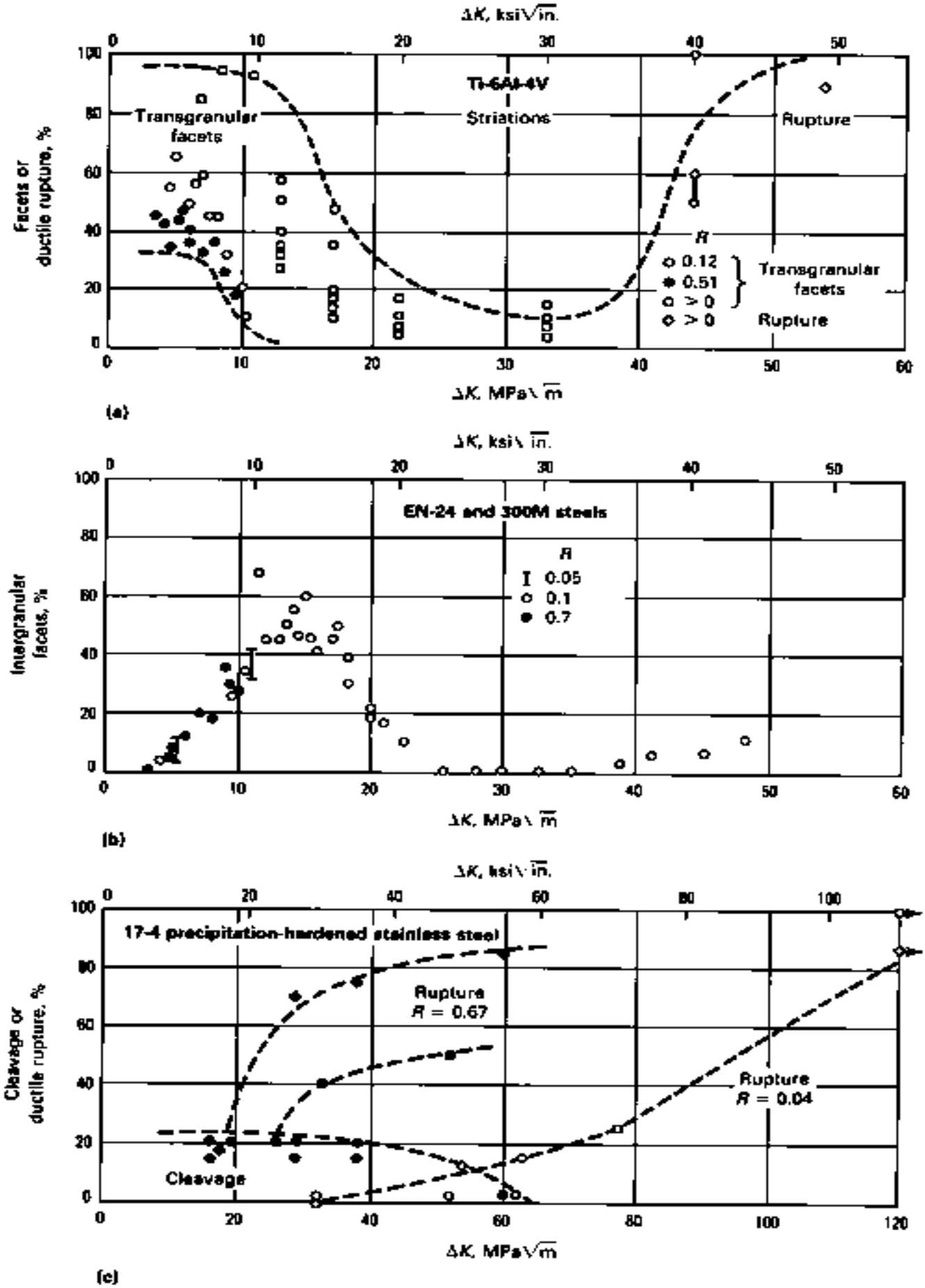


Fig. 43 Effect of Δ*K* on fatigue fracture mechanisms. (a) Alpha-beta titanium alloy. (b) EN-24 and 300 M steels. (c) 17-4 PH stainless steel. Source: Ref 31

The following sections describe the typical microscopic processes of fatigue crack growth for the three regions of the Δ*K*-*da*/*dN* curve (i.e., low, intermediate, and high values of Δ*K*). The microstructure plays a strong role in crack propagation mechanisms in regions I and III. It is often stated to have a minimum role in the Paris law region (II) of the curve, but see the additional information to follow. Mixed transgranular (TG) and IG fractures have been observed, and both ductile and brittle TG mechanisms have been reported. Some of the possible growth mechanisms are shown in Fig. 44. For single-phase materials, the transgranular and intergranular modes shown in Fig. 44(a), 44(b), 44(e), 44(f), 44(g), and 44(i) are possible, if the type of behavior observed in Fig. 43 is present.

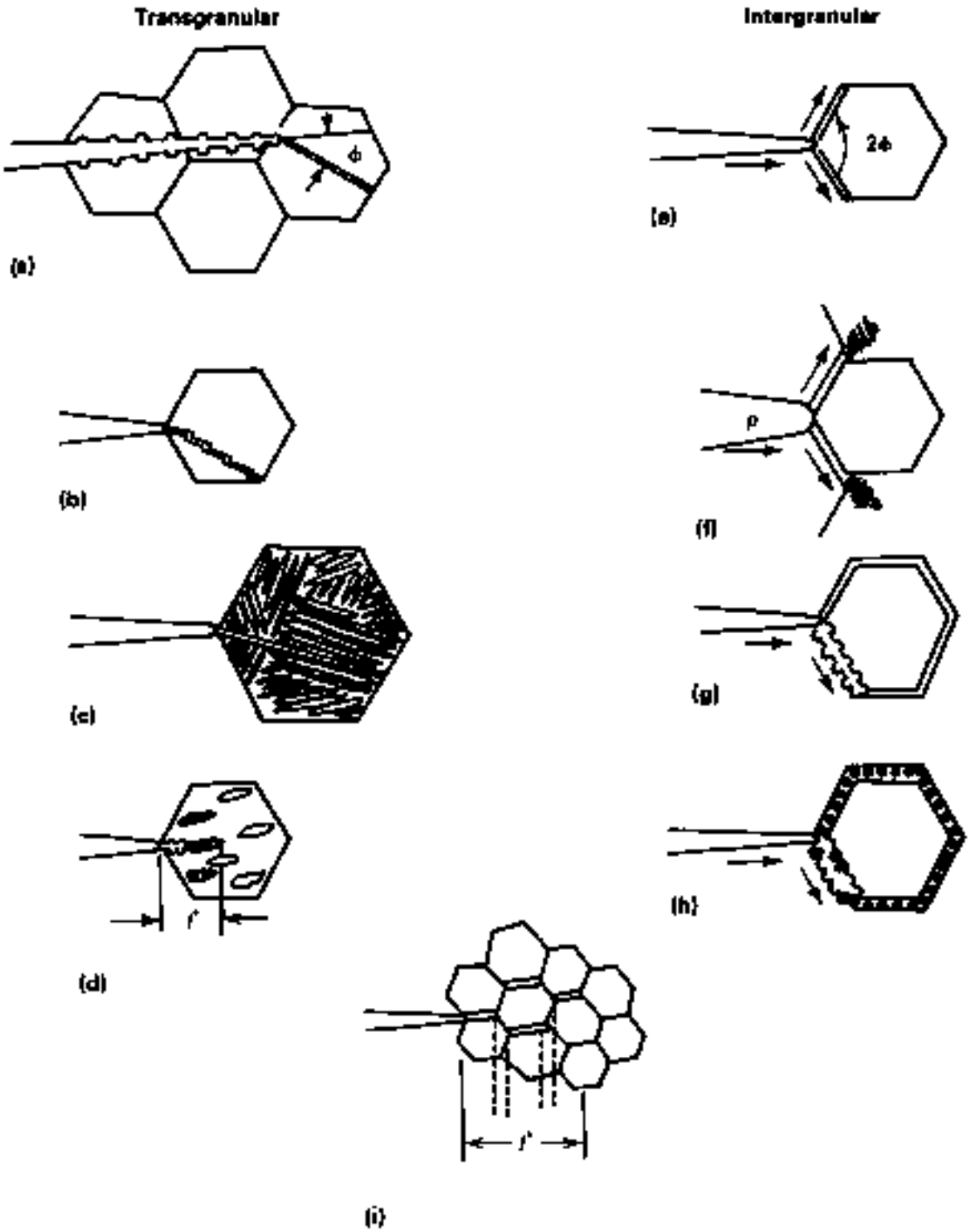


Fig. 44 Types of alternate microscopic fracture modes in fatigue. (a) Ductile striations triggering cleavage. (b) Cyclic cleavage. (c) Alpha-beta interface fracture. (d) Cleavage of alpha in an alpha-beta phase field. (e) Forked intergranular cracks in a hard matrix. (f) Forked intergranular cracks in a soft matrix. (g) Ductile intergranular striations. (h) Particle nucleated ductile intergranular voids. (i) Discontinuous intergranular facets. Source: Ref 31

Dual fracture modes may occur from the different properties of microstructural constituents such as second phases. For example, the optical micrograph in Fig. 45(a) and the fractograph in Fig. 45(b) clearly show this effect in an α-β titanium alloy (Ti-6A1-4V). The coarse alpha platelets tend to promote cyclic cleavage, while a more ductile fatigue process occurs in the Widmanstätten α-β structure between the alpha platelets. Another example of microstructural inhomogenity is regions of shear from deformation twinning. In one case, this prominent deformation led to fracture mechanisms of cyclic ductile decohesion that produced a more faceted fracture Fig. 46 (Ref 51), as described in the following section.

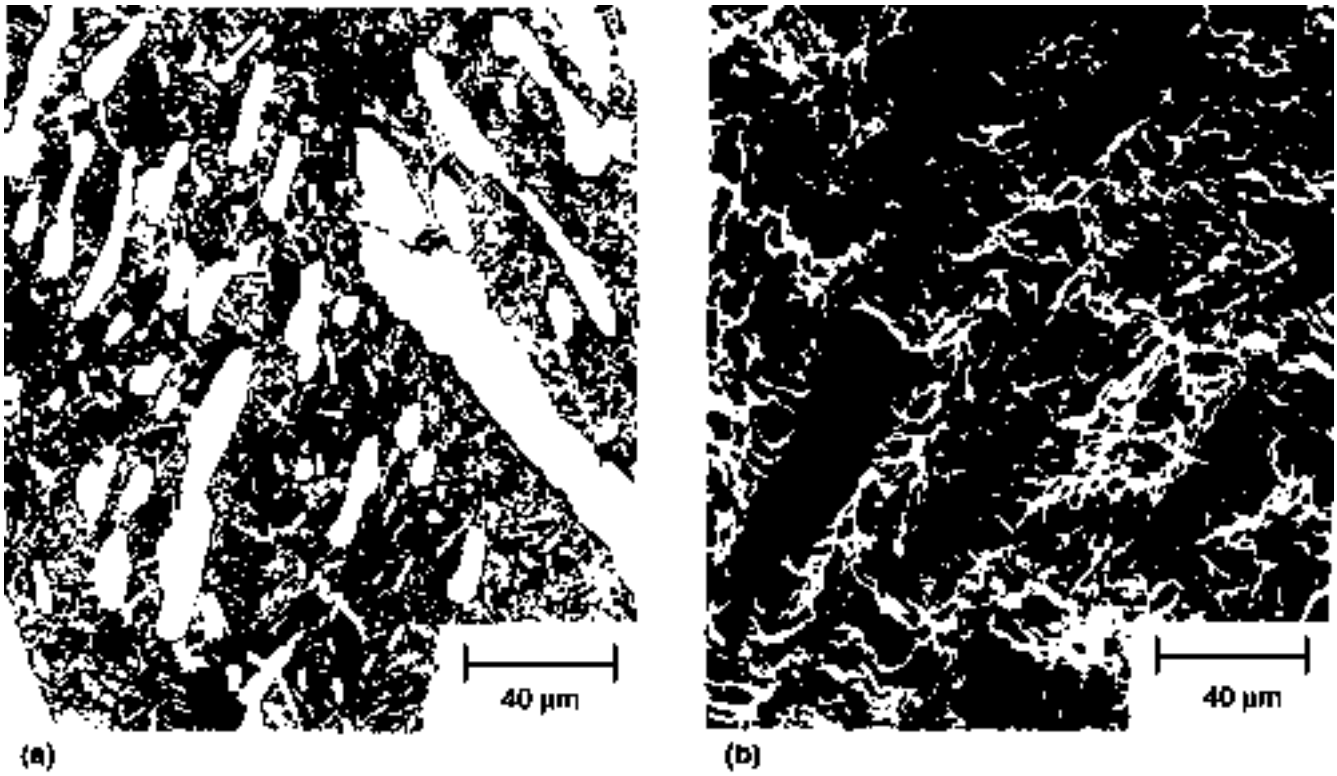


Fig. 45 Second-phase cleavage fracture in Ti-6Al-4V. (a) Light micrograph of polished and etched surface. (b) SEM of fracture surface. Source: Ref 10

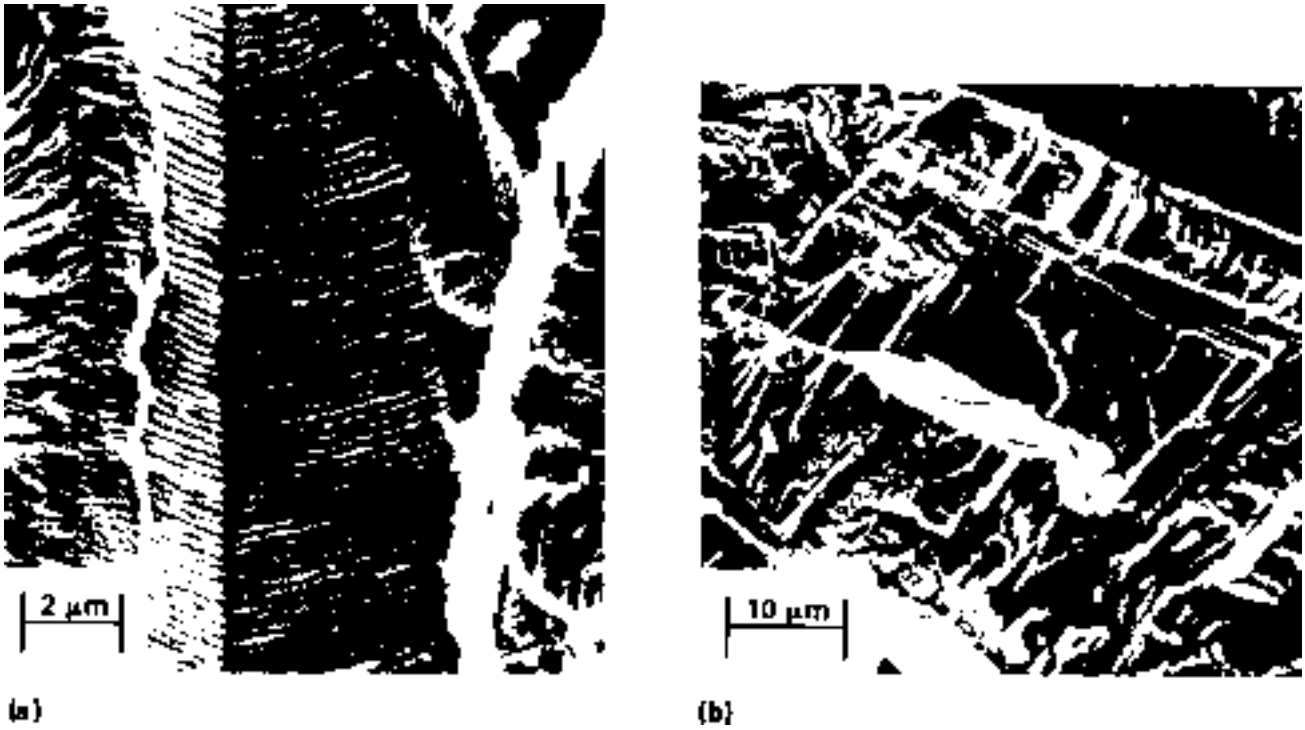


Fig. 46 Scanning electron micrographs of the fracture surfaces of Inconel 718 specimens tested at room temperature. (a) ΔK = 30 MPa$\sqrt{m}$ (27 ksi$\sqrt{in}$), striation spacing • 0.2 μm, and *da/dN* ~ 0.1 μm/cycle. Arrow indicates direction of crack propagation. (b) ΔK = 14 MPa$\sqrt{m}$ (13 ksi$\sqrt{in}$) and *da/dN* = 2 × 10^{-3} μm/cycle. Source: Ref 51

Appearances at Low ΔK. Fatigue fracture of metals with low stacking fault energy, which favors planar slip, can occur over extended distances along specific crystallographic planes, producing a faceted appearance. Metals such as austenitic stainless steel and aluminum fracture in this way at low ΔK values near threshold, but a cobalt-base alloy used in medical implants exhibited this feature in laboratory fatigue testing at all levels of applied load (see Fig. 39 in the article "Fatigue Fracture Appearances" in this Volume). This process has also been referred to as cyclic ductile decohesion or "crystallographic fatigue" due to the faceted appearances. Alternatively, etch-pitting techniques used on fracture surfaces have indicated that low ΔK fracture may occur on a cleavage plane and also result in a faceted appearance. Identification of the fracture plane as a "slip" plane or "cleavage" plane is further complicated by the observation that these two planes are identical in some space lattices (e.g., the hcp lattice) (see below).

Figure 46 is an example that appeared related to slip. Whereas the growth rate above 22 MPa$\sqrt{m}$ (20 ksi$\sqrt{in}$) occurs predominantly by the ductile striation process (Fig. 46a), it is "crystallographic" at lower ΔK values (Fig. 46b), showing a high density of crystallographic facets on {111} planes. As {111} fcc are slip planes, there is some evidence that this is an inhomogeneous cyclic-slip type of decohesion process; other evidence suggests that microstructural transformations (e.g., inhomogeneous shear by deformation twinning) may also be involved (Ref 51). Because mode II crack tip advance can occur along inhomogeneous shear bands, the local slip character of the microstructure may affect asperity contact during the unloading portion of the fatigue cycle. This depends on asperity size, degree of sliding, and the minimum load. Thus, a combination of both microstructure and type of load cycle may affect the effective stress intensity near threshold.

Cyclic brittle cleavage is another important microscopic fatigue mechanism. An example is shown in Fig. 47 (Ref 52), where a stepwise growth occurred on {100} cleavage planes. Similar features have been observed for high-strength, low-alloy steels and medium-strength steels (Ref 53). Such features have been observed both slightly above and below room temperature.

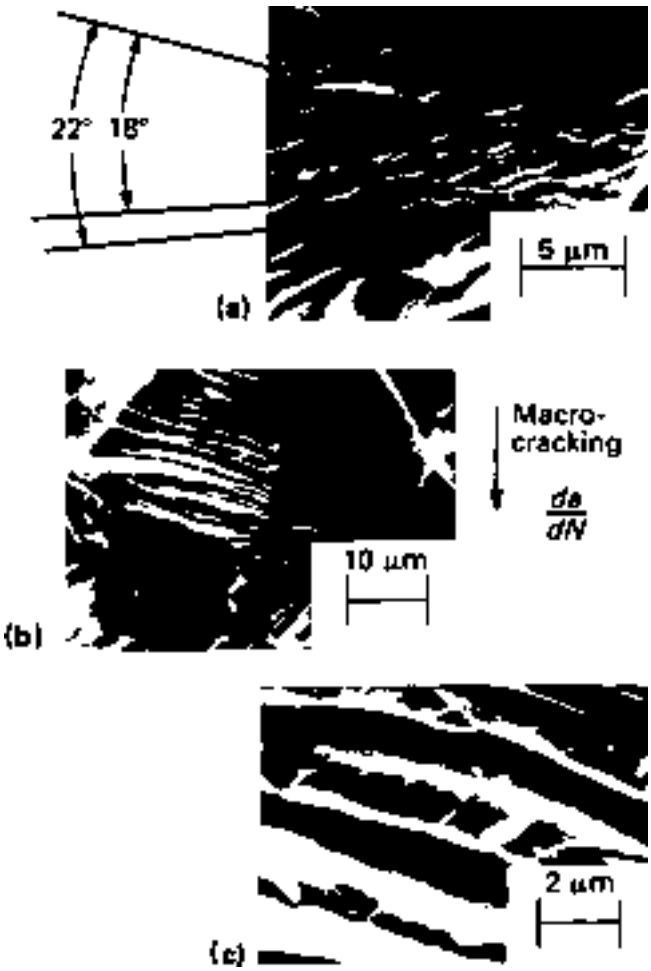


Fig. 47 Cyclic cleavage observations in Fe-4Ni (at.%) at 233 K, ΔK = 17 MPa$\sqrt{m}$ (15.5 ksi$\sqrt{in}$). (a) Jogs formed during cyclic stepping across grain. (b) Large number of cleavage rivers formed to accommodate the twist angle misorientation between two grains. (c) Cleavage striations superimposed on rivers during cyclic crack growth. Source: Ref 52

Cyclic cleavage in Fe-4Si (at.%) is shown in Fig. 48. At 75× magnification (Fig. 48a), this specimen appears to illustrate a brittle overload fracture. The predominant cleavage mode with patches of IG fracture suggests a catastrophic rapid fracture. However, this surface was produced under fatigue loading at 233 K (-40 °C, or -40 °F) at a ΔK of 18.4 MPa$\sqrt{m}$ (16.7 ksi$\sqrt{in}$). The macroscopic growth rate was only 5×10^{-9} m/cycle (1.9×10^{-7} in./cycle). At higher magnifications in Fig. 48(b), clear evidence of brittle fatigue striations on one of the major cleavage steps is visible.

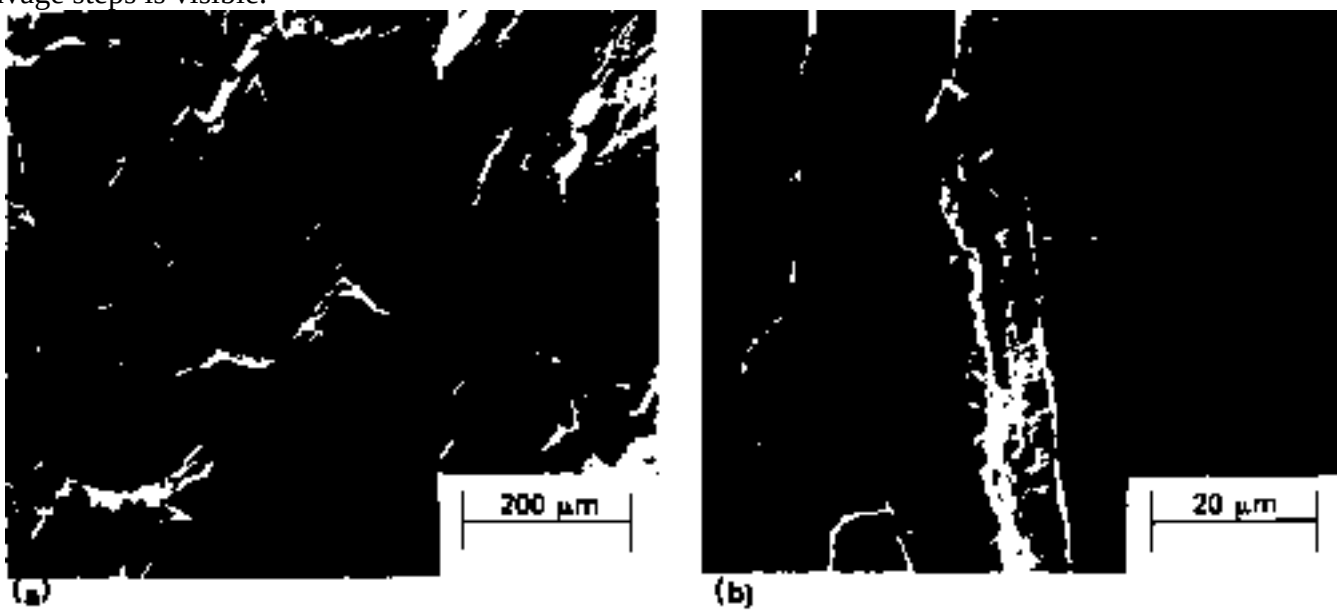


Fig. 48 Scanning electron micrographs of cyclic cleavage of Fe-4Si (at.%) at 233 K, ΔK = 18.4 MPa$\sqrt{m}$ (16.7 ksi$\sqrt{in}$). (a) Overload cleavage appearance at 75×. (b) Magnifications at 750× shows brittle striations on large cleavage river

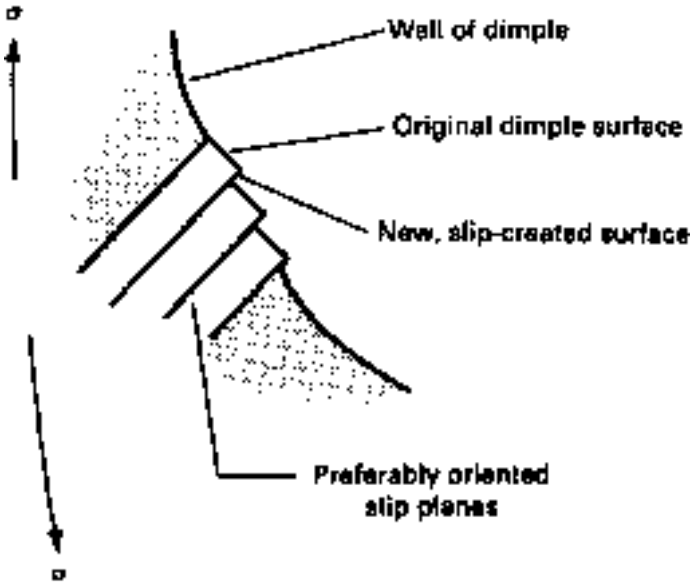
Appearances at Intermediate ΔK (Beach Marks and Striations). Beach marks and striations are, respectively, macroscale and microscale fractographic features associated with cyclic loading of metallic materials. Other features are ratchet marks and the change in surface roughness from fatigue region to overload region as previously discussed. Beach marks also may be visible in polymeric materials under visual or low-magnification examination.
Beach marks arise from two unrelated sources:

- Selective oxidation at the tip of the crack in the component when it is in the unloaded condition
- Rapid changes in ΔK

Beach marks created by oxidation at the crack tip are not expected for constant amplitude loading if the loading is uninterrupted. Beach marks from oxidation are created when the loading is intermittent.
When beach marks represent changes in crack growth rates, they may be associated with crack arrest and/or a discontinuous change in ΔK during the loading cycle. At the microscale, a sudden change in striation spacing indicates either a sudden large increase in ΔK to cause the crack increment Δa, or a sudden large decrease in ΔK. If the resolution of the microscope is inadequate, the much smaller and closer spaced striations at low ΔK may not be visible. One striation corresponds to one cycle. Whether a given cycle created a visible striation is a matter of resolution of the fracture surface features as well as crack tip stress field closure effects.
However, it is also well known that beach marks and striations are not always evident on the surfaces of fatigue fractures. Both macroscale (beach marks) and microscale (striations) indicators are less well defined or absent in cast products, and they may be obliterated in situations where the crack closes on itself during the loading cycle. For example, Fig. 49 is an example of clear *beach marks* in six regions of fatigue cracking caused by multiple initiation sites that are expected under the conditions of rotating bending fatigue. In Fig. 40, however, beach marks are less visible, as some rubbing on both sides due to crack closure has obliterated beach marks created during early stages of crack growth. If the crack closes on itself during cyclic loading (as it must, if the load goes compressive during a part of the cycle), surface rubbing may destroy a radial pattern and/or beach marks.

Fig. 49 Steel shaft (AISI 4817) carburized and hardened to HRC 60. Component loaded in rotating bending. Fracture initiation in six fillet areas around the periphery. Six initiation sites are visible as evidenced by multiple sets of beach marks (curved black lines). Source: Ref 29

Moreover, in some cases other fractographic features can look similar to fatigue striations. An example is wavy slip on an inclusion interface surface (Fig. 50a, b). Another example is when stress-corrosion cracking produces features that look like beach marks (e.g., see Fig. 16 in the article “Fatigue Failures”). The term “striation” is also used to describe other fractographic features that are not associated with fatigue loading, and perhaps a better practice in that case is to describe such a surface as a “striated surface” or “striated structure” and to use the term “fatigue striations” when cyclic loading is involved. Etching of the fracture surface can reveal whether the morphology of a second phase is responsible for the striated features (Fig. 51). However, since etching is irreversibly destructive, it must be one of the final steps in the analysis and it may be necessary to obtain permission.



(a)

(b)

Fig. 50 Wavy slip. (a) Schematic of slip band extrusions in the matrix at a matrix-inclusion interface or at the surface of a notch. (b) Wavy slip lines (arrow) in oxygen-free, high-conductivity copper

Fig. 51 Portion of a fracture surface in a Ti-6Al-4V alloy. (a) Low magnification (35.5×) image showing a striated structure in the facet near the center of the photograph. (b) Higher magnification (1420×) view of the facet. (c) Same region as in (b) after etching with 0.5% hydrofluoric. Striated structure runs from 7 to 1 o'clock and is not due to fatigue loading but rather to second-phase beta needles as revealed by etching. Source: Ref 21

Fatigue striations are most commonly seen when the microstructure is single phase and/or contains only a small volume fraction of second phases or constituents. Then the dominant crack propagation path is in a large volume-fraction matrix phase and is unaffected by crack nucleation at second phase interfaces or by cleavage cracking in second phases. Thus, fatigue striations may be revealed on the fracture surface of a hot rolled ferritic low carbon (0.1% C) steel, (Fig. 52) (Ref 54) but not in a structural grade steel where a higher carbon content results in volume fractions of pearlite that may easily approach more than 50%.

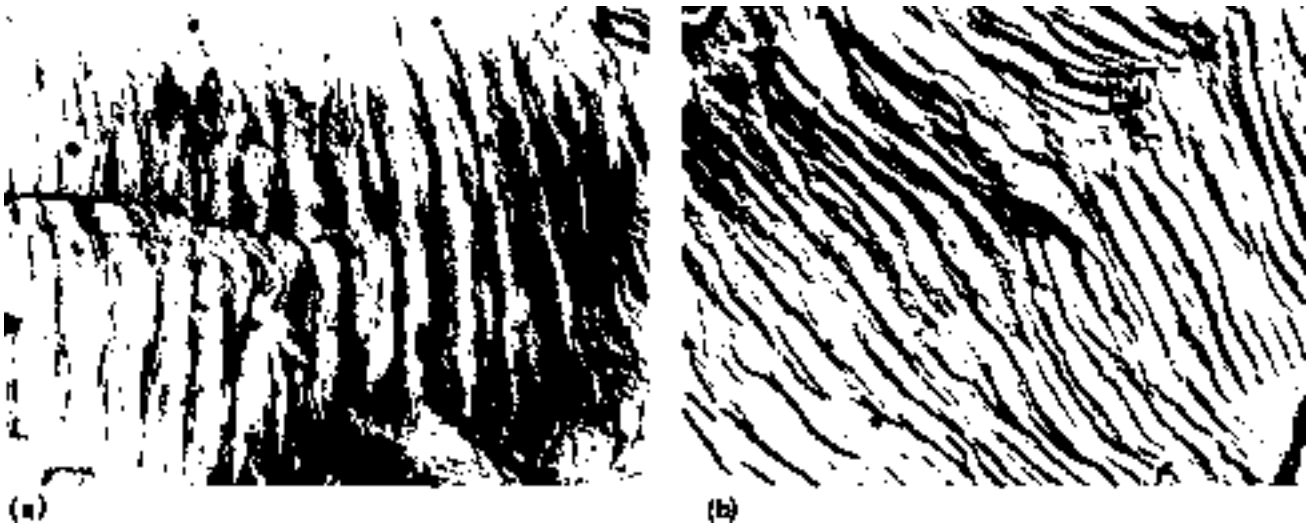


Fig. 52 Fatigue striations in a vanadium HSLA steel. (a) L-T orientation; ΔK = 32.3–34.3 MPa $\sqrt{m}$. da/dN = 3.3–3.8 × 10^{-5} cm/cycle. (b) T-L orientaton; ΔK = 24.3–25.5 MPa $\sqrt{m}$, da/dN = 9.4–11.2 × 10^{-6} cm/cycle. Source: Ref 54

Fatigue striations are most prominent in the "Paris law" regime (where the crack propagation rate versus change in stress intensity factor is linear on a log-log plot—i.e., $(da/dn) = C(\Delta K)^n$. The morphology of the striations changes with changes in ΔK and can disappear at high values and low values of ΔK because of a change in the crack propagation mechanism. At very high values of ΔK, it is common to see dimpled rupture. At low values of ΔK, a flatter brittle-like surface may be evident, as previously noted.

In summary, the presence of beach marks and/or striations typically identifies cyclic loading conditions, but their absence does not necessarily indicate monotonic loading conditions. In addition, other (nonfatigue) fractographic features also may appear like fatigue beach marks and striations, as previously noted. This

emphasizes the necessity of correlation of both macroscopic and microscopic fractographic features in conjunction with microstructural examination to identify a unique cause for failure. Unfortunately, damage to the fracture surface often occurs at the microscale (damage during failure, damage during improper storage, damage due to improper handling), sometimes making it difficult to obtain microscopic fractographic information.

Appearances at High ΔK. Cyclic microvoid coalescence growth typically occurs at high ΔK, but one exception is shown in Fig. 43(c). In this case, a high percentage of ductile rupture is plotted at high R values (stress ratio) where K_{max} approached K_c values. The assumption on the size of the plastic region for the K-zone and the crack length relative to the size of the K-zone are probably violated in this region of the curve so that LEFM procedures are not justified. An example of a fracture surface from cyclic microvoid growth at high ΔK is shown in Fig. 53 from an ultrahigh-strength D-6 AC steel.

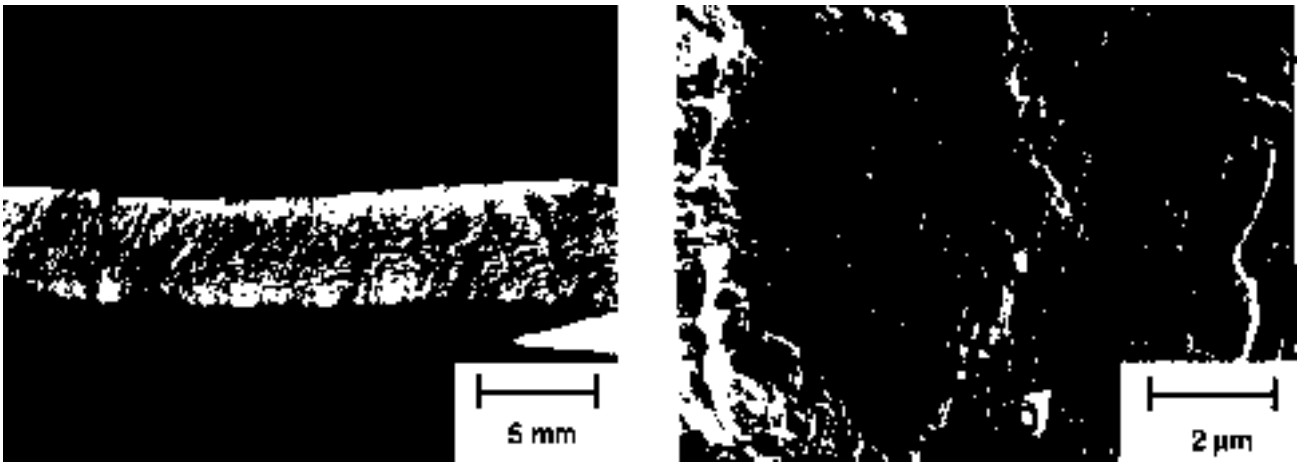


Fig. 53 High-cycle fatigue in D-6 AC steel at ΔK > 60 MPa$\sqrt{m}$ (54.6 ksi$\sqrt{in}$. (a) Macroscopic view of rapid fatigue growth under plane-stress conditions. (b) High-quality TEM replica showing elongated shear dimples from fatigue region

References cited in this section

10. *Fractography, Microscopic Cracking Processes,* STP 600, C.D. Beachem and W.R. Warke, Ed. ASTM, 1976

21. G. Powell, Shu-hong Cheng, C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys,* Battelle Press, Columbus, 1992

29. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

31. W. Gerberich, Microstructure and Fracture, *Mechanical Testing,* Vol 8, *Metals Handbook,* 9th ed., Mechanical Testing, American Society for Metals, 1985, p 476–491

50. D.A. Bales, Effects of Decarburization on the Fatigue Life of SAE 1042 Steel, master's thesis, University of Tennessee, 1977

51. M. Clavel and A. Pineau, Frequency and Waveform Effects on the Fatigue Crack Growth Behavior of Alloy 718 at 298 °K and 823 °K, *Metall. Trans. A,* Vol 9 (No. 4), April 1978, p 471–480

52. W.W. Gerberich and K. Jatavallabhula, Quantitative Fractography and Dislocation Interpretations of the Cyclic Cleavage Crack Growth Process, *Acta Metall.,* Vol 31, 1983, p 241–255

53. W. Yu, K. Esaklul, and W.W. Gerberich, *Metall. Trans A,* Vol 15A, 1984, p 889–900

54. R.W. Hertzberg and R.H. Goodenow, “Fracture Toughness and Fatigue Crack Propagation in Hot-Rolled Micro Alloyed Steel”, *Micro Alloying, '75 Proceedings,* Union Carbide Corp. p 503

W.T. Becker, University of Tennessee, Emeritus; S. Lampman, ASM International

Appendix: Modeling with Fracture Mechanics

Introduction

Modeling of the fracture process has been approached at both the macro- and microscales. When models reproduce experimental behavior, they become tools to use for quantitative evaluation of fracture in failure analysis. Multiple approaches have been used; some are based on continuum mechanics to predict behavior of the bulk assembly (without assumptions of the microscopic scale details and operative mechanisms of crack initiation and propagation), while others methods consider either a submicroscopic or microscopic description of structure to provide criteria to predict crack propagation and fracture based on microscopic processes. For example, dislocation theory has been used to develop models for the formation of crack nuclei. There also has been considerable interest in developing models for the strain at fracture of ductile materials based on the common fracture mechanism of microvoid coalescence. The latter is discussed in more detail in the Appendix of the article in this Section, “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals.”

Methods based on continuum mechanics are appealing in some ways, as the material is taken as a “black box” or continuum without specific assumptions about fundamental processes at the microstructural level (i.e., the micro-mechanics associated with fracture initiation mechanisms such as debonding at inclusions or cracking of inclusions). Such approaches are thus appealing in some ways, because no materials background is required for the analysis other than the ability to recognize and identify fractographic features. For example, fracture mechanics is an important example in the use of continuum mechanics in modeling and predicting of fracture. This includes the use of fracture mechanics to predict subcritical crack growths rates and/or the conditions for the onset of final (i.e., critical) fracture from a preexisting crack or cracklike discontinuity in a bulk assembly. Microstructural details are not required in such an analysis of critical fracture, although microstructural details may affect results, as described subsequently. Microstructural conditions may also be a factor in the analysis of subcritical (fatigue) crack growth, as described in the last section in this article, “Fatigue Crack Growth Rates.”

Crack Initiation

Cracks initiate at some region of stress concentration, where loading conditions cause the *local stress* to exceed the *local strength* of the material. Regions of stress concentrations may be due to macroscopic discontinuities (a geometric notch or other change in cross section) and/or microscopic discontinuities (an inclusion or second phase). Some imperfections introduced during primary or secondary metal working cause a local elevation of the stress, and therefore are common locations of crack initiation. Fabrication imperfections are discussed in more detail in the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals.”

The crack tip is a region of high stress concentration and in a real material having some ductility, the crack blunts due to plastic flow, thus redistributing the stress and causing a plastic zone to be created. Crack initiation by a ductile process thus involves some degree of plastic deformation, and the tip of a crack is a region with a plastic zone. The shape of the plastic zone in turn depends on how and if the material strain hardens.

On a macroscopic scale, fracture mechanics in conjunction with fractographic information has led to three different models for the macroscopic characterization of cracks in metals, polymers, and ceramics (Fig. 54) (Ref 55). Strain hardening in metallic materials is greater than in ductile polymeric materials in the cold drawing regime, leading to different shapes of the plastic zones. Ceramics are brittle, and if they do not contain

a distribution of microcracks prior to loading, cracks are created on loading. Under normal circumstances (the exception being a very large hydrostatic pressure), ceramic materials are inherently brittle, and it is usually assumed that these materials contain a random distribution of small cracks due to either prior processing or created by loading prior to net-section fracture.

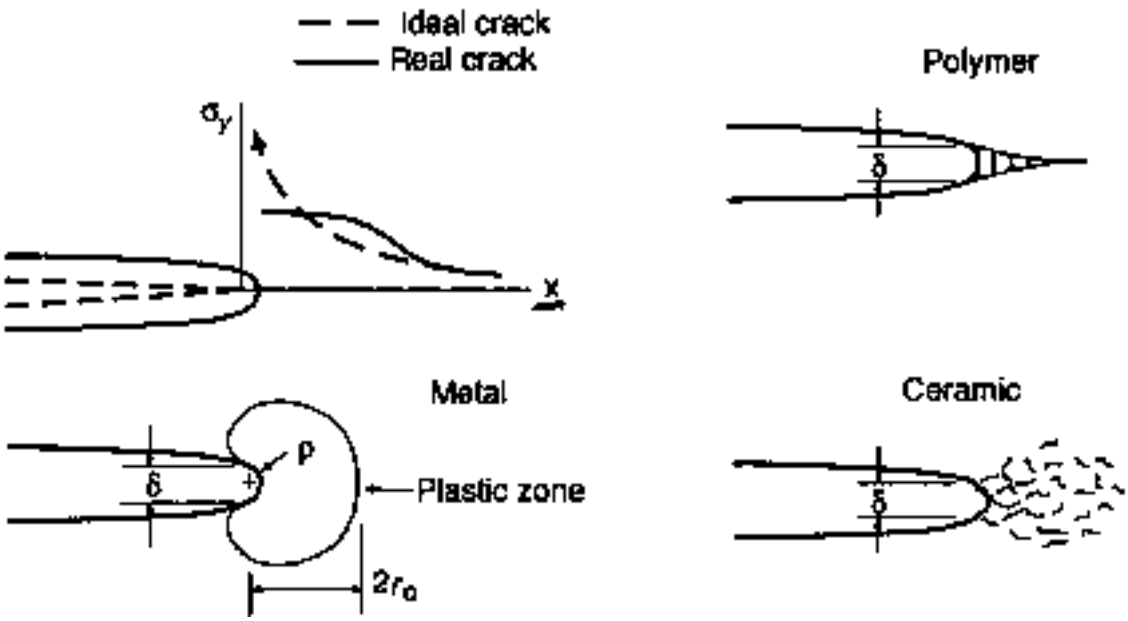


Fig. 54 Near crack-tip deformation and fracture processes. The stress singularity at the free surface of a crack is relieved by crack blunting and plastic flow or by microcracking, Shape of the plastic zone depends on the extent of strain hardening in the plastic zone. ρ, tip radius of the ellipse; δ, crack tip opening displacement. Source: Ref 55

Damage modeling has been applied to both ductile and brittle fracture in metals and ceramics, and strip necking has been used to model both metals and polymers. An introduction to damage modeling can be found in Ref 40. Damage mechanics (based on submicroscopic or microscopic description of structure to provide criteria to predict crack propagation and fracture based on microscopic processes) is a topic of ongoing interest that involves micro to macroscale solutions based on micro-mechanics and micro-fracture mechanics. Damage mechanics is not discussed here.

Fracture Mechanics of Crack Growth (Monotonic Loads)

As previously introduced in the article “Stress Analysis and Fracture Mechanics” in this Volume, fracture mechanics provides a quantitative way to predict the onset of rapid fracture for a stressed part with a crack. This is accomplished via Irwin's expression for fracture toughness:

$$K = s\sqrt{p\,a}[\,f(a/W)] \qquad \text{(Eq 1)}$$

Where K is the stress intensity factor, a is the crack length, $f(a/W)$ = geometrical correction factor for the loading conditions, part geometry, and crack length.

This concept is of importance in determination of root cause because it permits quantitative assessment of materials properties (strength and toughness) with imperfections to determine if the imperfection is in fact a defect. Fracture mechanic concepts are also helpful in identifying unobvious sources of stress such as thermal loading. When known loads cannot explain the calculated level of stress using these techniques, high stresses due to transient thermal gradients may have been present.

Any quantitative estimations of toughness or stress level require the geometric correction factor $f(a/W)$ to be known. A few common geometric factors are given in the article in this Volume “Stress Analysis and Fracture Mechanics” with reference to more complete compilations. General treatments of fracture mechanics are available in multiple sources including Dowling (Ref 55), Anderson (Ref 56), and an applications-oriented text

by Broek (Ref 34). A less comprehensive treatment of fracture mechanics but with more of a materials orientation is available in Hertzberg (Ref 25).

Plane-Strain (Linear Elastic) Conditions. If there are no shear lips on the failed component, or if they are small in comparison to the flat fracture region, plane strain loading conditions can be assumed so that the value of K at fracture is the plane strain fracture toughness (K_{Ic}). Equation 1 may then be utilized in three different ways depending on what is assumed for the independent variable: the critical value of K at fracture, the stress at fracture, or the crack length at fracture. Since the crack length at the time of overload can be measured on the failed component, either the stress or the fracture toughness may be calculated. Consistency must exist between published values of K_{Ic} and the yield stress or design strength and the measured crack length at failure.

For example, assume that the published values of yield strength and fracture toughness for a ferrous alloy in plate form are YS = 75 ksi, $K_{IC} = 100 Ksi\sqrt{in}$, that the geometric correction factor is 1.12, and that the measured crack length at overload is 1.0 inches. Assume that the design stress was $\frac{1}{3}$ of yield or 25 ksi. Then:

$$K = s\sqrt{p\,a}[f(a/W)] \qquad \text{(Eq 2)}$$

Substituting values for the crack length and the correction factor:

$$K = \sigma(1.985) \qquad \text{(Eq 3)}$$

If the stress is 25 ksi, the calculated toughness is 49.6 ksi, and if the toughness is 100 ksi, the calculated stress at the time of fracture was 50.4 ksi, so that the results are inconsistent. A hardness check can indirectly provide an approximate value of the yield strength, and an examination of the microstructure will usually show whether it is consistent with the toughness. The microstructure should be checked to determine the direction of crack propagation since toughness can be highly anisotropic. Literature values of fracture toughness values, unless otherwise specified, are for the longitudinal-transverse (L-T) orientation, where this two-letter designation system is per ASTM E 399 to indicate the direction in which the load is applied (L for longitudinal or rolling direction) and the direction in which the crack propagates (T for transverse). Considerable anisotropy may exist for crack propagation in the S, T, and L directions due to banding (compositional segregation) and fibering from the stringers (elongated oxides and inclusions that form during metalworking (see the article “Failures Related to Metalworking” in this Volume). Propagation in the rolling direction gives lower values of K_{Ic}. Consequently, if the crack propagated in the rolling direction, the toughness will be lower than the literature value, and may be an explanation for inconsistency in some calculations. Likewise, the values quoted from literature are typical for a hot-rolled pearlitic microstructure; examination of the microstructure may reveal any other possible anomalies, such as coarse pearlite, abnormally large grain size, etc. that can affect plane-strain toughness.

The microstructure should always be checked since it may indicate that the material has been processed differently from that inherent to the material selected in the design process. Recognition of a hot-rolled microstructure in a presumably quenched and tempered component is obvious. However, if the material is presumably quenched and tempered, it may be difficult to assess whether the processing was done correctly. Some deviations are obvious, such as large amounts of free ferrite, but if the total microstructure contains cartides from tempering, it may not be possible to determine the presence, for example, of temper embrittlement (although help can be provided by microscale fractographic examination since temper embrittlement results in IG fracture.).

Calculation Based on Shear Lip Width. The validity/accuracy of the use of plane-strain fracture toughness values in Eq 1 depends on the presence of minimal shear lips on the failed component, such that plane strain loading conditions may be assumed. It may still be possible to make an estimate of the stress intensity at fracture (and therefore, the applied stress) if shear lips are present. Hertzberg (Fig. 55) has suggested using shear lip width on plate specimens to estimate the stress responsible for fracture. From fracture mechanics, the radius of the plastic zone at the crack tip is:

$$r_y = \frac{1}{2p}\left(\frac{K}{s_g}\right)^2 \qquad \text{(Eq 4)}$$

From Fig. 55, the shear lip depth D is of the order of the shear lip radius because of plane stress conditions prevailing at the free surface and because the angle of the plane of maximum shear stress is 45°. Then:

$$D = \frac{1}{2p}\left(\frac{s^2 a\left[f(a/W)^2\right]}{s_g^2}\right)^2 \qquad \text{(Eq 5)}$$

where σ is the nominal stress, σ_Y is the yield stress, ***a*** is the crack length at failure, and ***f***(***a***/***W***) is the geometric correction factor. Hertzberg comments that this approach is approximate and seems to work best for high-strength aluminum alloys and some alloy steels. For low-strength steels, the model gives too low an estimate of the ***K*** prevailing at fracture; that is, that the lip is narrower than that expected from Eq 5.

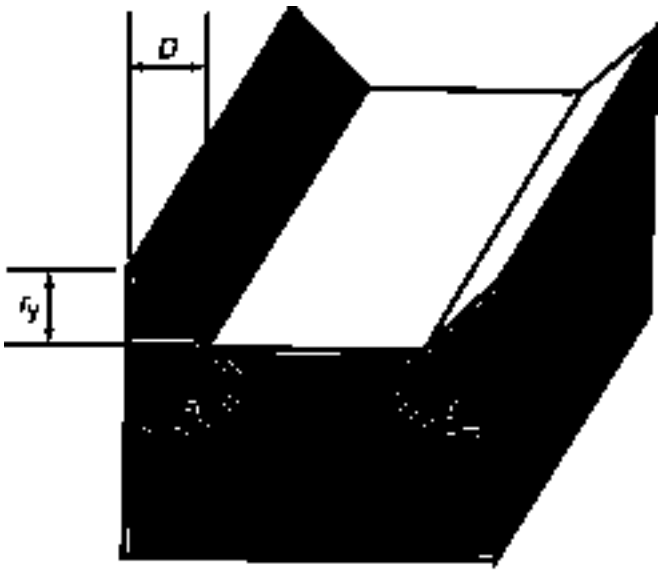


Fig. 55 Correlation of shear lip width with fracture toughness. The depth of the shear lip (*D*) is related to the plane-stress plastic zone size and then to the fracture toughness. See text for discussion Source: Ref 25

Crack Blunting and Stretch Zones. For many loading conditions, a preexisting macroscopic cracklike defect blunts as the load increases, and a new crack forms by ductile tearing and creates a ***stretch zone.*** This is true even when the preexisting crack is created by cyclic loading. Continued increase in loading may then cause the crack propagation mechanism to change to cleavage or quasi-cleavage. The length of the stretch zone δ correlates with ***J***-integral fracture toughness normalized by the elastic modulus as shown in Fig. 56 (Ref 25).

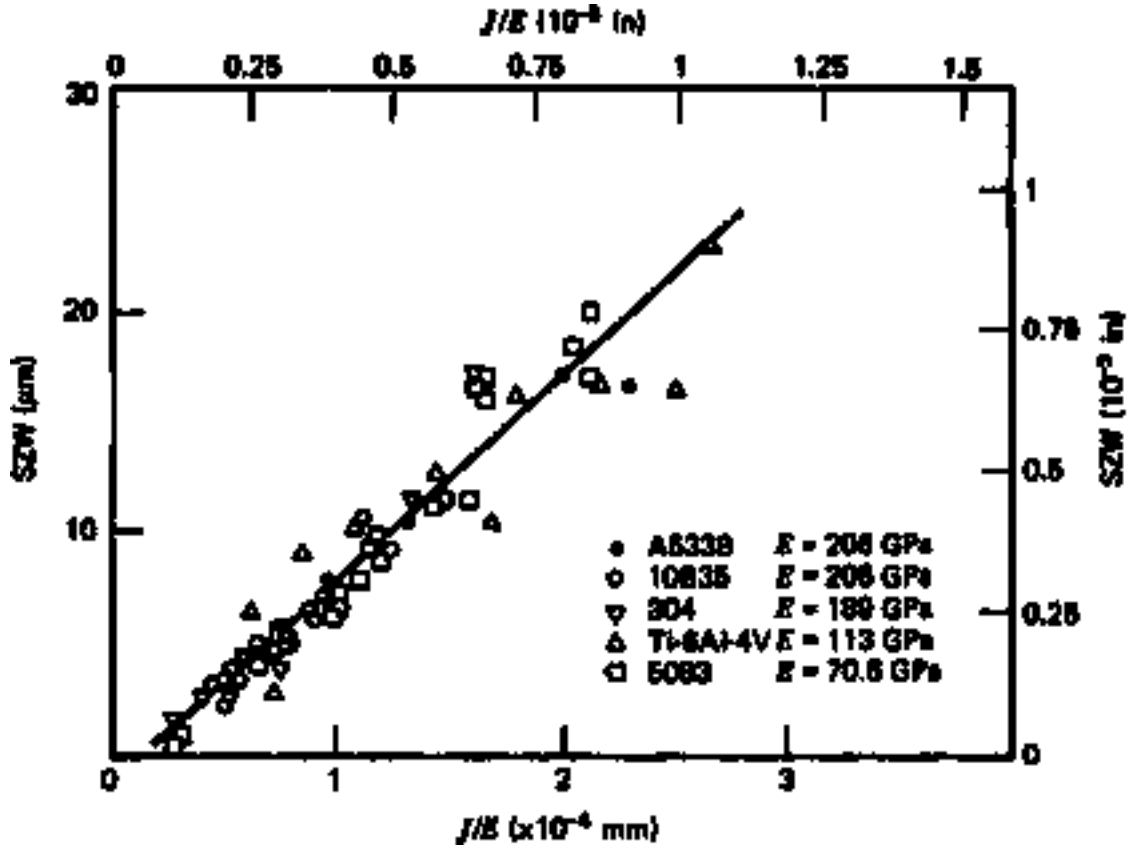


Fig. 56 The correlation between the stretch zone width (SZW or δ) and fracture toughness normalized by the elastic modulus, *E*. Source: Ref 25

The appearance of the stretch zone is shown in Fig. 57. Various workers have proposed models to utilize the stretch-zone width to determine toughness and stress levels (Ref 34, 25). A quantitative assessment of toughness and stress level based on the length of the stretch zone has been proposed. By measuring the length of the stretch zone and assuming the relationship between the length of the stretch zone and the blunted preexisting crack tip radius, it is possible to estimate the fracture toughness or stress level that caused the crack to grow (Ref 34).

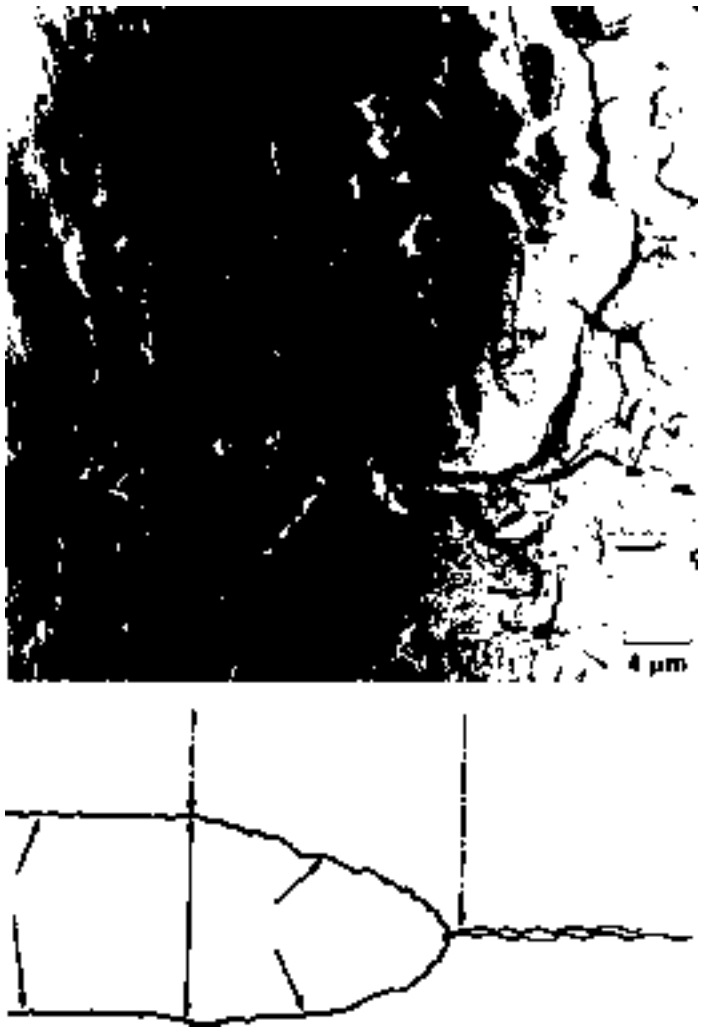


Fig. 57 The plastic stretched zone at the tip of a macroscale crack. Source: Ref 28

Accurate calculations require that the true length of the stretch zone be determined, not the projected length. That the length of this zone varies considerably across the crack front so that determination of toughness should be treated with caution has also been indicated. The crack tip opening displacement (CTOD) is related to the fracture toughness, and the CTOD is assumed to be $\frac{1}{2}$ of the stretch zone length. Then:

$$G_c = \frac{K_c^2}{E} = s_y(CTOD) \qquad \text{(Eq 6)}$$

or

$$J_c = \sigma_y(\text{CTOD}) \qquad \text{(Eq 7)}$$

Fatigue Crack Growth Rates

Fatigue Crack Growth Rates at Intermediate ΔK Values. As previously noted, fatigue crack growth rates (da/dN) at intermediate levels of ΔK (region 2 in Fig. 42) are described by a power law (the Paris equation) where:

$$\frac{da}{dN} = C(\Delta k)^n \qquad \text{(Eq 8)}$$

The Paris equation is only valid for a given stress ratio but can be generalized to other fatigue stress ratios by the Walker relation (see the section “Fracture Mechanics” in the article “Stress Analysis and Fracture Mechanics”).

The constants C and n in the Paris equation are generally available in the literature for many alloys (see for example Ref 55). It has been noted that microstructure has little influence on fatigue crack growth rates in stage 2. Thus generic constants for several alloy systems (structural steels, titanium alloys, and aluminium alloys) have been reported (e.g., Ref 57), and fatigue crack growth rates versus ΔK fall within a relatively narrow scatter band for alloys with a given base microstructure (e.g., see the article "Fatigue Failures" for da/dN - ΔK for ferrite-pearlite steels and martensitic steels).

However, the assumption about the insensitivity of the Paris exponent to microstructural conditions may require closer examination. Just as K_{Ic} is sensitive to orientation in fabricated material (because of banding and fibering), so are the Paris constants. Data in ASTM STP 794 (Ref 58) show that the constants can vary considerably with orientation and deoxidation practice. For the steels, calcium ladle additions reduce the degree of anisotropy to raise toughness in the through-thickness direction and also, in general, decrease the Paris exponent.

The Bates-Clark approximation (Ref 59) is another procedure to utilize the form of Eq 8 where:

$$\frac{da}{dN} = 6\left(\frac{\Delta K}{E}\right)^2 \qquad \text{(Eq 9)}$$

The analytical procedure to estimate life relies on obtaining the initial crack length possibly visible as an oxidized fingernail, and the crack length at overload from a macroscale fractograph. Additional assumptions about cycles to crack initiation must be made, because the Paris law assumes the presence of a cracklike imperfection (i.e., the crack has initiated). It is often assumed that for smooth-surface components, the initiation life is long relative to the propagation life; vice-versa, crack propagation is long relative to initiation for components containing a cracklike stress riser. Stress-life or strain-life data may be used to predict initiation life (not discussed here).

Then, rearranging Eq 8:

$$\int_{N=N_{initiation}}^{N=N_{failure}} dN = \int_{a_0}^{a_{crit}} \frac{da}{A(\Delta K)^m} = \frac{1}{A}\int_{a_0}^{a_{crit}} \frac{da}{\left[(\Delta\sigma)\left(\sqrt{\pi a}\right) f(a/W)^m\right]} \qquad \text{(Eq 10)}$$

Measurements of the initial and overload crack lengths set the integration limits on the right hand side. Finding a solution to the calculation then depends on several factors.

Geometric correction depends on crack length, so that if the crack grows appreciably prior to overload, it is inappropriate to assume that this term remains constant. That significantly complicates the integration since the mathematical form of $f(a/W)$ usually does not lend itself to integration. In such an instance, it is better to numerically integrate a solution using a spreadsheet. An increment of load cycles is taken as the independent variable, and an analytical form of the correction factor can be used in the summation. The increment size should be selected so that a change in the increment size does not change the life estimation. Cycle summation is halted when the summed crack length exceeds the value measured on the fractured specimen.

Striation Spacing. If fatigue striations are present on the fracture surface, the striation spacing at multiple crack lengths can be measured. The striation spacing is substituted into Eq 8 to calculate a stress intensity range, and from that the stress range is calculated using the measured crack length. Multiple measurements of the striation spacing are required in most cases because of the change in the geometric correction factor with crack length that changes the calculated K value and increases the incremental crack spacing. An inconsistency in calculated ΔK for various crack lengths indicates a poor estimation of the stress range. This might occur for nonconstant stress amplitude during the loading history, but that should be known by a careful examination of the striation spacing obtained from the microscale examination.

Results of this calculation are subject to several experimental variables. The true striation spacing is measured only if the surface containing the striations is viewed normal to the surface. Additionally, the crack propagation direction at the microscale is often different from the macroscale crack propagation direction, as affected by grain orientation. Experimental evidence (Ref 60) indicates that macroscale crack propagation rates are faster than indicated by microscale measurements.

References cited in this section

25. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* 4th ed., John Wiley, 1996

28. D. Broek, *Fatigue Design Handbook,* AE-10, 2nd ed., Society of Automotive Engineers, 1988

34. D. Broek, *Practical Use of Fracture Mechanics,* Kluwer, 1988, p 431, 436

40. J. D. Embry, Damage Accumulation in Fracture Processes, *Modeling the Deformation of Crystalline Solids,* T.C. Lowe, A.D. Rollett, P.S. Follansbee, and G.S. Daehn, Ed., TMS, 1991

55. N.E. Dowling, *Mechanical Behavior of Materials,* 2nd ed., Prentice-Hall, 1998, p 290

56. T.L. Anderson, *Fracture Mechanics,* 2nd ed., CRC Press, 1995

57. W.G. Clark, *Metall. Eng. Quart.,* May 1969, p 380

58. A.D. Wilson, Comparing the Effect of Inclusions on Ductility, Toughness and Fatigue Properties, *Through Thickness Tension Testing of Steel,* R.J. Gladowski, Ed., STP 794, ASTM, 1983, p 130

59. R.C. Bates and W.G. Clark Jr., *Trans. Quart. ASM,* Vol 62, 1969, p 380

60. B.V. Whiteson, A. Phillips, V. Kerlins, and R.A. Rowe, Special Fractographic Techniques for Failure Analysis, *Electron Fractography,* STP 436, ASTM, 1968, p 151

Fracture Appearance and Mechanisms of Deformation and Fracture

W.T. Becker, University of Tennessee, Emeritus; S. Lampman, ASM International

References

1. *Microelectronic Failure Analysis Desk Reference,* 4th ed., R.J. Ross and C. Boit, Ed., ASM International, 1999

2. R.W. Honeycombe, *The Plastic Deformation of Metals,* Edward Arnold and ASM, 1984

3. *Fracturing in Metals,* American Society for Metals, 1948

4. *Fracture,* B.L. Averbach, D.K. Fellbeck, G.T. Hahn, and D.A. Thomas, Ed., John Wiley, New York, 1959

5. *Fracture of Solids,* D.C. Drucker and J.J. Gilman, Ed., Gordon and Breach Science Publishers, 1962

6. C.D. Beachem, *Interpretation of Electron Microscope Fractographs,* NRL Report 6360, U.S. Naval Research Laboratory, Washington, D.C., 21 Jan 1966

7. *Electron Fractography,* STP 436, ASTM, 1968

8. G. Henry and J. Plateau, *La Microfractographie,* Institut de Recherches de Siderugie Francais, 1967

9. *Application of Electron Microfractography to Materials Research,* STP 493, ASTM, 1971

10. *Fractography, Microscopic Cracking Processes,* STP 600, C.D. Beachem and W.R. Warke, Ed. ASTM, 1976

11. D. Hull, *Fractography,* Cambridge University Press, 1999

12. *Cleavage Fracture,* K.S. Chan, Ed., TMS, 1996

13. *Recent Advances in Fracture,* R.K. Mahidhara et al., Ed., TMS, 1997

14. L. Engel, H. Kingele, G.W. Ehrenstein and H. Sharper, *An Atlas of Polymer Damage,* Carl Hanser Verlag, 1978

15. *Failure Analysis and Prevention,* Vol 11, *ASM Handbook,* 9th ed., ASM International, 1986

16. *Fractography and Atlas of Fractographs,* Vol 9, *Metals Handbook,* 8th ed., American Society for Metals, 1974

17. *Fractography,* Vol 12, *ASM Handbook,* 9th ed., ASM International, 1987

18. J.E. Masters and R.W. Hertzberg Ed., *Fractography of Modern Engineering Materials; Composites and Metals,* STP 948, ASTM, 1987

19. *IITRI Fracture Handbook,* S. Bhattacharya, et al., Ed., IIT Research Institute, Chicago, 1979

20. G. Henry and D. Horstmann, *De Ferri Metalligraphie,* Vol 5, *Fractography and Microfractography,* Verlag Stahleisen, mbH. Dusseldorf, 1975

21. G. Powell, Shu-hong Cheng, C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys,* Battelle Press, Columbus, 1992

22. D.A. Ryder, *The Elements of Fractography,* Advisory Group for Aerospace Research and Development, Paris, Nov 1971; also available from National Technical Information Service (NTIS), U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA

23. *SEM/TEM Fractography Handbook,* MCIC-HB-06, Battelle Columbus Labs, Columbus, 1975; also available as AFML-TR-75-159

24. L. Engel and H. Klingele, *An Atlas of Metal Damage,* Prentice-Hall, Inc., 1981

25. R.W. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* 4th ed., John Wiley, 1996

26. H. Rogers, The Fundamental Aspects of Fracture in Deformation Processing, *Deformation Processing,* Syracuse University Press, 1964, p 199

27. G. Fett, Testing for Fatigue, *Adv. Mater. Process.,* May 1999, p 37

28. D. Broek, *Fatigue Design Handbook,* AE-10, 2nd ed., Society of Automotive Engineers, 1988

29. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

30. W.A. Spitzig, G.E. Pellister, C.D. Beachem, A.J. Brothers, M. Hill, and W.R. Warke, A Fractographic Analysis of the Relationships Between Fracture Toughness and Surface Topography in Ultrahigh-Strength Steels, *Electron Fractography,* STP 436, ASTM, 1968, p 17

31. W. Gerberich, Microstructure and Fracture, *Mechanical Testing,* Vol 8, *Metals Handbook,* 9th ed., Mechanical Testing, American Society for Metals, 1985, p 476–491

32. H.L. Ewalds and R.J.H. Wanhill, *Fracture Mechanics,* Edward Arnold Ltd., London, 1984

33. R. Paprino, Ductility in Structural Design, *Ductility,* American Society for Metals, 1968

34. D. Broek, *Practical Use of Fracture Mechanics,* Kluwer, 1988, p 431, 436

35. W.S. Lin, The Ductile-Brittle Fracture Transformation: A Comparison of Macro and Microscopic Observation on Compact Tension Specimens, master's thesis, University of Tennessee, 1974

36. I.M. Ward, *Mechanical Properties of Solid Polymers,* 2nd ed., John Wiley, New York, 1983

37. R.P. Kambour and R.E. Robertson, Mechanical Properties of Plastics, *Polym. Sci.,* A.D. Jenkins, Ed., North-Holland Publishing Co., 1972, p 778

38. T.H. Courtney, *Mechanical Behavior of Materials,* McGraw-Hill, 1990

39. M.F. Ashby, C. Gandhi, and D.M.R. Taplin, *Acta Metall.,* Vol 27, 1979, p 1565

40. J. D. Embry, Damage Accumulation in Fracture Processes, *Modeling the Deformation of Crystalline Solids,* T.C. Lowe, A.D. Rollett, P.S. Follansbee, and G.S. Daehn, Ed., TMS, 1991

41. H.H. Kausch, *Polymer Fracture,* Springer Verlag, 1978

42. D.R. Johnson, “Toughness of Tempered Bainitic Microstructures in a 4150 Steel”, master's thesis, University of Tennessee, 1990

43. C.D. Beachem, “Interpretation of Electron Fractographs,” NRL Report 6360, Naval Research Laboratory, Washington D.C., 1966, p 49

44. G. Bodine, “The Effect of Strain Rate Upon the Morphology of High Purity 26% Chromium, 1% Molybdenum Ferritic Stainless Steel”, master's thesis, University of Tennessee 1974

45. C.R. Brooks, and A. Choudhury, *Failure Analysis of Engineering Materials,* McGraw-Hill, New York, 2001, p 136

46. C.D. Beachem, “Microscopic Fracture Processes” *Fracture,* Vol 1, Academic Press, 1968

47. C.S. Barret and R. Bakish, Trans. AIME, Vol 212, 1958, p 122

48. R.E. Reed-Hill, *Physical Metallurgy Principles,* Van Nostrand, 1973, p 755

49. G. Dieter, Introduction to Ductility, *Ductility,* American Society for Metals, 1968, p 26

50. D.A. Bales, Effects of Decarburization on the Fatigue Life of SAE 1042 Steel, master's thesis, University of Tennessee, 1977

51. M. Clavel and A. Pineau, Frequency and Waveform Effects on the Fatigue Crack Growth Behavior of Alloy 718 at 298 °K and 823 °K, *Metall. Trans. A,* Vol 9 (No. 4), April 1978, p 471–480

52. W.W. Gerberich and K. Jatavallabhula, Quantitative Fractography and Dislocation Interpretations of the Cyclic Cleavage Crack Growth Process, *Acta Metall.,* Vol 31, 1983, p 241–255

53. W. Yu, K. Esaklul, and W.W. Gerberich, *Metall. Trans A,* Vol 15A, 1984, p 889–900

54. R.W. Hertzberg and R.H. Goodenow, "Fracture Toughness and Fatigue Crack Propagation in Hot-Rolled Micro Alloyed Steel", *Micro Alloying, '75 Proceedings,* Union Carbide Corp. p 503

55. N.E. Dowling, *Mechanical Behavior of Materials,* 2nd ed., Prentice-Hall, 1998, p 290

56. T.L. Anderson, *Fracture Mechanics,* 2nd ed., CRC Press, 1995

57. W.G. Clark, *Metall. Eng. Quart.,* May 1969, p 380

58. A.D. Wilson, Comparing the Effect of Inclusions on Ductility, Toughness and Fatigue Properties, *Through Thickness Tension Testing of Steel,* R.J. Gladowski, Ed., STP 794, ASTM, 1983, p 130

59. R.C. Bates and W.G. Clark Jr., *Trans. Quart. ASM,* Vol 62, 1969, p 380

60. B.V. Whiteson, A. Phillips, V. Kerlins, and R.A. Rowe, Special Fractographic Techniques for Failure Analysis, *Electron Fractography,* STP 436, ASTM, 1968, p 151

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Mechanisms of Deformation and Fracture

At homologous temperatures low enough that creep deformation does not contribute to strain prior to or accompanying crack propagation (T_H < ~0.4), the mechanism of permanent deformation in metallic materials is by slip or deformation twinning. The microscale brittle fracture mechanism is by cleavage. In metallic materials, slip and twinning deformation processes thus compete with the brittle fracture process of cleavage. All mechanisms may not operate in a given lattice (material), and the activation of a given mechanism depends on the temperature, loading rate, and degree of constraint.

In an ideal material containing neither inclusions nor second phases, ductile fracture would be expected to occur by slip and possibly twinning and result in complete reduction in area (Fig. 1). Ref 1 Alternately, cleavage across a grain on a single plane would be expected to result in smooth fracture surfaces. Such results are sometimes observed in high-purity single-crystal specimens but are seldom seen in commercial engineering materials. Commercial engineering materials contain both a distribution of inclusions and often second-phase particles and constituents as well as grain boundaries, all of which impact the fracture nucleation and growth process.

Fig. 1 Single-crystal chisel point. Source: Ref 1

Early work using metallic materials of low symmetry (rhombohedral bismuth, hexagonal close-packed, or hcp zinc) as well as nonmetallic materials (NaCl) seemed to indicate that cleavage could be predicted by a critical normal stress law (Sohnke's law) (Ref 2) dating from 1869. References 3, 4, 5 contain data and discussion of this early work. Similarly, plastic deformation by slip seemed to obey a critical resolved shear stress law (Schmid's law). Single crystals could be made to extensively twin rather easily. Today, Sohnke's law is no longer assumed to be valid (although a normal stress facilitates cleavage), and no critical shear stress law based on external loading has been accepted for the onset of twinning deformation in polycrystalline material. A critical resolved shear stress law for the onset of plastic deformation by slip is still accepted in both single crystals and polycrystalline material.

References cited in this section

1. D. Hull, *Fractography*, Cambridge University Press, 1999

2. L. Sohnke, Normal Stress Law for NaCl, *Poggendorfs Ann.*, Vol 137, 1869, p 177

3. E. Schmid and W. Boas, *Plasticity of Crystals*, 2nd ed., F.A. Hughes & Co., 1959, and Chapman and Hall Ltd., 1968

4. R.W.K. Honeycombe, *The Plastic Deformation of Metals*, 2nd ed., Edward Arnold and American Society for Metals, 1984

5. D. MacLean, *Mechanical Properties of Metals*, John Wiley & Sons, 1962

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Background

Extensive experimental work in Germany and England in the early part of the 20th century identified much of the geometry and mechanics of slip, twinning, and cleavage, primarily using single crystals. That work has been summarized in the classic text of Schmid and Boas (Ref 6), which subsequently first became available in an English translation in 1950 and currently as Ref 3. A later review of the behavior of hcp materials is available in Ref 7. Some results from the pre-1940 work remain valid today, while other results, although perhaps valid for single crystals, must be modified to account for observations in polycrystalline materials.

Much of the work accomplished to 1940 was subsequently reexamined in light of rapidly developing dislocation theory as well as rapidly developing concepts in fracture mechanics, starting with Griffith's model for brittle fracture. First attempts at rationalizing the problem of brittle fracture by normal separation across a plane used Griffith's criterion for brittle fracture. This model inherently assumes that the material contains "flaws." The flaws/imperfections lower the stress required to cause fracture, and, by implication, fracture initiates and propagates from the flaws. Application of the Griffith model to metallic materials was not very successful, presumably because these materials were too ductile for a defect to sufficiently constrain or limit plastic deformation. The result was Orowan's modification of the Griffith model:

$$s_{critical} = \left[\frac{2Egr}{p\,ca}\right]^{1/2}$$

where E is the elastic modulus, ρ is the crack tip radius, γ is the specific surface energy of the crack faces, c is the crack length, and a is the interatomic spacing of the fracture planes. This expression becomes equal to the Griffith criterion when the crack tip radius becomes equal to the interplanar spacing.

References cited in this section

3. E. Schmid and W. Boas, *Plasticity of Crystals*, 2nd ed., F.A. Hughes & Co., 1959, and Chapman and Hall Ltd., 1968

6. E. Schmid and W. Boas, *Kristallplastizitaet mit Besonderer Bericksichtigung der Metalle*, Springer-Verlag, Berlin, 1935

7. P.G. Partridge, The Crystallography and Deformation Modes of Hexagonal Close Packed Metals, *Metals and Mater.*, Vol 1, 1967

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Single-Crystal Cleavage Models

It is best to first examine what is meant by "cleavage" in crystalline materials. A figure often used to illustrate the differences in mechanisms for slip and cleavage is that shown in Fig. 2. A shear stress applied to the material parallel to the slip plane causes alternate rupture and self-healing of the interatomic bonds, resulting in

permanent deformation. A normal stress applied to the body causes only continued separation of atoms and therefore brittle cleavage fracture.

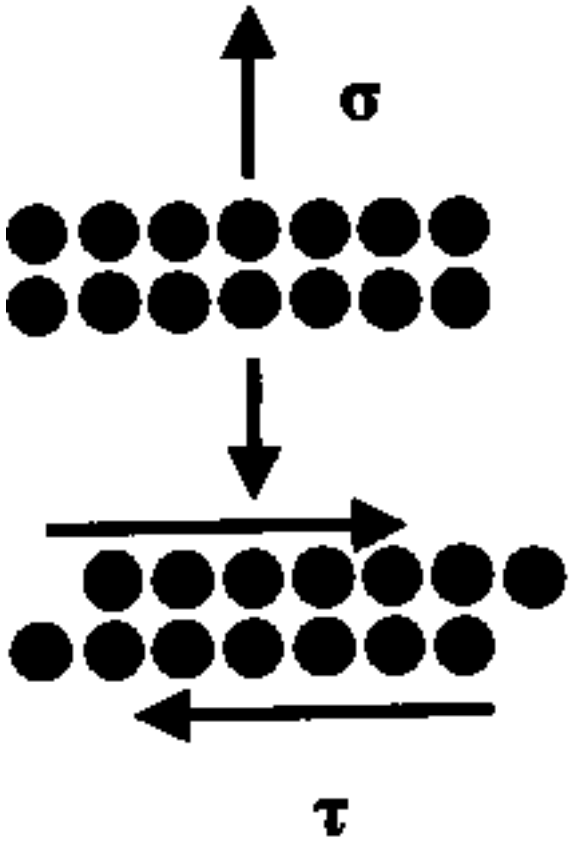


Fig. 2 Schematic figure showing the effect of a normal stress, σ, and a shear stress, τ, on a crystalline material. Application of a normal stress increases the interplanar distance and ultimately results in fracture. Application of a shear stress causes the planes of atoms to slide over each other and results in permanent deformation rather than fracture.

Much of the evidence for the processes of ductile slip deformation and cleavage fracture to support critical shear and normal stress laws was obtained prior to 1935 by macroscale observation. Photographs are available in Schmid and Boas (Ref 3, 6). Slip band extrusions on the surface of smooth single crystals were resolvable and photographed, and they implied evidence of slip deformation. Similarly, when cleavage fracture occurred, the resulting fracture surface was sufficiently smooth that optical light provided high light reflectivity from the crystal scaled facets.

That view of deformation versus fracture must be reexamined, considering the limits of resolution during examination of specimens as well as how much plastic deformation is required to conclude that plastic flow has preceded fracture. It is tempting to associate slip deformation—that is, the glide of dislocations—with ductile behavior and to associate no dislocation motion (i.e., no plastic deformation) with brittle cleavage behavior. The problem, of course, is one of scale. If there is no macro- or microscale evidence of deformation, it is tempting to assume brittle behavior. However, considerable evidence suggests that that is not a workable distinction.

Estimates of the yield strength of materials by simultaneous rupture of all of the bonds across a plane quickly were determined to be in serious disagreement with experimental observations and indicated that a different model had to be created to explain the discrepancy between observed and measured yield strengths. The contradiction between experiment and theory was solved by the use of dislocation models to argue that only local bonds had to be broken to permit successive movement of a dislocation along the slip plane. Similarly, estimations of a cleavage fracture stress versus experimentally determined values indicated need for a model other than simultaneous rupture of all bonds.

A contemporary view of cleavage fracture is a sequential tearing across the fracture plane rather than a simultaneous rupture of all bonds. Cleavage fracture that occurs by simultaneous bond rupture or by progressive bond rupture is described as “ideal cleavage” or “perfect cleavage.” The question, then, is what is the mechanism by which this occurs?

Understanding of brittle fracture in metals came, in part, during the 1950s and 1960s (although some of this work was done using nonmetallic materials that were essentially intrinsically brittle). Several models for cleavage fracture using dislocation theory were developed and are briefly discussed subsequently. It is important to note that they all involve dislocation motion. Whether experimental (including fractographic) evidence of that dislocation motion can be observed is still a matter of interest. Reed-Hill (Ref 8) has shown that cleavage fracture in zinc occurs on the basal plane whether the plane is flat or distorted in bending. Some recent results discussed subsequently (Ref 9) suggest a difference in the appearance of cleavage fracture initiated by "elastic" stress fields in nondeformed material (pure cleavage) versus cleavage fracture in material plastically deformed in the yield zone in front of a macroscale crack.

Important work from the period of the 1950s and 1960s included that found in Ref 10, 11, 12, 13. Additionally, there is an interesting discussion of identifying what is meant by "cleavage" in metallic materials in Ref 14. A summary of models and experimental evidence for brittle fracture is available in Ref 4, together with multiple references. Figure 3 is a compilation of the different dislocation models. Figure 3(a) is an elastic crack considered to be a pileup of edge dislocations. Figure 3(b) is simply the pileup of multiple edge dislocations at some barrier in the lattice, so that the local Orowan (modified Griffith) criterion can be met (see the section "Background" in this article). This model is used to predict cleavage cracking of grain-boundary carbides in low-carbon steels. This model was examined (Ref 12) and it was determined that the maximum normal stress at the pileup occurred on a plane rotated 70.5° from the slip plane. Using this model, it can be shown that a crack is nucleated when the shear stress concentrated by the pileup becomes equal to approximately 75% of the shear modulus. This requires between 100 and 1000 dislocations, a not unreasonable number in either bulk material or in second phases in a two-phase microstructure.

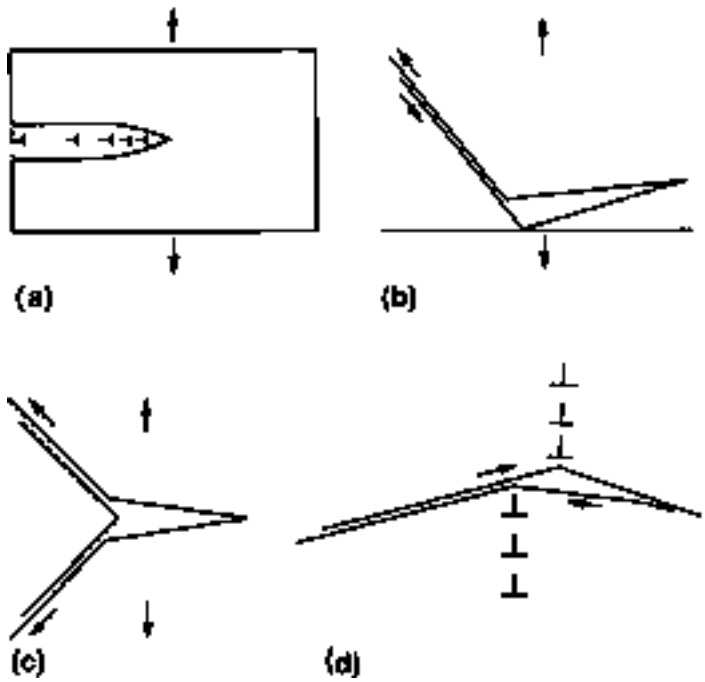


Fig. 3 Dislocation models for cleavage fracture. (a) Elastic crack regarded as a pileup of edge dislocations. (b) Pileup against a boundary forming a crack. (c) Crack forming by movement of dislocations on two slip planes. (d) Crack formation at tilt boundary. Source: Ref 4

A second model (Ref 15, 16) for cleavage crack initiation involves the interaction of two dislocations on intersecting slip planes, creating a sessile dislocation and an incipient crack nucleus. Trailing pairs of dislocations on the same planes going through the same reaction enlarge the crack nucleus. The reaction, which is energetically favorable (self-energy of reactants > self-energy of product) is of the form:

$$\left(\frac{a}{2}\right)[-1,-1,1]_{(1,0,1)} + \left(\frac{a}{2}\right)[1,1,1]_{(-1,0,1)} \rightarrow a[0,0,1]_{(1,0,0)}$$

Two glissile dislocations on different {1,1,0} planes interact along the intersection of the two slip planes (a .0,1,0. direction) to create a pure edge sessile dislocation on a {1,0,0} plane, which is the incipient crack

nucleus. Experimental work obtained results consistent with Cottrell's model using MgO (Ref 17) and in iron (Ref 18). As Ref 4 indicates, the work reported in Ref 13 supports such a model of dislocation interaction to create a cleavage nucleus. Specimens tested at 77 K could be made to flow plastically or to fail in a brittle manner, depending on specimen orientation. Specimen axes removed from .1,1,0. failed in a brittle manner, but specimens oriented near .1,1,0. did not fail in a brittle manner. This implies that duplex slip (simultaneous slip on more than one system) favored by the harder orientations promotes brittle fracture.

The third model shown in Fig. 3(d) is due to research discussed in Ref 19. In this model, a portion of the edge dislocations making up a tilt boundary are assumed to glide, creating the crack nucleus. Experimental evidence for this mechanism is shown in Fig. 4 for zinc. Note in this case that the slip plane is also the cleavage plane.

Fig. 4 Basal plane cracking in zinc at a subgrain (tilt) boundary. The basal plane is both a slip plane and a cleavage plane in this material. Source: Ref 19, as cited in Ref 4

Deformation (mechanical) twinning, in some instances, can also create cleavage crack nuclei and is discussed subsequently.

The preceding, then, is the case for which cleavage cracking would occur in a material that behaves in an ideally brittle manner for whatever reason (perhaps inherently brittle). In the more realistic case, when the requirements for ductile and brittle crack propagation are not very different, specimen geometry, imperfection geometry, loading rate, and temperature can cause different micro-scale events to occur that alter the macroscale appearance.

This is especially true in the body-centered cubic (bcc) materials (and perhaps some hcp materials), which show a strong temperature dependence of the yield strength. Fracture mechanics indicates that important variables in geometry-induced constraint are section thickness, crack length, and crack tip radius. For some combinations of these variables, constraint is incomplete across the full width of the specimen, so that the crack propagates on both a plane of maximum normal stress and planes of maximum shear stress, creating macroscale shear lips on the sides of the specimen. Examination of the fracture surface at the microscale reveals that crack propagation in the flat fracture area can occur by either or both cleavage and MVC, and that fracture of the shear lip regions occurred by MVC. Close examination of the morphology of the dimples on the shear lip shows that they are elongated rather than being equiaxed, and that if dimples are present in the flat region, they can also be elongated or equiaxed. These cases are discussed in more detail following a discussion of the effects of deformation twinning on cracking mechanisms and fracture surface morphology.

The face-centered cubic (fcc) materials do not generally cleave, as indicated in the preceding article in this Volume, “Fracture Appearance and Mechanisms of Deformation and Fracture.” However, Ref 20 does report brittle behavior in iridium and rhodium. Additionally, nitrogenated austenitic stainless steels and austenitic stainless steels in stress-corrosion cracking conditions may possibly cleave. Stresses to cause cleavage in hcp materials vary with purity, and there is some question as to whether cleavage can occur with sufficient purity.

References cited in this section

3. E. Schmid and W. Boas, *Plasticity of Crystals*, 2nd ed., F.A. Hughes & Co., 1959, and Chapman and Hall Ltd., 1968

4. R.W.K. Honeycombe, *The Plastic Deformation of Metals,* 2nd ed., Edward Arnold and American Society for Metals, 1984

6. E. Schmid and W. Boas, *Kristallplastizitaet mit Besonderer Bericksichtigung der Metalle,* Springer-Verlag, Berlin, 1935

8. R.E. Reed-Hill, *Physical Metallurgy Principles,* Van Nostrand, 1973 p 753

9. A. Vasilev, S. Firstov, and A. Koval, Cleavage Mechanism Nucleation and Fracture Toughness, *Recent Advances in Fracture,* R.K. Mahidhara et al., Ed., TMS, 1997, p 315

10. J.J. Gilman, *Trans. AIME,* Vol 200, 1954, p 621

11. A. Deyurterre and G.B. Greenough, *J. Inst. Met.,* Vol 84, 1956, p 337

12. A.N. Stroh, *Proc. R. Soc. A.,* Vol 233, 1954, p 404

13. N.P. Allen, B.E. Hopkins, and J.E. McClennan, *Proc. R. Soc. A.,* Vol 234, 1956, p 221

14. C.D. Beachem, Microscopic Analysis of Cleavage Mechanisms, *Special Volume of the Proceedings of the Second International Conference on Mechanical Behavior of Materials,* American Society for Metals, 1976, p 1188

15. A.H. Cottrell, *Fracture,* B.L. Averbach, B.L. Felbeck, D.K. Hahn, and D.A. Thomas, Ed., John Wiley & Sons, 1959

16. A.H. Cottrell, *Trans. AIME,* Vol 212, 1958, p 196

17. E.R. Parker, *Fracture,* B.L. Averbach, D.K. Felbeck, and G.T. Hahn, Ed., John Wiley, 1959, p 181

18. R. Honda, *Acta Metall.,* Vol 9, 1961, p 969

19. J.J. Gilman, *Dislocations and Mechanical Properties of Crystals,* John Wiley, 1957

20. T.H. Courtney, *Mechanical Behavior of Materials,* McGraw-Hill, 1990, p 392

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Slip, Twinning, and Cleavage in Polycrystals

Although critical stresses for activation of twinning and cleavage cannot be reproducibly measured in polycrystalline material, it is possible to make some general statements about the onset of each of these mechanisms as well as slip deformation in polycrystalline material. Deformation twinning is not commonly seen in the fcc lattice, although there is some evidence in the literature that these materials can, in fact, form deformation twins (Ref 21, 22). The fcc materials mechanically twin only with considerable difficulty. Alternatively, hcp materials mechanically twin more easily; so easily in fact, that sometimes mechanical polishing of a metallographic specimen can introduce artifacts. There are considerable differences, however, in the ease of twinning among alloys having a hcp structure, just as there are differences in the stress necessary to cause cleavage.

Deformation twinning does occur in materials having a bcc lattice, with twinning more likely at low temperature and elevated strain rates. (Strain rates commonly encountered in tensile testing of the order of 10^{-3} to 10^{-2} per minute do not typically cause deformation twinning at room temperature, but a hammer blow will cause twinning.) When deformation twinning does occur during plastic straining, it is usually activated after prior deformation by slip (but see subsequent information). That is, single crystals may be oriented to yield by twinning, but polycrystalline material must strain harden sufficiently in order to favor twinning deformation over slip deformation at room temperature.

Twinning density is typically not high near room temperature but increases as the temperature is decreased. In a series of iron-molybdenum polycrystalline alloys (67 grains across the cross section) no twinning was found during tensile straining until the temperature was decreased to 119 K, but then the number of grains having twins increased rapidly, reaching 60% at 83 K (Ref 23). Additionally, a not-quite-linear correlation was found between the lower yield strength and the number of grains containing twins, leading to the argument that twinning initiates at the yield stress with no further increase in the number of twins. Note that this is different from twinning initiated due to high strain rate deformation at room temperature, where plastic strain must usually accumulate by slip prior to the onset of twinning.

Twinning Deformation in Fracture Analysis. Twinning deformation is of importance with respect to deformation and fracture in three ways. On one hand, the change in shape of an arbitrarily oriented grain in a polycrystalline body requires activation of five independent deformation systems. In materials of low symmetry, this condition may not be met, such as if slip is restricted to the basal plane in hcp materials. Twinning can cause activation of additional independent deformation systems and therefore produce (limited) ductile behavior. This probably influences the deformation and fracture behavior of ordered intermetallic alloys. Alternatively, twinning can, in some instances, provide mechanisms for cleavage crack initiation, either by the interaction of strain fields created at the intersection of twins or at the intersection of a twin and a grain boundary, or by dislocation reactions favoring the formation of twins. (See Fig. 29 in the preceding article in this Volume, “Fracture Appearance and Mechanisms of Deformation and Fracture.”)

Fig. 5 Tongues on an iron fracture surface. Some additional secondary fine-scale cracking adjacent to tongues is also visible. Two-stage cellulose-acetate replica. Source: Ref 24

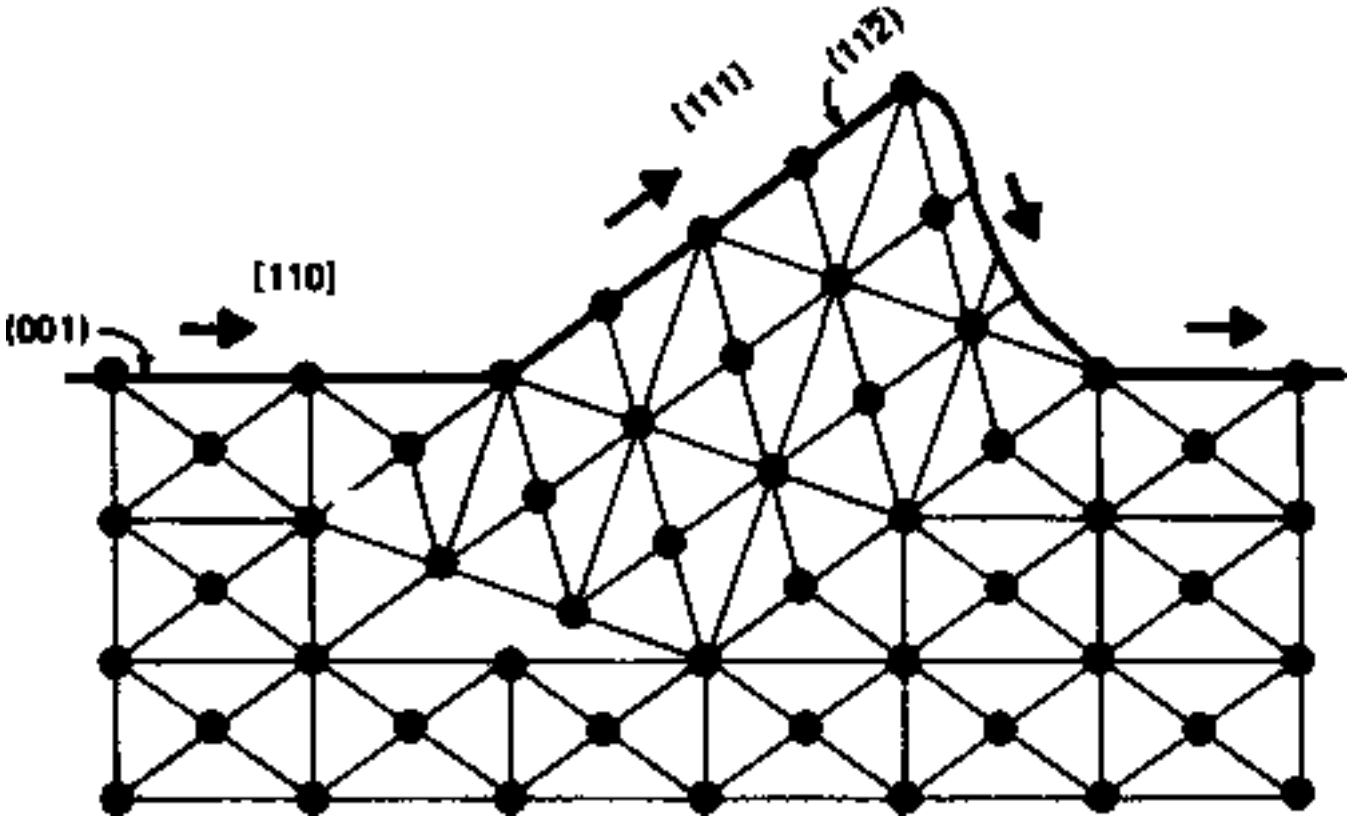


Fig. 6 Orientation relationships for {1,1,2} • 1,1,-1• twinning in the bcc lattice. It is assumed that a crack is propagating on the (0,0,1) cleavage plane and then on the (1,1,-2) K_1 plane in the [111] direction. (The required shear direction for simple shear twinning on the (1,1,-2) plane is the [1,1,1] direction.) Source: Ref 25

Fig. 7 Optical micrographs of the fracture surface of a Fe-3.25% Si alloy showing cleavage facets with river lines, tongues, and traces of two K_1 planes of the {1,1,2} family. (a) Lower-magnification view. Crack propagation is from left to right. (b) Detail of the area labeled "B" in (a). Source: Ref 1

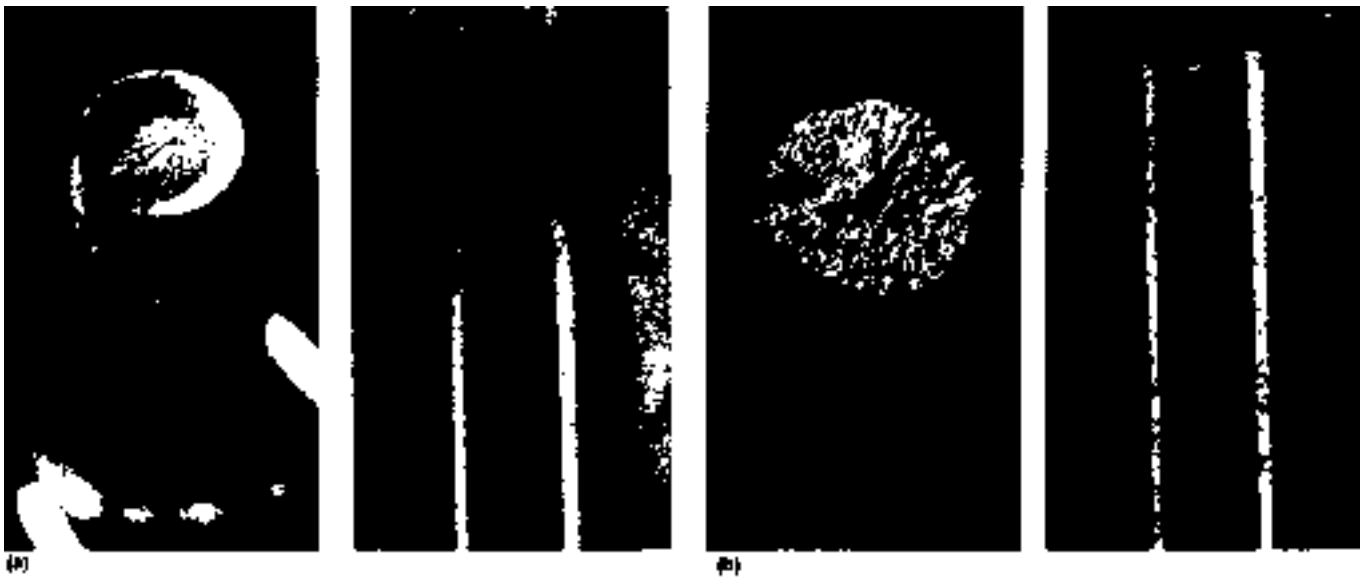


Fig. 8 Appearance of (a) ductile and (b) brittle tensile fractures in unnotched cylindrical specimens. Courtesy of George Vander Voort

The third point is that high strain rates favor initiation of twinning deformation. A crack propagating by a cleavage mechanism propagates at a high strain rate, and it is quite common to see evidence of twinning deformation (i.e., tongues) on a fracture surface dominated by cleavage. (See the preceding article in this Volume, "Fracture Appearance and Mechanisms of Deformation and Fracture.") From a failure analysis perspective, it is sometimes of benefit to know if the fractured component was subjected to high strain rate loading. Both cleavage fracture and deformation twinning are facilitated by high strain rate loading. The presence of tongues on a fracture surface showing cleavage does not necessarily support a conclusion of high strain rate loading, because the high strain rate may have been created by the crack propagation conditions, not the nominal loading conditions. However, if metallographic examination shows evidence of twinning well removed from the fracture surface, there is a higher likelihood that the loading was, in fact, at a high strain rate. Additionally, Ref 23 suggests that twins formed in front of a propagating crack are considerably finer than twins formed at the onset of yield.

Tongues and Deformation Twinning. As indicated in the introduction of this article, a rapidly propagating crack not only can nucleate cleavage fracture but also can nucleate deformation twinning in front of the advancing crack. One fractograph illustrating twinning and tongues is shown in Fig. 9 of the preceding article in this Volume, "Fracture Appearance and Mechanisms of Deformation and Fracture." Another example (Fig. 5) is shown here at higher magnification. As previously indicated, a tongue is created when the propagating cleavage crack turns from the cleavage plane ({0,0,1}) in bcc materials and runs on the composition or K_1 twinning plane (i.e., {1,1,2}). Figure 6 shows the atom arrangements and crack direction for tongue creation. Figure 7 is another example of tongues created in a Fe-3.25% Si alloy. Visible in the optical fractograph are tongues, river lines, and traces of two {1,1,2} family K_1 planes.

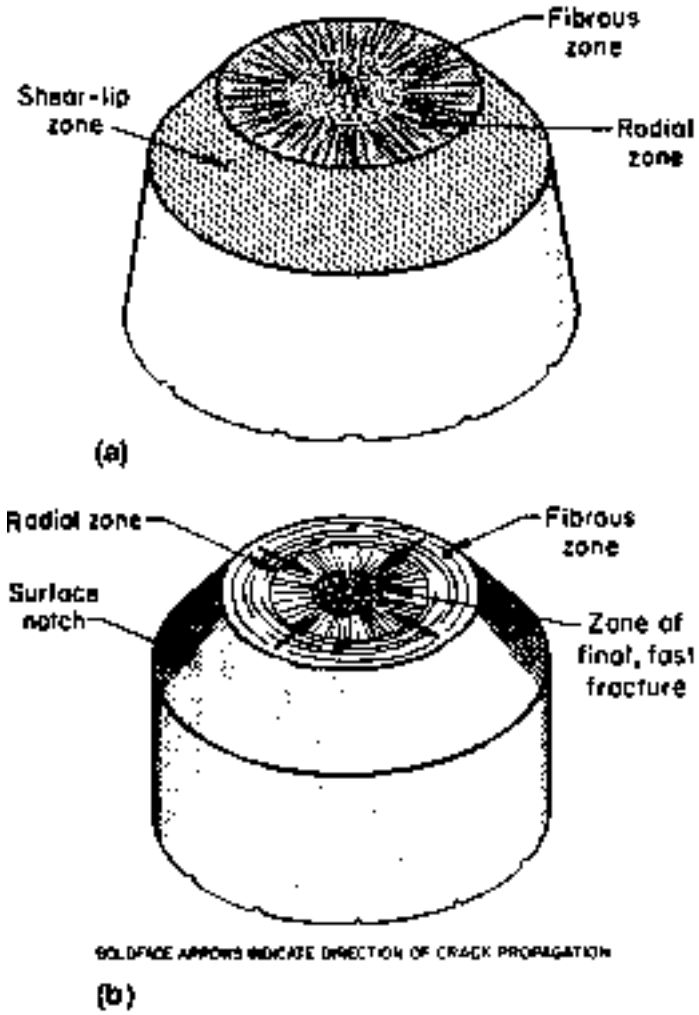


Fig. 9 Schematic of fracture surface regions in cylindrical tension-test specimens. (a) Surface from cone portion of fractured unnotched tensile specimen. (b) Surface of fractured notched specimen. Unlike the fracture surface for an unnotched specimen, the fracture surface for the notched specimen (b) does not have shear lips, because the fracture initiates near the root of the notch (and completely around the specimens in this idealized case without additional stress raisers).

Twinning and Crack Nucleation. Several workers (Ref 4) and several articles in Ref 26 and 27 have indicated that the presence of deformation twins in the microstructure can provide locations for subsequent crack nucleation. Reference 27 shows micrographs of cracking associated with twinning. Figure 8 in the preceding article ("Fracture Appearance and Mechanisms of Deformation and Fracture" in this Volume) shows cleavage cracking associated with deformation twins in an iron-chromium alloy after explosive (high-rate) forming.
Figure 10 shows postulated cleavage crack nucleation mechanisms at the intersection of two twins. Pairs of twinning dislocations, *n* in number, interact according to:

$$n\left(\frac{a}{6}[-1,1,1]\right)+n\left(\frac{a}{6}[1,-1,1]\right)\rightarrow\frac{1}{3}na[0,0,1]$$

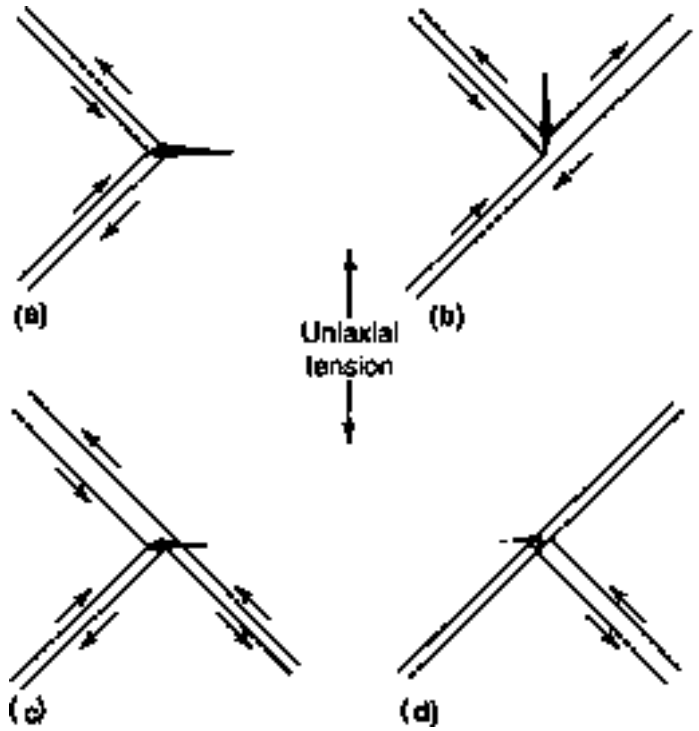


Fig. 10 Microcrack formation at twin intersections. (a, b, c) Incipient crack nucleation by dislocation reactions at the intersection of mechanical twins. (d) Incipient crack nucleation by strain concentration created when a growing twin intersects a previously existing twin. The direction of the stress is indicated. Source: Ref 28

Because the crack nucleus in Fig. 10(b) is parallel to the nominal stress, it would have to be activated by the local state of stress within a grain. However, the stress is normal to the cleavage crack nucleus for the other orientations. Figures 10(a–c) assume that both twins are nucleated at the same time. Figure 10(d) shows a different situation, where the cleavage crack is nucleated by strain concentration in the matrix created by secondary twinning. The new twin is blocked by a preexisting twin. An example of cleavage cracking initiated at the intersection of two twins at a grain boundary is discussed in Ref 23, which also suggests that grain boundaries are a common location for transgranular crack initiation, especially if a particle is present in the grain boundary.

As pointed out in Ref 28, the dislocation mechanisms responsible for nucleation of a twin are not very much different from those for cleavage crack nucleation. As a result, there may not be very much difference in energy required to cause either event. Twinning and cleavage processes are both temperature dependent, so that there may be different behavior in different temperature ranges; neither twinning nor cleavage at higher temperatures decreased deformation by slip and competition between cleavage and twinning at lower temperatures.

References cited in this section

1. D. Hull, *Fractography,* Cambridge University Press, 1999

4. R.W.K. Honeycombe, *The Plastic Deformation of Metals,* 2nd ed., Edward Arnold and American Society for Metals, 1984

21. R.W. Cahn, Survey of Recent Progress in the Field of Deformation Twinning, *Deformation Twinning,* R.E. Reed-Hill, J.P. Hirth, and H.C. Rogers, Ed., Goordon and Breech, 1964, p 1

22. G.Y. Chin, Formation of Deformation Twins in fcc Crystals, *Acta Metall.,* Vol 21, 1973, p 1353

23. J. Spreadborough, Informal Discussion on Fundamental Brittle Fracture Research on Unirriadiated and Irradiated Steels, *Proceedings of the Colloqium on Brittle Fracture and Selected Problems in Nuclear Pressure Vessel Steels,* J. Sebille, Ed., (Euratom, Brussels) European Atomic Energy Community, 1966

24. C.D. Beachem, "The Interpretation of Electron Microscope Fractographs," NRL Report 6360, Naval Research Laboratory, Washington, D.C., 21 Jan 1966

25. J.L. McCall and P.M. French, Ed., *Metallography in Failure Analysis,* Plenum Press, 1977, p 7

26. R.E. Reed-Hill, J.P. Hirth, and H.C. Rogers, Ed., *Deformation Twinning,* Gordon and Breech, 1964

27. D. Hull, Twinning and Nucleation of Cracks, *Fracture of Solids,* D.C. Drucker and J.J. Gilman, Ed., Gordon and Breech, 1963

28. R. Priestner, The Relationship between Brittle Cleavage and Deformation Twinning in bcc Metals, *Deformation Twinning,* R.E. Reed-Hill, J.P. Hirth, and H.C. Rogers, Ed., Gordon and Breech, 1964, p 321

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Ductile Fracture and Microvoid Coalescence

The hallmark indication of ductile failure, possibly first reported by Tipper (Ref 29), is the microscale presence of dimples on the fracture surface. The dimples form by a process referred to as MVC, where voids nucleate (initiate), grow, and coalesce to develop the final fracture surface that is dimpled. The most common process is one of pore (incipient crack) formation at or in a particle, followed by either void growth and linkage or by localized shear band slip deformation in the intervoid ligaments. These mechanisms of void nucleation and void coalescence are briefly described, along with associated microscale and macroscale appearances. This section does not include detailed dislocation models or complex mathematical treatment but rather describes mechanisms to the extent that relates to fracture-surface appearances. A more detailed treatment of ductile fracture is in the Appendix to this article.

Ductile Crack Nucleation

The dominant and common mechanism for creation of a (ductile) crack nucleus is a combination of debonding at particle interfaces and cracking of second-phase particles. There may be alternative crack nucleation mechanisms in high-purity single-phase materials containing a minimum number of very small particles, but the most common site of pore (incipient crack) formation is at second phases and/or inclusions in the matrix after varying amounts of plastic deformation.

Two types of particles may be involved: inclusions and second phases. In poly-phase alloys, the presence of eutectic or eutectoid structures also affects crack nucleation and propagation. In cast alloys, the presence of both microscale interdendritic porosity and macroscale shrinkage cavities provide "holes" for crack nucleation. There is little difference in void growth from shrinkage porosity and a hole containing an unbonded inclusion.

The incipient crack can form in four ways:

- Hole formation around an often weakly bonded particle
- Cracking of a particle
- Growth from a precracked particle (for example, from prior working operations)
- Cracking in a matrix of limited ductility in which faceted particles cause stress concentration in the matrix (faceted nitride inclusions in a high-strength, low-alloy, or HSLA, steel)

Inclusions tend to have a low-strength interface bond (if any) with the matrix and have mechanical properties different from those of the matrix. In steels, these inclusions fall into three general groups:

- Stringer types that are sulfides and perhaps contain additional phases
- Spheroidal types (oxides and sulfides)
- Faceted types (nitrides)

The sulfide inclusions are ductile and low strength, while the oxides and nitrides are brittle. Similarly, second-phase particles have mechanical properties different from those of the matrix, but experimental evidence suggests that there is often some bond strength (partial coherency) between these particles and the matrix (i.e., carbide particles in a low-carbon steel).
As a material is subjected to a load, debonding occurs in directions of maximum strain in the matrix, with the more weakly bonded materials debonding first. For unidirectional loading, debonding occurs at the interface generally normal to the applied load (Fig. 11a, b). An example is shown in Fig. 12. When the applied loading creates a tensile-hydrostatic stress (often poorly referred to as a “triaxial” stress-state condition) in the material, debonding occurs all around the particle (Fig. 11c). Figure 13 shows an example of volumetric debonding at a MnS inclusion. If there is sufficient load transfer to an elongated particle, cracking is also observed at approximately half the length of the particle (Fig. 11b).

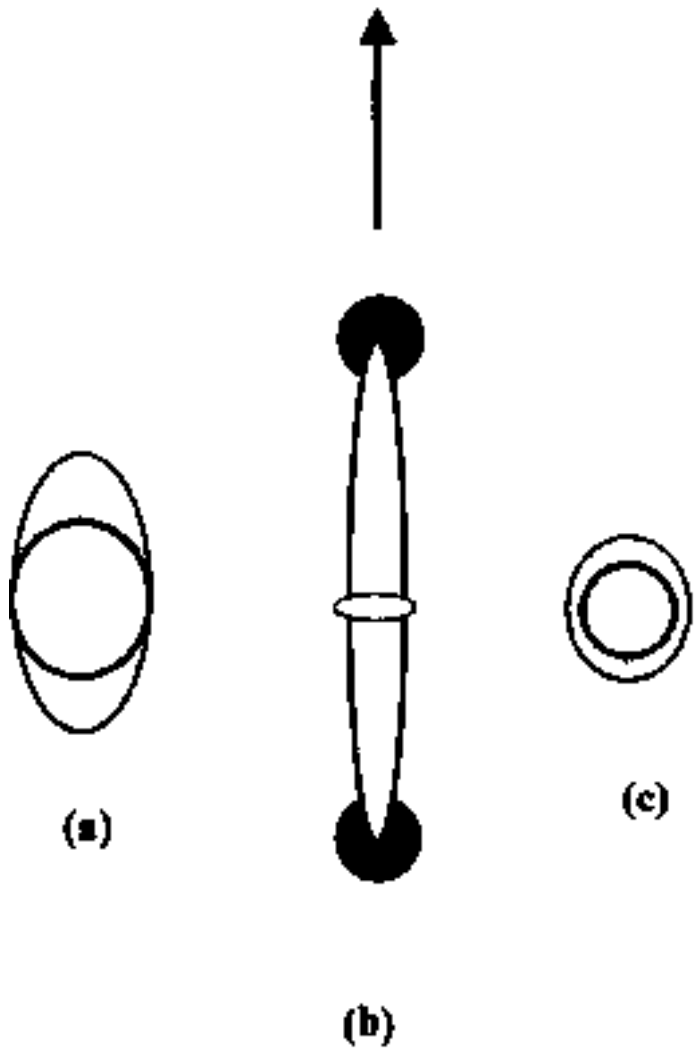


Fig. 11 Schematic of debonding at a matrix-particle interface with unidirectional stress. (a) Plane stress loading of an inclusion (no interfacial bond) causes debonding at the particle caps. (b) Debonding and fracture of high-aspect-ratio particles (elongated inclusion) due to shear transfer. Fracture is indicated at half-length. (c) More general (uniform) debonding from a large triaxial (i.e., tensile-hydrostatic) stress

Fig. 12 Debonding associated with a MnS inclusion in a bainitic microstructure. Loading direction indicated. Source: Ref 30

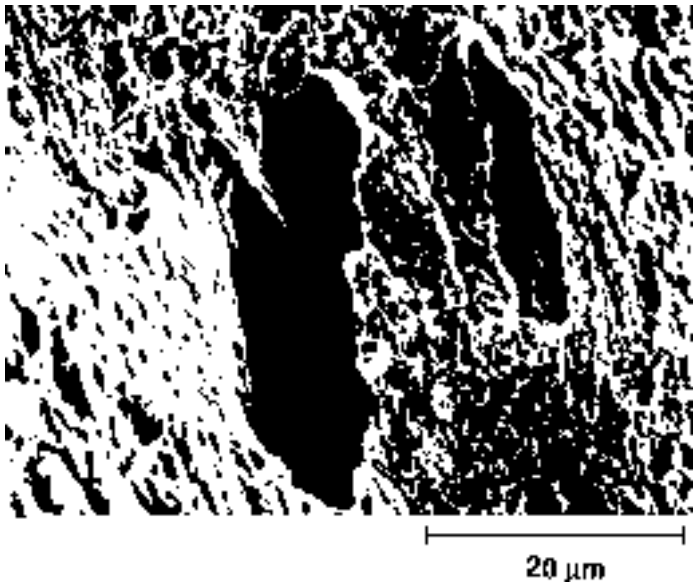


Fig. 13 Debonding of a MnS particle in a 4140 quenched and tempered steel due to a bending stress. The component was forged, and considerable directionality (banding and fibering) existed in the material. Crack propagation from bottom to top in the photograph. Courtesy of Michael West, University of Tennessee

The common observation for steels is that fracture is initiated by (relatively large) sulfide stringers. If the material is quenched and tempered, debonding and fracture of carbide occurs at a greater strain. Particles that are totally or partially coherent with the matrix (e.g., carbides in steel) require development of substantial stress for decohesion to start at the particle-matrix interface. Careful work using steels containing small, spheroidized carbides and (typical) MnS inclusions shows that debonding occurs first around the larger and more weakly bonded MnS inclusions and at a greater strain by debonding at the carbide particles. Fracture has also been associated with clusters (bead stringers) of oxide particles.

When material has been previously cold worked, it is often observed that not only is there debonding at matrix-particle interfaces but also the particles themselves are cracked (Fig. 14). These cracked particles then provide additional crack nucleation sites. Elongated particles, such as MnS, crack at approximately half-length due to load transfer from the matrix to the particle. Large second-phase brittle particles may crack by cleavage caused

by dislocation pileup. Similarly, particle cracking is also observed metallographically when the material is initially in the annealed condition and the component is examined metallographically after fracture.

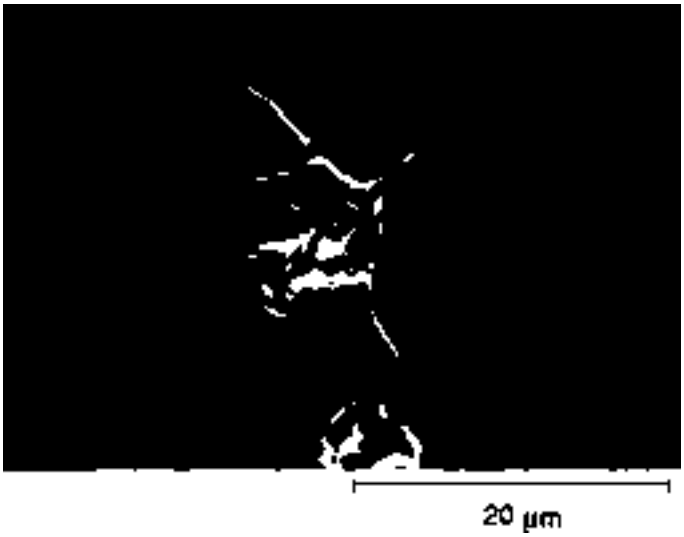


Fig. 14 A cracked cementite particle in a cold-rolled low-carbon steel (approximately 0.1% C). A high magnification view of a cracked cementite particle showing multiple cracks and shattering. Courtesy of Richard Holman, University of Tennessee

In materials containing a dual distribution of particle sizes (for example, MnS inclusions and spheroidal carbides in a quenched and tempered steel), a dual-dimple size is often seen on the fracture surface. In this case, the MnS inclusions are larger than the carbide particles and have lower bond energy to the matrix than do the carbide particles. Fracture initiation is initially by debonding and void growth at the MnS-matrix interface (large dimples). After a larger strain, debonding then occurs at the smaller carbide particles (resulting in smaller dimples than the voids from MnS-matrix debonding) (Fig. 15).

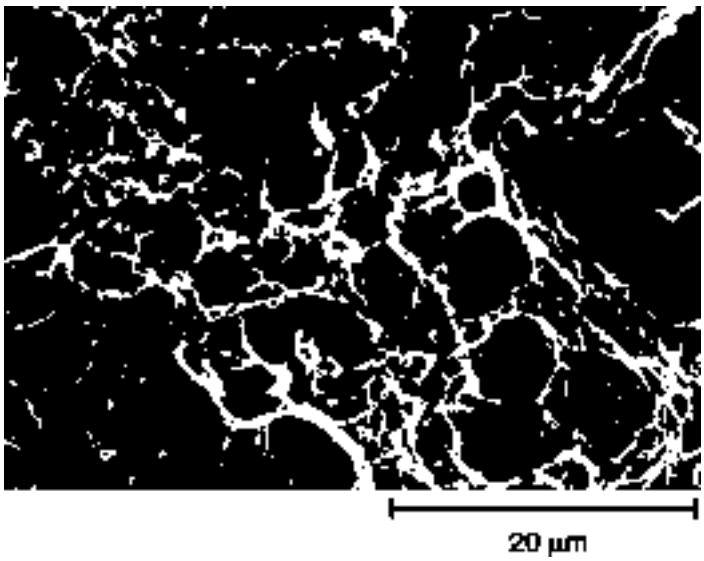


Fig. 15 A dual-dimple size observed in a 4150 steel. Material was isothermally transformed at 190 °C (375 °F) and was not tempered. Tested as a Charpy V-notch specimen at 0 °C (30 °F). Source: Ref 30

There is also a particle size distribution in high-strength age-hardening aluminum alloys, but different behavior in an aluminum casting alloy has been reported (Ref 31). The iron- and/or silicon-rich phases are the inclusions that are smaller than other phases created by solute additions for age-hardening response. In this case, no debonding was observed for any of the second-phase particles. Fracture initiation was by cleavage in the larger solute-addition-rich phases. The smaller second phases may or may not subsequently crack, depending on their size, but no debonding at the interface of second phases with the matrix was observed.

Models of Void Nucleation. Modeling the void initiation process is difficult because of the variety of void nucleation origins. Sources such as pores, inclusions, or cracked second-phase particles are effectively voids even before plastic flow is initiated. Soft inclusions, such as MnS, are virtually voids with little or no bond between the particles and the matrix surrounding them. The size and distribution of inclusions affects the ductility of a metal, with smaller and round particles offering more ductility than large, irregular particles.

Models to predict the effect of porosity or inclusions on ductility are given in the Appendix to this article. Inclusions and second-phase particles are the primary sources of voids, and two parameters of importance in these models are particle spacing and particle size. One of the simpler models predicts a strain to fracture (Fig. 16) (Ref 32):

$$e_f \cong 2\ln\frac{L}{w} \approx \ln\frac{1}{r}$$

where L is the void spacing, w is the void size, and ρ is the void volume fraction.

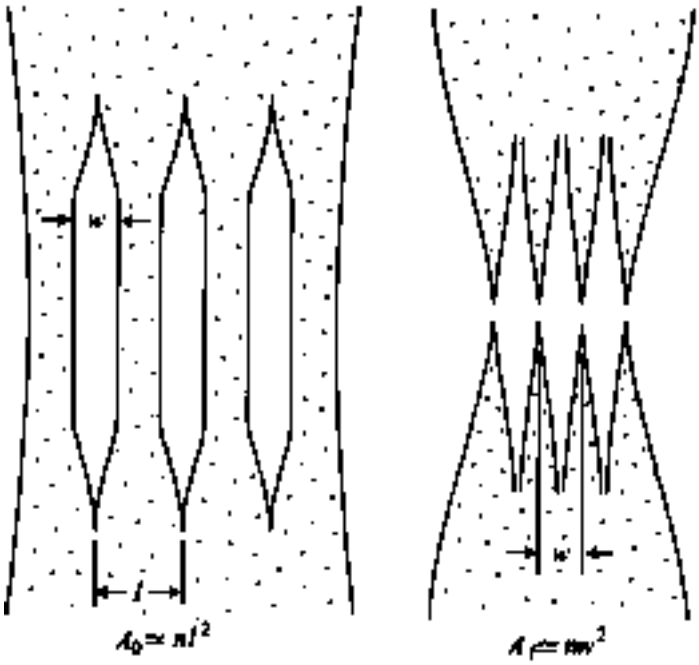


Fig. 16 Backofen's model for fracture from holes. See text for discussion. Source: Ref 32

Models of void formation have been based on dislocation interaction with the particle. One such model was developed by Brown and Stobbs (Ref 33) and is discussed in the Appendix to this article. Another model, developed by Le Roy (Ref 34), takes into account that different sources of void formation lead to a dual population of voids.

As noted, particles that are totally or partially coherent with the matrix (e.g., a carbide compared to a MnS inclusion) require development of more stress at the particle-matrix interface for decohesion to form a void. The parameters affecting the strain or stress at which voids initiate at coherent particles vary within a single part. For example, the particle-matrix interface strength (degree of coherency) affects void formation, as does the thermal expansion mismatch. Yet, reasonable models (see Appendix) have been developed for simple geometries, such as round particles and rods (particles with high aspect ratios).

Comments on Void Nucleation without Inclusions Present. The preceding discussion considers ductile crack nucleation due to initial void formation at particles. There is little debate that this is the dominant mechanism, especially in steels and in most other engineering materials. The question is whether this is the only mechanism for crack nucleation at the microscale. The debate arose at one time because examination of fracture surface did not always indicate the presence of particles in the microvoids. Questions arose as to whether particles could be seen, even if they were present (adequate resolution, statistical probability of seeing a particle at very small volume fractions) (Ref 35). This question has not been examined recently, but it has been commented that dimple rupture occurs without void nucleation in alloys that deform by planar slip (Ref 36).

There has also been a report of void nucleation without particle debonding in an aluminum-lithium alloy (Ref 31). In this alloy, a variation in particle size was present, ranging from very small iron- and silicon-rich

particles (the inclusions) to larger particles present because of solute additions for age-hardening response. In this study, examination showed that the fracture process was initiated by cleavage cracking in larger solute-rich phases. Smaller second phases may or may not subsequently crack, depending on their size, but no debonding at the interface of second phases with the matrix was observed. Crack nucleation, then, is only by cracking in (larger) second phases.

Other incipient crack nucleation mechanisms have been proposed. It has been suggested that grain-boundary triple points can create sufficient constraint to nucleate a crack (Ref 37). Crack nucleation at dislocation pileups has been suggested (Ref 38). Additionally, it has been suggested that holes (porosity) in the matrix may act as crack nuclei (and therefore, shrinkage porosity in a casting) as well as faceted inclusions. In the latter case, stress concentration at the corners of the inclusion (in a not very ductile matrix) can cause crack nucleation in the matrix, for example, nitrides in HSLA steels.

When the particles are small and the total volume fraction is small, other mechanisms may be factors. In a more recent study of fracture in high-purity single and polycrystalline copper, high-purity copper (99.9998%) and notched high-strain-rate-loaded specimens were used in order to prevent necking (Ref 39). The results have apparently confirmed the premise that the fracture mechanisms is by void nucleation, even when no particles can be detected in the voids. These results indicate that when grain boundaries are present, fracture initiates in the grain boundary, and when the boundaries are absent, fracture initiates within high-dislocation-density cell walls. An in situ transmission electron microscopy (TEM) study of ductile fracture in high-purity silver single crystals has provided experimental evidence of dislocation reactions that can provide hole nucleation in front of an advancing crack tip (Ref 40).

Void Growth and Coalescence

The incipient cracks (voids) associated with the particles may grow by two mechanisms, depending on particle size and spacing:

- Growth of an incipient crack by enlargement and linkage of the voids
- Slip band deformation in ligaments between the voids

For large particles relatively closely spaced, and especially in the presence of constraint, void coalescence typically occurs on the plane of maximum normal stress. This mechanism of void coalescence is referred to as pure ductile tearing (Fig. 17a).

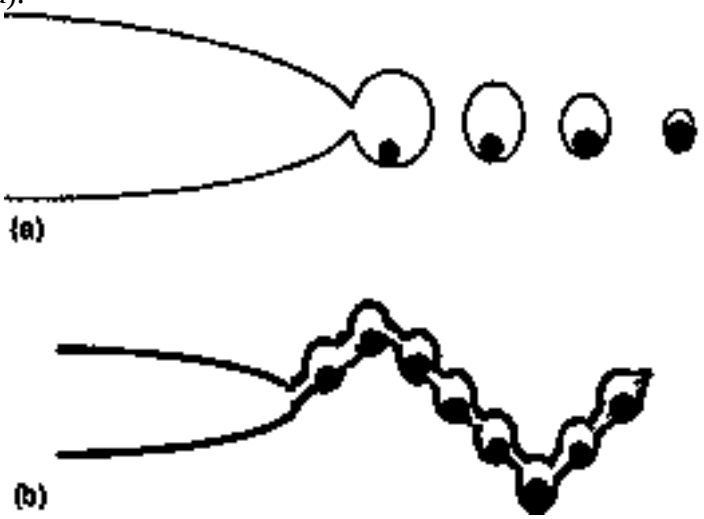


Fig. 17 Schematic of (a) ductile tearing along plane normal to normal stress and (b) zig-zag path of void sheet fracture along shear planes

When the particles are more widely spaced, the process of void coalescence can involve slip deformation in the interligament region between the voids. Deformation of the ligaments may occur from localized shear bands or by necking of the ligaments between the voids. A good example of the former is the mechanism known as void sheet formation, which is a common fracture mode in steels containing dual-size particles (i.e., larger MnS

inclusions and smaller spheroidal carbides). Fracture initiates at (large) MnS inclusions, followed by microscale strain localization on planes of high shear stress. These planes randomly contain smaller carbide particles, which subsequently nucleate voids. This is referred to as void sheet fracture and results in a characteristic zig-zag pattern of void growth, shown schematically in Fig. 17(b). Figure 18(a) (Ref 41) shows and example of void sheet formation between MnS inclusions.

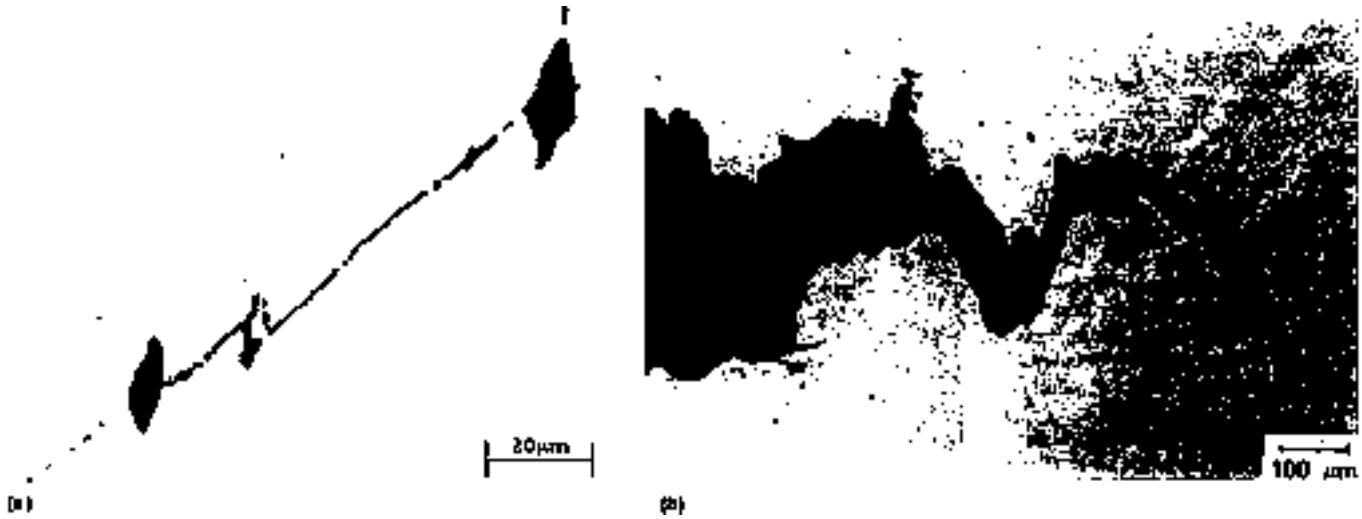


Fig. 18 Examples of ductile fracture on shear planes. (a) Void sheets from propagation of a crack between widely spaced inclusions within a shear band in a 4340 steel. Stress axis is vertical. Source: Ref 41. (b) Ductile crack growth in HSLA steel (A710). Source: Ref 77

In addition, necking can occur in the ligaments between the particles. With sufficient plastic flow, the interparticle ligaments can show sufficient ductility to display necking with a large height similar to that seen in single crystals. An example (Fig. 19a, b, c) shows ductile MVC fracture in an aluminum-silicon alloy, which is quite different in appearance for MVC in steels. In the aluminum-silicon alloy shown, fracture in the matrix is initiated by cleavage in the silicon particles.

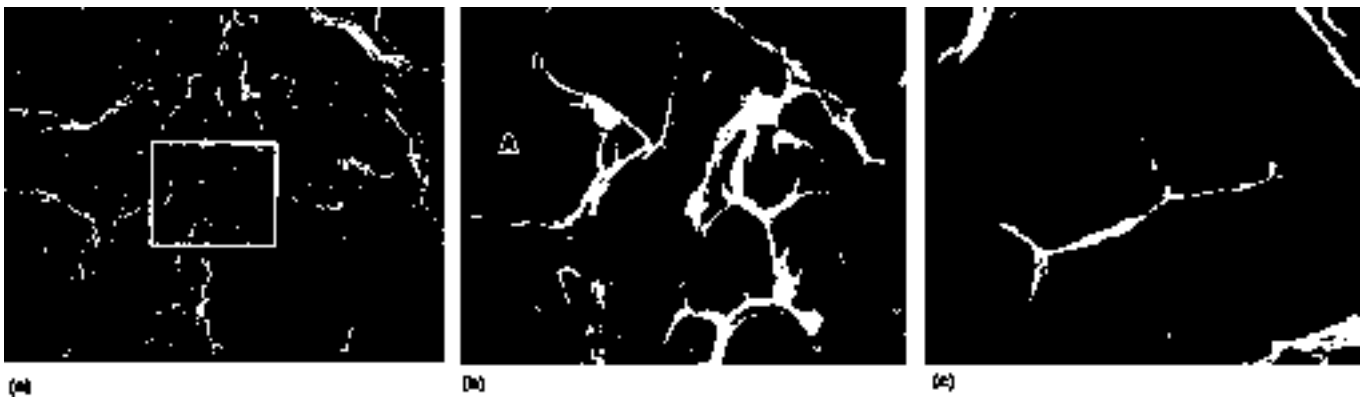


Fig. 19 Microvoid coalescence in an aluminum-silicon alloy (A380) loaded in tension. (a) Fracture surfaces consist of cleaved particles (i.e., silicon) and ridged fracture of the aluminum. 200×. (b) Higher-magnification (1440×) view of boxed region. (c) A fractured aluminum ligament surrounded by intermetallic particles at higher magnification (2880×). Source: Ref 42

In summary, two kinds of microscale ductile crack propagation are recognized, depending on whether growth of incipient cracks occurs on a plane of maximum normal stress or on a plane of high shear stress:

- Pure ductile tearing by void coalescence, usually on the plane of maximum normal stress
- Zig-zag slip deformation on planes of high shear stress between voids

However, the relation between these two microscale mechanisms and the macroscale of crack growth is less clear. For example, the zig-zag path from void sheet formation (Fig. 18a) does not necessarily result in macroscale shear bands. The macroscale fracture surface may be normal to the load, while the microscale paths of fracture may occur along the shear planes with a small step height.

Nonetheless, the direction of crack growth at higher scales may also occur on either a plane of maximum normal stress or on a plane of maximum shear stress, depending on the degree of constraint (and thus component geometry), the size, and the spacing of particles that initiate incipient cracks. For example, Fig. 18(b) shows a zig-zag path of ductile fracture at a higher scale than the microstructural scale of void sheeting shown in Fig. 18(a). Ductile fracture along shear planes can also occur on the macroscale. The factors that influence the macroscale direction of crack growth during ductile fracture are discussed in more detail in the Appendix of this article and in sections on cylindrical and prismatic geometries.

It is important however, to recognize that microstructural factors and/or localized deformation (such as microscale or macroscale shear bands) can influence macroscale direction. For example, even though the mechanism of void sheet formation does not normally result in a macroscale zig-zag path of fracture, macroscale shear bands can influence the direction of crack growth, as shown in Fig. 20. There is no apparent flat, central fibrous region, and there are five macroscale shear bands and additional small, unresolved bands across the specimen. Fracture initiation is reported to occur in the center of the specimen. The material in this case is slow-cooled HY-100 steel with a severely banded microstructure (alternating bands of soft ferrite and hard “granular” ferrite). The test specimen orientation was perpendicular to the banded microstructure, which is reported to facilitate shear banding. Other examples of macroscale shear banding are discussed in the Appendix of this article.

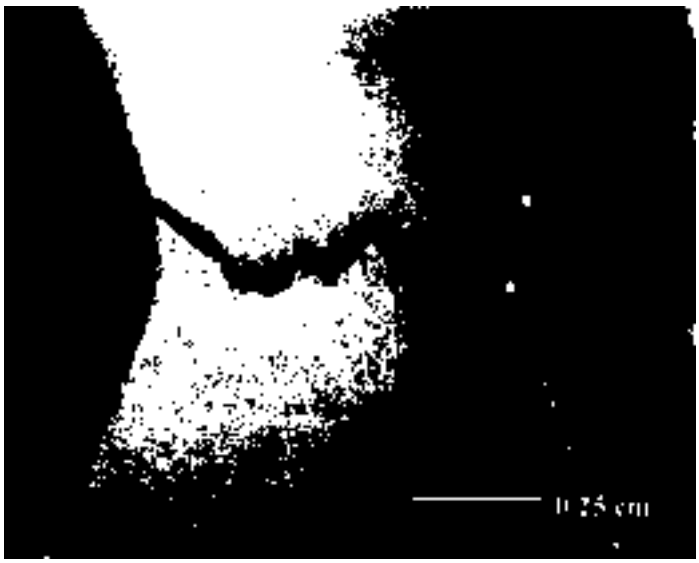


Fig. 20 Profile of fracture path in a notched-bar specimen taken from the long-transverse direction of HY-100 steel plate with a severely banded microstructure (alternating bands of ferrite and “granular” ferrite). Loading direction was perpendicular to the bands. Fracture is reported to initiate near the centerline. Note the apparent absence of a central, flat fibrous zone. Source: Ref 43

Dimple Shape and Size

The following sections describe some of the different morphologies of MVC dimples. Sometimes a striated-type surface is seen on the fracture surface, especially on smooth walls that were interfaces between inclusions and the matrix (see Fig. 21 in the preceding article, “Fracture Appearance and Mechanisms of Deformation and Fracture,” in this Volume). The lines are resolved slip packets intersecting a free surface of the material and should not be confused with fatigue striations. One differentiation is that these lines often contain many branches and are of nonuniform spacing, in contrast to fatigue striations that have few branches and are of consistent spacing (within a small area).

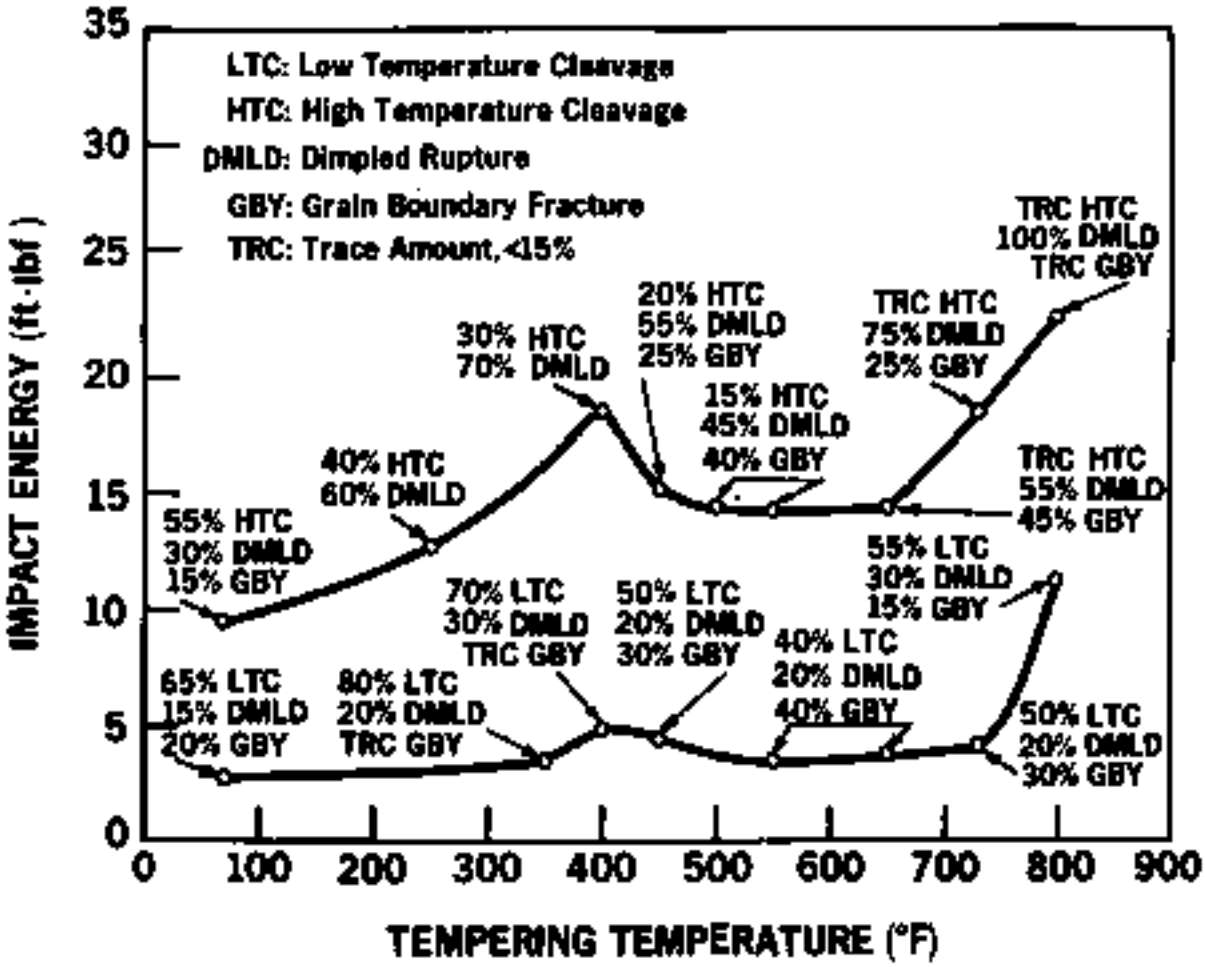


Fig. 21 Fractographic analysis of failed Charpy specimens of 4340 steel tempered to various strength levels. The upper curve represents room-temperature specimens; the lower curve represents specimens broken at -196 °C (-320 °F). All percentages are estimated. Source: Ref 69

Dimple Morphology with Loading Conditions. Dimple morphology is influenced by the loading condition, as shown in Fig. 22, for the three modes of crack displacement (Fig. 23):

- Dimples are equiaxed for axial loading (mode I, opening mode)
- Dimples are elongated in shear loading (mode II or mode III)
- Dimples are elongated and point in the same direction on both sides of the fracture surface from bending loading (Fig. 22)
- Dimples are either elongated or equiaxed in combined axial and bending

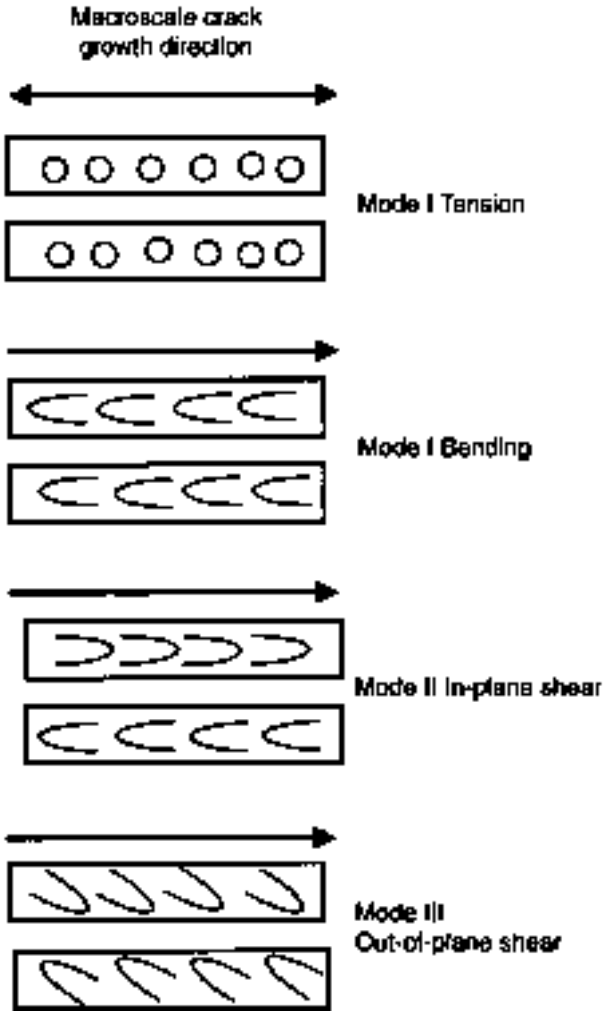


Fig. 22 Effect of loading conditions on dimple shape for loading modes I, II, and III. Mode I can be axial loading, bending loading, or a combination of bending and axial loading. Axial loading creates equiaxed dimples. Bending loading creates elongated dimples that face in the same direction on both halves of the fracture surface. Mode II (in-plane shear) loading creates dimples that point in opposite directions on the two halves of the fracture surface. Mode III (out-of-plane shear) loading creates dimples that are rotated with respect to the macroscale direction of crack growth.

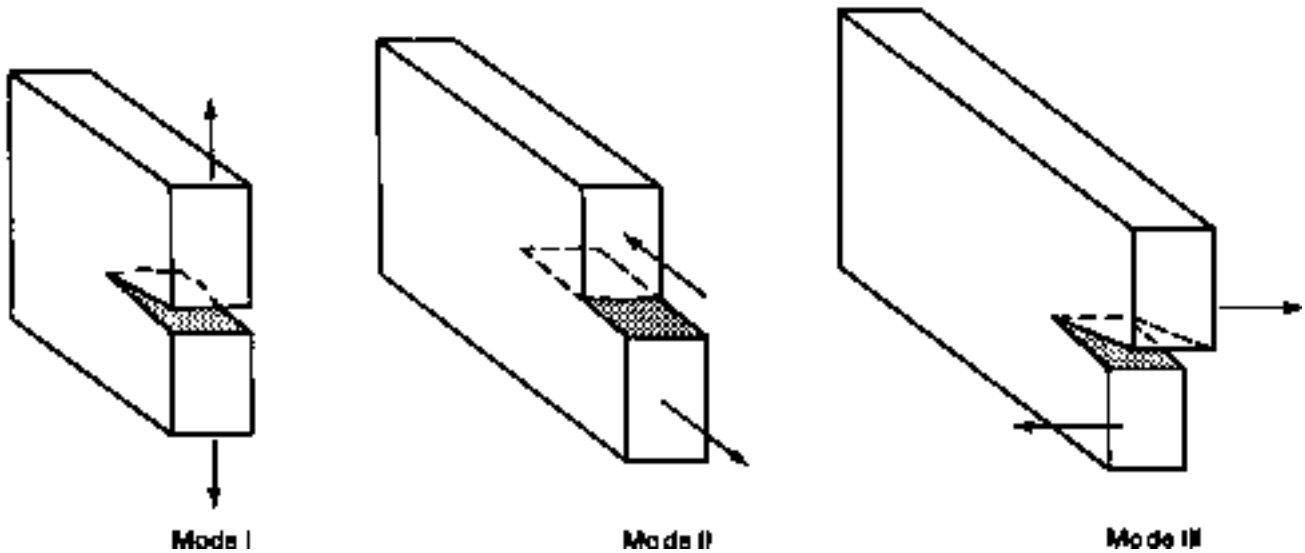


Fig. 23 Schematic illustration of loading modes for crack displacement: mode I (opening mode), mode II (in-plane shear), and mode III (out-of-plane shear, tearing)

By noting that the dimples can point in different directions on each half of the fracture surface, it is possible to identify loading conditions at the microscale, for example to distinguish between bending and torsion or shear loading. When compact tension-type specimens are loaded, the combined bending and axial loading does not result in very much, if any, elongation of the dimples parallel to the crack plane.

Pure mode II and III deformation lead to contact between crack asperities on each side of the fracture surface during propagation and may prevent identification of dimple morphology due to contact or rubbing damage. Although it is conceptually possible to identify loading conditions from microscale dimple morphology and is an important tool in failure analysis, in practice it is not always possible to obtain this information due to rubbing contact of the two surfaces. (However, as discussed elsewhere in this section, the macroscale swirl rubbing pattern on a ductile torsion surface is a strong indicator of that loading.) If the fracture surface is quite rough, as is common in combinations of ferrite and pearlite in hot-rolled or annealed steels, the points of contact may be smeared, but there are often regions below the rubbed surface that can still be viewed to obtain loading condition information.

Dimple Size. If more than one type of inclusion or a second phase and a distribution of inclusions are present, there is a distribution of dimple sizes that correlate qualitatively with particle size (Fig. 24). Dimple size can be correlated with particle size; with inclusion spacing, as noted previously; or with fracture strain. High-strength low-ductility materials show shallow dimples. Figure 24 shows the correlation with void size and inclusion spacing. A correlation between dimple depth (and therefore, relative ductility) in a steel as a function of temperature has been shown (Ref 45). A series of tensile specimens of a titanium alloy (B-120CVA) that were aged for various times prior to testing are shown in Ref 46. For short aging times, large shear dimples were observed on the fracture surface. For increased aging times, the shear dimple size decreased, concurrent with a decrease in fracture toughness. For the longer aging times, the fracture surface consisted of increasing amounts of intergranular fracture together with the much smaller dimples (Fig. 24b).

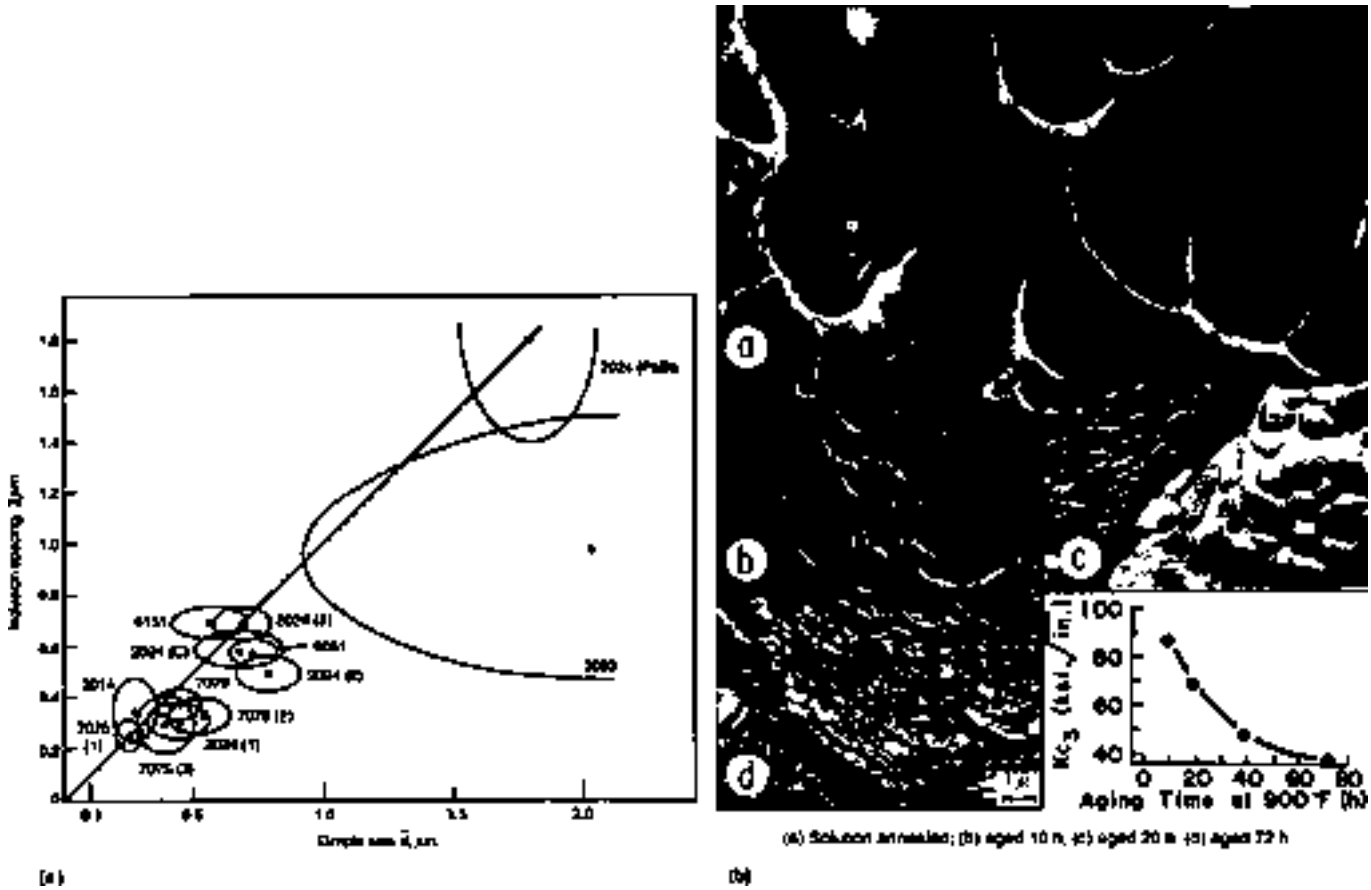


Fig. 24 (a) The correlation of particle spacing with dimple size. Data are from several aluminum alloys, as indicated. Ellipse size indicates scatter. Source: Ref 44. (b) Change in dimple size and onset of intergranular cracking as a function of aging time for a precipitation-hardened titanium alloy (B-1208CA). Source: Ref 46

As noted previously, a dual-dimple size is often observed in carbide-containing steels. (Fig. 15). A distribution of dimple sizes is also seen in aluminum alloys. However, in these alloys, if large second phases are present in the microstructure, cleavage cracking of these phases provides the early incipient cracks (Fig. 25).

Fig. 25 Micrograph of specimen shown in Fig. 17. Cleaved second-phase particles are visible in the microstructure, and no debonding at second-phase/matrix interfaces is visible. Source: Ref 42

References cited in this section

29. C.F. Tipper, *Metallurgia,* Vol 39, 1949, p 133

30. D.R. Johnson, “Toughness of Tempered Bainitic Microstructures in a 4150 Steel,” master's thesis, University of Tennessee, 1990

31. H. Tada, T. Kobayashi, and A. Takahashi, Micromechanism of Fracture in Wrought Aluminum Alloy Containing Coarse Inclusion Particles, *Alum. Trans.,* Vol 1 (No. 1), 1999, p 109

32. W.A. Backofen, *Deformation Processing,* Addison-Wesley, 1972

33. L.M. Brown and W.M. Stobbs, *Philos. Mag.,* Vol 334, 1976, p 351

34. G. Le Roy, J.D. Embury, G. Edwards, and M.F. Ashby, *Acta Metall.,* Vol 29, 1981, p 1509

35. A.W. Thompson, Modes of Ductile Fracture, *Metall. Trans. A,* Vol 18, 1987, p 1877

36. K.S. Chan, Near Tip Fracture Processes in Ductile Materials, *Recent Advances in Fracture,* TMS/AIME, 1998, p 137

37. H.C. Rogers, The Effect of Material Variables on Ductility, *Ductility,* American Society for Metals, 1968, p 31

38. C.J. Beevers and R.W.K. Honeycombe, *Philos. Mag.,* Vol 7, 1962, p 763

39. G.A. Lange, The Fracture Behavior of Pure Face-Centered Cubic Metals with Prevention of Macroscopic Deformation, *Scanning Electron Microscopy, Part III,* 1985

40. R.L. Lyles, Jr. and H.G.F. Wilsdorf, Microcrack Nucleation and Fracture in Silver Crystals, *Acta Metall.,* Vol 23, 1975, p 269

41. R.H. Van Stone and T.B. Cox, "Use of Fractography and Sectioning Techniques to Study Fracture Mechanisms," ASTM STP 600, *Fractography: Microscopic Cracking Mechanisms,* C. Beachem and W. Warke, Ed., ASTM, 1976 p 18

42. G. Powell, S.-H. Cheng, C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys,* Battelle Press, 1992

43. D. Chae, D.A. Koss, A.L. Wilson, and P.R. Howell, The Effect of Microstructural Banding on Failure Initiation of HY-100 Steel, *Metall. Trans. A,* Vol 31, 2000, p 995

44. D. Broek, *Engineering Fracture Mechanics,* Vol 5, 1973, as cited in R. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* 4th ed., John Wiley, 1996

45. A.J. Edwards, "Depth Measurements on Fracture Surfaces," report of NRL progress, Naval Research Laboratory, 1963, as cited in A.J. Brothers and S. Yukawa, "Engineering Applications of Fractography." ASTM STP 436, *Electron Fractography,* ASTM, 1968 p 179

46. B.R. Banerjee, J.J. Hauser, and J.M. Caperos, "Relation of Strength and Toughness to Fine Structures in a Beta Titanium Alloy," ASTM STP 432, *Applications Related Phenomena in Titanium Alloys,* ASTM, 1968

69. J.H. Bucher, G.W. Powell, and J.W. Spretnak, *Application of Fracture Toughness Parameters to Structural Metals,* H.D. Greenberg, Ed., Gordon and Breach Science Publishers, 1966, p 323

77. T.A. Anderson, *Fracture Mechanics,* 2nd ed., CRC Press, 1995

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Factors Affecting Ductility

For ductile material, crack growth can be stable, requiring an ever-increasing load to propagate a crack until a geometric instability is attained. This level of ductility normally requires gross deformation that is readily visible. For macroscale brittle materials, increasing loads can be stable up to a point when a crack can grow quickly with no prior gross deformation. This sudden crack growth is reported to be the predominant fracture process in commercial products (Ref 47).

Even microscale ductile material can develop critical size flaws (cracks) that result in rapid crack growth. The critical size flaw can be very large for ductile materials at low stress levels. For low-strength ductile materials, critical crack sizes tend to be large enough that visual inspection can detect them before they lead to a rapid crack growth. However, at stresses near the yield point, low-ductility materials can have toughness levels that result in critical flaw sizes (a_c) on the order of manufacturing and processing imperfections and even significantly smaller, as in quenched and tempered steels. The critical flaw size for macroscale (plane-strain) brittle fracture is:

$$a = 2.5\left(\frac{K_{IC}}{s_{YS}}\right)^2$$

Imperfections such as cold laps, hot tears, quench cracks, and tool marks can be severe enough to lead to fracture. Fatigue cracks and stress-corrosion cracks also may grow in ductile material without showing any gross deformation. Thus, various imperfections or stress raisers can be the source of crack initiation, stable propagation at subcritical size, and perhaps (if not avoided) catastrophic crack growth when cracks reach critical size.

This section briefly reviews the general geometric factors and materials aspects that influence the stress-strain behavior and fracture of ductile metals. This is followed by the section “Fracture Appearances,” which reviews appearances for various types of geometries and loading conditions. This is then followed by sections of “Fracture at or near Stress Raisers” and “Fracture from Manufacturing Imperfections,” both of which may influence the initiation and propagation of cracks. This is then followed by a case history.

Reference cited in this section

47. D. Broek, *The Practical Use of Fracture Mechanics,* Kluwer Academic Publishers, 1989, p 1

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Geometric Limits of Ductility

Ductile fractures are characterized by tearing of metal accompanied by appreciable gross (albeit possibly localized) plastic deformation and expenditure of work during the fracture process. On some scale, plastic deformation is always a localized event. Any matrix discontinuity, such as inclusions, second-phase particles, or even grain boundaries, leads to localized plastic deformation. On a specimen-scale, localized deformation may be forced by geometric constraints, for example, grooves, notches, holes, and cracks. The total absence of second-phase particles would possibly result in fracture by 100% reduction of area, but this rarely occurs in service failures. In components without geometric constraints, where uniform deformation should occur, macroscale localized deformation, known as shear bands, also can develop. Shear bands are discussed in the Appendix of this article.

Smooth (Unnotched) Cylindrical Specimen. An informative first approach to understanding constraints on ductility is to examine the failure of a smooth round bar loaded in uniaxial tension (tensile test). The failure process starts with a geometric instability in the form of a diffuse neck (described as diffuse because the load-elongation curve near maximum load stays at a high load for a large amount of deformation). Materials that do not develop a neck before fracture of a round tensile bar are generally considered macroscale brittle, although there is not a general dividing line between ductile and brittle fracture in terms of macroscale necking in a tensile-test specimen (see the discussion in the preceding article, “Fracture Appearance and Mechanisms of Deformation and Fracture,” in this Volume). For a brittle material, fracture occurs at a location where the local stress reaches a critical value dependent on the size and shape of the largest imperfection. The imperfection that initiates fracture can be internal, or it can be at an exterior surface and may be either geometric or metallurgical.

In ductile fracture during tension testing, necking generally begins at maximum load during the tensile deformation of ductile metal. In an ideal plastic material in which no strain hardening occurs, it would become unstable in tension and begin to neck as soon as yielding occurred. However, an actual metal (not heavily cold worked) undergoes strain hardening, which tends to increase the load-carrying capacity of the specimen as deformation increases.

The size of the neck and the extent to which it is visible depends primarily on strain hardening and strain-rate hardening (when at temperatures below approximately 0.4 of the melting temperature). Lowering the temperature below room temperature generally increases the strain-hardening exponent, *n*. This increases the strain to neck formation. However, the flow stress is also higher, so the resulting hydrostatic stress is higher at the same strain. For fcc material, the tensile elongation increases as the temperature decreases (Ref 48), but the reduction in area (RA) is affected less, with little change as the temperature is lowered.

No matter how long the sample, the remaining deformation and failure are confined to the neck that initially has a length dimension about twice the diameter (depending on the strain-hardening exponent and the strain-rate hardening exponent). The amount of elongation is predominantly controlled by the strain to the onset of necking, especially if strain-rate hardening is small. When strain-rate hardening is small, the elongation that

occurs after neck formation is small, compared to the elongation before necking. For this reason, reduction in area is more closely related to the true logarithmic tensile strain, $\varepsilon = \ln(A_i/A_f)$, at the fracture location than is total elongation.

The resulting necked region is, in effect, a mild notch. The formation of a neck in the tensile specimen introduces a complex state of stress that has a large tensile-hydrostatic component. It has been shown that a neck induces radial and circumferential stresses in addition to the axial stresses of the applied load (Ref 49). This triaxial (tensile-hydrostatic) stress is highest in the center of the specimen and causes growth of voids. Thus, a common observation is to note crack initiation near the center of the necked region. This central region of the fracture surface is typically flat at the macroscale. The plane of the flat-face fracture is normal to the direction of the largest tensile stress, and microscopic examination of flat-face tensile fractures in ductile materials magnifications of approximately 100× and greater reveals equiaxed dimples that are formed by microvoid coalescence. Fracture of the remaining section typically reveals an annulus of slant fracture, which is less well understood but probably occurs by a crack growing circumferentially around the specimen under plane-stress conditions. The ratio of the area of the flat-face region to the area of the shear lip usually increases with section thickness.

Similarly, notched samples create a stress state with a tensile-hydrostatic component during the onset of loading and during any subsequent plastic deformation. The tensile-hydrostatic component occurs immediately during loading of a notched bar, while it occurs after the onset of necking in a smooth bar. For the same specimen minimum diameter and load, a notched round bar has a much higher tensile-hydrostatic component than a necked unnotched bar. General deformation before fracture is therefore greatly limited in a notched bar compared to a smooth bar. The stress state in a notched bar is also different than that of a necked bar (Fig. 26). Unlike the round necked bar, which has the maximum tensile-hydrostatic (mean) stress component in the center, the hydrostatic stress of a notched bar is the highest just below the notch surface, with the magnitude depending on the depth of the notch and the notch radius.

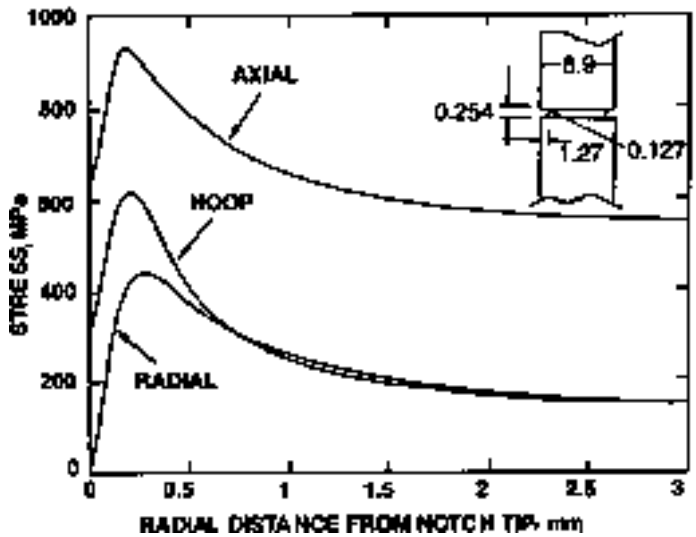


Fig. 26 The distribution of principal stresses beneath the notch root of a notched-bar specimen. Specimen dimensions in millimeters. Source: Ref 50

The magnitude of the tensile-hydrostatic component can alter the fracture initiation location. Figure 27 shows the effect of dramatically different notch radii for two notched bars of the same material. For the sample with the larger notch-root radius (2.5 mm in Fig. 27a), the fracture initiates with a fibrous texture at the center of the bar, develops a radial pattern as the fracture surface moves rapidly outward, and changes to a slant fracture before connecting to the outside surface. For the sample with smaller notch-root radius (0.25 mm in Fig. 27b), the fibrous fracture starts just below the notch, gradually grows outward, and connects to the outside surface by a very narrow slant surface. It also proceeds inward. As the fibrous fracture surface starts to move rapidly inward, it develops a radial pattern as it moves toward the center of the sample. As the notch-root radius is reduced, there is a progressive reduction of the amount of slant fracture.

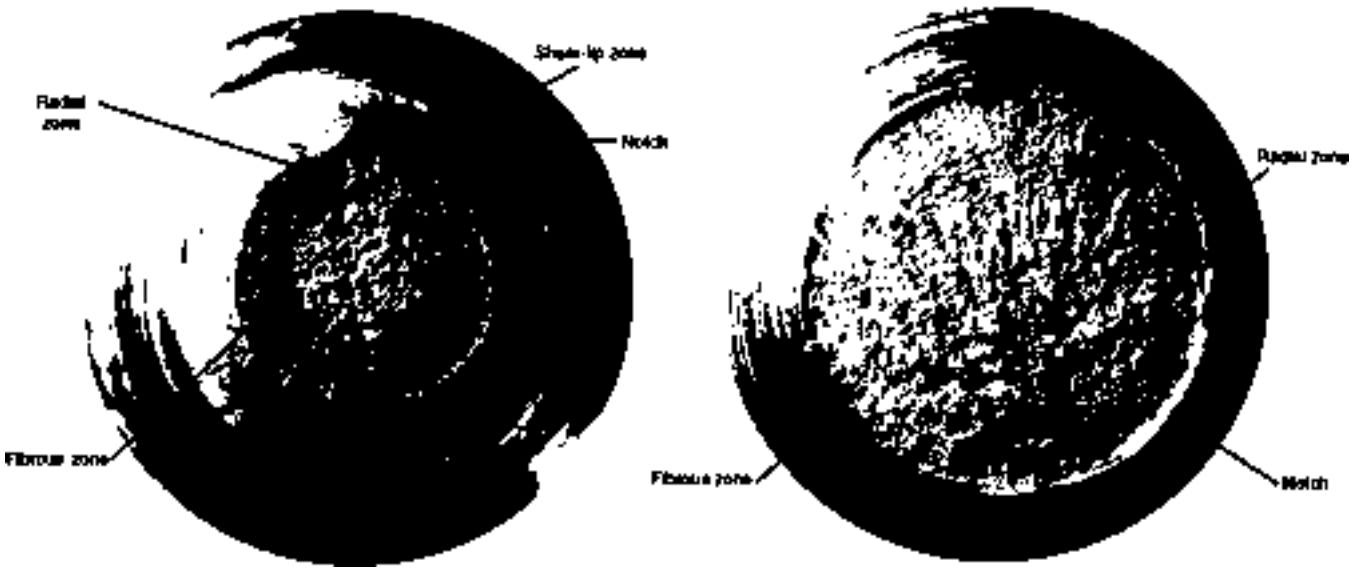


Fig. 27 Fracture surfaces of notched round specimens (4340 steel) from tensile overload at -40 °C (-40 °F). (a) Specimen with a mild notch with a root radius of 2.5 mm (0.1 in.) produced a fracture similar to an unnotched bar (i.e., central fibrous zone with shear lips). Tensile strength at test temperature was 1544 MPa (224 ksi). (b) Specimen with a sharper notch with a root radius of 0.25 mm (0.01 in.) altered the stress state, caused the fracture to initiate near the root of the notch, and produced an initial ring of circumferential ridges and a central radial tone. Notched tensile strength at test temperature was 1760 MPa (255 ksi). Tensile strength at room temperature, 1150 MPa (167 ksi)

Local Necking in Prismatic Specimens. When the component geometry is prismatic rather than cylindrical, both uniaxial and multiaxial loading permit the development of plane-strain deformation, resulting in the formation of a local neck. Visually, the local neck consists of a narrow band of material (of the order of the section thickness) in which plane-strain deformation conditions exist such that within the neck, axial extension occurs by a decrease in thickness. Figure 28 shows this behavior in copper and 400 Monel. The neck is described as local because of its narrow width, in contrast to diffuse necking in which the neck grows a considerable distance (controlled by strain and strain-rate hardening) along the length of the specimen. The term *diffuse* arises from the small change in load with large length changes near the maximum load where the neck initiates.

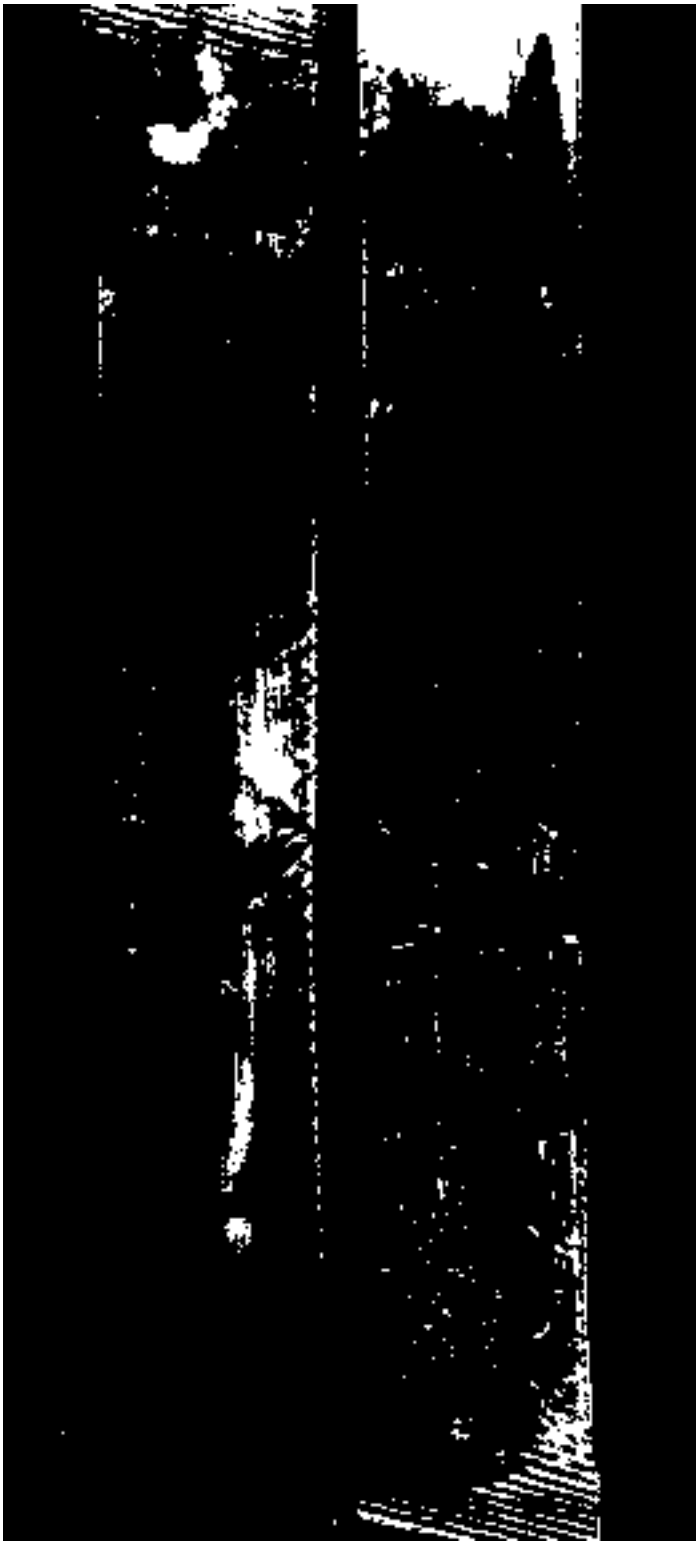


Fig. 28 Local and diffuse necking in copper and in 200 nickel. Width-to-thickness ratio ~24:1. Note that fracture initiation has occurred in the local necks at the edges of the specimens.

If the width-to-thickness (w/t) ratio is large, two necks are seen in the specimen (assuming some strain hardening), one described as a diffuse neck and the second as a local neck (Fig. 29). The diffuse neck is seen to form first and is analogous to the neck observed in cylindrical specimens. A plasticity analysis predicts the formation of a local neck within the diffuse neck (Ref 52), because a plane of zero extension develops. The local neck is predicted to form at a strain equal to twice the strain-hardening exponent, $2n$ (see the Appendix of this article).

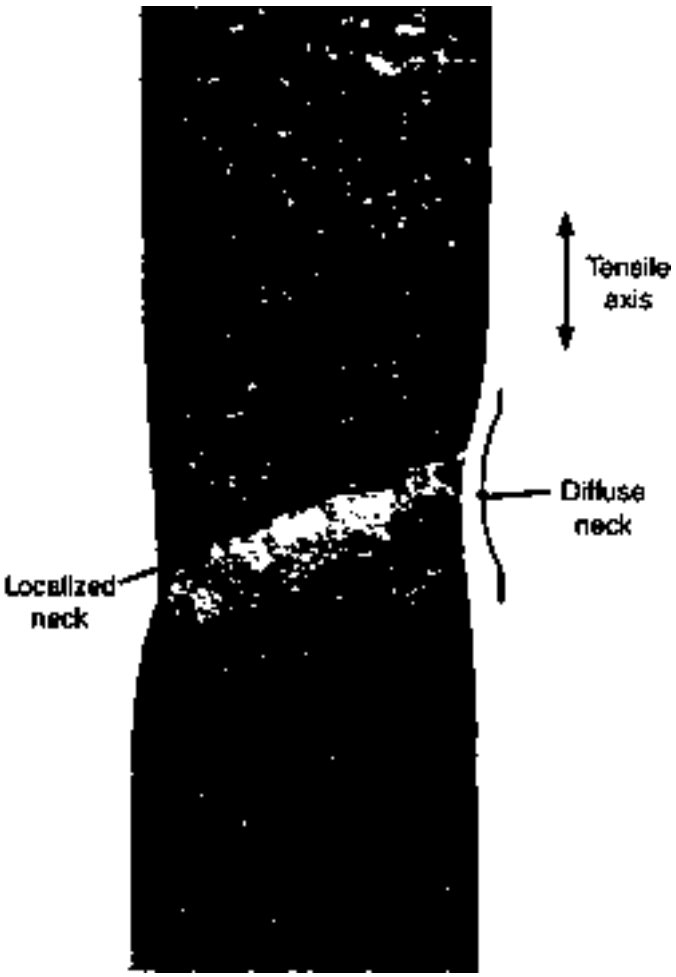


Fig. 29 Diffuse and localized necks in an 1100 aluminum sheet tensile specimen. Source: Ref 51

The length (in the load direction) of the diffuse neck is approximately twice the width of the sample, while the width of the local neck is approximately twice the thickness, so long as the specimen has adequate length so that grip constraint is not a factor and so long as the material shows typical strain and strain-rate hardening. A strain analysis using Mohr's circle predicts that the local neck should form at an angle of approximately 55° to the axis of the sample in isotropic material. The angle of the local neck does vary with texture in the material and can be predicted via Hill's anisotropic yield criterion. Although not treated here, local necking also occurs for some biaxial loading conditions and is the dominant failure mode is sheet metal stretching operations (see, for example, Ref 32, p 207, and the Appendix of this article).

The actual strain at which the local neck forms and the angle at which it forms is influenced by the size of the diffuse neck, because the neck has a nonuniform distribution of strain. For example, the length of the sample should be eight times its width to avoid end effects that can alter the state of strain and stress in the neck.

Effect of Size. Large structures generally have less ductility than small ones. There are two reasons for this observation. First, necking in larger structures represents more geometric constraint that tends toward a plane-strain condition with higher tensile-hydrostatic stress component, which reduces the strain required to initiate fracture. Second, the probabilistic nature of imperfections and lack of homogeneity in commercial alloys suggests that larger structures contain a wider range of imperfection size. Smaller components usually undergo more deformation during processing and would have, on average, smaller and more uniformly distributed imperfections. Predicting the effect of size on final structure properties is difficult. Most material properties are determined using relatively small samples. Knowing the actual size and distribution of inclusions and second-phase particles in a very large structure is not an easy assessment. There has been no reasonable prediction of reduction in ductility due to increased part size. In general, the reduction is small.

References cited in this section

32. W.A. Backofen, *Deformation Processing*, Addison-Wesley, 1972

48. H. Conrad, *High-Strength Materials,* V.F. Zackay, Ed., John Wiley & Sons, Inc., 1965, p 436

49. A. Nadai, *Theory of Flow and Fracture of Solids,* 2nd ed., McGraw-Hill, 1950

50. M.G. Mendiratta, R. Goetz, D.M. Dimiduk, and J.J. Lewandowski, *Met. Mater. Trans. A,* Vol 26, 1995, p 1772

51. S.L. Semiatin and J.J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization,* American Society for Metals, 1984

52. P.F. Thomason, *J. Inst. Met.,* Vol 96, 1968, p 360

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Materials Factors Affecting Ductility

For single crystals oriented in an optimal direction, the neck in a tensile test can reduce to a point, that is, 100% RA (Fig. 1). Because single crystals are not isotropic, the cross section does not remain round, but the true strain at the point of separation approaches an infinite value. Similar ductility in tension can be attained in polycrystalline materials at elevated temperatures and in very pure metals. More often, a failure occurs well before 100% RA, with typical values in the range of 40 to 70% RA for commercial alloys.

The development of a “chisel” point at the site of final separation has not been reported in commercial alloys. A noncircular cross section in the neck causes an unusual shape to the final separation. Noncircular final fracture areas that are square, star-shaped, and elliptical can be caused by the presence of a strong texture in single-phase alloys. Texture in polycrystalline materials can lead to unusual geometric changes during tensile deformation. Texture causes the sample to act more similar to a single crystal, sometimes resulting in noncircular cross sections.

Effect of Grain Size. The well-known Hall-Petch relationship describes the effect of grain diameter on yield stress:

$$\sigma_y = \sigma_o + kd^{1/2}$$

where σ_y is lower yield stress, σ_o and k are constants, and d is the grain diameter. A similar response for the flow stress as plastic deformation proceeds has been found, that is, smaller grain material has a flow stress higher than larger grain material at all levels of plastic strain. Some investigators have found that steel with smaller grain size also tends to have greater total elongation to fracture, but this is not a universally confirmed response. There is consistent data supporting the claim that grain size does measurably affect the ductile-brittle transition temperature in steels. A good review summarizing the effect of grain size on mechanical properties can be found in (Ref 53).

The effect of grain size is relatively small for the range of sizes found in typical commercial alloys. However, when grain size becomes very large, crystallographic slip leads to continuity problems at grain boundaries. This can lead to brittle fracture at the grain boundary. The effect of large grains has been noted to influence the reproducibility of mechanical properties.

A material with large grain size also has a much lower ductility, if significant levels of impurities are present. The explanation for this is that grain boundaries tend to be a sink for impurities. When the total surface area becomes small (large grains), the local concentration at the grain boundaries becomes critically large. This can lead to both macroscale and microscale brittle failures. On a microscale, the fracture is at the grain boundaries; that is, intergranular. Because the grain size is often very large the crystalline, shiny appearance of the grains may be visible to the naked eye in some cases, and there may be virtually no ductility associated with the failure. However, Fig. 30 shows a fractured tensile specimen of brass in which there was extensive tensile

elongation, little transverse necking, and no cup-and-cone fracture. This alloy would fail by necking and cup-and-cone failure if the grain size were smaller. Also apparent in this figure is the rough “orange-peel” surface texture on the outside of the sample, which is another macroscale indication of a very large grain size. The rough surface and large grains may lead to cracks initiating at the outside surface.

Fig. 30 Two tensile bars of cartridge brass: one tested to failure, showing low ductility; other not tested. Rough “orange-peel” surface texture on the outside of the tested sample is another macroscale indication of a very large grain size.

Reference cited in this section

53. T. Gladman and F.B. Pickering, The Effect of Grain Size on Ferrous Materials, *Yield, Flow, and Fracture of Polycrystals,* T.N. Baker, Ed., Applied Science Publishers, 1983

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Fracture Appearances

Macroscale and microscale appearances of ductile and brittle fracture are influenced by:

- External loading conditions together with the constraint introduced by the presence of the imperfection (i.e., the stress-intensity, or *K*, field)

- Metallurgical structure

This section introduces macroscale and microscale appearances of ductile and brittle fracture for common specimen geometries (smooth cylindrical or prismatic) and loading conditions (e.g., tension, compression, bending, torsion). This is then followed by a review of fractures from manufacturing imperfections and stress raisers.

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Cylindrical Specimens in Tension

Figure 9(a) shows schematically the typical surface features for a ductile fracture of a smooth (unnotched) cylindrical tensile-test section in comparison to the typical fracture surface for a notched tensile bar specimen (Fig. 9b). For a smooth-bar specimen, the classic cup-and-cone geometry is typical for ductile fracture (Fig. 8a) as opposed to the flat surface from brittle fracture of an unnotched bar specimen (Fig. 8b).
In the absence of severe surface or subsurface imperfections or stress raisers, the ductile fracture process in a smooth bar starts with the development of internal voids or cracks near the center of the neck in the region of high mean normal stress, creating a somewhat flat surface that is nominally perpendicular to the tensile axis. This internal fracture surface grows and finally connects to the outside surface through a change in the mode of fracture to a shear or a slant fracture that is at an angle to the tensile axis. In contrast, fracture in a notched specimen may initiate near the root of the notch, depending on the severity of the notch (e.g., Fig. 27).

Ductile Fracture (Unnotched Bar)

The classic cup-and-cone fracture surface typically has three distinct surface textures: the central geometry fibrous zone, radial marks, and shear lips. The center area that represents the gradual internal crack formation is fibrous. The slant surface or shear lip that forms the cup wall is usually considered a smooth shear surface. In some cases, the slant or shear lip does not always form a complete cup (e.g., see Fig. 31). The transition to a cup surface or side wall on one half of the fracture surface to the other is usually sharp.

Fig. 31 Tensile fracture of maraging steel showing transition from cup feature from one half of the fracture to the other half, that is, not a complete cup-and-cone

Lowering the temperature increases the size of the flat region and results in a reduction of the surface area of the slant-fracture region (Ref 54). Body-centered cubic (bcc) materials and hexagonal close-packed (hcp) materials also undergo a dramatic ductility reduction in the ductile-brittle transition temperature (DBTT) region, which can cause a change in fracture appearance. In the DBTT region, the fibrous fracture-surface area decreases, and radial patterns (discussed subsequently) either grow (if they are already present) or start to appear, if they are not initially present above DBTT.

In rare instances, fracture surfaces have been reported to consist of another feature, where a circumferential ridge surrounds the fibrous zone. In one case of a cylindrical aluminum 1100 specimen, the fracture surface does not have radial marks but does show a small height circumferential ridge surrounding the very small fibrous zone, followed by another small region of macroflat fracture and then a macroscale shear region at the surface. In this aluminum specimen, the intervening zone between the radial region and the near-surface shear lip cannot be cleavage. In another case (Ref 55), the central fibrous region is encircled by a reflective (brittle) region and finally, a small shear lip. Figure 27(b) also shows circumferential ridging. In a ferritic or other bcc material, it is possible to trigger cleavage cracking depending on the degree of constraint, test temperature, loading rate, and grain size.

Radial Marks. Under the right conditions, the central fibrous region is sometimes surrounded by another region containing radial lines or ridges. This radial pattern, sometimes called a radial shear, star, or rosette, is a region that is generally associated with rapid or unstable crack propagation. Radial marks are considered to be perpendicular to the crack front and, as such, may provide the round-sample equivalent appearance of chevrons that appear on sheet or plate samples.

The radial marks, or radial shear marks, are visually distinct from the central fibrous region. Figure 32 shows the radial patterns for an isothermally transformed bainitic steel (4150) with different heat treatments. Figure 32(a) shows a clear boundary between the fibrous center region and the large ridge pattern of the radial marks. Figure 32(b) shows shallower radial marks and a slightly larger fibrous zone from a heat treatment that results in more ductility. Figure 32(c) shows very weak or shallow radial marks that develop further from the center.

Fig. 32 Radial marks on tensile test specimen of Society of Automotive Engineers (SAE) 4150 steel isothermally transformed to bainite, quenched to room temperature, and then tempered.

Specimen	Hardness, HV	Yield strength	Ultimate tensile strength	Reduction of area, %	Charpy V-notch impact energy	Fibrous zone as percentage of total area			
		MPa	ksi		MPa		ksi	J	ft · lbf
(a)	285	0.73	0.106	0.83	0.120	66	163	120	**~25**
(b)	258	0.65	0.094	0.79	0.115	67	174	128	**~31**
(c)	**301**	**0.81**	**0.117**	**0.97**	**0.141**	**49**	**27**	**20**	**~44**

Figure 33 is a schematic of radial shear marks that appear as very high ridges with a slant surface on one side and a vertical surface (aligned with the sample axis) making up the other side. The vertical portion of the fracture surface has been called delamination by some investigators (Ref 56). These clear differences in radial

marks suggest that the mechanism by which radial patterns are created may not be the same for all radial marks. This delamination is visible in Fig. 32(a).

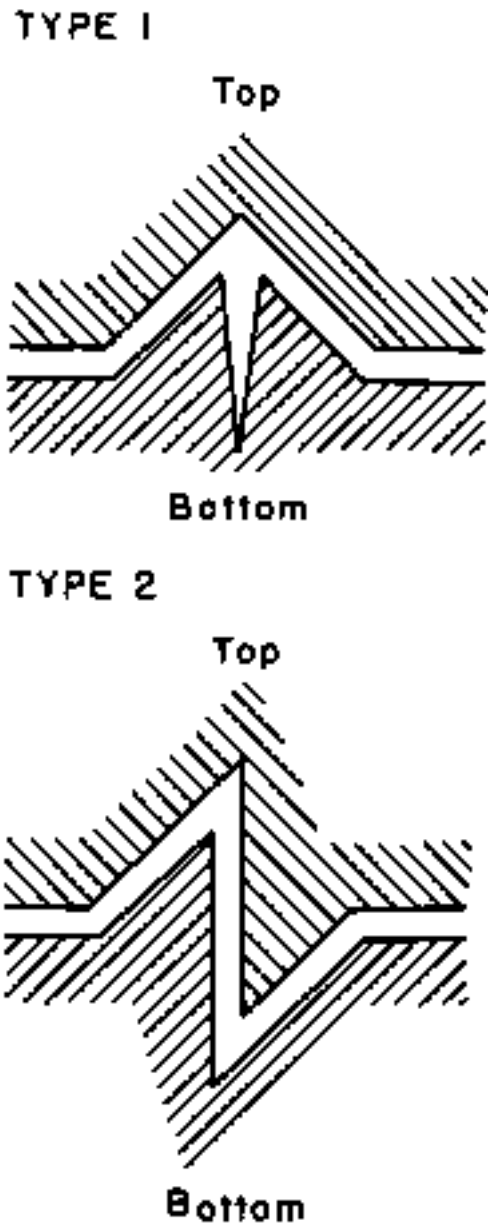


Fig. 33 Sketch of two types of radial shear elements that would lead to a radial mark. Source: Ref 54

The general use of the term *delamination* has been applied to material with aligned inclusions that are responsible for a crack along the length of a tensile specimen but not necessarily star patterns. The delamination surface of a radial shear element is generally not associated with inclusions but is attributed to low transverse ductility and the high hoop stress created by constraint in the neck. Large-scale delamination has been observed in metals and is not associated with a radial pattern nor attributed to material quality but rather reported to be dependent on mean normal stress (hydrostatic tension) and level of strain (Fig 34). This type of delamination predominately occurs in high-strength material containing void-nucleating particles.

Fig. 34 Delamination (longitudinal crack) in a tensile specimen. Source: Ref 56

For some materials and test conditions, the flat portion can consist entirely of a radial marked surface with no fibrous region. This suggests that a small imperfection can cause unstable, rapid crack growth. This condition also results in a very small shear lip. Radial marks have been observed on notched samples made from material that does not have radial marks on the fracture surface of smooth samples. The fracture surface of a notched pearlitic steel bar tested at 43 °C (110 °F) has been found to have areas of shiny brittle failure, while a smooth round bar of the same material tested at the same temperature and strain rate had a ductile cup-and-cone fracture.

Shear Lip. It is generally accepted that the shear lip of a cup-and-cone fracture forms at a high strain rate, when the elastic energy stored in the sample and the test machine cannot be reduced by slow propagation of the fracture surface. Thus, the size of the shear-lip region is influenced by machine stiffness, with stiffer testing machines creating less slant-fracture surface (Ref 57). The fracture toughness of the material dictates the size of the internal crack before it becomes unstable (critical). However, even for very brittle materials that have radial marks on all the flat surface, a shear lip may form at the tip of an advancing crack before it intersects an outside surface. These shear lips are very narrow for a brittle material. In general, the presence of a shear lip indicates a level of ductility, with large shear lips indicating greater ductility. A large shear lip is also associated with a small fibrous zone. Because the size of the fibrous zone is significantly affected by the volume fraction of inclusions, a large shear lip is also associated with cleaner material.

A universally accepted explanation for the transition of an advancing internal crack to a shear lip has not been developed. The primary questions are: does a crack change directions to an internal slant surface and grow to the outside? Or, does a gross, specimen-scale localized shear band develop and lead to failure more or less uniformly on the entire surface? It has been suggested that the shear lip may be a result of an unstable, nonuniform shear deformation process known as adiabatic shear (Ref 58). This process is discussed in more detail in the next section. The shear lip is not a flat plane, but rather it is curved in two dimensions or a hyperbolic surface in three dimensions, with the angle between the surface and the tensile axis ranging from 25 to 60° (Fig. 35a).

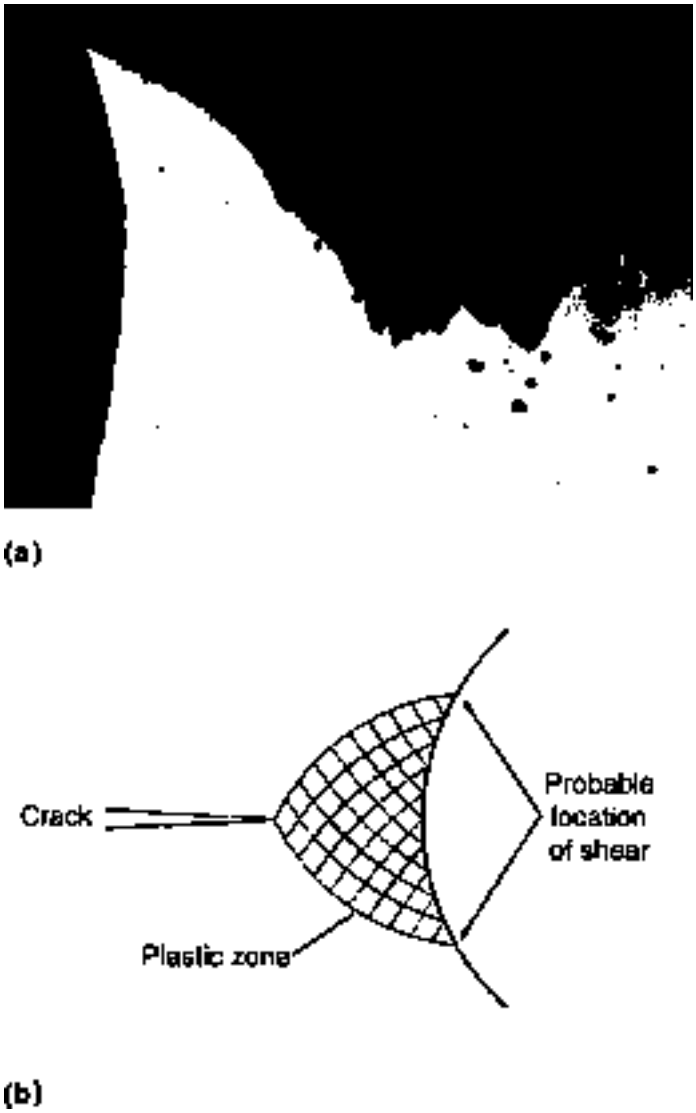
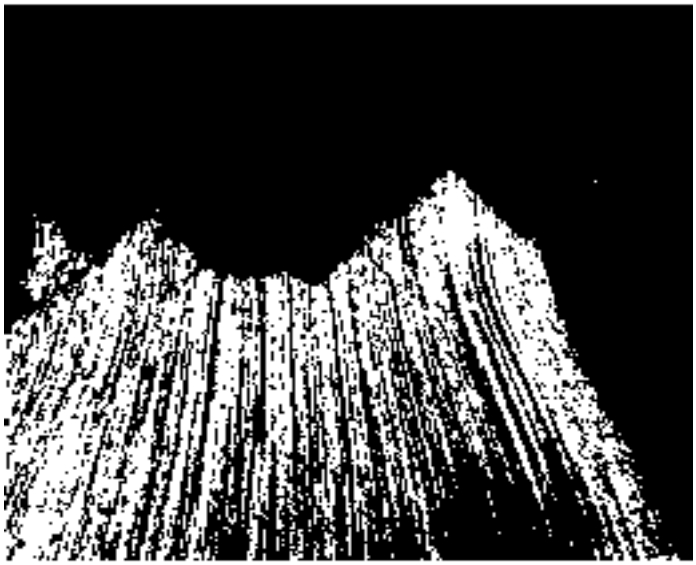


Fig. 35 Shear-lip formation. (a) Cross section of fractured tensile specimen of 1020 steel. (b) Sketch of estimated slip-line field at the tip of a central crack as it approaches the exterior surface of a necked tensile specimen

If the crack-growth mode changes to shear as the crack approaches an external surface, two equally probable planes of shear should develop. These planes would consist of a symmetrical pattern that would require the formation of a shear plane above and below the central crack plane. As previously noted, numerous tensile fractures are not full cup-and-cone; in some areas, a shear lip forms to the “top” partway around the bar and, in the remaining volume around the central crack, forms a shear lip on the “bottom” (Fig. 31). Most steel samples do not have exterior surface indications that shear had developed both above and below the crack. However, Fig. 36(a) shows significant local deformation in the neck region opposite the fracture shear lip. A section of this sample reveals localized deformation on the opposite side of the central crack (Fig. 36b).

(a)



(b)

(c)

Fig. 36 Local deformation in shear-lip formation. (a) UNS C71500 (70Cu-30Ni) steel tensile specimen showing localized deformation at the exterior surface of the necked region. (b) Cross section of sample in (a) showing shear nature of localized deformation in a region opposite the slant fracture. (c) Higher-magnification view of (b). Courtesy of FTI/Anamet Laboratory

The state of stress in the neck of a round bar is complex. Because of the three-dimensional character of the neck, planes of maximum shear may appear, as seen in Fig. 37. If a crack develops in the center region, the stress state becomes more complex. Figure 35(b) depicts possible planes of shear deformation at the tip of a crack. These lines represent equally probable shear planes for a plane-strain condition. With a large central crack, the circumferential direction is constrained, thus creating a plane-strain condition. The curvilinear, theoretical planes of pure shear of Fig. 35(b) are consistent with the actual shear-lip plane, as depicted in Fig. 35(a); that is, the fracture surface intersects the free surface at 45°.

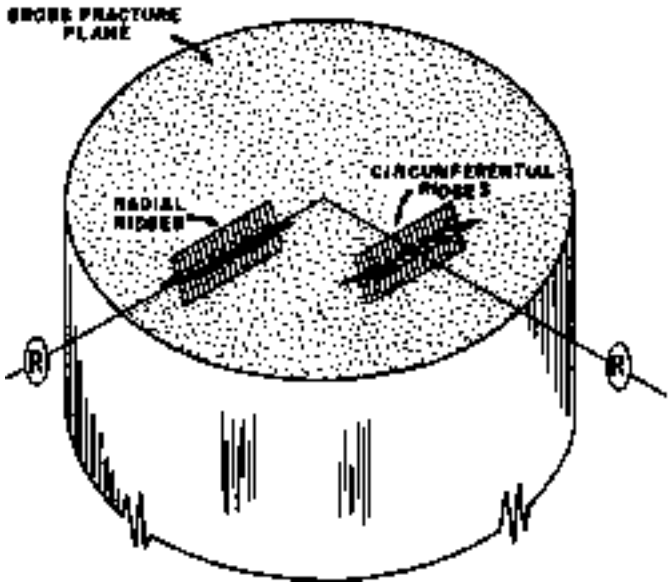


Fig. 37 Maximum shear stress planes for radial and tangential stresses created by necking. Source: Ref 54

Some necked round tensile specimens have not failed by cup-and-cone but have demonstrated complex shear fractures (e.g., Fig. 38). Multiple shear bands are visible on the surface of the specimen. Crack initiation is visible at the intersections of the shear bands (thin black lines). There are many of them on the specimen, and multiple (very) small cracks can be seen at the intersection of these bands. Also, these bands can be extended to the top fracture surface to show that cracking did occur within these bands. Figure 39 shows a shear fracture plane in a complex cup-and-cone fracture that makes a well-defined angle at the outside surface, similar to those seen in Fig. 38. This is in contrast to the surfaces connecting upper and lower shear lips shown in Fig. 31.

Fig. 38 Tensile fracture of a 1020 steel showing slanted fracture intersecting the outside surface at an angle

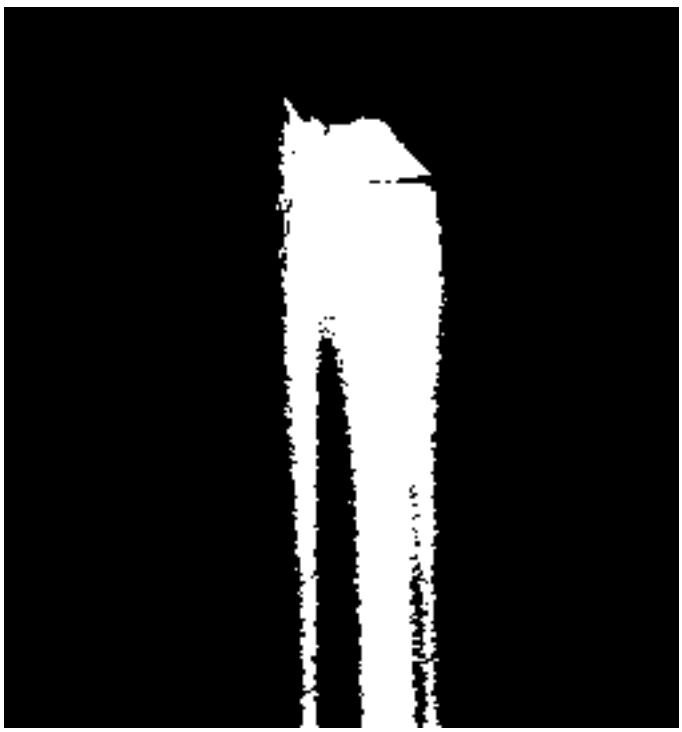


Fig. 39 Fractured 1020 steel showing an angled connection between a cup portion on one half of the fractured bar and a cup portion on the other half

The axially symmetric shear lip of a cup-and-cone fracture may represent only a unique set of active shear planes. The shear planes depicted in Fig. 37 show that shear planes can contain a circumferential element. This suggests that, at some combination of material properties and stress state, radial shear can extend to the outside surface. Predicting the dominance of one of these shear planes over the other has not been addressed in detail.

Effect of Strain Rate. Data from tensile impact tests of round bars can be combined with standard tensile-test machine test data generated at low strain rates to demonstrate the effect of strain rate. There is no clear trend in

the change in tensile elongation over a wide range of strain rates in tension testing of steel (Ref 59). Fracture surface appearance is likewise not apparently affected by strain rate for unnotched round samples tested at strain rates typical of tensile testing (10^{-3} to 10^{-2} min^{-1}). However, this may not be the case at higher strain rates. This is also not the case for other specimen geometries, especially steel specimens with cracklike imperfections. The section "Bending" in this article describes the effect of strain rate for notched bend specimens. The state of stress at a notch more closely simulates large engineering structures with imperfections than does a smooth round tensile specimen. For steel in the ductile-to-brittle temperature range, higher strain rates reduce the expected ductility.

Tensile Shear Failure. In general, necking instability develops for macroscale ductile materials. In rare cases, early failure can occur by shear localization. This is due to a flow stress instability that develops early in the deformation process. The mathematical arguments for adiabatic shear formation can be used at low strain rates to analyze this deformation localization process (see extended discussion in the Appendix of this article).

For most materials, increased strain causes increased flow stress (work hardening), increased strain rate causes increased flow stress (positive *m* value), and increased temperature causes lower flow stress (thermal softening). So, thermal softening is typically the only variable that can cause net softening. However, some aluminum alloys have negative strain-rate sensitivities and low strain-hardening exponents at room temperature. At higher homologous temperature, dynamic recrystallization or other metallurgical changes may occur that cause a loss in strain and strain-rate hardening. As a consequence, tensile tests of round samples of alloys with a negative strain-rate sensitivity exponent can fail by through-sample shear before necking (e.g., Ref 60). A very narrow shear plane of this failure is unique for round bars, and its appearance limited to rare cases of very low or negative strain-rate sensitivity together with a low strain-hardening capacity (see Appendix). Through-sample shear of round bars does occur in other unusual conditions, such as applied hydrostatic pressure (Ref 61) and large grain material (Ref 62), but the shear plane is more diffuse in these cases.

The temperature range for which through-sample shear occurs in aluminum alloys is limited. Figure 40 shows that the strain-rate sensitivity is negative over a very small range of temperature. The same material that demonstrates through-sample shear at room temperature (Ref 60) was found to have greater ductility and to develop a cup-and-cone fracture at cryogenic temperatures, a temperature where the strain-rate sensitivity is positive.

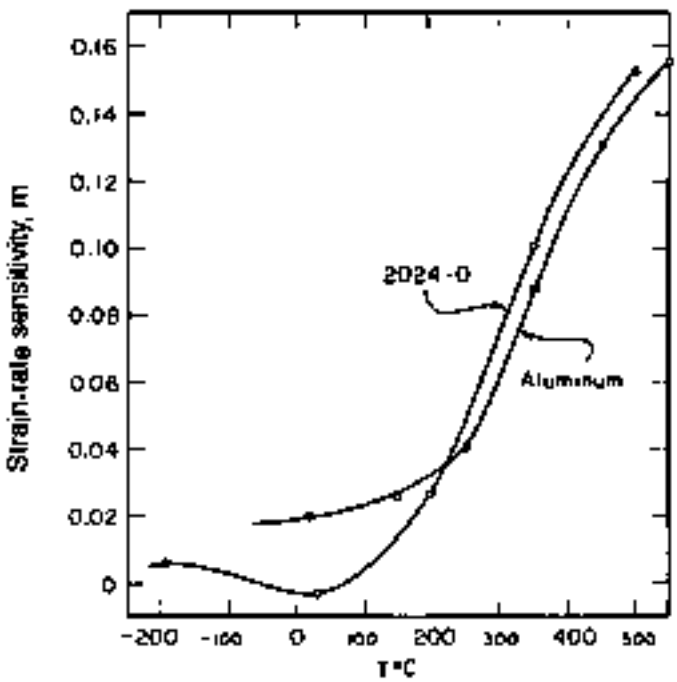


Fig. 40 Temperature dependence of the strain-rate sensitivity of pure aluminum and alloy 2024. Note 2024-O becoming negative near room temperature. Source: Ref 63

Footnotes

* (a) Lower bainite, isothermally transformed at 300 °C (570 °F) for 1 h, tempered at 600 °C (1110 °F) for 48 h.

** (b) Lower bainite, isothermally transformed at 375 °C (705 °F) for 1 h, tempered at 600 °C (1110 °F) for 48 h.

*** (c) Upper bainite, isothermally transformed at 450 °C (840 °F) for 24 h, as-quenched

References cited in this section

54. F.R. Larson and F.L. Carr, Tensile Fracture Surface Configurations of a Heat Treated Steel as Affected by Temperature, *Trans.* American Society for Metals, Vol 55, 1962, p 599–611

55. G.T. Hahn, B.L. Averbach, W.S. Cohen, and M. Cohen, *Conference on Fracture,* (Swampscott, MA), John Wiley & Sons, 1959, p 4-1

56. G.R. Speich and W.A. Spitzig, *Metall. Trans. A,* Vol 13, 1982, p 2250

57. A.W. Thompson and P.F. Weihrauch, *Scr. Metall.,* Vol 10, 1976, p 205

58. Y.W. Kim and D.L. Bourell, *Metall. Trans. A,* Vol 19, 1988, p 2041

59. M.J. Manjoine, *J. Appl. Mech. (Trans. ASME),* Vol 11 (No. 4), 1944

60. N. Chung et al., *Acta Metall.,* Vol 25, April 1977, p 377–381

61. P.F. Weinrich and I.E. French, Influence of Hydrostatic Pressure on the Fracture Mechanisms of Short Tensile Specimens of Cu and Brass, *Acta Metall.,* Vol 24, 1976, p 317

62. K.E. Puttick, *Philos. Mag.,* Vol 4, 1959, p 964

63. W.A. Backofen, *Trans. ASM,* Vol 51, 1959, p 946

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Prismatic Specimens in Tension

The same three zones present in the cylindrical specimen (fibrous, radial, shear lips) may be present on the fracture surface of a smooth prismatic tension-test section. The macroscale fracture appearances of unnotched sections are shown schematically in Fig. 41 for sections from square to sections with relatively high width-to-thickness (w/t) ratios. The possibility of crack initiation at the surface in still larger w/t ratios, is discussed subsequently and is not shown in Fig. 41. The w/t ratio also influences the ability of the sample to maintain a unidirectional state of stress. Very wide and thin samples develop a constraint to strain in the width direction. This tends to a condition of plane strain that dramatically alters the fracture process, the appearance of the fracture, and the total deformation to failure of the part.

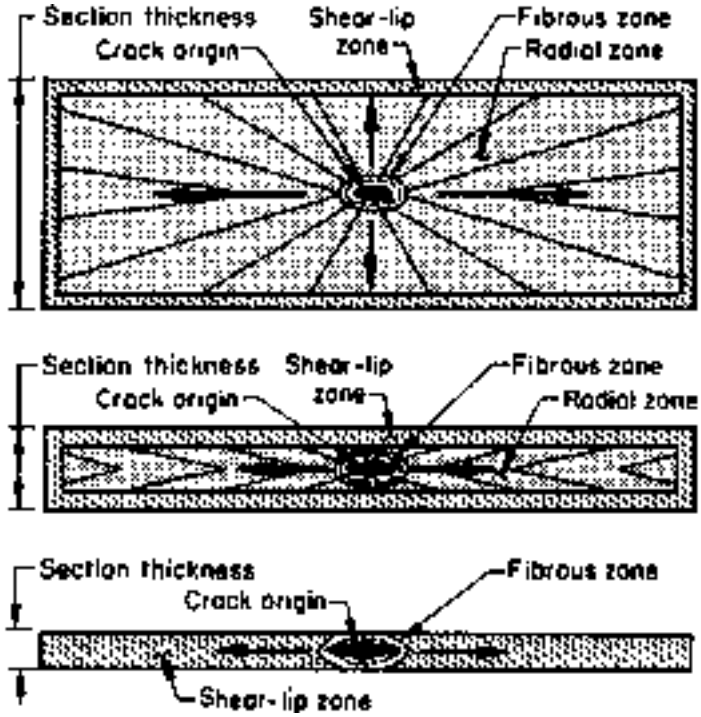


Fig. 41 Typical fracture appearances for unnotched prismatic tension-test sections

For an incompletely constrained steel edge- or side-notched rectangular section (Fig. 42), a fibrous zone may be generated from the notch and extend for a short distance. Outside of the fibrous zone is a region of either chevron marking or radial markings near the fibrous zone, which is surrounded at the specimen surface by a shear lip. The appearance of chevrons or radial marks near the fibrous zone depends in part on whether the crack-growth velocity at the surface is greater or less than that below the surface. If crack-growth velocity is at a maximum at the surface, radial marks have a fan-shaped appearance (Fig. 43). If crack-growth rate is greatest below the surface, the result is chevron patterns (Fig. 44). There is extended coverage of the cracking processes that can occur with a notch of varying degrees of constraint in "Microscale Details of Initiation and Propagation" in this article.

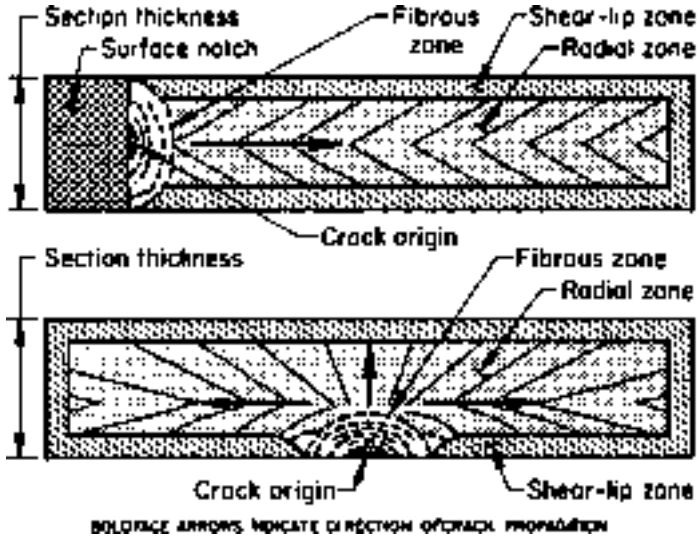


Fig. 42 Typical fracture appearances for edge- and side-notched rectangular tension-test sections. Note the shear lips when the fracture approaches the edge of the specimen.

Fig. 43 Radial marks typical of crack propagation that is fastest at the surface (if propagation is uninfluenced by the configuration of part or specimen)

Fig. 44 Chevron patterns typical when crack propagation is fastest below the surface. It is also observed in fracture of parts having a thickness much smaller than the length or width (see middle illustration in Fig. 40).

Unnotched Specimens

Analysis of an axially loaded tensile sample of plate and sheet material is a direct approach to understanding ductile tensile failure of prismatic components. As with round bars, the elongation is limited by strain instability that leads to neck formation. If the *w*/*t* ratio is small (for example, *w*/*t* = 1 to 1 or 1 to 2), instability leads to neck formation (Fig. 45a, b) and growth, followed by fracture, as in cylindrical specimens, although transverse contraction varies across the width and thickness directions, caused by constraint at the corners of the cross section. The same three zones present in the cylindrical specimen (fibrous, radial, shear) may be present.

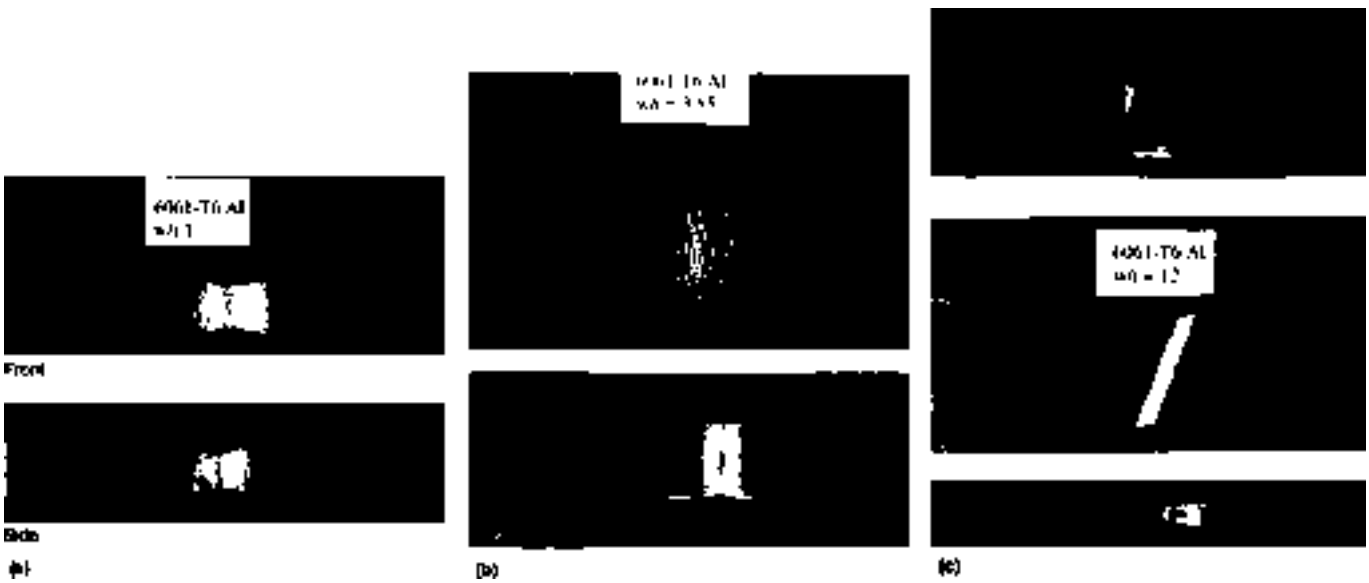


Fig. 45 Tensile fractures of aluminum alloy (6061-T6) sections with various width-thickness (*w*/*t*) ratios: (a) Diffuse necking (*w*/*t* = 1). (b) Diffuse necking with *w*/*t* = 3.85. (c) Local necking superimposed on a diffuse neck with *w*/*t* = 12 with magnified side view 1.5×. Courtesy of Patrick Werner, University of Tennessee

Increasing the *w*/*t* ratio causes:

- A change in the relative size of the three potential zones (fibrous, radial, shear)
- A change in the fracture surface orientation
- A change in the relative amount of transverse necking strain prior to fracture

As noted previously, there is constraint at the corners of the specimen, causing a resistance to transverse deformation in these regions. Transverse strain is at a maximum at half-width and half-thickness, but transverse strain is also decreased as a dimension is increased. Necking strain in the width direction decreases as the width direction increases, and similarly for the thickness direction. There is a change in the size of the shear lips intersecting the width and thickness direction. As the width increases relative to the thickness, the shear lip

emanating from the thickness dimension of the plate decreases. There may be either one or two slant fracture surfaces through the thickness direction, creating a V-fracture surface in the latter instance.
If the *w*/*t* ratio is large, two necks are seen in a specimen (Fig. 28, 29, 46a, b): one described as a diffuse neck and the second as a local neck. The diffuse neck is seen to form first and is analogous to the neck observed in cylindrical specimens. As is apparent in Fig. 28, necking strain in the width direction approaches zero or, more accurately, along the width direction of the band. Because there is no strain in this direction, the band forms under conditions of plane-strain deformation. A strain analysis for this case suggests that the angle of the band should be at approximately 55° to the load for an isotropic material.

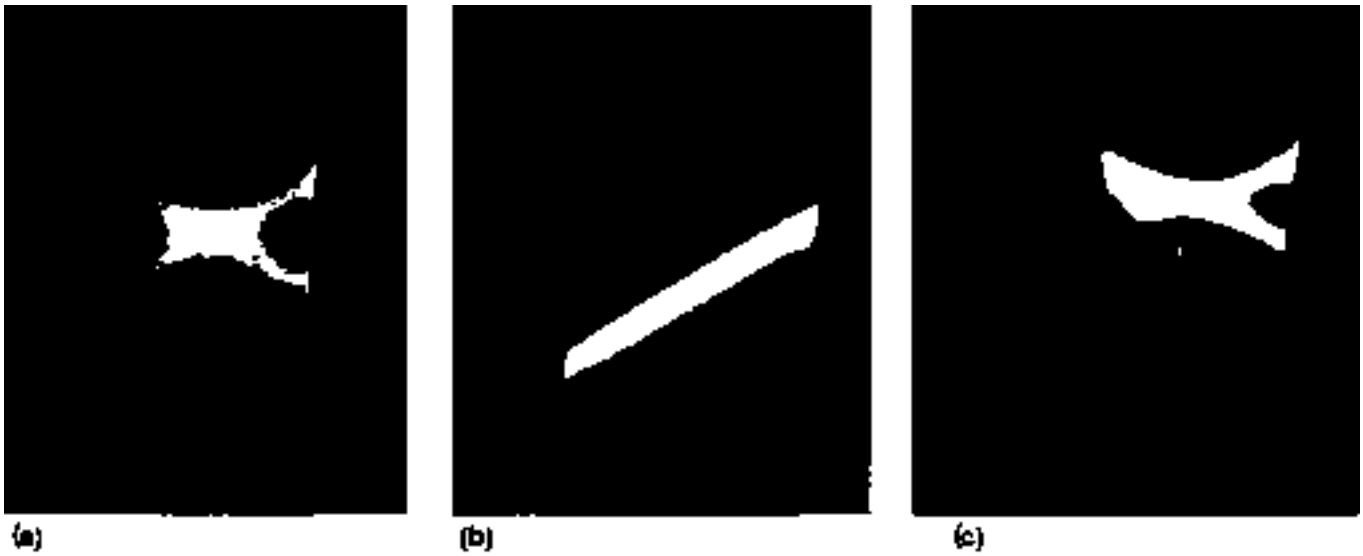


Fig. 46 Sheet samples with an initial *w*/*t* ratio of 6. (a) Single local neck (sample with tensile strength, 1586 MPa, or 230 ksi). (b) Two local necks (sample tensile strength, 827 MPa, or 120 ksi). (c) Fracture of specimen in (a)

As indicated in the Appendix, the strain at the onset of local necking is predicted to occur at twice the magnitude of the strain-hardening exponent for a material that parabolically hardens, while diffuse necking initiates when the axial strain becomes equal to the strain-hardening exponent. The local necking band is enclosed within the region of diffuse necking. In some instances, the necking strain in the thickness direction inside the band is very large, creating a knife-edge fracture. In other instances, thickness strain in the band is terminated earlier, resulting in a V-shaped fracture on the thickness face (see inset in Fig. 47).

Fig. 47 Crack propagation in shear bands in a 7075-T6 plate specimen. Shear banding has occurred on four planes of high shear stress (two containing the width direction and two containing the thickness direction). Crack initiation has occurred in multiple locations, including the edge of the specimen and along the centerline.

Figures 28 and 46(b) show that there is relative shear parallel to the length of the band of the two parts of the specimen. This is in spite of the fact that this creates a bending moment in the specimen and should be resisted by the applied load. Figure 28 (copper, nickel specimens with w/t = 23.5) shows that crack initiation can now occur at the edge of the specimen.

Similar crack initiation at the surface of the specimen occurs for the 7075-T6 plate specimen shown in Fig. 47. However, this specimen shows four different fracture planes of high shear stress and has mechanical properties considerably different (smaller strain-hardening and negative strain-rate hardening exponents) from the copper and nickel specimens shown in Fig. 28.

The length (in the load direction) of the diffuse neck is approximately twice the width of the sample, while the length of the local neck is approximately twice the thickness, so long as the specimen has adequate length so that grip constraint is not a factor, and so long as the material shows typical strain hardening. The 7075-T6 plate material (Fig. 47) has very little strain-hardening capacity and minimal or negative strain-rate hardening capacity. Additionally, the edges of the plate are rough saw-cut edges, so that small-scale stress concentrators are present. The absence of appreciable strain-hardening capacity implies that there is little diffuse neck growth, and none is visible in the specimen. Closer examination of the fracture surface along the intersection of the shear planes shows that there is no region near the centerline of the specimen that is normal to the applied load; that is, the central fibrous region is absent.

The angle of the local neck does vary with texture in single-phase sheet material and can be predicted via Hill's anisotropic yield criterion. Although not treated here, local necking also occurs for some biaxial loading conditions and is the dominant failure mode in sheet metal stretching operations. See, for example, the figure on page 208 of Ref 32 (also in the Appendix of this article), which shows the stress ratios for which local necking can and cannot develop.

Bird (see Appendix) has postulated that the local neck initiates at the intersection of the diffuse necked region and the uniformly straining region (a region of stress concentration). If so, there are four locations for local neck initiation, and the necks grow inward. This creates two possibilities for the location of crack initiation:

- At the four intersections of the shear bands with the surface
- In the interior of the specimen, where two or more propagating shear bands of the local necks intersect (Fig. 46a) (and where the strain is twice as large)

If the w/t ratio is decreased from that discussed previously for the copper and nickel w/t = 23.5 specimens, a small flat region begins to appear at the centerline and increases in the width direction as the w/t ratio decreases, but the two regions of the fracture on each side of the center flat region are inclined to the load at approximately 55° (isotropic material). This is visible in Fig. 45(c) and Fig. 46(a).

Notched Specimens

Sample size can result in a less ductile condition, and Table 1 lists the effect on ductility for maintaining proportional specimen dimensions (constant w/t and width/length ratios) while increasing the sample width (all samples had square holes located in the center of the panel and sides 45° to the longitudinal axis). The width strain becomes proportionally smaller than the thickness strain as the width is increased. This tendency to a plane-strain condition is evident, because the fracture surface changes from macroscale shear (ductile) to macroscale brittle fracture. Plane strain creates more of a triaxial (tensile-hydrostatic) stress that promotes cleavage in ferrous (bcc) alloys.

Table 1 Tensile test results for geometrically similar specimens containing a center through-thickness square hole

Specimen dimensions												
Width		**Length**		**Thickness**		**Test temperature**		**Type of fracture**	**Nominal stress at maximum load**		Reduction in thickness, %	
mm	**in.**	**mm**	**in.**	**mm**	**in.**	**°C**	**°F**	**kPa**	**psi**	**Maximum**	**Minimum**	
76	3	229	9	4.8	$\frac{3}{16}$	-30	0	Slant shear and matte	330	47.9	30.0	**18.9**
						0	32	Slant shear and matte	312	45.2	26.7	**15.6**
						23	74	Slant shear and matte	329	47.7	30.0	**18.5**
						23	74	Slant shear and matte	316	45.8	33.3	**(a)**
152	6	457	18	9.5	$\frac{3}{8}$	0	32	Flat and reflective	314	45.5	22.3	**9.0**
						10	50	Mixed (slant shear and flat)	305	44.2	23.1	**7.8**
						21	70	Slant shear and matte	307	44.5	25.7	**15.1**
						32	90	Slant shear and matte	306	44.4	25.6	**17.7**
305	12	914	36	19	$\frac{3}{4}$	0	32	Flat and reflective	282	40.9	16.2[(b)]	**1.4**
						21	70	Flat and reflective	275	39.9	17.9[(b)]	**1.4**
						39	**102**	**Flat and reflective**	**270**	**39.1**	**19.7[(b)]**	**1.7**

(a) Not completely broken.
(b) At base notch.
Source: Ref 64

Fracture toughness (K_c) and appearance is also influenced by section thickness. For thin material, the fracture is a shear lip, and the fracture toughness may initially increase slightly as the thickness increases (Fig. 48). Tests of notched flat samples show a maximum in toughness as the thickness is increased. As the thickness is increased further, a maximum in fracture toughness is reached, and it declines to a constant value at a large thickness. The optimal thickness is known as critical thickness (t_c) in Fig. 48. Continued width increases above t_c do not improve toughness, and the decline eventually levels off to the limit of plane-strain fracture toughness condition. Thus, shear fracture surface is associated with plane-stress condition and higher toughness fractures. At thin widths, the fracture surface is 100% shear.

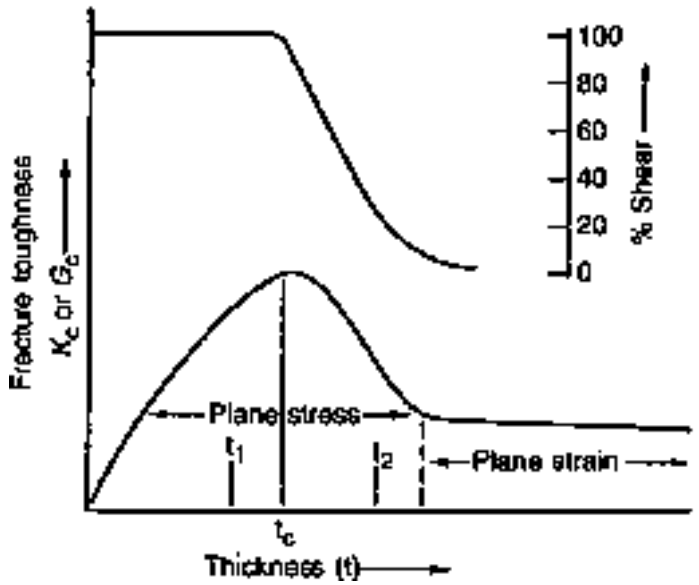


Fig. 48 Effect of section thickness on fracture toughness. Source: Ref 65

References cited in this section

32. W.A. Backofen, *Deformation Processing,* Addison-Wesley, 1972

64. E.R. Parker, *Fracturing of Metals,* American Society for Metals, 1948, p 87

65. R.N. Katz and K.H. Abbott, *Application of Fracture Toughness Parameters to Structural Metals,* H.D. Greenberg, Ed., Gordon and Breach Science Publishers, 1966, p 69

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Compression Failure

Compression failures of ductile materials are unusual during service in common structural parts but can be quite common during metalworking and fabrication. There are differences in crack-initiation location and fracture planes depending on whether the component is loaded in plane stress (cylindrical sections) or plane strain (plate-type specimens). For short (small height-to-width) cylindrical parts, a compressive load that leads to plastic deformation causes the cross-sectional area to increase and stabilize the loading condition. For columns that have large height-to-width ratios, buckling is the mode of mechanical instability leading to bending deformation and failure.

For those unique conditions that lead to extensive plastic deformation in compression, cracking often originates at or near the center because of “barreling” or bulging at the surface at midheight of the part. Platen friction leads to bulging that, in turn, creates a tensile hoop strain (ε_θ), which is maximum at half-height on the exterior surface. At relatively low strains, the mean stress can become positive because of the hoop stress (σ_θ). Cracking initiates at half-height on the hoop plane and grows in both the longitudinal and radial directions.

It has been shown that macroscale ductile fractures can occur along planes of pure shear (zero extensional strain) and that the angle the planes make with the compression axis is dependent on the platen friction coefficient (Ref 66). Fully lubricated compression results in the shear planes at ±32°, while a sticking condition at the platens results in a ductile fracture at 90°, the same as a brittle material. Most test conditions result in angles between 45 and 55°.

In plane-strain slab compression tests, shear cracks may initiate in a shear band starting at the platen surface. It may also initiate along the centerline at the intersection of two shear bands. This latter case, however, may subsequently heal with additional deformation. This appears in metals and in polymers loaded in compression as the width changes—two bands to four or more bands as the width increases. Alternatively, if there is considerable barreling in the unconstrained directions, cracking may also imitiate at the surface, as in the plane-stress loaded specimen.

Brittle fracture in compression can occur in multiple ways. Some brittle materials, such as cast iron, often show a spalling fracture when end loaded. Material at the edge of the specimen and in contact with a platen spalls off at approximately 45°. When the load is concentrated for end-loaded specimens, secondary tensile stresses again cause cracking, but this typically starts in the center of the specimen and results in vertical crack growth and splitting of the material. When there is a through hole or a central cavity in the end-loaded specimen, crack initiation is typically at the surface of the hole or cavity, and the crack(s) again grow vertically. Line-loaded (i.e., side-loaded) cylindrical sections show cracking that originates near the axis of the cylinder and grows vertically.

Reference cited in this section

66. P.F. Thomason, *Ductile Fracture of Metals*, Pergamon Press, 1990, p 159

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Bending

One common bending test sample is a Charpy V-notch bar that is often used to measure toughness and ductility. A second common laboratory specimen is the edge-notched specimen used for fracture-mechanics-based fracture toughness determination in which both axial and bending loads are present. The macro- and microscale fracture appearance of these specimens may be similar in some instances and quite different in other instances. The Charpy-type specimen is discussed here, and the fracture-mechanics-type specimen geometry is discussed in the section “Fracture at or near Stress Raisers.”

Impact energy absorption values of a Charpy test are used to characterize the ductility or the toughness of a material. Macroscopically measurable features are also used to describe the high-strain-rate bending failure associated with pendulum impact testing. The fraction of fracture surface that is shear lip is sometimes reported as percent shear. Also, the amount of lateral expansion at the fracture plane is frequently measured and reported. These measurements trend together; that is, increases in impact energy are accompanied by increases of both the percentage of shear lip and lateral expansion. Figure 49 shows this correlation for several grades of structural steel.

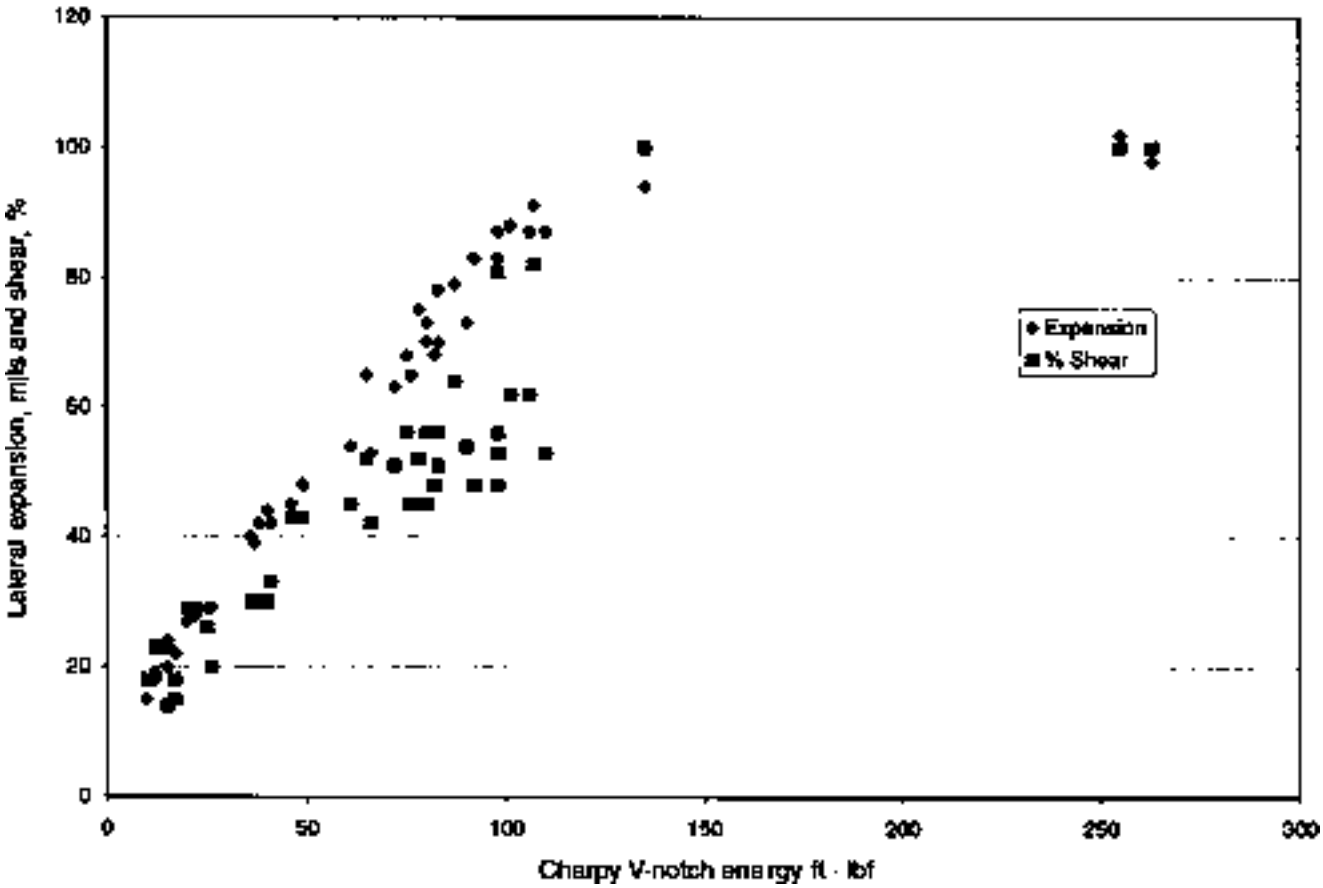


Fig. 49 Correlation between Charpy impact energy, lateral expansion, and percentage shear fracture for construction-grade steels. Courtesy of FTI/Anamet Laboratory

For large structures, Charpy bars are sometimes taken aligned in the transverse direction. This is very useful in detecting the effect of large inclusion stringers that can significantly reduce toughness in this direction, so that the crack propagates in the longitudinal (rolling) direction—that is, a long transverse-longitudinal (T-L) or short transverse-longitudinal (S-L) orientation. (See the article "Stress Analysis and Fracture Mechanics" in this volume for specimen designations). Charpy impact values are low when the crack runs in the longitudinal direction, as it would for a T-L specimen. Highest toughness is measured for a longitudinal-long transverse (L-T) specimen for the same grade of steel. For example, tests of A36 steel have ratios of impact energy for crack propagation in the T-L orientation compared to the L-T orientation ranging from 95 to 20%.

There is a central flat region that appears fibrous or dull for materials with high fracture energy. When the sample is brittle, the amount of shear lip is much less or none at all, and the center appears reflective or similar to rock candy. Cleavage or quasi-cleavage microscopic fracture surfaces are often found during microscopic observations of a shiny, reflective surface. A shiny granular or rock-candy appearance can also be associated with an intergranular fracture. Intergranular failures initiate with little plastic deformation and are macroscopically brittle. The appearance of microscale brittle and ductile fractures in parts failed in bending are not different than those of tensile or other loading conditions. However, as discussed in following sections of this article, the sequence of the zones can be more complex than in tensile specimens.

A bend failure of wrought material typically has some characteristic macroscale features that may be absent in cast materials (see the following). The fracture plane at the fracture origination area is often perpendicular to the axis of maximum stress but in some instances can show a zig-zag pattern, depending on toughness level and notch acuity. As the bending fracture proceeds, the mean path of the fracture typically curves for materials of higher toughness. This curvature of the fracture surface is a defining characteristic of bend failures of parts with slight or no grooves to confine the fracture path. As fracture progresses to the compression side of a bending failure, the stress state changes, and the crack deflects from the perpendicular plane (sometimes called compression curl, Fig. 50). Note also the radial pattern emanating from the black "fingernail" as well as the general roughening of the fracture surface as the crack propagates. In some cases of thinner specimens having a large width, buckling may occur during crack propagation.

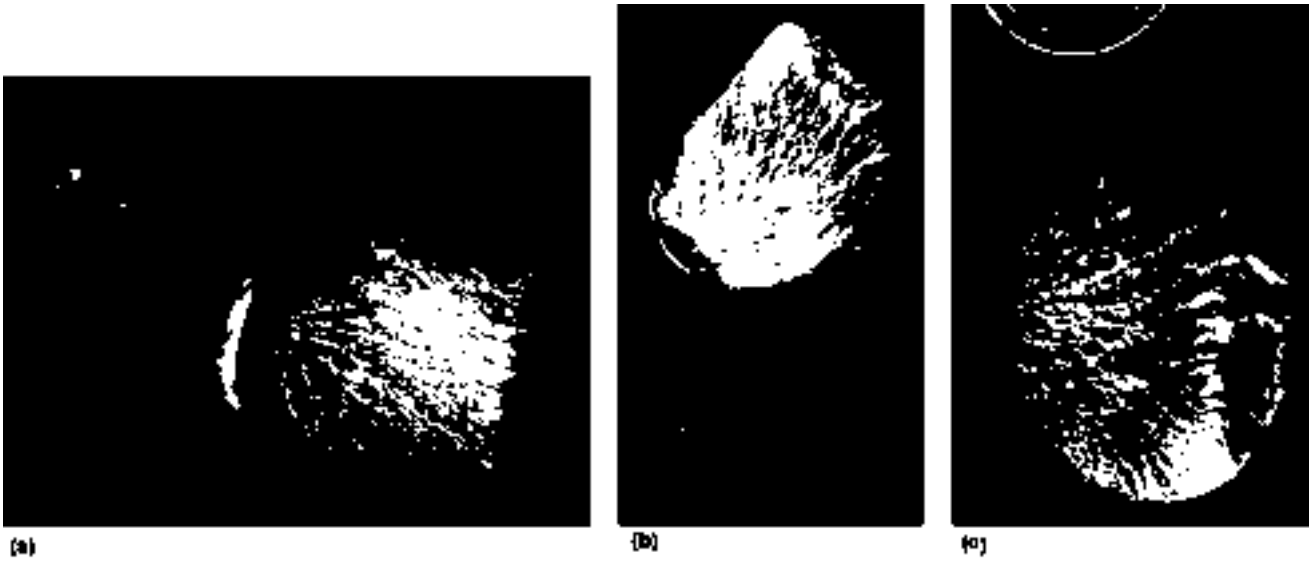


Fig. 50 Compression curl of bend fracture surface. (a, b) Angle views. (c) Top view showing radial marks emanating from the origin

The microscopic observations of a fractured bend specimen are not unique to the notched bending load condition. That is, the shear lips on specimen flanks consist of elongated dimples. The interior fibrous region consists of equiaxed dimples. In some cases, the crack-initiation region at the bottom of the notch can show signs of localized deformation. This has been shown for a V-notched duplex steel (Ref 67). Other investigators (Ref 68) have found elongated dimples at the root of a U-notched sample on a plane oriented at 45° to the root. This suggests crack initiation on a plane of pure shear that is a requirement for the formation of a shear instability (Ref 68). A similar fracture progression is shown in a tempered 4150 steel Charpy specimen (see subsequent text).

Because notched bend specimens approach a plane-strain condition, and impact testing is a high strain-rate test, this type of testing is frequently used to evaluate material toughness or propensity for brittle behavior. Body-centered cubic and hcp materials have a dramatic loss of ductility at lower temperatures that has been extensively studied using impact and slow loading bend testing. The sometimes sharp reduction of toughness as the temperature is lowered is accompanied by macroscopic and microscopic fracture surface feature changes. A detailed, microscopic analysis of the fracture surface was conducted that depicts the relative amounts of ductile and brittle features (Fig. 21). These results depict the gradual loss of ductility as the temperature is lowered. The transition is not sharp for carbon contents typically encountered in structural steels and heat treated machine components, but occurs by a gradual transition of a mixed-mode fracture, with the relative amounts of ductile and brittle microscale fracture surface varying with the temperature.

The effect of strain rate on fracture has been clearly illustrated using notched bend specimens. For pearlitic steel, the temperature at which the material goes through the ductile-to-brittle transition is found to be strain-rate sensitive. Well above the DBTT there is little effect of strain rate on ductility. Well below the DBTT there is not a significant effect of strain rate, but near the DBTT the toughness is strongly affected by strain rate. The magnitude of the shift of the DBTT to higher temperatures with higher strain rates decreases as the strength increases for quenched and tempered steels. An extensive study of the DBTT shift with strain rate in terms of yield strength shows that the shift generally decreases as the yield strength increases, disappearing (for the materials in the study) at a yield strength of 965 MPa (140 ksi) (Ref 70).

The effect of sample size in bending is similar to the effect in tension. A larger sample size develops plane-strain conditions and limits the ductility of a part. Figure 51 depicts the effect on the load versus deflection response of samples of the same material but at different widths. The plot is normalized to account for the linear load increase with width and depth of the sample and the strain increase that increases with the specimen depth at the same deflection. The plot in Fig. 51 is comparable to a stress-strain plot. Thus, the area under the plots is a measure of the energy density to fracture. The reduction in energy to fracture with increasing size is significant.

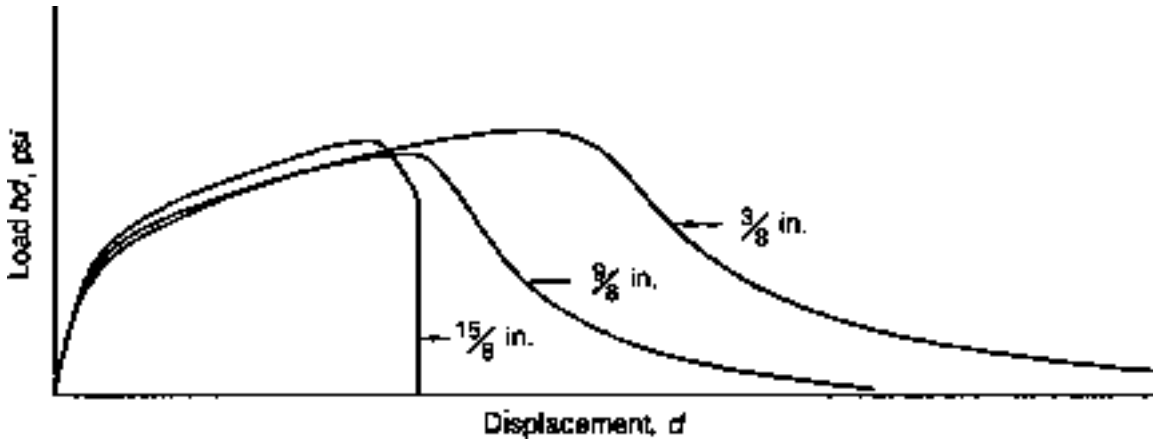


Fig. 51 Effect of sample size on the fracture of notched samples tested in slow bending. Data are normalized to the sample size for comparison. *b*, specimen width; *d*, retained depth at bottom of notch

References cited in this section

67. H.W. Hayden and S. Floreen, *Acta Metall.,* Vol 17, 1969, p 213

68. D.L. McGarry, “The Effect of Strength Level and Strain Rate on the Mode of Fracture Initiation in AISI 4340 Steel,” master's thesis, The Ohio State University, 1972, p 80

70. S.T. Rolfe and J.M. Barsom, *Fracture and Fatigue Control in Structures—Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Bending

One common bending test sample is a Charpy V-notch bar that is often used to measure toughness and ductility. A second common laboratory specimen is the edge-notched specimen used for fracture-mechanics-based fracture toughness determination in which both axial and bending loads are present. The macro- and microscale fracture appearance of these specimens may be similar in some instances and quite different in other instances. The Charpy-type specimen is discussed here, and the fracture-mechanics-type specimen geometry is discussed in the section “Fracture at or near Stress Raisers.”

Impact energy absorption values of a Charpy test are used to characterize the ductility or the toughness of a material. Macroscopically measurable features are also used to describe the high-strain-rate bending failure associated with pendulum impact testing. The fraction of fracture surface that is shear lip is sometimes reported as percent shear. Also, the amount of lateral expansion at the fracture plane is frequently measured and reported. These measurements trend together; that is, increases in impact energy are accompanied by increases of both the percentage of shear lip and lateral expansion. Figure 49 shows this correlation for several grades of structural steel.

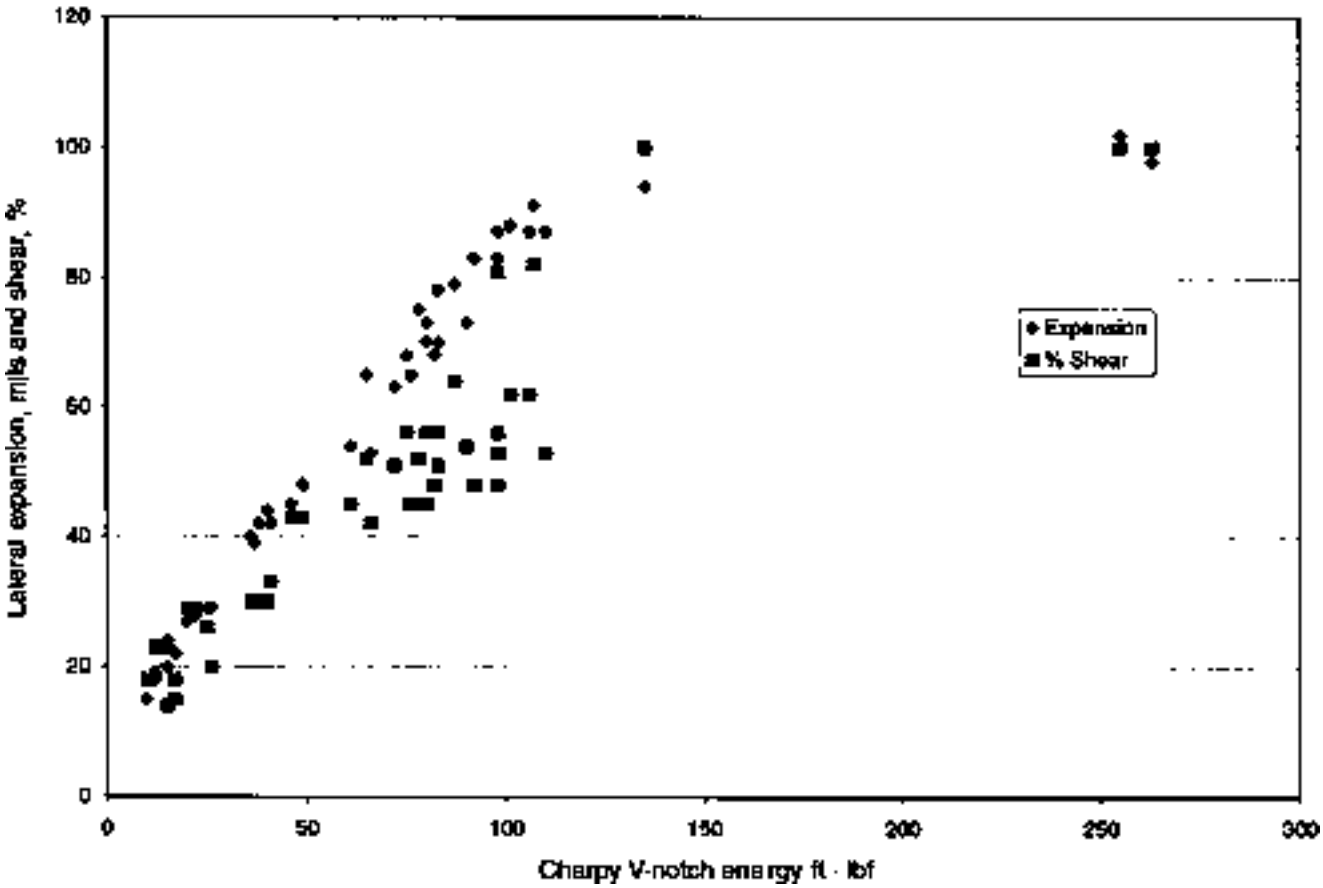


Fig. 49 Correlation between Charpy impact energy, lateral expansion, and percentage shear fracture for construction-grade steels. Courtesy of FTI/Anamet Laboratory

For large structures, Charpy bars are sometimes taken aligned in the transverse direction. This is very useful in detecting the effect of large inclusion stringers that can significantly reduce toughness in this direction, so that the crack propagates in the longitudinal (rolling) direction—that is, a long transverse-longitudinal (T-L) or short transverse-longitudinal (S-L) orientation. (See the article "Stress Analysis and Fracture Mechanics" in this volume for specimen designations). Charpy impact values are low when the crack runs in the longitudinal direction, as it would for a T-L specimen. Highest toughness is measured for a longitudinal-long transverse (L-T) specimen for the same grade of steel. For example, tests of A36 steel have ratios of impact energy for crack propagation in the T-L orientation compared to the L-T orientation ranging from 95 to 20%.

There is a central flat region that appears fibrous or dull for materials with high fracture energy. When the sample is brittle, the amount of shear lip is much less or none at all, and the center appears reflective or similar to rock candy. Cleavage or quasi-cleavage microscopic fracture surfaces are often found during microscopic observations of a shiny, reflective surface. A shiny granular or rock-candy appearance can also be associated with an intergranular fracture. Intergranular failures initiate with little plastic deformation and are macroscopically brittle. The appearance of microscale brittle and ductile fractures in parts failed in bending are not different than those of tensile or other loading conditions. However, as discussed in following sections of this article, the sequence of the zones can be more complex than in tensile specimens.

A bend failure of wrought material typically has some characteristic macroscale features that may be absent in cast materials (see the following). The fracture plane at the fracture origination area is often perpendicular to the axis of maximum stress but in some instances can show a zig-zag pattern, depending on toughness level and notch acuity. As the bending fracture proceeds, the mean path of the fracture typically curves for materials of higher toughness. This curvature of the fracture surface is a defining characteristic of bend failures of parts with slight or no grooves to confine the fracture path. As fracture progresses to the compression side of a bending failure, the stress state changes, and the crack deflects from the perpendicular plane (sometimes called compression curl, Fig. 50). Note also the radial pattern emanating from the black "fingernail" as well as the general roughening of the fracture surface as the crack propagates. In some cases of thinner specimens having a large width, buckling may occur during crack propagation.

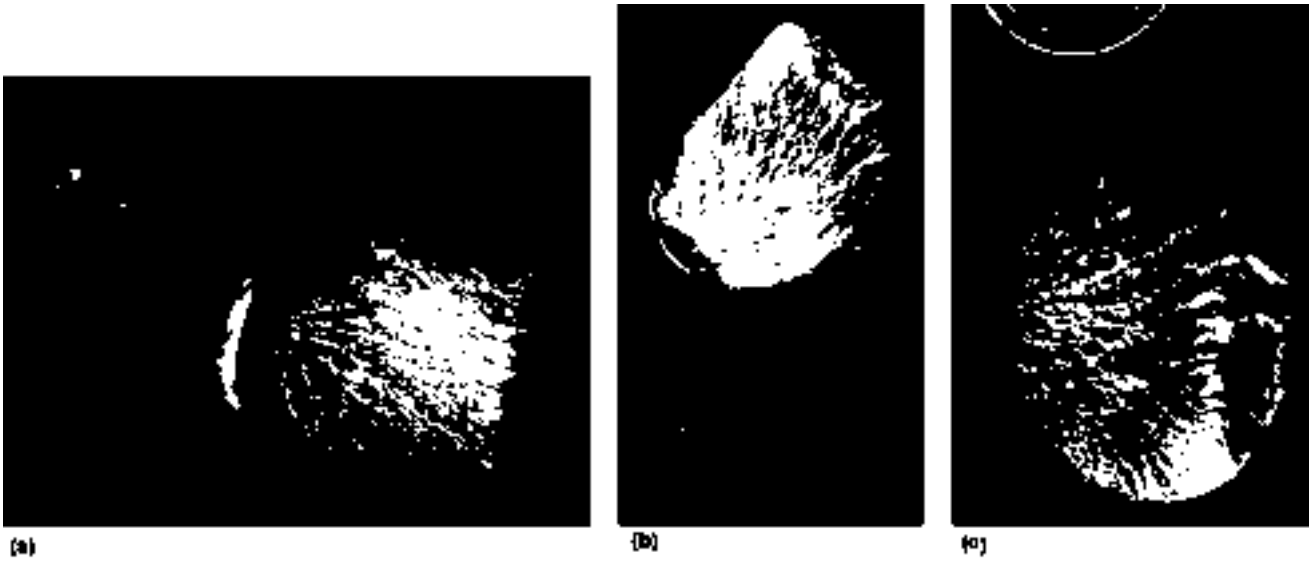


Fig. 50 Compression curl of bend fracture surface. (a, b) Angle views. (c) Top view showing radial marks emanating from the origin

The microscopic observations of a fractured bend specimen are not unique to the notched bending load condition. That is, the shear lips on specimen flanks consist of elongated dimples. The interior fibrous region consists of equiaxed dimples. In some cases, the crack-initiation region at the bottom of the notch can show signs of localized deformation. This has been shown for a V-notched duplex steel (Ref 67). Other investigators (Ref 68) have found elongated dimples at the root of a U-notched sample on a plane oriented at 45° to the root. This suggests crack initiation on a plane of pure shear that is a requirement for the formation of a shear instability (Ref 68). A similar fracture progression is shown in a tempered 4150 steel Charpy specimen (see subsequent text).

Because notched bend specimens approach a plane-strain condition, and impact testing is a high strain-rate test, this type of testing is frequently used to evaluate material toughness or propensity for brittle behavior. Body-centered cubic and hcp materials have a dramatic loss of ductility at lower temperatures that has been extensively studied using impact and slow loading bend testing. The sometimes sharp reduction of toughness as the temperature is lowered is accompanied by macroscopic and microscopic fracture surface feature changes. A detailed, microscopic analysis of the fracture surface was conducted that depicts the relative amounts of ductile and brittle features (Fig. 21). These results depict the gradual loss of ductility as the temperature is lowered. The transition is not sharp for carbon contents typically encountered in structural steels and heat treated machine components, but occurs by a gradual transition of a mixed-mode fracture, with the relative amounts of ductile and brittle microscale fracture surface varying with the temperature.

The effect of strain rate on fracture has been clearly illustrated using notched bend specimens. For pearlitic steel, the temperature at which the material goes through the ductile-to-brittle transition is found to be strain-rate sensitive. Well above the DBTT there is little effect of strain rate on ductility. Well below the DBTT there is not a significant effect of strain rate, but near the DBTT the toughness is strongly affected by strain rate. The magnitude of the shift of the DBTT to higher temperatures with higher strain rates decreases as the strength increases for quenched and tempered steels. An extensive study of the DBTT shift with strain rate in terms of yield strength shows that the shift generally decreases as the yield strength increases, disappearing (for the materials in the study) at a yield strength of 965 MPa (140 ksi) (Ref 70).

The effect of sample size in bending is similar to the effect in tension. A larger sample size develops plane-strain conditions and limits the ductility of a part. Figure 51 depicts the effect on the load versus deflection response of samples of the same material but at different widths. The plot is normalized to account for the linear load increase with width and depth of the sample and the strain increase that increases with the specimen depth at the same deflection. The plot in Fig. 51 is comparable to a stress-strain plot. Thus, the area under the plots is a measure of the energy density to fracture. The reduction in energy to fracture with increasing size is significant.

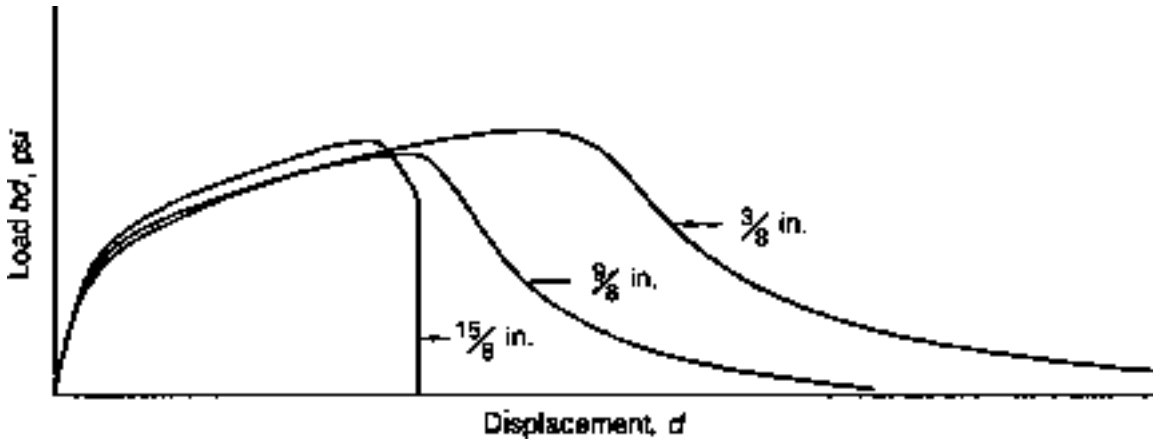


Fig. 51 Effect of sample size on the fracture of notched samples tested in slow bending. Data are normalized to the sample size for comparison. *b*, specimen width; *d*, retained depth at bottom of notch

References cited in this section

67. H.W. Hayden and S. Floreen, *Acta Metall.,* Vol 17, 1969, p 213

68. D.L. McGarry, “The Effect of Strength Level and Strain Rate on the Mode of Fracture Initiation in AISI 4340 Steel,” master's thesis, The Ohio State University, 1972, p 80

70. S.T. Rolfe and J.M. Barsom, *Fracture and Fatigue Control in Structures—Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Fracture Appearances in Cast Materials

Fractographic appearance of cast products, either in the as-cast or heat treated condition, is sometimes more complex and difficult to analyze than for wrought material, perhaps in part only because fractographs for wrought material were the first to appear in the literature and are more familiar. For example, the transition from fatigue crack propagation to overload crack propagation apparently often occurs over a larger distance in the castings than it does in wrought material. This may be due in part to cleavage cracking in second phases in front of the advancing crack tip. For the same reason, fatigue striations may be more difficult to locate, because they are expected to be present only in the dendritic matrix and not in the brittle second phases. Compared to wrought material, significant microstructural features that affect the fracture appearance are predominantly those associated with dendritic solidification (porosity, columnar grains, and the interdendritic location of relatively large, brittle second phases).

Cleavage cracking of either graphite in cast iron or second phases in the aluminum alloys is a major feature of their fractographic appearance. The second phases are typically larger than the inclusions present in the casting alloys and play an important role in fracture progression. At the microscale, graphite flakes may or may not be visible on the fracture surface when examining the fracture surface in the scanning electron microscope, although graphite nodules are often apparent. Similarly, second phases in aluminum alloys (such as silicon particles) may or may not be visible. Both phenomena are due in part to similarities in electron beam absorption for the two phases and therefore a loss in contrast in the fractograph. In some instances, the ability to identify constituents on the fracture surface can be improved by etching the fracture surface. This does alter the surface

from the as-received condition. If the failed component is, or potentially may be, involved in litigation, it is prudent to obtain formal permission from all parties before etching the surface, because the surface is then altered and not in its original condition. In some situations, even such relatively innocuous steps as cleaning or performing hardness tests may be considered destructive. Thus, it is advisable to allow all interested parties the opportunity to examine the evidence of a failure in its as-received condition before changing it in any way.

An alternative technique (for example, when phase contrast in poor) is to examine the fracture surface in edge view in a metallographic specimen to better understand the role of the microstructure in controlling the fracture progression (see the article “Metallographic Techniques in Failure Analysis” in this Volume.)

Macroscale fractographs of castings do not reveal as much detail on the fracture surface as do wrought materials. Chevrons and ridge or radial patterns are generally poorly defined or totally absent. An example is shown in Fig. 58 for gray cast iron loaded in tension. There is no chevron or radial pattern, nor is there a ridge pattern visible on the fracture surface to indicate the direction of crack propagation. Some information regarding crack propagation can be obtained when loading is in bending. Examination of the juncture of the fracture surface with the side of the test specimen (Fig. 59) shows that the fracture is significantly rougher on the tension side of the bend than on the compression side. Both microscale interdendritic shrinkage porosity and macroscale shrinkage voids are common in the casting alloys, especially in aluminum-base alloys due to the large volume change during solidification in those alloys. These imperfections are common sites for crack initiation, as discussed subsequently.

Fig. 58 As-cast gray cast iron loaded in tension. It is common to find little or no macroscale information visible on the fracture surface of many casting alloys to indicate the fracture origin and direction of crack propagation. There are exceptions to this, including the nodular irons and some aluminum alloys. Source: Ref 42

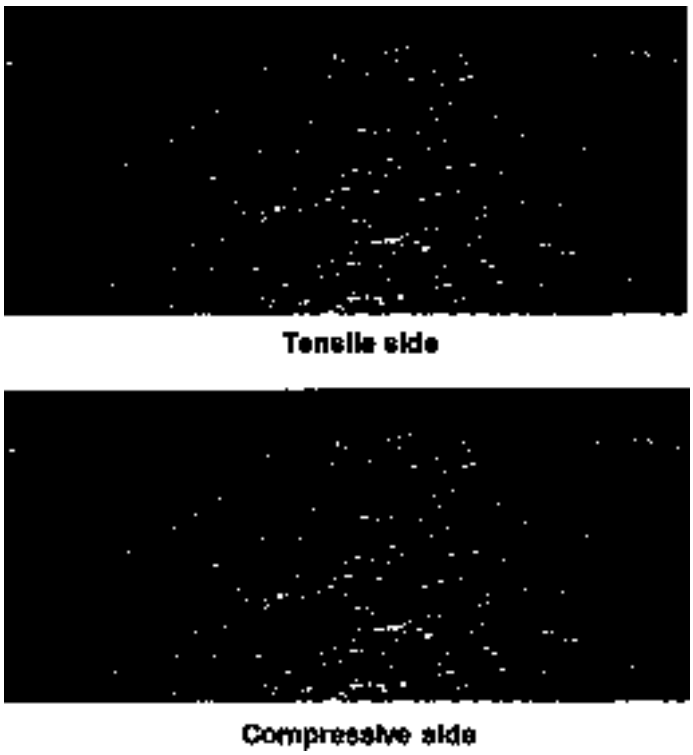


Fig. 59 Tension and compression side of a gray iron loaded in bending. The fracture surface is devoid of any markings to indicate the fracture origin or direction of crack propagation. However, examination of the edge of the component at the fracture surface indicates that the tension side of the bend is rougher than the compression side. Both 5×. Source: Ref 42

Figure 60(a) illustrates crack initiation at dendritic shrinkage. In this case, dendrites are visible within the shrinkage cavity, and the lower-magnification macroscale view in Fig. 60(b) shows some indication of a ridge pattern emanating from the shrinkage porosity.

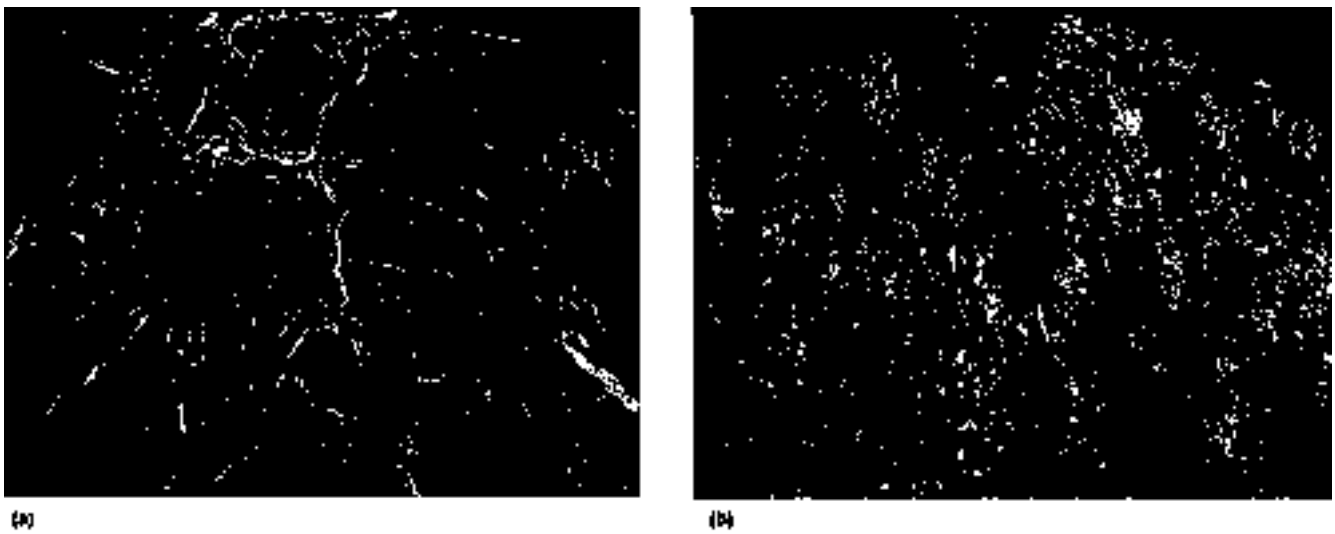


Fig. 60 Dendritic shrinkage porosity in aluminum alloy A356. Shrinkage porosity is a common imperfection in cast components and also a common location for fracture initiation. (a) Fracture surface from a fatigue specimen. 30×. (b) Same specimen as in part (a) but at lower magnification (13×). Surface in (a) is located at the top of surface in (b). There is some indication of the fatigue region (flatter and smoother), and there is some indication of a faint radial pattern near the top of the photography pointing to the region of shrinkage porosity. Source: Ref 42

Figure 61 shows the joining of a sleeve on a section of rebar by welding. Fracture initiated at or near the surface of the material within the weld metal and propagated across the section. The columnar structure associated with

solidification of the weldment is obvious on the fracture surface. (There is also one and possibly two beach marks visible, indicating cyclic loading.)

Fig. 61 Fracture of a heavy section of rebar. A sleeve was welded to the rebar. Fracture initiated in the bright, shiny region, which shows indication of the columnar structure of the weld. There is also one beach mark just below the end of the bright fracture surface, indicating cyclic loading. There is a faint radial pattern and an increase in surface roughness in the overload region. The critical (overload) crack length is very short in this material.

In some instances, the macroscale appearance of the fracture surface of a casting may reveal the dendritic solidification pattern, as in Fig. 62. Higher-magnification views in Fig. 63 are given here to again illustrate that a striated structure is not always due to cyclic loading. Although not reproduced here, additional features at the microscale included regions of MVC and considerable cleavage cracking of second phases.

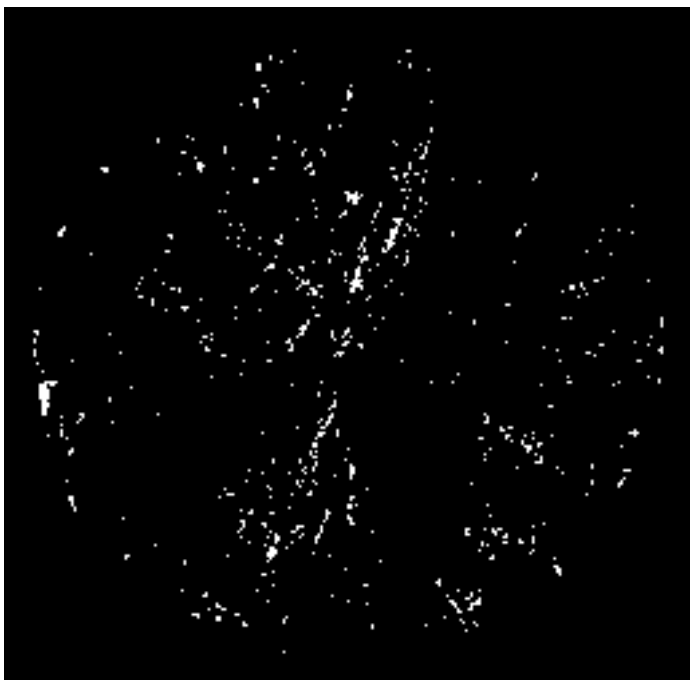


Fig. 62 Fracture surface of as-cast Inconel 713C tension-test section showing evidence of dendritic solidification. No fracture initiation location is evident. Tensile strength, 1048 MPa (152 ksi); yield strength, 827 MPa (120 ksi); total elongation, 6%; hardness, 36 HRC. Source: Ref 74

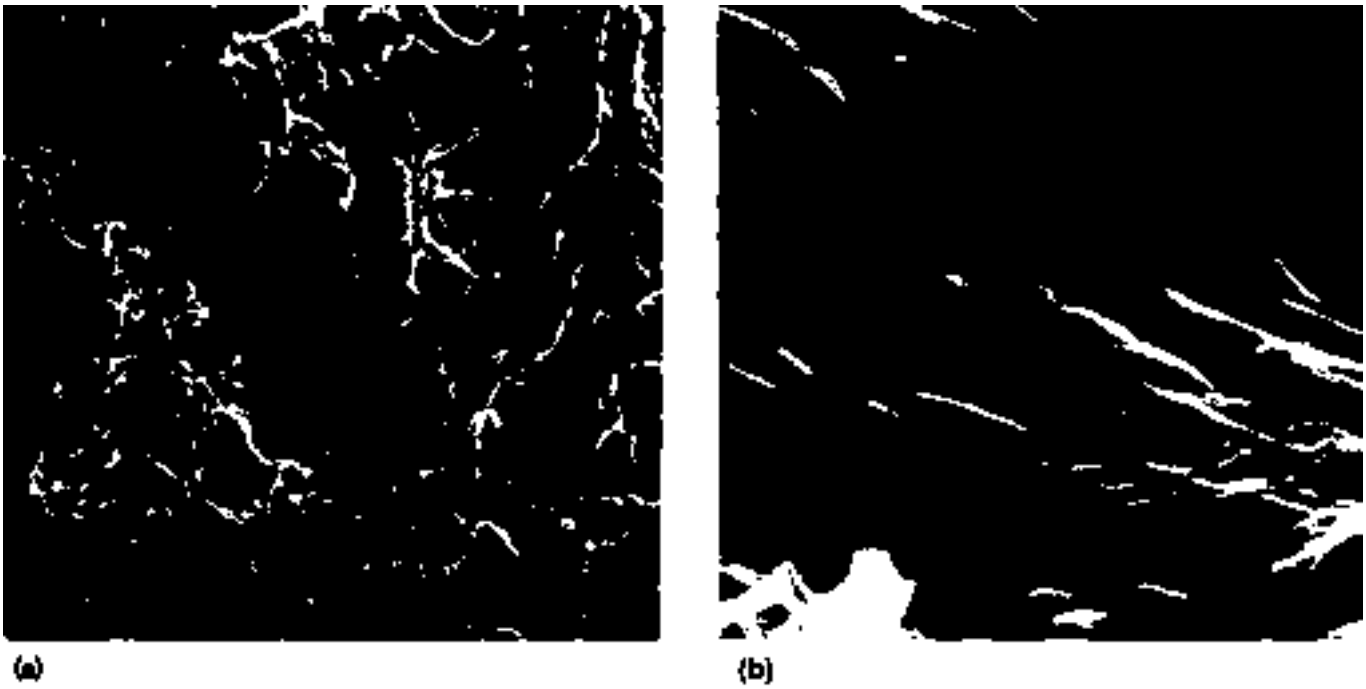


Fig. 63 Higher-magnification views from Fig. 62 showing a striated structure. The cast material was loaded monotonically, not cyclically. The striated structure is presumably due to slip band extrusions. There are many branches in the ridges, which is not typical of fatigue. (a) 770×. (b) 7700×. Source: Ref 74

References cited in this section

42. G. Powell, S.-H. Cheng, C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys,* Battelle Press, 1992

74. S. Bhattacharya, et al., Ed., *IITRI Fracture Handbook,* IIT Research Institute, 1979

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Fracture at or near Stress Raisers

Many fractures occur due to crack propagation from the vicinity of a macroscale stress raiser. Several different appearances are possible at both the macro- and microscales. Factors involved include:

- The degree of constraint created by the stress raiser and loading rate. Geometric constraint is increased by increased section thickness, increased crack length, and decreased radius of the stress concentrator.
- Loading mode (I, II, or III)
- Component dimensions relative to the size of the stress intensity (K) field
- Possible deformation and fracture mechanisms in the material, specifically whether cleavage is or is not possible, and whether deformation twinning is possible at the stresses associated with ductile fracture mechanisms

- The temperature (because this controls, in part, the likelihood of cleavage and deformation twinning)
- Grain size, because it affects the DBTT
- The presence of a chemically aggressive environment (not discussed in this article)

It is convenient to examine this behavior by consideration of a compact tension-type specimen, that is, a component containing a single edge notch and subjected to combined axial and bending loading. The results are equally applicable to other component and imperfection geometries. The macroscale appearance for combined mode I and mode II loading is discussed in the preceding article, "Fracture Appearance and Mechanisms of Deformation and Fracture," in this Volume (Fig. 33 of that article).

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Macroscale Appearance

Consider a load applied to an edge-notched component in mode I loading. As indicated in the Appendix of the preceding article in this Volume, "Fracture Appearances and Mechanisms of Deformation and Fracture," a plastic zone is created behind the notch, unless the material is inherently brittle. The resulting macroscale and microscale behavior is then controlled by the loading conditions, component and imperfection geometry, and the material properties. The macroscale appearance of fracture paths can be either a flat (with an average propagation direction parallel to the macroscale plane of maximum normal stress), slanted, or zig-zag macroscale appearance.

Shear Lips and Slant Fracture. Assume first a moderate-length crack and a moderate crack tip radius and mode I loading. Both the macro- and microscale fracture appearance depend on the degree of constraint at the notch and on material properties. Constraint is at a maximum at midthickness, slightly behind the notch.

As previously noted, the fracture surface orientation is influenced by section thickness, where thin sections result in full shear (plane-stress) fracture (Fig. 48) with a fracture surface on a single shear plane or with a V-shaped surface of two planes (see Fig. 15 in preceding article in this Volume, "Fracture Appearance and Mechanisms of Deformation and Fracture."). At large section thickness, the fracture surface is essentially flat and normal to the applied load, and the fracture toughness approaches the minimum plane-strain value. Intermediate values of thickness produce mixed-mode fracture consisting of a central flat region bounded by shear lips.

There is a transition from full shear fracture (plane stress) for thin sections and/or short crack lengths to complete flat fracture for heavy sections and/or long crack lengths. Radial marks may or may not be present at the location of crack initiation, and chevrons may or may not be present as the crack propagates further across the width. Both of these macroscale features require "fast fracture" and limited macroscale ductility.

High-magnification examination shows that the shear lips fail by MVC. The flat fracture region may fail by all MVC, all cleavage, or by both mechanisms. Two situations are possible: an intimate mixture of regions of cleavage and MVC, and alternating regions of the same features. The former case is often observed in quenched and tempered steels and has been described as quasi-cleavage (Fig. 64). However, the term *quasi* is somewhat inaccurate, because it implies that the fracture is not actually cleavage. The latter alternating-regions case results in crack arrest marks (see Fig. 18 in the preceding article in this Volume, "Fracture Appearances and Mechanisms of Deformation and Fracture").

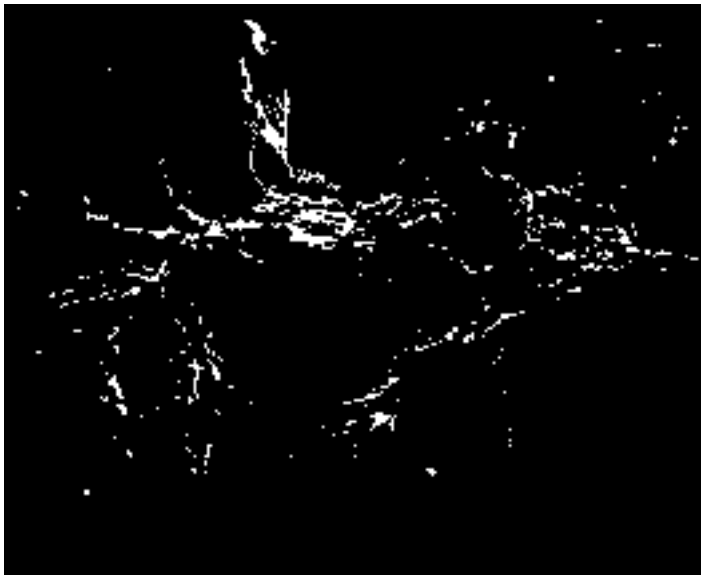


Fig. 64 Quasi-cleavage fracture in an O1 tool steel. Source: Ref 75

The specimen geometry necessary to produce essentially flat fracture (i.e., no shear lips) for through-thickness cracks can be predicted according to:

$$a, B \geq 2.5\left(\frac{K_{IC}}{S}\right)^2$$

where a is crack length, B is section thickness, K_{Ic} is plane-strain fracture toughness, and S is nominal stress.

This criterion is conservative. Plane-strain conditions are met for these conditions but may also be met for shorter cracks and/or thinner sections.

When the fracture surface is essentially flat with minimal shear lips present, it is possible to assume plane-strain loading conditions and quantitatively determine the fracture toughness or the magnitude of the nominal stress required for fracture. If shear lips are present, the fracture toughness is greater than K_{Ic}, and the assumption of plane-strain loading is less accurate. See the discussion in the Appendix to the preceding article in this Volume, "Fracture Appearance and Mechanisms of Deformation and Fracture."

When shear lips are present on notched-plate specimens, together with a flat fracture region, the shear lips can be used to identify the direction of crack propagation. Crack propagation occurs parallel to the flat portion of the fracture, with a parallel slant fracture creating the lip (Fig. 42). Crack propagation from a cracklike imperfection does not normally initiate and grow on a shear lip (but see subsequent information). Crack initiation is typically on a flat spot of the fracture surface. Shear lips can also be present on relatively smooth cylindrical sections loaded in bending. Crack initiation is at the surface of the component in the location of maximum bending stress, and the macroscale appearance is equivalent to that for prismatic sections. There has been some attempt to quantify the stress causing fracture from the width of the shear lips (see Appendix in the preceding article in this Volume, "Fracture Appearance and Mechanisms of Deformation and Fracture").

When a load is applied to the notched specimen, a yield zone is created behind the notch, unless the material is inherently brittle. This yield zone causes the crack tip to be blunted. If there is less initial constraint and the material is inherently ductile and tough, extensive crack-tip blunting can occur during loading and may be sufficient to cause crack propagation on a plane of maximum shear stress rather than on the plane of maximum normal stress (e.g., see Fig. 65 of section from a Charpy V-notch specimen).

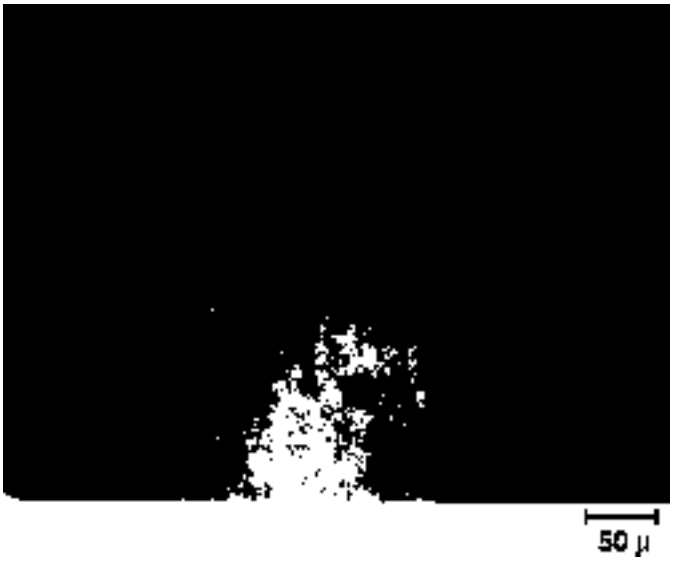


Fig. 65 Charpy V-notch specimen (notch area outlined) with crack path along shear plane due to crack-tip blunting

In order for constraint at the crack tip to control the direction of crack propagation, the *K* field (that region in which stresses predicted by linear elastic fracture mechanics analysis applies) must be well enclosed within (much smaller than) the volume of the specimen. When the specimen is insufficiently wide or insufficiently tall for this to occur, crack plane curvature results, as shown in Fig. 66.

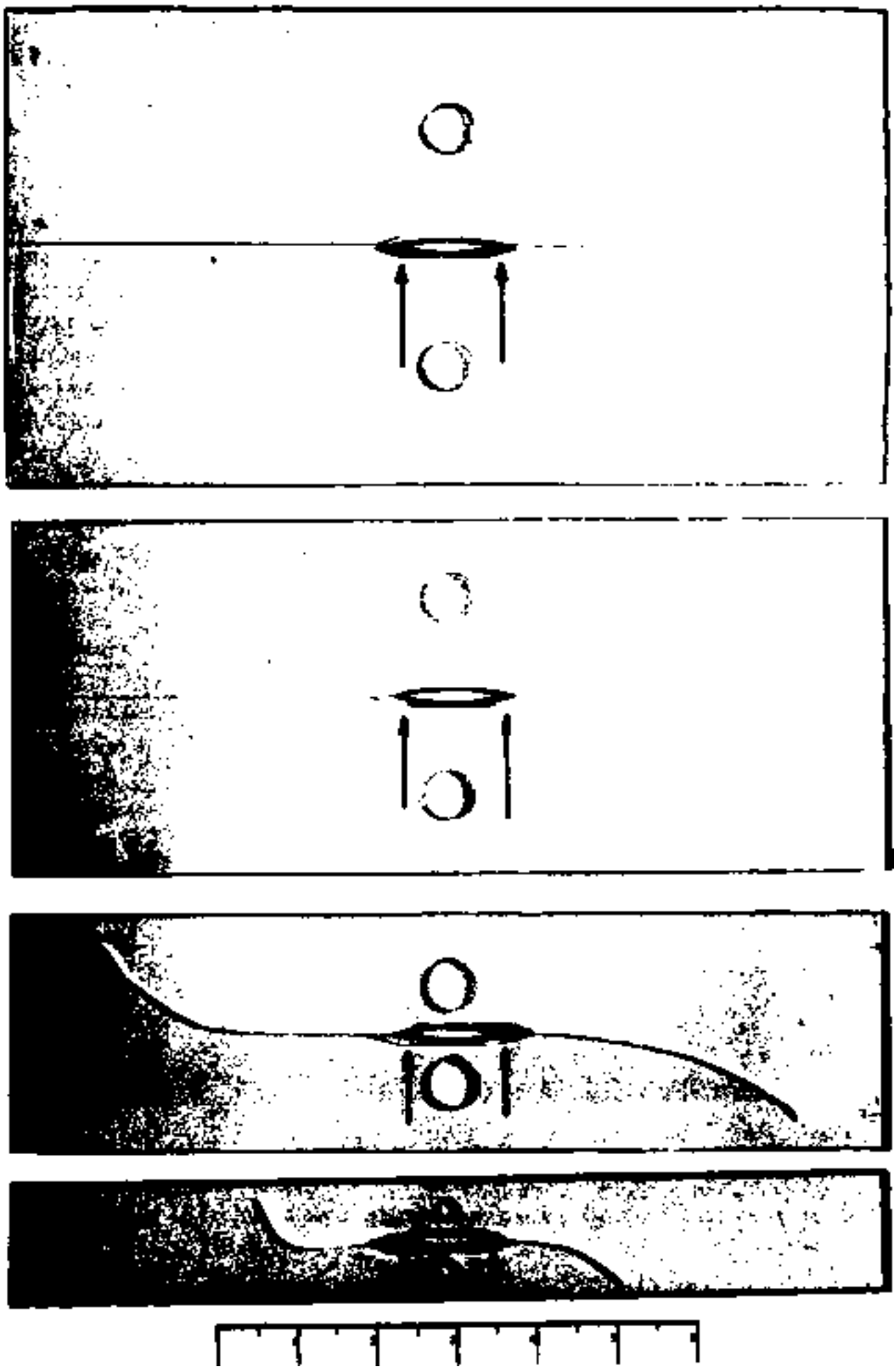


Fig. 66 When there is insufficient height or width of a component containing a crack, the *K* field created by the cracklike defect is altered by the free surfaces of the component. (And, in this case, linear elastic fracture mechanics is not valid.) The figure shows the effect of inadequate height of the specimen and the resulting curvature of the crack plane. Source: Ref 76

References cited in this section

75. W.S. Lin, "The Ductile Brittle Fracture Transformation: A Comparison of Macro and Microscopic Observations on Compact Tension Specimens," master's thesis, University of Tennessee, 1974

76. A.G. Atkins and Y.-W. Mai, *Elastic and Plastic Fracture,* Halsted Press, 1985

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Microscale Details of Initiation and Propagation

Cracks propagating from a preexisting stress raiser may propagate totally in plane stress with net-section yielding, totally in plane strain, or there may be a fracture transition as the crack propagates. In some instances, buckling may also occur. One common macroscale visible case is when the fracture surface changes from flat to slant as the crack propagates due to a loss in constraint and net-section yield in the uncracked ligament. A less obvious case at the macroscale, but quite apparent at the microscale, is that in which constraint develops as the crack propagates (i.e., subcritical crack growth). Subcritical crack growth could be due to fatigue, creep, or during current loading. The latter is considered here.

Cracking is initially by a ductile mechanism (with three types of appearances: a tear zone, MVC, or a tear zone followed by MVC) but may change to cleavage or quasi-cleavage as the crack reaches some (small) critical length, depending on the temperature, loading rate, and grain size. This is common in bcc ferrous materials. It is best understood by examination of the model for crack initiation and propagation from a stress raiser, as discussed subsequently.

It has been indicated that crack-tip blunting occurs and a plastic zone is created behind the notch (Fig. 54 in preceding article in this Volume, "Fracture Appearance and Mechanisms of Deformation and Fracture"). Additionally, material at the surface of the notch tip is loaded in plane stress rather than plane strain (no stress is applied at the tip of the notch, so the internal stress state must approach zero at both the notch tip and also at the transverse surfaces of the specimen for the same reason), so that constraint increases from the surface to the midplane and also increases rapidly from the notch tip in the width direction. There are therefore two locations for crack initiation: at the location of maximum constraint below the notch, or at the surface of the notch.

Unless the material is inherently brittle (for example, due to strain-age embrittlement or stress-corrosion cracking), the likelihood of brittle crack initiation at the free surface is not high, but see the subsequent information. This is true even when a crack of very small radius (as, for example, a fatigue crack) exists at the surface. In steels, three types of crack initiation are found, depending, in part, on the temperature. In the first case, crack-tip blunting followed by ductile crack propagation by tearing initiates at the free surface. In the second, quasi-cleavage fracture and/or ductile MVC fracture initiates at the location of maximum constraint. In the less-common third case, the material has sufficient toughness and sufficient crack blunting occurs that ductile fracture occurs on a shear plane at the crack tip. If blunting occurs, the peak stress is reduced, and the stress falls off more gradually behind the notch.

A simple, accepted model for ductile crack initiation in conjunction with crack-tip blunting is shown in Fig. 67. The crack tip blunts due to both elastic and plastic strain, and the stress field relaxes. When crack propagation occurs in front of the blunted crack, the fracture surface typically takes on a wavelike or ridged pattern (Fig. 68), which sometimes has been mistakenly interpreted as an indication of fatigue. After the crack propagates a

small distance by this mechanism, the increase in crack length at increased crack-tip sharpness can lead to a transition to plane-strain quasi-cleavage fracture in steels, in part depending on the grain size of the material (because of its effect on the DBTT), the section thickness, the test temperature, and the loading rate.

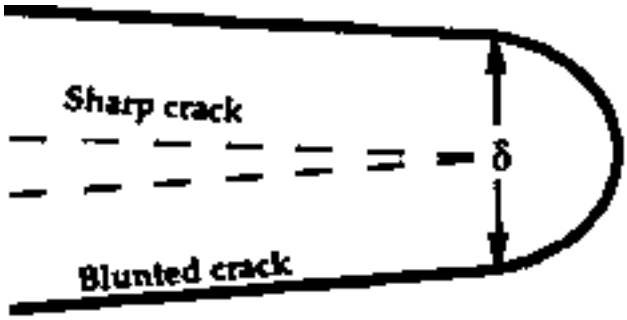


Fig. 67 A model for crack-tip blunting. As yield occurs in front of an initially sharp crack, the crack-tip radius increases, a plastic zone is created in front of the blunted crack, and the high elastic stresses decrease to the work-hardened flow stress. Source: Ref 77

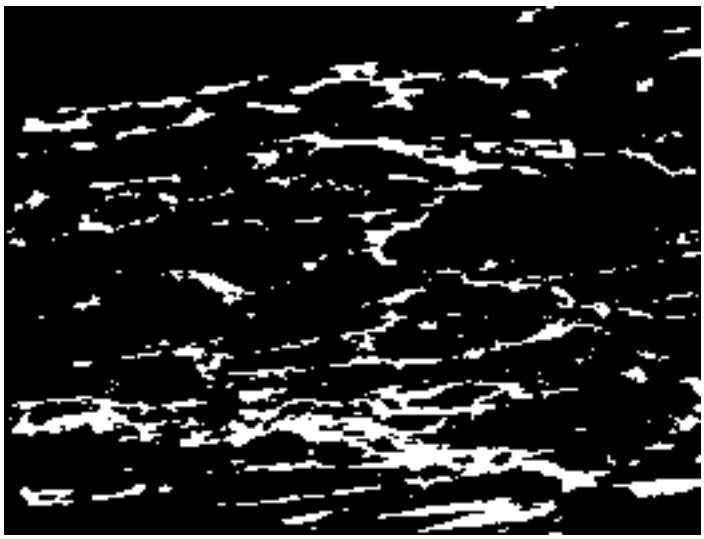


Fig. 68 An example of crack-tip blunting and ductile tearing on plane of maximum normal stress in a quenched and tempered O1 tool steel. Source: Ref 75

Figure 69 shows schematically the differences in cracking behavior with temperature. The presence of the stretch zone can be used to quantitatively estimate the magnitude of the nominal stress and the fracture toughness, as described in the Appendix of the preceding article in this Volume, “Fracture Appearances and Mechanisms of Deformation and Fracture.” The length of the stretch zone is assumed to depend on the crack-tip opening displacement, and by measurement of the length of the stretch zone, the nominal stress and toughness of the material can be calculated (see the Appendix in the preceding article in this Volume, “Fracture Appearances and Mechanisms of Deformation and Fracture”).

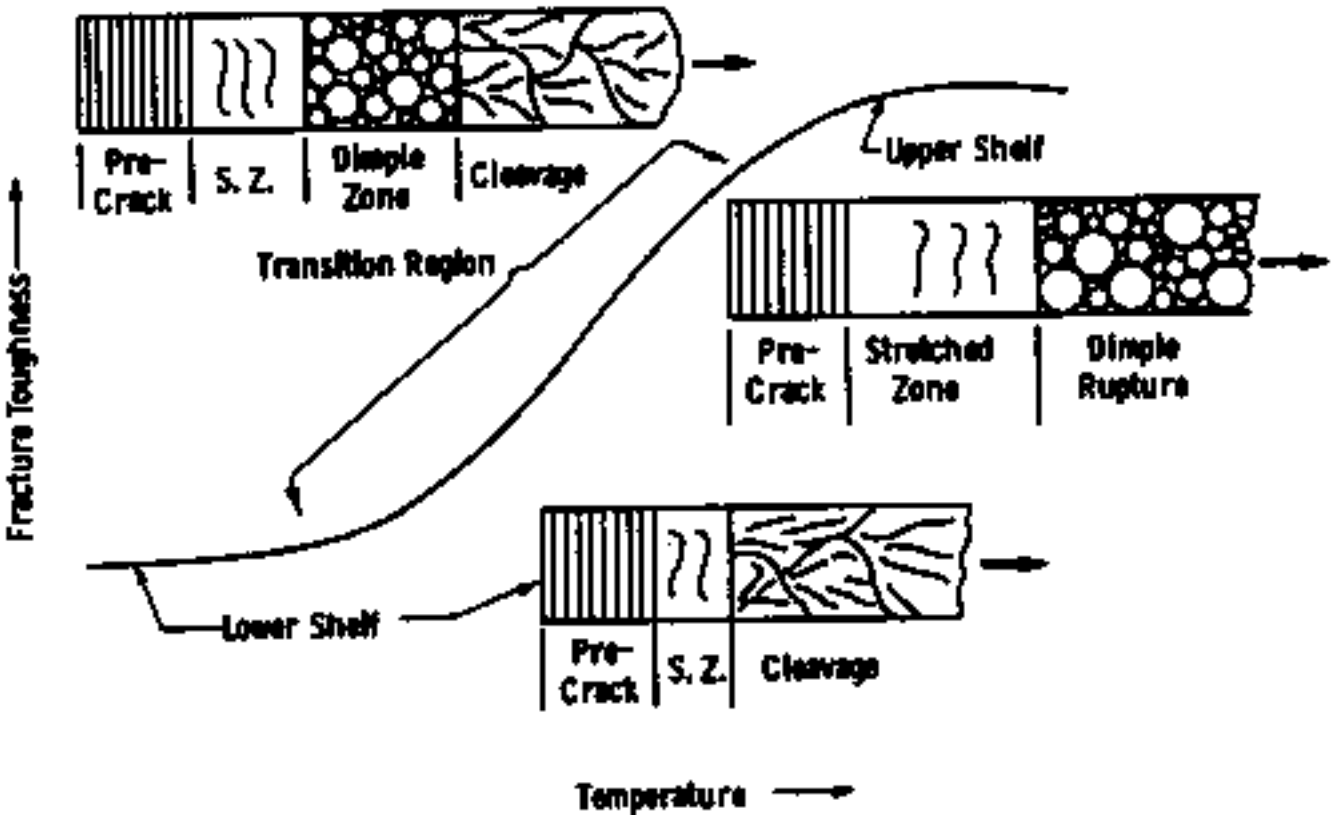


Fig. 69 Schematic figure of the brittle-to-ductile fracture transition. The relative area on the fracture surface of the three microscale fracture mechanisms (stretch zone, dimple zone, and cleavage zone) are indicated. Source: Ref 78

When cleavage or quasi-cleavage cracking is initiated at the location of maximum constraint, it can be initiated by one of the dislocation glide mechanisms discussed previously. It may initiate within the region of high constraint and may initiate at more than one location, in part depending on grain orientation. The specific location(s) can be identified by the convergence of river lines or fan marks. River lines should not be confused with flutes (see subsequent information). Experimentally observed initiation sites are typically at inclusions or second phases. One common observation in low-carbon steels is initial cracking of intergranular carbides, followed by grain-scale cleavage crack initiation at the cracked carbide particles.

The crack must grow through grains of variable orientation and therefore in grains having different resolved microscale normal stress on the cleavage plane and resolved shear stress on the slip systems. As shown in the preceding article in this Volume, “Fracture Appearnace and Mechanisms of Deformation and Fracture,” a cleavage crack can propagate across tilt boundaries but not across twist boundaries. In order for a grain-scale cleavage crack to continue to grow, it must reinitiate as it comes to a grain boundary of arbitrary orientation. This indicates that it is unlikely to have cleavage facets larger than the grain size, and that, in fact, facets are likely to be smaller. Figure 70 (Ref 30) shows multiple facets within larger blocks, which probably were prior austenite grains.

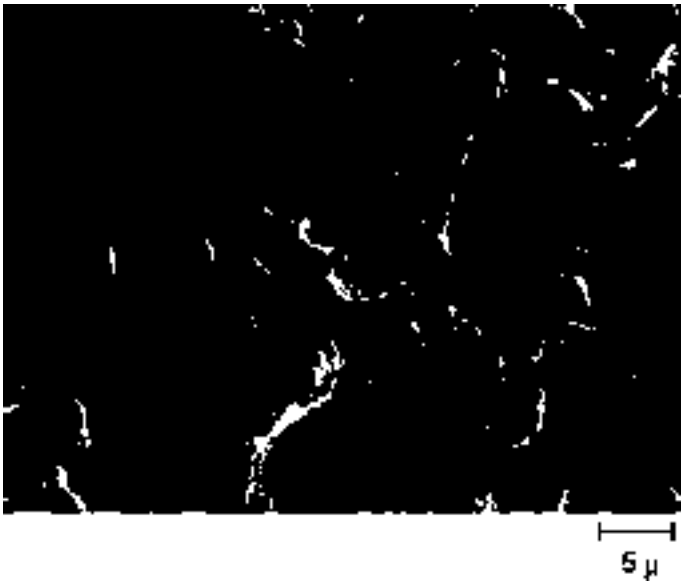


Fig. 70 Brittle fracture in a 4150 steel. Material has been isothermally transformed at 300 °C (570 °F) from a Charpy V-notched specimen fractured at room temperature. Brittle fracture facets within larger regions, which were probably prior austenite grains. Source: Ref 30

Additionally, cleavage facet size in bainitic and lath martensitic structures correlates with the packet size, not the grain size. There can then be multiple directions in which the microcracks grow, especially as the stress fields from these cracks overlap. Secondly, when subsurface crack initiation occurs, the increasing length crack not only propagates in the direction predicted by the external bending load but also back toward the macroscale notch that created the constraint.

When a river pattern results from multisite initiation, the fracture propagation mechanism on the ligament wall of the river line can develop from either slip or cleavage. (Note that if the material is "ideally" brittle, slip cannot occur, and some cleavage mechanism must operate to form the connecting ligament between the two planes). Several mechanisms to create river lines have been proposed (Ref 1), one of which is the intersection of a propagating cleavage crack with a screw dislocation (e.g., see Ref 1). River lines are also present in cleavage fracture surfaces created by torsion (mode III) loading, as shown in Fig. 57.

There are, then, three characteristic features of cleavage cracking:

- A highly reflecting set of flat facets representing local cleavage cracking
- A river pattern resulting from multiple crack reinitiation at grain and subgrain boundaries and/or bainite or martensite packet boundaries
- A similar, related feature known as a fan, in which river lines point to an initiation site

The river marks coalesce as the crack propagates "down stream," and so these lines and the microscale radial pattern of any fans can be used to determine the local direction of crack propagation.

It was indicated previously that cleavage initiation at the plane-stress loaded crack tip is not likely, and that ductile crack propagation occurs normal to the nominal stress or along shear planes in a zig-zag path (Fig. 17). It has been argued that cleavage crack initiation can occur at the crack tip, and that there is a difference in appearance between cleavage that occurs due to the elastic stress in the original material and cleavage in the plastically deformed zone created by the *K*-field yield zone (Ref 79).

Secondary Cracks. Sometimes, secondary cracks parallel to the width plane of the component are referred to as transverse cracks. This cracking is caused by the lateral stress created by constraint at the crack tip. Cracking is enhanced by the presence of elongated inclusions or bead strings oriented parallel to the direction of crack propagation (i.e., a T-L or S-L orientation). For example, Fig. 71 shows transverse splitting at MnS stringers. This is from a fracture of a forged 4140 toroidal link, where fracture initiated by fatigue in the forging flash region of the link. In this region, metallographic examination showed banding and fibering essentially parallel the direction of crack growth.

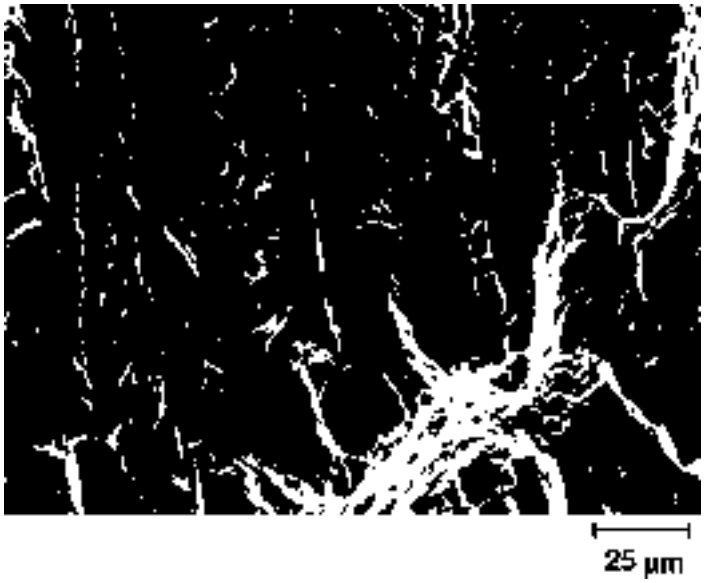


Fig. 71 Transverse splitting at MnS stringers in the overload region of a fractured 4140 forged toroidal link. The crack propagation direction is vertical. Courtesy of Michael West, University of Tennessee

Crack Arrest Lines. If loading and geometry conditions do not provide sufficient constraint at a stress concentrator to cause the elastic stored energy to completely drive a crack across the width of the component via cleavage or quasi-cleavage, crack arrest lines may appear on the fracture surface. These lines, described in the preceding article in this volume, "Fracture Appearance and Mechanisms of Deformation and Fracture," have the following characteristic:

- Initiate at greater distances from the stress concentrator as specimen width increases
- Become pinched off along the centerline as section thickness increases, representing more constraint
- Become thinner as constraint increases

When conditions required for quasi-cleavage are not met in mode I loading, ductile crack initiation can also occur subsurface by MVC, due to the increased constraint at that location. This depends, in part, on the random chance of an inclusion or second-phase particle being present at the location of concentrated stress.
Flutes are another feature sometimes present on a fracture surface (e.g., see Fig. 39 in the article "Modes of Fracture" in *Fractography*, Volume 12 of *ASM Handbook*). Flutes are created by a ductile tunneling crack mechanism that may occur in transgranular stress-corrosion cracking conditions. (See the article "Stress-Corrosion Cracking" in this Volume and Ref 80.) They bear some similarities in appearance to several other fractographic features, including river lines, a lamellar plate structure (e.g., pearlite) on the fracture surface, and fatigue striations.

References cited in this section

1. D. Hull, *Fractography*, Cambridge University Press, 1999

30. D.R. Johnson, "Toughness of Tempered Bainitic Microstructures in a 4150 Steel," master's thesis, University of Tennessee, 1990

75. W.S. Lin, "The Ductile Brittle Fracture Transformation: A Comparison of Macro and Microscopic Observations on Compact Tension Specimens," master's thesis, University of Tennessee, 1974

77. T.A. Anderson, *Fracture Mechanics*, 2nd ed., CRC Press, 1995

78. R.C. Bates, Modeling of Ductile Fracture by Microvoid Coalescence, *Fracture*, J.M. Wells and J.D. Landes, Ed., TMS, 1984

79. A. Vasilev, S. Firstov, and A. Koval, Cleavage Mechanism Nucleation and Fracture Toughness, *Recent Advances in Fracture,* R.K. Mahidhara et al., Ed., TMS, 1997, p 315

80. J.C. Scully, Ed., *The Theory of Stress Corrosion Cracking in Alloys,* NATO, Scientific Affairs Div., Brussels, 1971

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Fracture from Manufacturing Imperfections

Imperfections introduced during primary or secondary metalworking cause a local elevation of the stress and therefore are common locations of crack initiation. Crack initiation from manufacturing imperfections is most likely along the centerline or at the surface of the specimen (with some exceptions noted subsequently). Additionally, residual stresses may be distributed in the component as a result of prior mechanical/thermal processing. Residual stresses are added algebraically to the stress created by applied loads.

Centerline imperfections include incompletely closed macroscale shrinkage porosity due to inadequate working, or due to poor feeding in a heavy section region of a casting or ingot, and centerline cracking in forging and swaging operations. In the absence of surface imperfections, cracking in a smooth cylindrical section under axial load is expected to initiate along the centerline in the necked region. If cracking initiates in another location, it has to do so because the local stress (residual and/or applied) was higher than along the centerline. Similar arguments can be used to predict initiation sites for various kinds of bending loading and torsion loading. For example, for a three-point loaded beam, cracking is expected to initiate at the location of maximum bending moment. If it is not at that location, the implication is that a geometric or material imperfection has moved the location of the local maximum stress. For more details on stress analysis, see the article “Stress Analysis and Fracture Mechanics” in this Volume.

Imperfections along the centerline or at the surface are the most likely locations of crack initiation, but some exceptions include the random distribution of fine quench cracks in steels at locations where martensite formed after quenching, the cracking of divorced cementite in the grain boundaries of low-carbon steel after cold working, and the extensive microcracking common in ceramics. Surface imperfections include machining marks, draw marks, laps, seams, gouges and scratches, end grain in forging operations, and rolled-in debris in rolling operations. Subsurface crack initiation can occur in ferrous surface-treated components (induction hardened, carburized, nitrided) due to an inadequate case depth. Hertzian stresses developed in contact between gear teeth or when cylinders roll over flat surfaces may also move a crack initiation site from the surface to slightly subsurface.

The distinction between a manufacturing imperfection and a manufacturing flaw is critical in determination of root cause. Manufactured components typically contain geometric and material imperfections, but whether the imperfection caused failure and therefore a defect should be determined in many situations by a quantitative analysis. Figure 72 shows a section from a railroad coupler that failed in bending. The dark region near the top edge is a fairly large region of porosity. However, there is also a black “fingernail” on the opposite bottom side of the coupler. Close examination indicates that there has been some fatigue crack growth in this region (not visible in the photograph). Additionally, there is a well-defined chevron pattern that points back to the fingernail. Clearly then, the casting porosity did not initiate the failure and is therefore not a root-cause defect. However, the porosity may be a defect, using an alternate criterion based on permissible levels of porosity as specified by in-house or external specifications.

Fig. 72 Fracture surface of a railroad coupler that failed in bending. The dark, irregular region is a region of shrinkage porosity. The black "thumbnail" is the region of fatigue crack initiation. Chevrons (pointing to the black "fingernail") are visible in the overload region.

Castings have several features that can make them more prone to failure than wrought alloys of similar composition and heat treatment. Large structures undergoing slow solidification have a large grain size and/or dendrite arm spacing, with reduction in strength and ductility. A second important issue is the poor grain shape and the interdendritic segregation of second phases. The large grain size does have a secondary effect in that it is more difficult to homogenize the alloy in a solution heat treatment.

A common cause for near-surface crack initiation in castings is interdendritic (microscale) shrinkage porosity, which is easily detected by examination of an unetched metallographic specimen or on the fracture surface, as in Fig. 60. This is perhaps more noticeable in aluminum alloys because of the large volume change from liquid to solid (4 to 5%). In addition, second-phase particles tend to be larger in castings (as opposed to wrought material), because castings do not go through deformation that breaks down large particles. Second-phase particle size in castings can be reduced by heat treatment, but large castings may not be amenable to heat treatment requiring rapid temperature changes.

Castings are also prone to voids, such as gas porosity and shrink cavities that are healed during hot working processes for wrought materials. The involvement of these imperfections in the fracture process can usually be discerned at high magnification. There are few distinguishing differences between gas porosity and shrink porosity, so it is often hard to recommend corrective action even if voids are found. However, when shrinkage porosity is on the fracture surface, it can show the presence of dendrites. If entrapped gases came out of solid solution after solidification, there should be no evidence of dendrites.

As previously noted in this article, the fracture surfaces are different from that of wrought alloys. Because of limited ductility in many casting alloys, macroscale indicators of fracture characteristics are limited. The multiple mechanisms of crack propagation (Cracking in lamellar structures, cleavage cracking in large second-phase particles) together with the relatively small volume fraction of matrix phase or matrix constituent are likely responsible, there is evidence, however, that dendritic shrinkage porosity is a location of crack initiation.

A noteworthy observation was made about the exterior surface appearance of a bend failure of a brittle casting material (Ref 42). The intersection of the fracture surface with the outside surface on the compression side of the sample had a macroscale appearance of being flat and smooth compared to the tensile-side intersection that was rougher and more irregular (Fig. 59).

Wrought products produced by metalworking operations can introduce surface and subsurface imperfections that can act as initiation sites in both monotonic and cyclic loading. Imperfections such as forging folds or cold shuts can entrap oxides or create a discontinuity that acts as a crack. Cold shuts (or folds) can develop if two external surfaces fold onto each other and are not bonded (welded) in subsequent forging operations either, because the stress is not high enough or an oxide layer on them prevents welding. These imperfections tend to be at or just below the surface. Figure 73 shows a fold in a forged medium-carbon steel that led to cracking.

Fig. 73 Forging fold in a medium-carbon steel. Arrow indicates forging folds that initiated the crack. 20×

Forging operations may, at times, produce undesirable grain flow in the material. Stresses created in service loading that act normal to this flow may result in unexpected failure. Additionally, closed-die forging typically results in the formation of “ears.” Even if these ears are subsequently ground off of the forging, the grain flow at the location of the ears creates a poor microstructure (e.g., elongated inclusions), with exposed end grain. Additional problems can develop if, in hot forging, hot spots develop in the forging or if the nominal temperature control is not adequate. Overheating can result in excessive grain growth and, in more severe cases, partial melting. Resolidification of the melted material results in shrinkage cavities. Still another adverse effect of overheating is the solution of inclusions that then reprecipitate during cooling in grain boundaries, which can result in fracture in these regions.

Cold Forming. Because the cold-forming operation typically involves severe stresses and deformation, internal defects (large inclusions, remnants of secondary pipe, or internal bursts) can also be troublesome with cold-formed materials. Cold-forming operations introduce cracking of hard particles and/or debonding that may lead to cracks and breaks that propagate to the surface of the parts being made, leading to their rejection. Laps, seams, and other surface defects formed during manufacture of the blanks can also lead to major problems during cold forming of the actual parts. Once any surface defect is exposed to the atmosphere, its surfaces become oxidized, providing an initiation point for failure during future operations. When such cracks form early in the steel manufacturing process, the surfaces of such breaks can become completely decarburized, forming what is known as ferrite fingers. Even if the original crack about which they formed is then removed by conditioning, the remaining ferrite is low-strength and often splits during the metal movement necessary during cold forming.

Rolling operations can cause centerline split, edge cracking, and alligatoring. These imperfections can be present in final products if they are not removed. Rolling can cover up surface imperfections, with only close inspection able to uncover them, unless further forming reveals the flaws. Figure 74 shows large surface cracks that opened up during a bending operation. The origin of these cracks was subsurface inclusions, shown in Fig. 75. Cold rolling operations may also cause extensive cracking in second-phase particles and brittle inclusions.

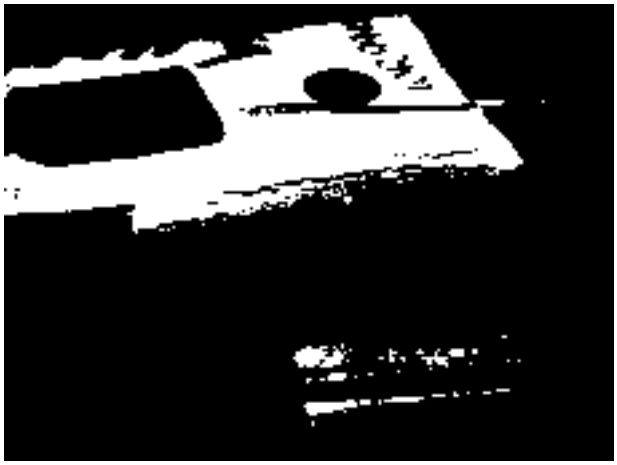


Fig. 74 Deep surface cracks that developed during a bending operation of a low-carbon steel. Courtesy of Worthington Industries Inc.

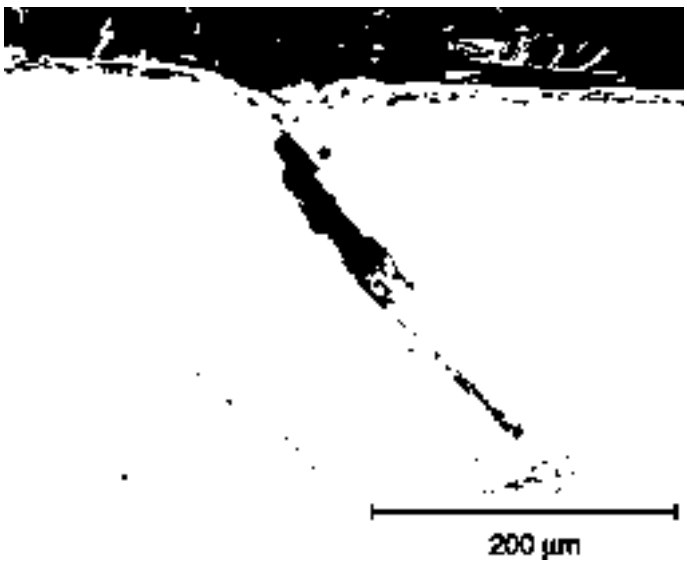


Fig. 75 Micrograph of surface area of material before bending and cracking (shown in Fig. 74). Note the deep rolled-in oxide connecting to the outside surface.

Extrusion operations can likewise produce hidden flaws. The limits of ductility can be attained in this process and result in centerline cracking. These hidden flaws are not as easily detected as surface flaws, and they can result in unexpected failures (see the article “Failures Related to Metalworking” in this Volume).

Anisotropy and Principal Strain Direction (Effect of Directionality). All forming operations can impart a nonuniform distribution of solute distribution, imperfections, and second-phase particles to a wrought material as well as crystallographic texture. Preferred crystal orientation (texture), banding (composition variation that may result in banding of constituents, such as pearlite in steels), and inclusion stringers or “bead strings” of spheroidal inclusions all represent effects of metal processing that result in an anisotropic response to loading. A similar effect occurs in working of austenitic stainless steel containing ferrite stringers. As a result of these effects, ductility, toughness, and fatigue crack propagation rates can be dramatically influenced by the direction of loading. Figure 76 shows how strong this effect can be for a Ti-6A1-4V alloy. The effect on ductility is significant.

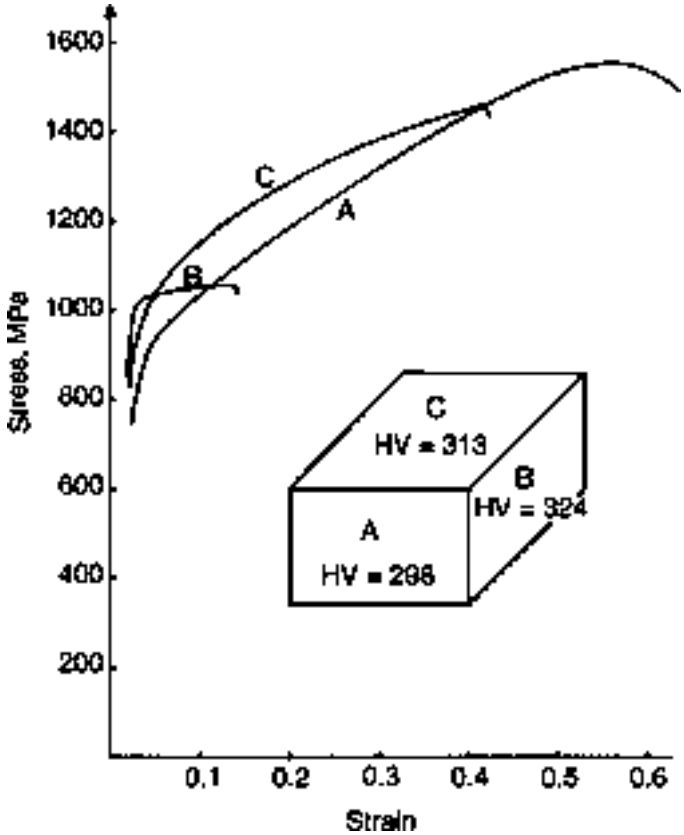


Fig. 76 Stress-strain plots for samples of Ti-6Al-4V, with sample axes aligned in the three principal directions of the material. Vickers hardness values appear in the schematic of a block of material. Source: Ref 81

The macroappearance of the fracture surface can also be affected by stringer direction. Figure 77 shows the banded nature in the fibrous portion of quenched and tempered American Iron and Steel Institute (AISI) 4340 steel tested with the tensile axis perpendicular to the rolling direction. The stringers are easily detected in region A (the flat, fibrous region) but are not detectable in the shear lip, region B.

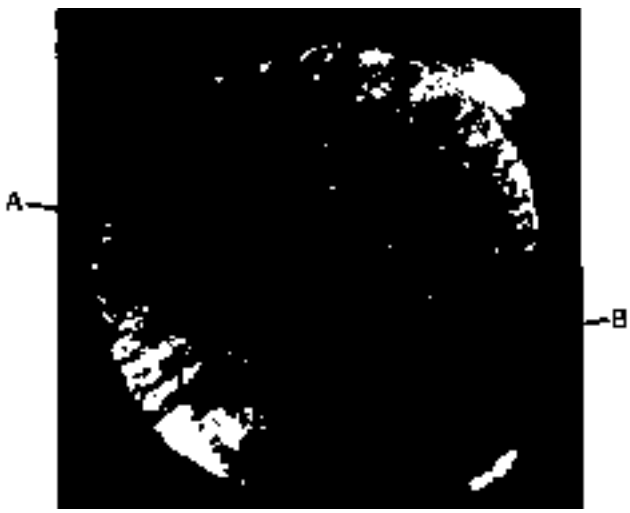


Fig. 77 End view of a fractured tensile bar of 4340 steel with specimen axis perpendicular to the rolling direction. A, aligned fibrous region in the center; B, slant fracture surface. 9×. Source: Ref 74

References cited in this section

42. G. Powell, S.-H. Cheng, C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys*, Battelle Press, 1992

74. S. Bhattacharya, et al., Ed., *IITRI Fracture Handbook,* IIT Research Institute, 1979

81. R.L. Woodward, *Metall. Trans. A,* Vol 10, 1979, p 569

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Case History

To illustrate some of the fractography concepts discussed previously, the following is provided as an example of an in-depth, comprehensive root-cause failure analysis. While an initial inspection would have concluded that the part was simply overloaded, more careful examination and analysis of the various fracture surfaces discovered material embrittlement.

Example 1: Brittle Failure in a Grain Storage Bin (Ref 82). A silo (22 m diameter, 25 m high) was constructed by bolting corrugated steel panels to vertical “U” channel risers and to an “L” channel anchored to a concrete pad. The steel panels were bolted to both the vertical uprights and to the base angle iron. When the silo failed, a crack started at a bolt hole at a connection to the base angle iron and propagated vertically for some distance. Figure 78 shows the fracture at the bottom bolt hole. Necking is visible, and the crack from the bottom surface to the bolt hole is a slant fracture. Examination of this ligament at higher magnification (Fig. 79) shows two more shear planes, and at still higher magnification, Fig. 79(b) shows mode II MVC.

Fig. 78 Fracture of a bolt hole in a grain storage bin. (a) Close view of bottom bolt hole that was in line with horizontal angle on concrete pad. Necking is visible in the ligament below the bolt hole, and the angle of the fracture is at approximately 45° to the circumferential hoop stress. (b) Rotated view of (a). Flat fracture is visible above the bolt hole, with no apparent lateral contraction. The flat fracture region contained poorly defined chevrons (not visible in photograph). Source: Ref 82

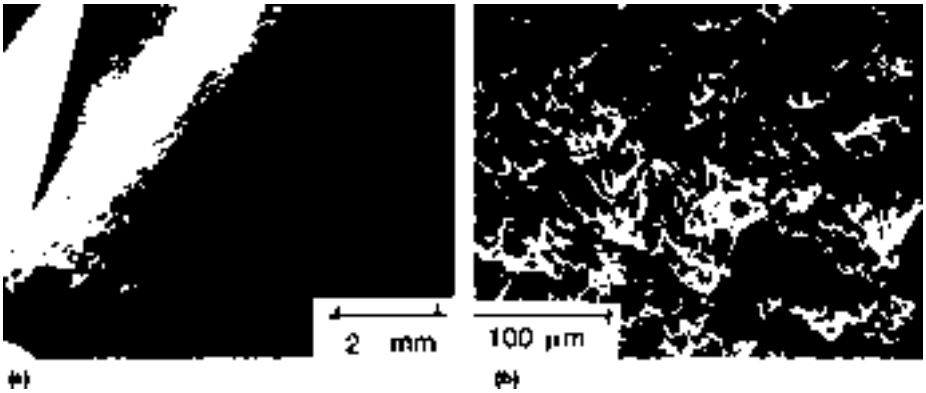


Fig. 79 Ligament below the bolt hole shown in Fig. 78. (a) Low-magnification view. Two more planes of slant fracture are visible. (b) Higher-magnification view of (a) showing elongated shear dimples. Source: Ref 82

All of these features are self-consistent and indicate ductile overload fracture in this ligament. Figure 80 shows two additional bolt holes, the one at a vertical riser being torn out. The bottom bolt hole is distorted into an ellipse, and the major axis of the ellipse is in line with the torn-out bolt hole. This indicates that this sheet of material shifted down and to the right when failure occurred. Examination of the inside of the vertical connection bolt hole shows thread imprinting from the bolts (which were threaded to the head).

Fig. 80 Section of the silo showing the bottom two bolt holes associated with the circumferental angle bolted to the concrete pad, and the first bolt hole associated with the vertical "U" channels. The bottom right-hand bolt hole is distorted, and the upper bolt hole has been torn out. The tear-out direction and the elliptical distortion of the bottom bolt hole are aligned, indicating that this panel moved down and to the right when it failed. Source: Ref 82

Macroscale examination of the material above the bottom bolt hole shows that the fracture is flat, and there is no observable through-thickness contraction. Both of these observations suggest plane-strain fracture. This is surprising, because from a fracture-mechanics perspective, the material does not begin to meet the requirements of section thickness and a crack length to obtain plane-strain loading conditions. The temperature at the time of failure was also taken into account, as discussed subsequently.

Higher-magnification examination of the flat portion of the fracture shows a radial pattern coming from the roots of the thread imprints, and one region that is oxidized (Fig. 81). This implies multiple crack-initiation sites at the local scale stress-concentrating thread imprints and also that partial failure occurred at some earlier time. Also visible at low magnification are chevrons pointing back to the bolt hole (Fig. 82a) Examination of the flat fracture region at high magnification (Fig. 82b) shows that the fracture mechanism is cleavage. Essentially no tear ridges of MVC were found in this part of the fracture, which extended several feet up the side of the silo.

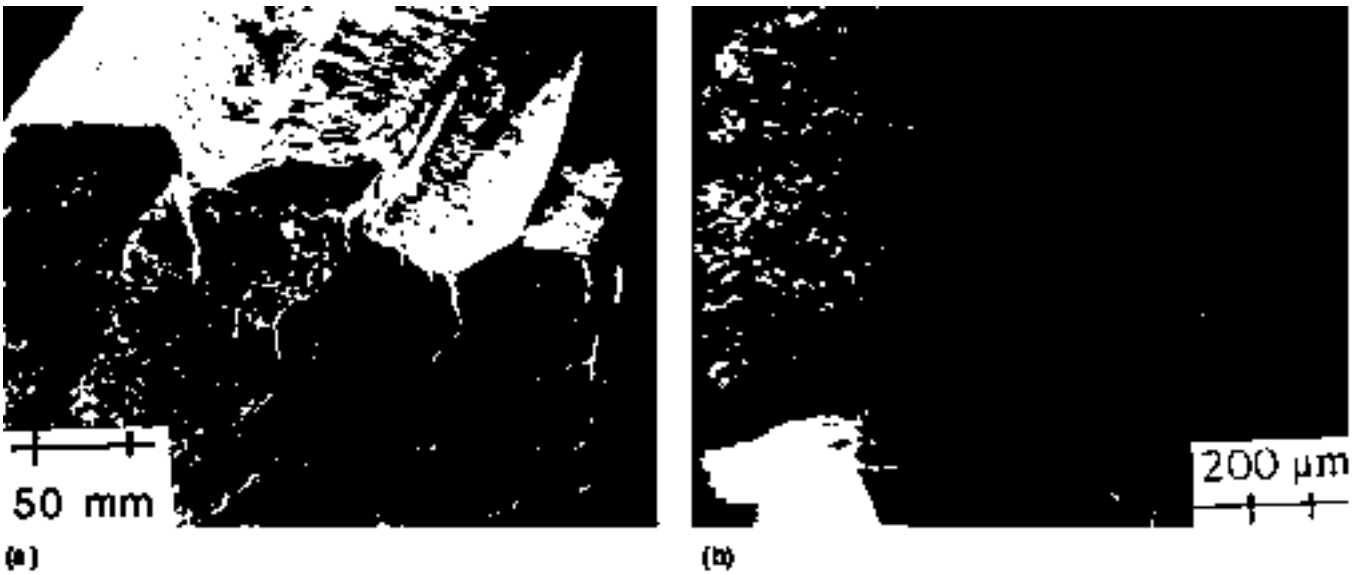


Fig. 81 Two views of upper bolt hole in Fig. 78. (a) Low-magnification view. Thread imprints are visible. Faint radial patterns point to each thread root. Not visible because of the lighting angle is that the material near the left edge is heavily oxidized. (b) Higher-magnification view of left side of (a) showing oxidized region (on right in this picture) and a radial pattern coming from the thread root in the oxidized region. Source: Ref 82

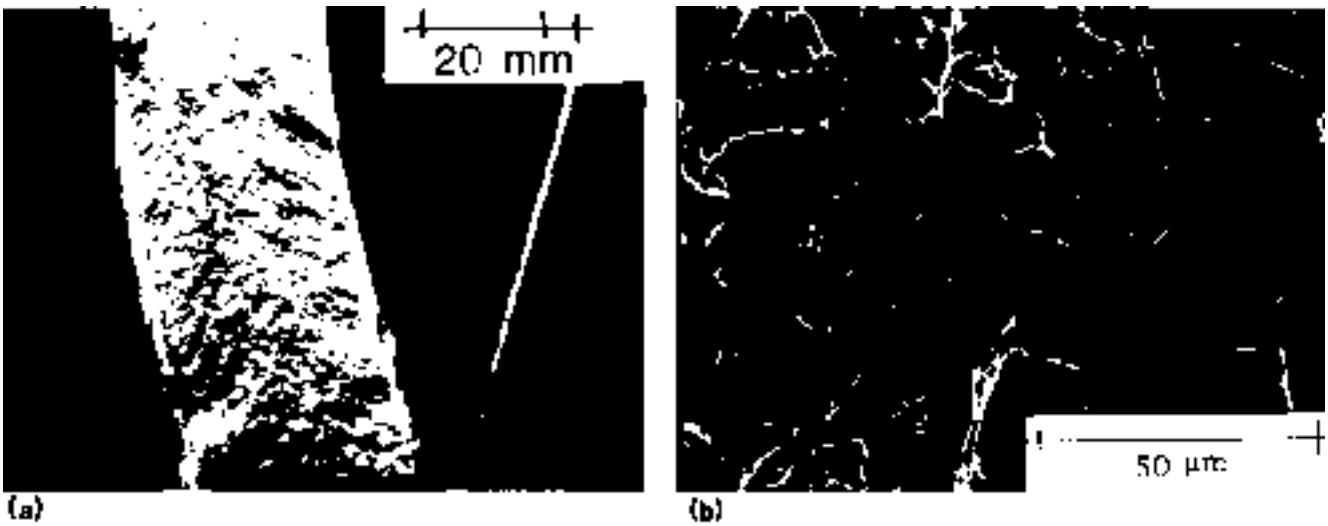


Fig. 82 Two views of material slightly above the bolt hole. (a) Low-magnification view. Chevrons point to the bolt hole (located at the top of the photograph). Location is slightly above the bolt hole. Note the absence of lateral contraction and two very thin shear lips. (b) Higher-magnification view of (a). Essentially total cleavage fracture. Source: Ref 82

Microstructural examination (Fig. 83) indicated an atypical appearance of the pearlite, and examination of the unetched microstructure revealed what was considered to be a typical volume fraction and size of MnS stringers. Subsize Charpy specimens prepared by another party indicated considerable difference in absorbed energy of this material compared to presumably identical material removed from another silo. Service temperatures at the time of the accident were not considered to be low enough to induce the observed cleavage fracture, and again, the section thickness was inadequate to generate sufficient geometric constraint to induce cleavage fracture. Chemical analysis performed by another party indicated an abnormally high nitrogen content of the steel. It was concluded that the cause for failure (and the appearance of the cleavage fracture) was consistent with strain-age embrittlement.

Fig. 83 Microstructure of the nominal 0.2% C steel. Pearlite is very coarse, with thick cementite plates in some locations, and there is a large amount of free cementite in the grain boundaries. Source: Ref 82

Reference cited in this section

82. W.T. Becker, Brittle Failure in a Large Grain Storage Bin, *Handbook in Case Histories of Failure Analysis,* Vol 2, ASM International, 1993, p 470

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

Appendix: Modeling of Ductile Plastic Flow

Introduction

Modeling of the fracture process has been approached at both the macro- and microscales. When models reproduce experimental behavior, they become tools to use for quantitative evaluation of fracture in failure analysis. Multiple approaches have been used; some are based on continuum mechanics to predict behavior of the bulk assembly (without assumptions of the microscopic scale details and operative mechanisms of crack initiation and propagation), while other methods consider either a submicroscopic or microscopic description of structure to provide criteria to predict crack propagation and fracture based on microscopic processes. For example, dislocation theory has been used to develop models for the formation of crack nuclei.

There also has been considerable interest in developing models for the strain at fracture of ductile materials, based on the common fracture mechanism of MVC. The following provides an overview of ductile flow modeling for metals and thermoplastic polymers. Implications on fracture appearance are also considered.

Macroscopic Plastic Flow and Instability

Whether ductile fracture relates to either the mechanism of fracture or the fracture-surface appearance, the fracture process is associated with strain accumulation. Fracture is preceded by macroscale instability in load-extension behavior for tensile loading and can also occur for compression and torsion loading. The following discussion is primarily limited to tension loading. Discussion of instability for other loading conditions can be found in Ref 51.

Several variables are involved in predicting the onset of necking, including strain hardening, strain-rate hardening, thermal softening, preferred orientation, and so on. It has been reasoned that the strain-hardening response of the material controls the onset of necking (Ref 83). For materials that show parabolic hardening, necking begins when the true strain (ε) equals the strain-hardening exponent (n), that is, $\varepsilon = n$ for a material, as predicted by the Holloman relationship between true stress and true strain ($\sigma = k\varepsilon^n$). Later, the analysis was refined to include the material property of strain-rate sensitivity, m, where $m = (d \ln\sigma)/(d \ln\dot{\varepsilon})$ (Ref 84). This analysis predicts the onset of necking when $\varepsilon = n/(1 - m)$. This correction is generally not large for room-temperature loading and in the absence of deformation heating (and therefore also for slow strain-rate loading), because most values of m for common structural materials at room temperature are small, less than 0.1.

The second approach to evaluate ductility is to consider the diameter reduction in the neck at fracture, %RA, reduction in area = $100((A_i - A_f)/A_i)$. This parameter is more closely related to the true tensile strain, $\varepsilon = \ln(A_i/A_f)$, at the fracture location than is total elongation. The total true strain is a measure of the strain of a tensile test plus any prior strain. Components can be processed with substantial cold work in compression, significantly raising the state of strain. This is why highly cold-worked wire starts to neck immediately after yielding; it has an inherent strain that exceeds the value of n. The extent of diameter reduction depends not only on prior work hardening but on a variety of physical and mechanical properties as well as testing conditions, such as strain rate, test temperature, state of stress, test machine stiffness (the size, shape, and volume fraction of inclusions and second-phase particles), and the specimen size. For example, a material that would normally be considered ductile could contain enough inclusions to make it macroscale brittle.

The simplest instability criteria is for uniaxial loading of an isotropic material and to consider only strain hardening as the independent variable. The change in load, P, with strain is given by:

$$\frac{dP}{de} = \frac{d(s\,A)}{de} \qquad \text{(Eq 1)}$$

The flow stress increases with axial strain due to strain hardening, and the cross section decreases with axial strain. At the onset of instability, the slope of the load-strain curve goes to zero, so $dP/d\varepsilon$ is set equal to zero in Eq 1. This leads to the well-known prediction of necking when the instantaneous hardening rate becomes equal to the flow stress for a constant volume material:

$$\frac{ds}{de} = s \qquad \text{(Eq 2)}$$

which can be extended to multiaxial loading conditions by substituting the effective stress and the effective strain for the uniaxial values:

$$\frac{ds_{eff}}{de_{eff}} = s_{eff} \qquad \text{(Eq 3)}$$

The effective stress for metallic materials is taken as (for isotropic materials):

$$\frac{ds_{eff}}{de_{eff}} = s_{eff} = \frac{1}{\sqrt{2}}\left[(s_2 - s_3)^2 + (s_3 - s_1)^2 + (s_1 - s_2)^2\right]^{1/2} \qquad \text{(Eq 4)}$$

The effective strain is taken as:

$$de_{eff} = \left[\left(\frac{2}{3}\right)\left(de_1^2 + de_2^2 + de_3^2\right)\right]^{1/2} \qquad \text{(Eq 5)}$$

More complex expressions are available for anisotropic metallic materials (cold rolled and/or annealed sheet materials) in Ref 51 and for polymeric materials in Ref 85.

Polymeric materials are often not isotropic. They behave in a more complex manner than metallic materials, because they show sufficient compressibility that the yield expression must include both a shear term and a

normal stress term. Additionally, forming operations typically produce preferred alignment of the carbon backbone chair and therefore strong macroscale orientation of the weak van der Waals bond. For uniaxial loading, the expression is:

$$\tau = \tau_c - \mu\sigma_N \qquad \text{(Eq 6)}$$

(constants τ_c and μ), which becomes:

$$\tau = \tau_d - \alpha\sigma_h \qquad \text{(Eq 7)}$$

for multiaxial loading with constants. There are, again, two constants: τ_d, the value of the deviatoric shear stress at yield, and α. τ is the applied deviatoric shear stress, and σ_h is the hydrostatic stress from the applied loads. Equation 6 might be recognized as the form of the Mohr-Coulomb model initially used to predict brittle fracture, and it is still sometimes used for that purpose with cast iron. The effect of nonconstant volume and Eq 7 is to shift the center of the yield ellipse off of the origin in principal stress space coordinates.

Assuming that strain hardening obeys a parabolic hardening law (Holloman behavior), which is not always the case, instability is predicted to occur when the true strain (logarithmic strain) becomes equal to the magnitude of the strain-hardening exponent (again, assuming isotropy):

$$\varepsilon_{neck} = n \qquad \text{(Eq 8)}$$

This has the interesting implication that in order to see the neck (and therefore conclude that the fracture was ductile at the macroscale), there must be some strain hardening in the specimen. It is convenient to think of change in behavior with varying degrees of hardening, just as it is convenient to consider the change in fracture mechanisms with volume fraction second-phase particles. In the limit, if there is no hardening, the neck does not grow along the specimen and in the limit would have no length and therefore not be visible. Real metallic materials do strain harden, but the strain-hardening exponent can be small, as in the case of age-hardening alloys and other high-strength materials. It is also small in previously and heavily cold-worked material.

Rate and Temperature Effects. The effect of other variables, for example, strain-rate hardening and local deformation heating, can be accounted for by assuming a relationship of the form:

$$\sigma = \sigma(\varepsilon, \dot{\varepsilon}, T) \qquad \text{(Eq 9)}$$

and again writing the instability criterion given by Eq 1, which now becomes:

$$-\left(\frac{1}{A}\right)\left(\frac{dA}{de}\right) = 1 = \left(\frac{1}{s}\right)\left[\frac{\partial s}{\partial e} + \frac{\partial s}{\partial \dot{e}}\left(\frac{\partial \dot{e}}{\partial e}\right) + \frac{\partial s}{\partial T}\left(\frac{\partial T}{\partial e}\right)\right] \qquad \text{(Eq 10)}$$

The strain and strain-rate hardening terms are positive. The term $d\dot{e}/d\varepsilon$ can be negative or zero, depending on whether loading is at constant extension rate or constant true strain rate. The term $dT/d\varepsilon$ represents the heat given up during plastic working, and the term $\partial\sigma/\partial T$ represents the change in flow stress with increased temperature (often negative but can be positive). For constant strain rate, the strain-rate term disappears ($d\dot{e}$ = 0), and for conditions of constant extension rate, the strain-rate term becomes equal to m. Equation 10 also indicates the importance of the thermal term in that it is the only (commonly) negative term in the expression and therefore plays a dominant role in creating instability.

The major effect of m is after necking. In this case, the area inside the neck is smaller than the area outside the neck, so that the strain rate inside the neck is greater. Therefore, the flow stress inside the neck is also increased, so long as m is positive. The result is that the neck grows along the specimen length.

It is often incorrectly assumed that metallic materials are loaded under isothermal conditions because of their high thermal conductivity. Such assumptions should not be made for high strain-rate loading conditions or for modeling of rapid crack propagation. Approximately 95% of the work of deformation is transformed to heat generation, and only approximately 5% of the work is used to create structural changes. Polymeric materials have lower thermal conductivities, and thermal softening plays a more important role in neck formation at even slower strain rates. There is a second difference for the polymeric materials in that there is a significant structural change (orientation hardening) associated with neck formation. The instability at yield in a polymeric material is similar to that observed in a metallic material when strain localization occurs in the form of Lüder band formation.

For isothermal conditions, instability at constant extension rate (not constant strain rate) axial loading is predicted when:

$$e = \frac{n}{1-m} \qquad \text{(Eq 11)}$$

See Ref 51 for an extended discussion including other loading conditions. Thus, strain-rate effects are typically small for metals at room temperature, because m is small (say, less than 0.1 and usually smaller than that) but increases with increased temperature.

In metals, thermal softening may arise from several conditions, including dynamic recovery and/or recrystallization at elevated temperature, strengthening mechanism reversal, strong preferred orientation, and deformation heating. For these cases, thermal softening is included by a temperature-compensated hardening term of the form:

$$\frac{\partial s}{\partial T}\frac{\partial T}{\partial e} \qquad \text{(Eq 12)}$$

Changes in Strain Path. Plastic-strain accumulation is path dependent, so that constitutive equations must be written in incremental form. Plastic-strain increments ($\delta\varepsilon_{ij}$) are proportional (proportionality constant $d\lambda$) to the gradient of the yield surface (∇ [$f(\sigma_{ij})$]) according to:

$$\delta\varepsilon_{ij} = d\lambda[\nabla f(\sigma_{ij})] \qquad \text{(Eq 13)}$$

so that particular loading paths create two-dimensional (plane-strain) deformation (i.e., when the normal to the yield surface is parallel to one of the principal stress axes) (Fig. 84). This occurs, for example, during biaxial loading of an isotropic material at stress ratios of $\frac{1}{2}$ and 2.

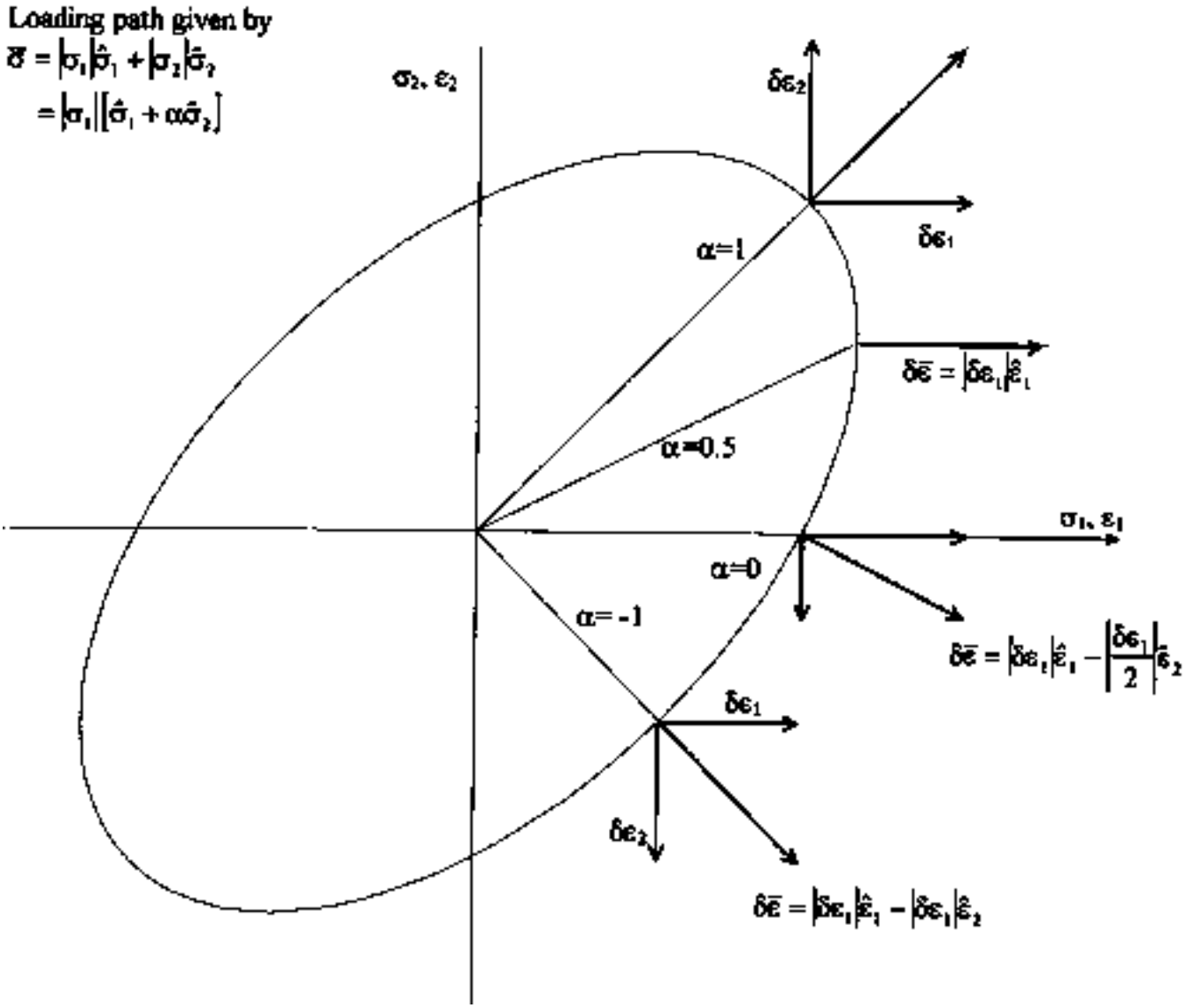


Fig. 84 The yield surface for an isotropic material in principal stress space with strain axes superimposed on the stress axes. Strain increments obtained on yielding for various loading paths are indicated. The resultant strain increment in the σ_1-σ_2 plane is normal to the yield surface. Plane-strain deformation corresponds to the strain increment lying parallel to one of the principal stress axes.

Equation 13 derives from continuum mechanics, and the model does not always predict the behavior of real materials. A change in loading path may result in a decrease in fracture strain, say, by first loading in torsion, followed by axial loading, as has been demonstrated (Ref 32). Although a cup-and-cone fracture is expected for ductile fracture in tensile loading, prestraining in torsion can result in a different fracture appearance, as shown in Fig. 85, which is more characteristic of brittle torsion fracture.

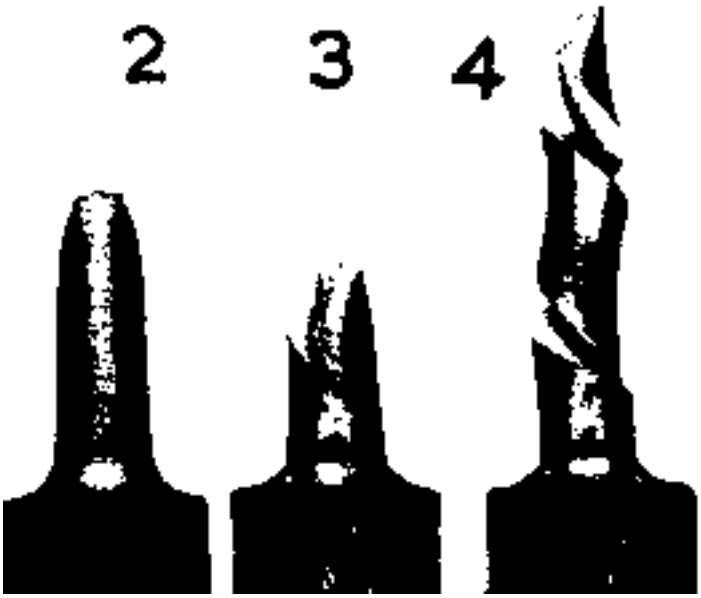


Fig. 85 Fracture obtained by first plastic straining in torsion and then straining in tension. The fracture appearance becomes more characteristic of the first strain increment as the first strain increment increases in magnitude. Source: Ref 4

Multiaxial Loading. Considering only strain hardening, instability is predicted when:

$$\frac{ds_{eff}}{de_{eff}} = \frac{s_{eff}}{Z} \qquad \text{(Eq 14)}$$

where $Z = 1$ for uniaxial plane stress loading and $Z = 2$ for local necking (see subsequent information) in plane-strain deformation of sheet or plate. For multiaxial loading and for anisotropic materials, Z depends on the analytical expression taken for the yield surface See Ref 51 and 86.

Postnecking Strain-Rate Effects. It was noted previously that in order to see the presence of a neck in ductile fracture, it was necessary for strain hardening to spread the neck along the length of the tensile specimen. However, even if there is no strain hardening, strain-rate hardening may cause the neck to grow along the specimen length. The strain rate:

$$\frac{1}{L}\frac{dL}{dt} = \frac{-1}{A}\frac{dA}{dt} \qquad \text{(Eq 15)}$$

is greater inside the neck than it is outside the neck, because the neck area is smaller. So long as strain hardening is positive, an increase in strain rate forces the neck to grow along the length of the specimen in the same way that strain hardening forces the neck to grow along the length. However, if strain-rate hardening is negative, and especially if there is little strain-hardening capacity, then strain localization is extremely intense. Strain-rate hardening is negative in some aluminum alloys at room temperature (Fig. 40) and presumably all materials at elevated temperature. Then, a diffuse neck is not visible in either plate or cylindrical plate specimens. For positive strain and strain-rate hardening material in cylindrical sections, the most common macroscale fracture appearance is that of the well-known cup-and-cone or double-cup fracture appearance. However, if strain localization occurs because of negative strain-rate hardening in association with little strain hardening, the cylindrical tensile specimen can show no cup-and-cone and a single slant fracture appearance (Fig. 86).

Fig. 86 Fracture on essentially one plane of high shear stress in a 7075-T6 cylindrical tensile specimen. There is a small flat region in the center of the specimen (not visible in photograph) that does not extend to the surface of the specimen. No fracture surface markings exist to indicate the fracture origin.

Local Necking in Prismatic Sections. As previously noted, when the component geometry is prismatic rather than cylindrical, both uniaxial and multiaxial loading permits the development of plane-strain deformation, resulting in the formation of a local neck. For an isotropic material subjected to uniaxial loading, the strain at the onset of local necking is predicted to occur when (neglecting strain-rate hardening):

$$\varepsilon_L = 2n$$

which is at a strain twice as large as for diffuse necking. That is, the sequence is from homogeneous flow to diffuse necking to local necking to fracture (see Ref 51 and 86). For uniaxial loading, use of Mohr's circle for strain shows that the angle of the local neck to the load axis is approximately 55° for isotropic material. For multiaxial loading, local necking can occur for only certain in-plane stress ratios that can be predicted by examination of normals to the yield surface, because not only must the thickness strain be negative, but one of the in-plane strains must also be negative (Fig. 87). Measurement of the orientation of the local neck band to the sheet geometry can be used to determine the state of stress on the material and is therefore potentially helpful in failure analysis but is not often used in a quantitative way.

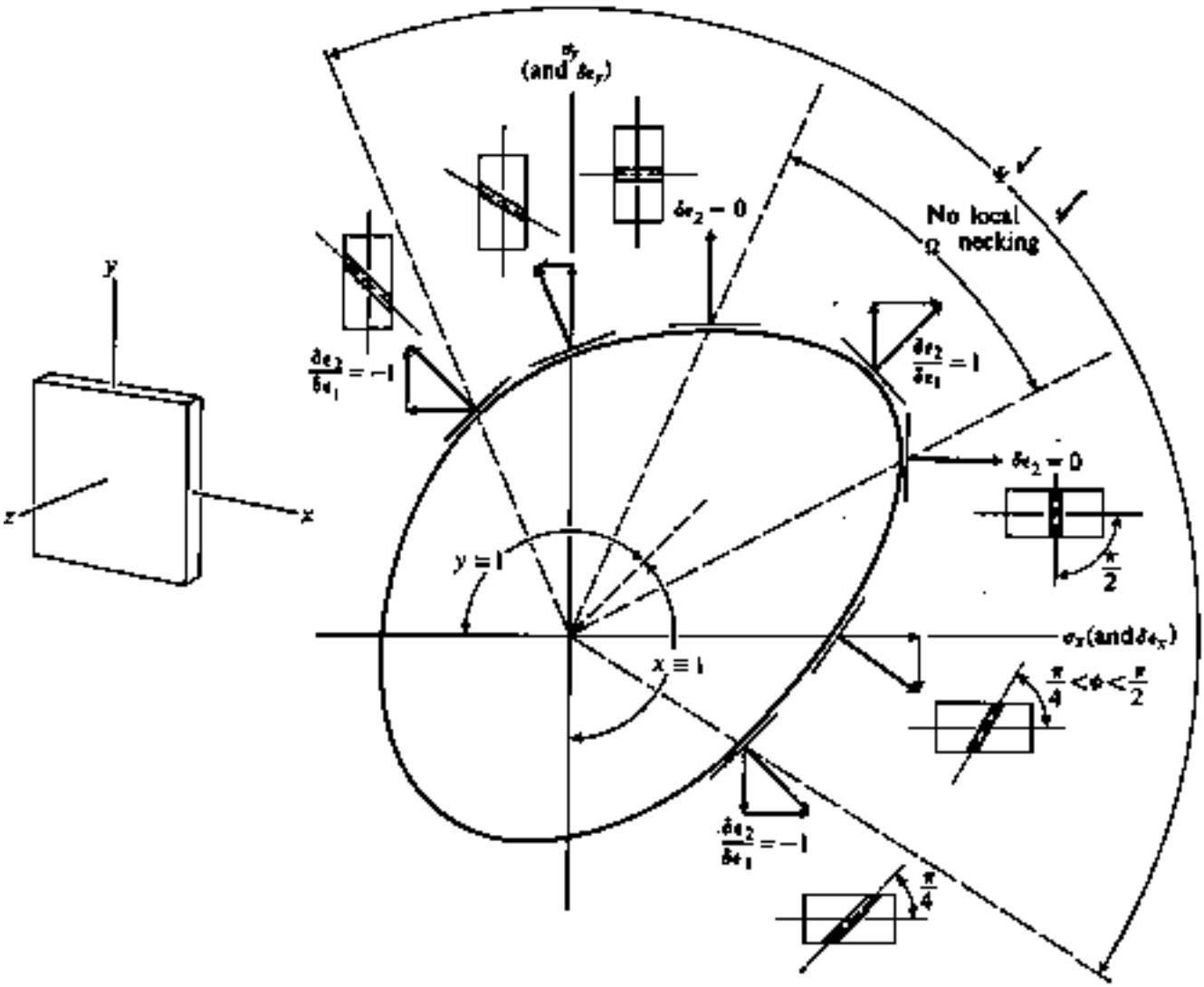


Fig. 87 Biaxial stress ratios that permit no necking, diffuse necking, and local necking. Source: Ref 32

Strain Localization

On some scale, plastic deformation is always a localized event. On an atomic scale, not all crystallographic slip planes are active. After extensive deformation, dislocation cells develop that represent regions of much higher dislocation density than other areas of the same crystal. Any matrix discontinuity, such as inclusions, second-phase particles, or even grain boundaries, lead to localized plastic deformation. On a specimen-scale, localized deformation is forced by geometric constraints, for example, grooves, notches, holes, and cracks.

In areas without geometric constraints, where uniform deformation should occur, macroscale localized deformation known as shear bands can develop. This localized deformation initiates within individual grains and, at first, is crystallographically controlled by the local stress field and active dislocations. These grain-length bands having micron-scale width then initially vary in orientation from grain to grain. However, at larger strains, the bands coalesce across specimen-scale widths and lose their grain-dependent orientation, changing from grain-scale bands to specimen-scale bands (Ref 87, 88). When the localized deformation is sample-scale, traversing many grains, it is called a shear band, as described further in this section.

Shear band formation is generally thought to limit the total overall deformation that would otherwise be expected. It represents an instability at which the material is undergoing dramatically increased strain rates. Investigators have taken several different approaches in attempts to predict the onset of localized plastic flow by this mechanism. One approach focuses on localized flow at high strain rates and the thermal softening associated with plastic deformation. As noted previously, approximately 95% of the work of plastic deformation is transformed into thermal energy. The formation of localized plastic flow at high strain rates is attributed to strain and strain-rate hardening being overcome by thermal softening. Investigators have labeled

this type of localized plastic deformation adiabatic shear or adiabatic shear band formation to describe the volume of material affected and to describe the shear displacement observed as the deformation mode.

Several investigators have developed two criteria for predicting the onset of adiabatic shear (e.g., see Ref 89 and the subsequent discussions). Critical strains to the instability condition and additional strain in a quasi-stable condition have been estimated (Ref 90, 91, 92, 93). These models have adequately predicted adiabatic shear at high strain rates in standard tests, and they have allowed investigators to rank material in terms to their susceptibility to adiabatic shear band formation.

Other investigators (Ref 94, 95) have found extensive shear deformation localization at low strain rates that cannot be attributed to thermal softening. This phenomenon is similar to Lüder's bands that occur at yielding in annealed steels. Similar to Lüder's bands, their formation is attributed to a dislocation breakaway mechanism. However, it does not develop until work hardening is “saturated” (for example, quenched and tempered steels rapidly strain harden, then harden at a relatively slow rate) for materials with positive strain-rate sensitivity. Some materials with negative strain-rate sensitivities can develop shear bands very early in the deformation process. If the alloy is subject to dynamic strain aging, a dislocation breakaway mechanism, localized shear deformation can operate as a shear deformation instability that leads to incipient failure by shear.

Localized shear deformation typically follows planes of pure shear. In the plane-strain condition, the planes of pure shear are defined as a slip-line field in plasticity theory of homogeneous, isotropic materials (Ref 96). These are planes of zero extension that intersect each other at 90° and intersect a free surface at 45°. This only describes the planes of shear and does not justify plastic flow localization in the shear mode. However, most observations of localized deformation on a scale larger than grain size report that it conforms to expected slip-line fields.

Predictive Model for the Onset of Strain Localization. Semiatin and Jonas (Ref 51) developed a predictive model for the onset of strain localization. First, define a strain localization parameter:

$$a = \frac{d \ln \dot{e}}{de} = \frac{g' - 1}{m} \qquad \text{(Eq 16)}$$

where γ' is a temperature-compensated strain-hardening parameter, and m is the strain-rate sensitivity exponent. That is:

$$g = \frac{1}{s}\left(\frac{ds}{de}\right)\Bigg|_{\dot{e}} \qquad \text{(Eq 17)}$$

so, for isothermal loading:

$$g = \frac{1}{Ke^{n}}\left(nKe^{n-1}\right) = \frac{n}{e} \qquad \text{(Eq 18)}$$

but for nonisothermal loading, the term:

$$g = \frac{1}{s}\left(\frac{\partial s}{\partial T}\right)\frac{\partial T}{\partial e}\Bigg| \qquad \text{(Eq 19)}$$

must be added to γ, giving γ'.

According to Ref 51, materials that strain harden or show little flow softening do not easily develop strain localization shear bands. For materials that do flow-soften, strain localization is still minimized, so long as the rate-hardening exponent is of the order of 0.2 or more. Note that this is not the case at room temperature for many engineering materials. Semiatin predicts strain localization in shear bands, based on extensive compression testing when the parameter α (Eq 16) exceeds 5.

Once strain localization initiates, further deformation outside of the band essentially ceases. The displacements that occur within the band consist of both strain components and rotation components. Assuming small strain deformation, the resulting expression is:

$$d_{ij} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s \\ 0 & 0 & 0 \end{bmatrix}}_{Displacement} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s/2 \\ 0 & s/2 & 0 \end{bmatrix}}_{Strain} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & s/2 \\ 0 & -s/2 & 0 \end{bmatrix}}_{Rotation} \qquad \text{(Eq 20)}$$

The resolved shear stress acting in the band increases, because elements within the band rotate into the direction of the band, which still further intensifies the strain. The magnitude of the rotation is as large as the strain.

Effects of Strain Localization on Crack Initiation and Crack Growth. Another consequence of strain localization is that it provides alternative crack-initiation sites and growth directions and therefore, alternative fracture surface orientations. Both macroscale structural observation and microscale examination of fracture surfaces of strained components often reveal the presence of strain localization within narrow bands of material in both metallic and polymeric materials (Fig. 88). Deformation within these bands has been shown to develop at early stages of deformation via TEM studies of dislocation configurations (Ref 98), but they are not visible at the micro- or macroscale until relatively large deformations have been achieved. They have been observed for axial tension and compression as well as for torsion loading.

Fig. 88 Shear band formation in oriented polyethylene terephthalate. Source: Ref 97

For some conditions, deformation is localized within small-scale shear bands in the ligaments between voids, leading to crack propagation on planes of high shear stress. The crack can then assume a zig-zag path back and forth across the mean plane normal to the applied load. Crack propagation is on planes of high shear stress at the microscale between particles, and the scale of the shear band is that of the microstructure.

Cracking during compression loading of a ductile material between two platens results in a hoop strain that varies with height (barreling). Maximum tensile hoop strain occurs near midheight. Two common crack-initiation sites are then at the surface of the specimen in the barreled region and along the centerline at the

development of a cavity. However, this incipient crack cavity may close with further deformation (Fig. 89). Once strain localization occurs, two additional fracture-initiation sites are possible: within and parallel to one of the shear bands, typically at the component-platen interface, and at the intersection of shear bands. References 99 and 100 provide an extensive discussion of the formation of these shear bands in tensile loading plate geometries, providing plane-strain deformation and references for experimental observation of each of the models. Void nucleation has been shown at the intersection of shear bands in aluminum single crystals that appear to be bounded by slip planes (first case) (Ref 101). Figure 38 of a mild steel tensile specimen shows crack initiation at the intersection of shear bands on the surface of a cylindrical section, and Fig. 46 shows shear band intersections in a plate specimen. Semiatin and Jonas (Ref 51) observed the first case in compression loading.

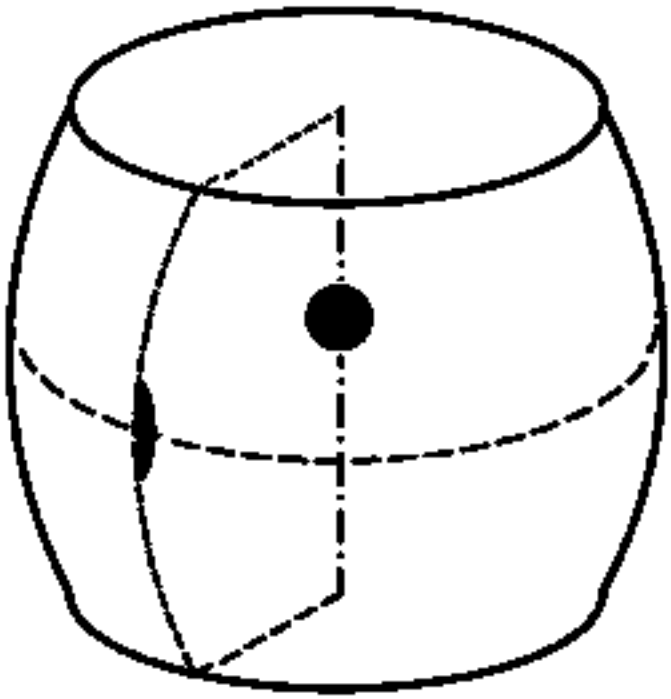


Fig. 89 Crack-initiation locations in compression loading. Crack initiation may occur along the centerline and at the surface on a hoop plane. Further deformation may close the centerline crack.

Void Nucleation Models

One of the first early models for ductile fracture by void coalescence was developed by several workers, starting in the late 1950s. (See the discussion and model in Ref 102.) As for this model and subsequent models, two parameters of importance are the particle size and spacing. If there is no bond between the particle and the matrix, the initial hole size is the same as the particle size. The model predicts failure when the ratio of particle spacing to hole size exceeds 2.7.

A later model to predict the strain at the onset of fracture by void nucleation and growth is due to McClintock (Ref 103). The model recognizes that ductile fracture is typically initiated by debonding and the growth of holes around second phases or inclusions. It neglects the debonding part of fracture and considers only hole growth. McClintock assumes that the holes are initially sufficiently far apart (spacing = *l*) that there is no interaction between the holes. He then considers the growth of a single hole within an element of material equal to the hole spacing. Using the von Mises yield criterion and considering a cylindrical specimen in which the cylinder diameter is equal to the hole spacing, McClintock concludes that the fracture strain is given according to:

$$e_f \approx \frac{(1-n)\ln(1/r)}{\sinh\left[(1-m)(\overline{s}_t / s_e\right]} \qquad \text{(Eq 21)}$$

where n is the strain-hardening exponent, ρ is the area fraction of voids on the plane of maximum normal stress, σ_t and σ_e are the mean transverse tension stress in the neck and the nominal effective stress. The equation is not easy to use because of the often unknown quantities σ_t and σ_e. However, Backofen (Ref 32) (Fig. 16) has

presented a simpler argument to conclude that the fracture strain is related to void size (w) and void spacing (L) according to:

$$e_f \cong 2\ln\frac{L}{w} \approx \frac{1}{r} \qquad \text{(Eq 22)}$$

where ρ is again the area fraction of voids. Backofen then argues that McClintock's model reduces to his model because of the small dependence of fracture strain on strain hardening. Neither Backofen nor McClintock address the case of zig-zag fracture.

Bates (Ref 104) approached the problem of pure ductile tearing in a different way, with an emphasis not on fracture strain but on fracture toughness. Using a slip-line approach to the problem and a nonhardening material, he was able to consider the case of a dual size distribution (and dual interfacial strength) common to sulfide-containing quenched and tempered steels. The results of his work predict conditions for both void nucleation and void growth as well as the important extension to quantitative fractography associated with ductile fracture.

A Dislocation Theory Model for Ductile Fracture. This model from Brown and Stobbs (Ref 33) relates the dislocation density (ρ_d) to the nominal principal strain (ε_1):

$$\rho_d = 1.7\varepsilon_1/(rb) \qquad \text{(Eq 23)}$$

where r is the particle radius and b is the Burgers vector.

The local elevated stress (σ_L) at the particle due to plastic deformation is given by:

$$\sigma_L = AG(\varepsilon_1 b/r)^{1/2} \qquad \text{(Eq 24)}$$

where A is a constant for a given material lying between 0.76 and 1.78, and G is the shear modulus.

The total stress (σ_T) acting to debond the particle from the matrix is a combination of the local elevated stress (σ_L), the mean (hydrostatic) stress (σ_m), and the deviatoric stress (von Mises stress) (σ_d) and is given by:

$$\sigma_T = \sigma_L + \sigma_m + \sigma_d \qquad \text{(Eq 25)}$$

The critical principal strain (e_1^c) at which decohesion occurs can be calculated from the mean stress and the critical total stress for decohesion (e_T^c):

$$e_1^c = Ar(s_T^c - s_m)^2 \qquad \text{(Eq 26)}$$

where A is a material constant related to the volume fraction of particles, and r is the particle radius. Note that the value of the critical strain is affected by the sign of the mean stress, so that the application of a hydrostatic pressure increases the strain to decohesion and, as Bridgeman (Ref 105) demonstrated, can result in complete avoidance of ductile fracture.

The critical stress for decohesion is not always known. Plots of measured critical strain for decohesion have been found experimentally for spherodized steel and generally fit this relationship shown in Ref 106. Values of the decohesion stress can be estimated from extrapolation of the measured data to a zero strain. It is in good agreement with other estimates for the strength of a carbide-steel interface (Ref 107).

Stable void growth is generally considered to be elongation of voids without dilation or coalescence. A model for the growth of a spherical void in both a nonhardening and a hardening matrix has been developed (Ref 108, 109). Prediction matches experimental data for a copper system. Particle elongation is the stable stage of MVC (carbides not considered). Void coalescence occurs on the weakest plane in a rapid manner. For hard, brittle particles, a dual-population model has been proposed to account for cracking of particles before reaching the critical decohesion strain. This model not only takes into account the size of the particles but also the volume of particles. Reference 110 shows that data fit the prediction very well for round particles.

Void Coalescence Microscale Models

As previously noted, the growth of an incipient crack may occur either by enlargement and linkage of the voids or by slip band deformation in ligaments between the voids. The dominant mechanism depends on particle size and spacing. For large particles relatively closely spaced, growth of an incipient crack is by enlargement and linkage of the voids, especially in the presence of constraint. This type of crack growth is referred to as pure ductile tearing, which usually occurs on the plane of maximum normal stress. A model for pure ductile tearing

is described first, followed by models of slip deformation in the inter-ligament region when nucleation occurs at wider-spaced particles.

Pure Ductile Tearing. Linkage of the voids often occurs on the plane of maximum normal stress. For large particles relatively closely spaced, and especially in the presence of constraint, void coalescence occurs on the plane of maximum normal stress, and so, the crack opens on this plane. If this process occurs in a prenotched specimen, the sequence to permit creation and growth of a new crack from the preexisting defect is that shown in Fig. 90 (Ref 104) and is often described as pure ductile tearing. Linkage of voids occurs (approximately) when the void height (D_y) is approximately equal to the intervoid ligament (L - D_x/2). Bates provides a discussion of various models for ductile crack growth based on slip-field analysis, together with several references.

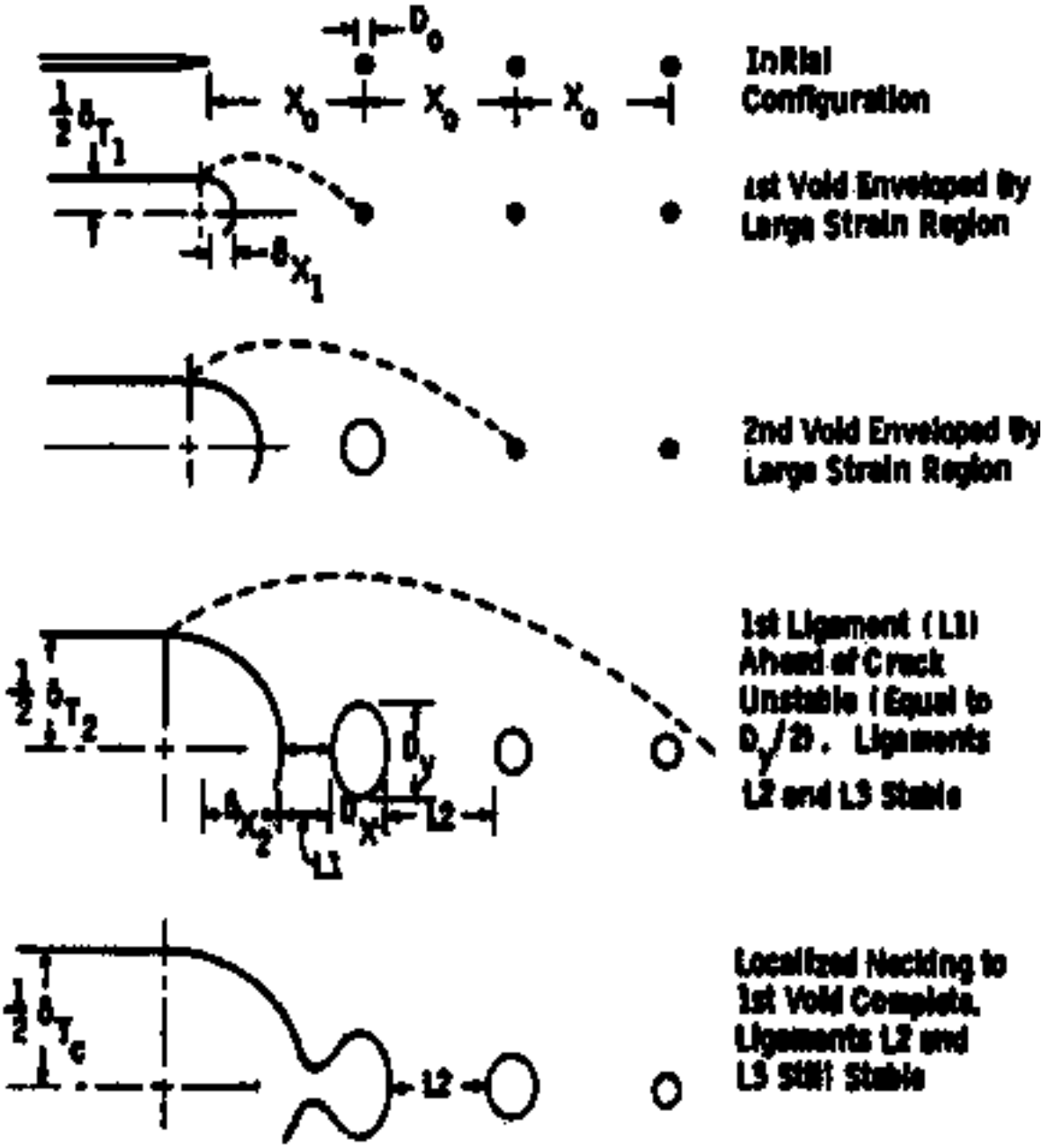


Fig. 90 A model for ductile fracture due to Rice and Johnson. The model considers pure tearing on the plane of maximum normal stress. The preexisting flaw is assumed to blunt during yield to create a large strain field in front of the flaw. Debonding and growth at a particle interface occurs. The resulting void growth permits linkage of the void with the preexisting flaw. Details are available in Ref 104. Source: Ref 104

Interligament Deformation. When void nucleation occurs at wider-spaced particles, slip deformation in the interligament region becomes more important in the process of void coalescence. Two general types of models

are considered: one based on localized shear deformation in the interligament region (void sheet fracture), and the other based on necking in the interligament region (the plastic load-limit model).

Void sheet fracture proposed by McClintock (Ref 111) and others (Ref 52, 112, 113, 114) involves localized shear bands connecting large pores. The model known as the Berg-Gurson model (Ref 112) uses plastic continuum mechanics to predict the critical condition based on a weakening instability of voids.

McClintock first presented experimental evidence for such a model. Fracture initiates at (large) MnS inclusions, followed by microscale strain localization on planes of high shear stress. These planes randomly contain smaller carbide particles that subsequently nucleate voids and thus is referred to as void sheet fracture. The carbides act as void-initiation sites only after considerable strain accumulation between the larger MnS sites. It is postulated that void sheet formation (but not necessarily shear band formation) can be prevented or delayed if there are no carbides present to nucleate the secondary voids (Ref 43).

Plastic Load-Limit Model. The plastic load-limit model is also based on plastic continuum mechanics but does not depend on shear localization to predict the growth and coalescence of voids. This model predicts the growth and coalescence of voids from necking of the material between voids when a critical stress is attained (Ref 52). Because the plastic flow is localized to the weakest plane and involves only a thickness of material on the order of particle spacing, a very small amount of macroscopic strain is needed to complete coalescence of voids. Thus, once a critical stress is attained, incipient coalescence (necking of the material between voids) occurs.

For this model, the local stress (σ_n) in the neck between the voids is higher than the mean normal stress (σ_m). The relationship for stress in a plane-strain condition is given by:

$$(s_n / k)\left(1/\sqrt{V_f}\right) = 1 + s_m / k \qquad \text{(Eq 27)}$$

where V_f is the void volume fraction, and k is the current shear stress (assumed to be the same in the bulk as it is in the intervoid region).

It is apparent that for a fixed condition of mean stress, raising the void volume fraction increases the stress in the neck, which would lead to a critical condition of void coalescence at an earlier strain.

In conditions of very large mean stress, for example, the stress state at a crack tip, a critical stress for void coalescence may be present at void nucleation. Thomason (Ref 52) has theorized that the value of the ratio ($\sigma_m/2k$) over 2.0 results in incipient necking at the point of void nucleation at a void volume fraction above 0.02. The value of this ratio is 2.07 for plane strain. It is only approximately 0.5 for plane stress and is approximately 0.7 for a round tensile specimen. At this level ($\sigma_m/2k = 0.7$), a particle volume fraction of 0.12 or higher would be required to initiate incipient pore coalescence in a round tensile bar.

Closed-form solutions to continuum plasticity depend on isotropic properties and uniformly distributed sources for pore formation. Attempts to include the natural variation and random nature of second-phase particles have been made using finite-element methods. One such approach concludes that heterogeneities are predicted to reduce the overall ductility of a tensile test (Ref 115).

References cited in this section

4. R.W.K. Honeycombe, *The Plastic Deformation of Metals,* 2nd ed., Edward Arnold and American Society for Metals, 1984

32. W.A. Backofen, *Deformation Processing,* Addison-Wesley, 1972

33. L.M. Brown and W.M. Stobbs, *Philos. Mag.,* Vol 334, 1976, p 351

43. D. Chae, D.A. Koss, A.L. Wilson, and P.R. Howell, The Effect of Microstructural Banding on Failure Initiation of HY-100 Steel, *Metall. Trans. A,* Vol 31, 2000, p 995

51. S.L. Semiatin and J.J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization,* American Society for Metals, 1984

52. P.F. Thomason, *J. Inst. Met.,* Vol 96, 1968, p 360

83. A. Considére, *Ann. Ponts Chaussées,* Vol 9 (Ser. 6), 1885, p 574

84. E. Hart, *Acta Metall.,* Vol 15, 1967, p 351

85. S.S. Sternstein, Yielding Modes in Glassy Polymers, *Polymeric Materials,* American Society for Metals, 1975

86. R. Hill, On Discontinuous Plastic States with Special Reference to Localized Necking in Thin Sheets, *J. Mech. Phys. Solids,* Vol 1, 1952, p 19

87. Y.W. Kim and D.L. Bourell, *Metall. Trans. A,* Vol 19, 1988, p 2041

88. K. Choo, Y.C. Chi, and L. Duffy, *Metall. Trans. A,* Vol 21, 1990, p 1161

89. H.C. Rogers, *Deformation, Processing, and Structure,* American Society for Metals, 1984, p 425

90. R.S. Culver, *Metallurgical Effects at High Strain Rates,* R.W. Rohde, B.M. Butcher, J.R. Holland, and C.H. Karnes, Ed., AIME, 1973, p 519

91. R.F. Recht, *J. Appl. Mech. (Trans. ASME),* June 1964, p 189

92. M.R. Staker, *Scr. Metall.,* July 1984, Vol 18, p 735

93. S.L. Semiatin, M.R. Staker, and J.J. Jonas, *Acta Metall.,* Vol 32 (No. 9), 1984, p 1347

94. R.H. Ernst and J.W. Spretnak, *Trans. Iron Steel Inst. Jpn.,* Vol 9, 1969, p 361

95. D.L. McGarry, "The Dynamic Ideally Plastic State in AISI 4340 Steel," Ph.D. dissertation, The Ohio State University, 1977

96. R. Hill, *The Mathematical Theory of Plasticity,* Oxford University Press, 1967, p 132

97. I.M. Ward, *Mechanical Properties of Solid Polymers,* 2nd ed., John Wiley & Sons, 1983

98. Y.W. Kim and D.L. Bourell, Microscopic Shear Band Localization in Nickel, *Metall. Trans. A,* Vol 19, 1988, p 2041

99. J.E. Bird and J.M. Carlson, Geometric Models of Internal Shape Change as Shear Bands Form during Plane Strain Extension, *Metall. Trans. A,* Vol 18, 1987

100. J.E. Bird and J.M. Carlson, Shear Band Formation during Sheet Forming, *J. Met.,* Nov 1986, p 47

101. K.E. Puttick, Necking and Fracture in Aluminum Crystals, *Acta Metall.,* Vol 11, 1963, p 986

102. *Mechanical Working of Steel,* P.H. Smith, Ed., Gordon and Breach, 1964, p 131

103. F.A. McClintock and A.S. Argon, *Mechanical Behavior of Materials,* Addison-Wesley, 1966, p 524

104. R.C. Bates, Modeling of Ductile Fracture by Microvoid Coalescence, *Fracture: Interactions of Microstructure, Mechanisms, Mechanics,* J.M. Wells and J.D. Landes, Ed., TMS, 1984, p 117

105. P.W. Bridgeman, *Studies in Large Plastic Flow and Fracture,* McGraw-Hill, 1952

106. G. Leroy, J.D. Embry, G. Edwards, and M.F. Ashby, *Acta Metall.,* Vol 29, 1981, p 1509

107. A.S. Argon, *J. Im. Met. Trans. 6A,* 1975, p 839

108. J.R. Rice and D.M. Tracey, *J. Mech. Phys. Solids,* Vol 17, 1969, p 201

109. J.D. Atkinson, Ph.D. dissertation, University of Cambridge, 1973, and L.M. Brown, Cambridge Fracture Conference, Churchill College Inst. of Physics, 1974

110. G. Leroy, J.D. Embury, G. Edwards, and M.F. Ashby, *Acta Metall.,* Vol 29, 1981, p 1509

111. F.A. McClintock, S.M. Kaplan, and C.A. Berg, *Int. J. Fract. Mech.,* Vol 2, 1966, p 614

112. A.L. Gurson, *J. Eng. Mater. Technol. (Trans. ASME),* Vol 99, 1981, p 2

113. D. Kwon and R.J. Asaro, *Metall. Trans. A,* Vol 21, 1990, p 117

114. V. Tvergaard, *Int. J. Fract.,* Vol 178, p 389

115. J. Besson and A. Pineau, *Recent Advances in Fracture,* TMS, 1997, p 125

Mechanisms and Appearances of Ductile and Brittle Fracture in Metals

W.T. Becker, University of Tennessee Emeritus; Dennis McGarry, FTI/SEA Consulting

---

References

1. D. Hull, *Fractography,* Cambridge University Press, 1999

2. L. Sohnke, Normal Stress Law for NaCl, *Poggendorfs Ann.,* Vol 137, 1869, p 177

3. E. Schmid and W. Boas, *Plasticity of Crystals,* 2nd ed., F.A. Hughes & Co., 1959, and Chapman and Hall Ltd., 1968

4. R.W.K. Honeycombe, *The Plastic Deformation of Metals,* 2nd ed., Edward Arnold and American Society for Metals, 1984

5. D. MacLean, *Mechanical Properties of Metals,* John Wiley & Sons, 1962

6. E. Schmid and W. Boas, *Kristallplastizitaet mit Besonderer Bericksichtigung der Metalle,* Springer-Verlag, Berlin, 1935

7. P.G. Partridge, The Crystallography and Deformation Modes of Hexagonal Close Packed Metals, *Metals and Mater.,* Vol 1, 1967

8. R.E. Reed-Hill, *Physical Metallurgy Principles,* Van Nostrand, 1973 p 753

9. A. Vasilev, S. Firstov, and A. Koval, Cleavage Mechanism Nucleation and Fracture Toughness, *Recent Advances in Fracture,* R.K. Mahidhara et al., Ed., TMS, 1997, p 315

10. J.J. Gilman, *Trans. AIME,* Vol 200, 1954, p 621

11. A. Deyurterre and G.B. Greenough, *J. Inst. Met.,* Vol 84, 1956, p 337

12. A.N. Stroh, *Proc. R. Soc. A.,* Vol 233, 1954, p 404

13. N.P. Allen, B.E. Hopkins, and J.E. McClennan, *Proc. R. Soc. A.,* Vol 234, 1956, p 221

14. C.D. Beachem, Microscopic Analysis of Cleavage Mechanisms, *Special Volume of the Proceedings of the Second International Conference on Mechanical Behavior of Materials,* American Society for Metals, 1976, p 1188

15. A.H. Cottrell, *Fracture,* B.L. Averbach, B.L. Felbeck, D.K. Hahn, and D.A. Thomas, Ed., John Wiley & Sons, 1959

16. A.H. Cottrell, *Trans. AIME,* Vol 212, 1958, p 196

17. E.R. Parker, *Fracture,* B.L. Averbach, D.K. Felbeck, and G.T. Hahn, Ed., John Wiley, 1959, p 181

18. R. Honda, *Acta Metall.,* Vol 9, 1961, p 969

19. J.J. Gilman, *Dislocations and Mechanical Properties of Crystals,* John Wiley, 1957

20. T.H. Courtney, *Mechanical Behavior of Materials,* McGraw-Hill, 1990, p 392

21. R.W. Cahn, Survey of Recent Progress in the Field of Deformation Twinning, *Deformation Twinning,* R.E. Reed-Hill, J.P. Hirth, and H.C. Rogers, Ed., Goordon and Breech, 1964, p 1

22. G.Y. Chin, Formation of Deformation Twins in fcc Crystals, *Acta Metall.,* Vol 21, 1973, p 1353

23. J. Spreadborough, Informal Discussion on Fundamental Brittle Fracture Research on Unirriadiated and Irradiated Steels, *Proceedings of the Colloqium on Brittle Fracture and Selected Problems in Nuclear Pressure Vessel Steels,* J. Sebille, Ed., (Euratom, Brussels) European Atomic Energy Community, 1966

24. C.D. Beachem, "The Interpretation of Electron Microscope Fractographs," NRL Report 6360, Naval Research Laboratory, Washington, D.C., 21 Jan 1966

25. J.L. McCall and P.M. French, Ed., *Metallography in Failure Analysis,* Plenum Press, 1977, p 7

26. R.E. Reed-Hill, J.P. Hirth, and H.C. Rogers, Ed., *Deformation Twinning,* Gordon and Breech, 1964

27. D. Hull, Twinning and Nucleation of Cracks, *Fracture of Solids,* D.C. Drucker and J.J. Gilman, Ed., Gordon and Breech, 1963

28. R. Priestner, The Relationship between Brittle Cleavage and Deformation Twinning in bcc Metals, *Deformation Twinning,* R.E. Reed-Hill, J.P. Hirth, and H.C. Rogers, Ed., Gordon and Breech, 1964, p 321

29. C.F. Tipper, *Metallurgia,* Vol 39, 1949, p 133

30. D.R. Johnson, "Toughness of Tempered Bainitic Microstructures in a 4150 Steel," master's thesis, University of Tennessee, 1990

31. H. Tada, T. Kobayashi, and A. Takahashi, Micromechanism of Fracture in Wrought Aluminum Alloy Containing Coarse Inclusion Particles, *Alum. Trans.,* Vol 1 (No. 1), 1999, p 109

32. W.A. Backofen, *Deformation Processing,* Addison-Wesley, 1972

33. L.M. Brown and W.M. Stobbs, *Philos. Mag.,* Vol 334, 1976, p 351

34. G. Le Roy, J.D. Embury, G. Edwards, and M.F. Ashby, *Acta Metall.,* Vol 29, 1981, p 1509

35. A.W. Thompson, Modes of Ductile Fracture, *Metall. Trans. A,* Vol 18, 1987, p 1877

36. K.S. Chan, Near Tip Fracture Processes in Ductile Materials, *Recent Advances in Fracture,* TMS/AIME, 1998, p 137

37. H.C. Rogers, The Effect of Material Variables on Ductility, *Ductility,* American Society for Metals, 1968, p 31

38. C.J. Beevers and R.W.K. Honeycombe, *Philos. Mag.,* Vol 7, 1962, p 763

39. G.A. Lange, The Fracture Behavior of Pure Face-Centered Cubic Metals with Prevention of Macroscopic Deformation, *Scanning Electron Microscopy, Part III,* 1985

40. R.L. Lyles, Jr. and H.G.F. Wilsdorf, Microcrack Nucleation and Fracture in Silver Crystals, *Acta Metall.,* Vol 23, 1975, p 269

41. R.H. Van Stone and T.B. Cox, "Use of Fractography and Sectioning Techniques to Study Fracture Mechanisms," ASTM STP 600, *Fractography: Microscopic Cracking Mechanisms,* C. Beachem and W. Warke, Ed., ASTM, 1976 p 18

42. G. Powell, S.-H. Cheng, C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys,* Battelle Press, 1992

43. D. Chae, D.A. Koss, A.L. Wilson, and P.R. Howell, The Effect of Microstructural Banding on Failure Initiation of HY-100 Steel, *Metall. Trans. A,* Vol 31, 2000, p 995

44. D. Broek, *Engineering Fracture Mechanics,* Vol 5, 1973, as cited in R. Hertzberg, *Deformation and Fracture Mechanics of Engineering Materials,* 4th ed., John Wiley, 1996

45. A.J. Edwards, "Depth Measurements on Fracture Surfaces," report of NRL progress, Naval Research Laboratory, 1963, as cited in A.J. Brothers and S. Yukawa, "Engineering Applications of Fractography." ASTM STP 436, *Electron Fractography,* ASTM, 1968 p 179

46. B.R. Banerjee, J.J. Hauser, and J.M. Caperos, "Relation of Strength and Toughness to Fine Structures in a Beta Titanium Alloy," ASTM STP 432, *Applications Related Phenomena in Titanium Alloys,* ASTM, 1968

47. D. Broek, *The Practical Use of Fracture Mechanics,* Kluwer Academic Publishers, 1989, p 1

48. H. Conrad, *High-Strength Materials,* V.F. Zackay, Ed., John Wiley & Sons, Inc., 1965, p 436

49. A. Nadai, *Theory of Flow and Fracture of Solids,* 2nd ed., McGraw-Hill, 1950

50. M.G. Mendiratta, R. Goetz, D.M. Dimiduk, and J.J. Lewandowski, *Met. Mater. Trans. A,* Vol 26, 1995, p 1772

51. S.L. Semiatin and J.J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization,* American Society for Metals, 1984

52. P.F. Thomason, *J. Inst. Met.,* Vol 96, 1968, p 360

53. T. Gladman and F.B. Pickering, The Effect of Grain Size on Ferrous Materials, *Yield, Flow, and Fracture of Polycrystals,* T.N. Baker, Ed., Applied Science Publishers, 1983

54. F.R. Larson and F.L. Carr, Tensile Fracture Surface Configurations of a Heat Treated Steel as Affected by Temperature, *Trans.* American Society for Metals, Vol 55, 1962, p 599–611

55. G.T. Hahn, B.L. Averbach, W.S. Cohen, and M. Cohen, *Conference on Fracture,* (Swampscott, MA), John Wiley & Sons, 1959, p 4-1

56. G.R. Speich and W.A. Spitzig, *Metall. Trans. A,* Vol 13, 1982, p 2250

57. A.W. Thompson and P.F. Weihrauch, *Scr. Metall.,* Vol 10, 1976, p 205

58. Y.W. Kim and D.L. Bourell, *Metall. Trans. A,* Vol 19, 1988, p 2041

59. M.J. Manjoine, *J. Appl. Mech. (Trans. ASME),* Vol 11 (No. 4), 1944

60. N. Chung et al., *Acta Metall.,* Vol 25, April 1977, p 377–381

61. P.F. Weinrich and I.E. French, Influence of Hydrostatic Pressure on the Fracture Mechanisms of Short Tensile Specimens of Cu and Brass, *Acta Metall.,* Vol 24, 1976, p 317

62. K.E. Puttick, *Philos. Mag.,* Vol 4, 1959, p 964

63. W.A. Backofen, *Trans. ASM,* Vol 51, 1959, p 946

64. E.R. Parker, *Fracturing of Metals,* American Society for Metals, 1948, p 87

65. R.N. Katz and K.H. Abbott, *Application of Fracture Toughness Parameters to Structural Metals,* H.D. Greenberg, Ed., Gordon and Breach Science Publishers, 1966, p 69

66. P.F. Thomason, *Ductile Fracture of Metals,* Pergamon Press, 1990, p 159

67. H.W. Hayden and S. Floreen, *Acta Metall.,* Vol 17, 1969, p 213

68. D.L. McGarry, "The Effect of Strength Level and Strain Rate on the Mode of Fracture Initiation in AISI 4340 Steel," master's thesis, The Ohio State University, 1972, p 80

69. J.H. Bucher, G.W. Powell, and J.W. Spretnak, *Application of Fracture Toughness Parameters to Structural Metals,* H.D. Greenberg, Ed., Gordon and Breach Science Publishers, 1966, p 323

70. S.T. Rolfe and J.M. Barsom, *Fracture and Fatigue Control in Structures—Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

71. E.R. Parker, Fracture, *Proc. Second Tewksbury Symposium,* Butterworth Co., 1969, p 11

72. T. Yokobori, A. Otsuka, and T. Takahashi, A Note on Cleavage Fracture of Low-Carbon Steel at Liquid-Nitrogen Temperatures, *Fracture of Solids,* Gordon and Breach Science Publishers, 1963, p 263

73. H.C. Rogers, Fundamental Aspects of Fracture in Deformation Processing, *Fundamentals of Deformation Processing,* Syracuse Univ. Press, 1964, p 199

74. S. Bhattacharya, et al., Ed., *IITRI Fracture Handbook,* IIT Research Institute, 1979

75. W.S. Lin, "The Ductile Brittle Fracture Transformation: A Comparison of Macro and Microscopic Observations on Compact Tension Specimens," master's thesis, University of Tennessee, 1974

76. A.G. Atkins and Y.-W. Mai, *Elastic and Plastic Fracture,* Halsted Press, 1985

77. T.A. Anderson, *Fracture Mechanics,* 2nd ed., CRC Press, 1995

78. R.C. Bates, Modeling of Ductile Fracture by Microvoid Coalescence, *Fracture,* J.M. Wells and J.D. Landes, Ed., TMS, 1984

79. A. Vasilev, S. Firstov, and A. Koval, Cleavage Mechanism Nucleation and Fracture Toughness, *Recent Advances in Fracture,* R.K. Mahidhara et al., Ed., TMS, 1997, p 315

80. J.C. Scully, Ed., *The Theory of Stress Corrosion Cracking in Alloys,* NATO, Scientific Affairs Div., Brussels, 1971

81. R.L. Woodward, *Metall. Trans. A,* Vol 10, 1979, p 569

82. W.T. Becker, Brittle Failure in a Large Grain Storage Bin, *Handbook in Case Histories of Failure Analysis,* Vol 2, ASM International, 1993, p 470

83. A. Considére, *Ann. Ponts Chaussées,* Vol 9 (Ser. 6), 1885, p 574

84. E. Hart, *Acta Metall.,* Vol 15, 1967, p 351

85. S.S. Sternstein, Yielding Modes in Glassy Polymers, *Polymeric Materials,* American Society for Metals, 1975

86. R. Hill, On Discontinuous Plastic States with Special Reference to Localized Necking in Thin Sheets, *J. Mech. Phys. Solids,* Vol 1, 1952, p 19

87. Y.W. Kim and D.L. Bourell, *Metall. Trans. A,* Vol 19, 1988, p 2041

88. K. Choo, Y.C. Chi, and L. Duffy, *Metall. Trans. A,* Vol 21, 1990, p 1161

89. H.C. Rogers, *Deformation, Processing, and Structure,* American Society for Metals, 1984, p 425

90. R.S. Culver, *Metallurgical Effects at High Strain Rates,* R.W. Rohde, B.M. Butcher, J.R. Holland, and C.H. Karnes, Ed., AIME, 1973, p 519

91. R.F. Recht, *J. Appl. Mech. (Trans. ASME),* June 1964, p 189

92. M.R. Staker, *Scr. Metall.,* July 1984, Vol 18, p 735

93. S.L. Semiatin, M.R. Staker, and J.J. Jonas, *Acta Metall.,* Vol 32 (No. 9), 1984, p 1347

94. R.H. Ernst and J.W. Spretnak, *Trans. Iron Steel Inst. Jpn.,* Vol 9, 1969, p 361

95. D.L. McGarry, "The Dynamic Ideally Plastic State in AISI 4340 Steel," Ph.D. dissertation, The Ohio State University, 1977

96. R. Hill, *The Mathematical Theory of Plasticity,* Oxford University Press, 1967, p 132

97. I.M. Ward, *Mechanical Properties of Solid Polymers,* 2nd ed., John Wiley & Sons, 1983

98. Y.W. Kim and D.L. Bourell, Microscopic Shear Band Localization in Nickel, *Metall. Trans. A,* Vol 19, 1988, p 2041

99. J.E. Bird and J.M. Carlson, Geometric Models of Internal Shape Change as Shear Bands Form during Plane Strain Extension, *Metall. Trans. A,* Vol 18, 1987

100. J.E. Bird and J.M. Carlson, Shear Band Formation during Sheet Forming, *J. Met.,* Nov 1986, p 47

101. K.E. Puttick, Necking and Fracture in Aluminum Crystals, *Acta Metall.,* Vol 11, 1963, p 986

102. *Mechanical Working of Steel,* P.H. Smith, Ed., Gordon and Breach, 1964, p 131

103. F.A. McClintock and A.S. Argon, *Mechanical Behavior of Materials,* Addison-Wesley, 1966, p 524

104. R.C. Bates, Modeling of Ductile Fracture by Microvoid Coalescence, *Fracture: Interactions of Microstructure, Mechanisms, Mechanics,* J.M. Wells and J.D. Landes, Ed., TMS, 1984, p 117

105. P.W. Bridgeman, *Studies in Large Plastic Flow and Fracture,* McGraw-Hill, 1952

106. G. Leroy, J.D. Embry, G. Edwards, and M.F. Ashby, *Acta Metall.,* Vol 29, 1981, p 1509

107. A.S. Argon, *J. Im. Met. Trans. 6A,* 1975, p 839

108. J.R. Rice and D.M. Tracey, *J. Mech. Phys. Solids,* Vol 17, 1969, p 201

109. J.D. Atkinson, Ph.D. dissertation, University of Cambridge, 1973, and L.M. Brown, Cambridge Fracture Conference, Churchill College Inst. of Physics, 1974

110. G. Leroy, J.D. Embury, G. Edwards, and M.F. Ashby, *Acta Metall.,* Vol 29, 1981, p 1509

111. F.A. McClintock, S.M. Kaplan, and C.A. Berg, *Int. J. Fract. Mech.,* Vol 2, 1966, p 614

112. A.L. Gurson, *J. Eng. Mater. Technol. (Trans. ASME),* Vol 99, 1981, p 2

113. D. Kwon and R.J. Asaro, *Metall. Trans. A,* Vol 21, 1990, p 117

114. V. Tvergaard, *Int. J. Fract.,* Vol 178, p 389

115. J. Besson and A. Pineau, *Recent Advances in Fracture,* TMS, 1997, p 125

Fatigue Fracture Appearances

Russell A. Lund, Consultant; Shahram Sheybany, Pacific Metallurgical Company

Introduction

FATIGUE failure of engineering components and structures results from progressive fracture caused by cyclic or fluctuating loads. The magnitude of each individual load event is too small to cause complete fracture of the undamaged component, but the cumulative action of numerous load cycles, often numbering in the hundreds of thousands and millions, results in initiation and gradual propagation of a crack or cracks. Complete fracture ensues when the crack reaches critical size. Fatigue is an important potential cause of mechanical failure, as most engineering components or structures are or can be subjected to varying loads during their lifetime.

This article focuses on fractography of fatigue. However, it should be noted that fractography is only part of the failure analysis of a fractured component. As with other fracture modes, proper identification of fatigue requires understanding of the fracture behavior of the particular material subject to failure analysis. At least some knowledge of environmental and service conditions is usually necessary. Evaluation of loading conditions, often in conjunction with fractography, also provides useful information regarding the fundamental cause of failure.

This article begins with an abbreviated summary of fatigue processes and mechanisms. Characteristic fatigue fracture features that can be discerned visually or under low magnification are then described. Typical microscopic features observed on structural metals are presented subsequently, followed by a brief discussion of fatigue in nonmetals. As described in this article, various macroscopic and microscopic features can be used in some situations to characterize the history and growth rate of fatigue. Because fatigue propagation occurs progressively over time, knowledge of inspections performed prior to complete failure can benchmark the time span over which cracking occurred (or can point out inspection deficiencies).

Correlation between mechanical analysis and fractography also can help pinpoint the fundamental cause of fatigue failure. Techniques for predicting fatigue crack initiation, fatigue crack propagation rate, and fatigue life, based on loading conditions in engineering components, are discussed in many texts on applied mechanical engineering such as Ref 1 and 2. The mechanisms and processes of fatigue are described in more detail in Ref 3. These topics are also addressed in *Fatigue and Fracture,* Volume 19 of *ASM Handbook.*

A number of publications provide extensive examples of representative fatigue fractographs (Ref 4, 5, 6, 7, 8, 9). This article is not intended to provide a comprehensive atlas of fatigue fracture appearances. While these reference texts are of immense value, the experience of the fractographer remains an important aspect of good failure analysis.

References cited in this section

1. R. Stephens, A. Fatemi, R. Stephens, and H. Fuchs, *Metal Fatigue in Engineering,* 2nd ed., John Wiley & Sons, 2001

2. J.A. Collins, *Failure of Materials in Mechanical Design; Analysis, Prediction and Prevention,* 2nd ed., John Wiley & Sons, 1993

3. S. Suresh, *Fatigue of Materials,* Cambridge University Press, 1991

4. *Fractography,* Vol 12, *Metals Handbook,* (9th ed.), ASM International, 1987

5. S. Bhattacharyya, V.E. Johnson, S. Agarwal, and M.A.H. Howes, Failure Analysis of Metallic Materials by Scanning Electron Microscopy, *IITRI Fracture Handbook,* IIT Research Institute, 1979

6. L. Engel and H. Klingele, *An Atlas of Metal Damage,* Prentice-Hall, 1981

7. G.W. Powell, S. Cheng, and C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys*, Battelle Press, 1992

8. J. Aleszka, Y. Kim, J. Scott, and A. Kumar, "Fracture Characteristics of Structural Steels: Reference Manual," U.S. Army Construction Engineering Research Laboratory Technical Report M-258, 1979

9. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage*, Prentice-Hall, 1981

Fatigue Fracture Appearances

Russell A. Lund, Consultant; Shahram Sheybany, Pacific Metallurgical Company

Fatigue Processes

Fatigue has been defined as "the process of progressive localized permanent structural change occurring in a material subjected to conditions that produce fluctuating stresses or strains at some point or points and that may culminate in cracks or complete fracture after a sufficient number of fluctuations" (ASTM E 1150-87, reapproved in 1993). From the perspectives of design, testing, and failure analysis, this fatigue process can be conveniently divided into three parts:

- Crack initiation
- Stable progressive crack propagation
- Final (fast) fracture

Design and testing aspects of fatigue are discussed in more detail elsewhere, such as in *Fatigue and Fracture*, Volume 19 of *ASM Handbook* and *Mechanical Testing and Evaluation*, Volume 8 of *ASM Handbook.* Volume 19 also includes more detailed discussion of the fatigue process than is presented in this article.
As illustrated in Fig. 1, fatigue failures have one or more fatigue origins (initiation sites), a region of progressive fatigue crack propagation, and a final fast overload fracture zone. Identification of the location and nature of origin sites is important in failure analysis of fatigue, as fatigue crack initiation is frequently the life-controlling step in the failure process. Discontinuities introduced during processing or manufacture (e.g., abnormally large inclusions, weld discontinuities, grinding cracks) or service (e.g., corrosion pits, impact damage) can result in fatigue crack initiation under circumstances where an unflawed component would have had acceptable fatigue life. Conversely, inadequate design, excessive load, or low material strength in fatigue can result in fatigue failure in the absence of a particular "defect" at the origin site. Fractography provides the basis for identifying the fatigue fracture mode and for locating origin sites from examination of fracture surfaces.

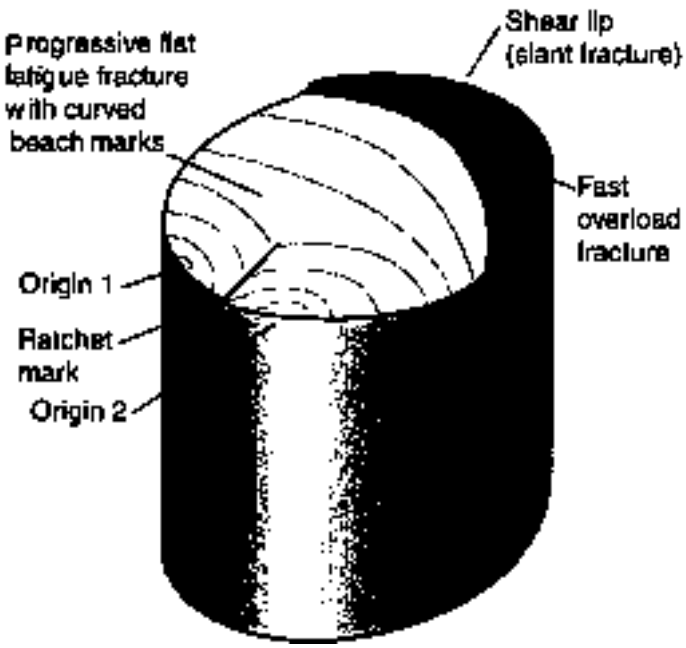


Fig. 1 General features of fatigue fractures.

Initiation of fatigue in common components made of commercial alloys occurs at—and is dominated by—material and geometric heterogeneities. Sites fostering fatigue crack initiation include inclusions, second phase particles, voids, machining marks and other surface flaws, and notches or other geometric variations. These sites can act as stress concentrators, or can rapidly produce microcracks by mechanisms such as inclusion debonding and particle fracture. The stress concentration effect permits local permanent plastic deformation at nominal stresses below yield. Initiation frequently occurs near free surfaces, as nominal stresses are often higher there (e.g., bending). Geometric variations at surfaces can locally elevate stresses and strains, and processes of damage accumulation (described below for pure unflawed materials) are also favored at surfaces.

Tests made to elucidate mechanisms of initiation frequently have been performed using pure single phase relatively homogeneous metals (without the heterogeneities described above). Based on observations from these tests, some researchers further separate the initiation stage into (Ref 3):

- Accumulation of local irreversible plastic deformation
- Creation of microscopic flaws
- Growth and coalescence of flaws to form one or more macroscopic cracks

Careful laboratory study of unnotched high-purity metals and alloys has shown that repeated plastic deformation can result in localization of permanent microstructural changes. For example, under inelastic cyclic loading, bands of intense slip known as *persistent slip bands* form in individual crystals of pure fcc metals such as copper. Fine protrusions and intrusions (peaks and troughs) of metal can form where these persistent slip bands intersect free surfaces (e.g., see Fig. 2 on page 64 in *Fatigue and Fracture*, Volume 19 of *ASM Handbook*). Continued surface roughening and damage accumulation ultimately results in crack nucleation at the surface. Cyclic damage can also accumulate at grain boundaries, particularly at elevated temperature or under conditions that cause grain boundary embrittlement, and at twin boundaries.

Fig. 2 Close-up view of ratchet marks between distinct surface origin sites in a low-alloy steel 18.4 cm (7.25 in.) shaft that failed in rotating bending fatigue. Ratchet marks (at arrows) are roughly radial steps formed where fatigue cracks initially propagating on different planes intersected. The ratchet mark at the middle arrow was damaged by postfracture contact.

Propagation. Fatigue crack propagation has been historically divided into two parts, designated stage I and stage II. Stage I propagation occurs on specific crystallographic planes with greatest resolved fluctuating shear stress. Stage II propagation occurs on planes normal to the fluctuating tensile stress. Certain metals (such as aluminum alloys) and loading conditions considered by early researchers investigating fatigue processes exhibited stage I propagation over a limited region (characteristically spanning no more than a few near-surface grains) adjacent to origin sites, followed by a transition to stage II propagation (e.g., see Fig. 16 on page 53 of *Fatigue and Fracture*, Volume 19 of *ASM Handbook*). Consequently, this division between stages of propagation was a natural one.

However, many commercial alloys (notably most steels) typically do not exhibit any detectable stage I propagation. After initiation in such alloys, growth immediately occurs by so-called stage II propagation. Notches, sharp corners, or preexisting cracks can also eliminate detectable stage I propagation in many metals. Similarly, some alloys (such as certain nickel base superalloys and cobalt base alloys) can display very extensive regions of propagation on specific crystallographic planes (so-called stage I propagation) with little or no discernable stage II type of growth prior to final fracture. Depending on loading and geometry, extensive propagation in these materials occurs on alternate crystallographic planes such that growth follows a zigzag or serrated faceted pattern, on average normal to the direction of fluctuating tension.

The stage I propagation mechanism is favored in materials that exhibit planar slip and in some alloys strengthened by coherent precipitates (such as age hardened aluminum) and under fluctuating loads low enough that the cyclic fracture process zone is small

relative to a characteristic microstructural scale parameter. Hence, stage I propagation is often favored by large grain size. Stage II propagation is favored in materials with easy cross-slip (wavy slip). In this article, fatigue propagation appearance is described in terms of morphology rather than the didactic separation into stages I and II.

From a practical failure analysis perspective, fatigue crack propagation regions are the most important for identifying the mode or type of failure. While fatigue propagation usually occupies a much larger area than the origin site(s), initiation processes may occupy a significant portion or majority of the overall time (duration) of the fatigue process.

Final fracture does not occur by a fatigue mechanism at all but is simply fracture caused by the last load application, resulting in complete separation. This final overload fracture may be more or less ductile or brittle, depending on notched or precracked properties of the material (see the articles “Overload Failures” and “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals” in this Volume).

Reference cited in this section

3. S. Suresh, *Fatigue of Materials*, Cambridge University Press, 1991

Fatigue Fracture Appearances

Russell A. Lund, Consultant; Shahram Sheybany, Pacific Metallurgical Company

Microscopic Appearance of Fatigue Fracture in Metals

Microscopic examination of fracture surfaces can disclose features related to mode and mechanism that cannot be discerned macroscopically. While readily identifiable microscopic features associated with fatigue of particular materials exist, not all materials display the same characteristics, and not all materials exhibit a unique fatigue morphology.

Examination of fracture features by light microscopy is limited by resolution and depth of field capabilities. Fine features cannot be resolved, and the height of important features on fracture surfaces often exceeds optical depth of field, especially at higher magnifications.

Electron microscopes have been used to study fracture of materials since the 1950s, and scanning electron microscopes (SEMs) have been commercially available for more than 35 years (Ref 22). Today, the SEM is standard equipment for any laboratory performing failure analysis and fractography. The SEM offers significantly superior resolution and depth of field compared to light microscopes, markedly increasing the useful magnification range. Compared to transmission electron microscopes, SEM permits examination of relatively large samples and produces images with a 3D character.

Striations are a classic microscopic feature of fatigue in metals. The dictionary defines striations as series of parallel grooves or lines. In this article, the narrower, commonly used definition of fatigue striation is used: a striation is a microscopic groove or furrow that delineates a local fatigue crack front after a single cycle of loading. Figure 32 illustrates the correlation between striations and loading. Determination that particular marks are striations, and consequent identification of characteristic appearance, has been accomplished by comparing striation spacing with macroscopic measurements of fatigue crack growth rate.

Fig. 32 Transmission electron fractograph of aluminum alloy laboratory spectrum loading fatigue test. Striation spacing varies according to loading, which consisted of ten cycles at a high stress alternating with ten cycles at a lower stress. The fracture surface exhibits bands of ten coarse striations alternating with bands of ten fine striations. Source: Ref 16

Well-defined fatigue striations form in certain materials under appropriate stress ranges, particularly in metals with multiple active slip planes. For example, well-defined striations are frequently observed at intermediate growth rates in metals such as wrought aluminum and stainless steel. Examples of fatigue striations are presented in Fig. 33, 34, 35, 36.

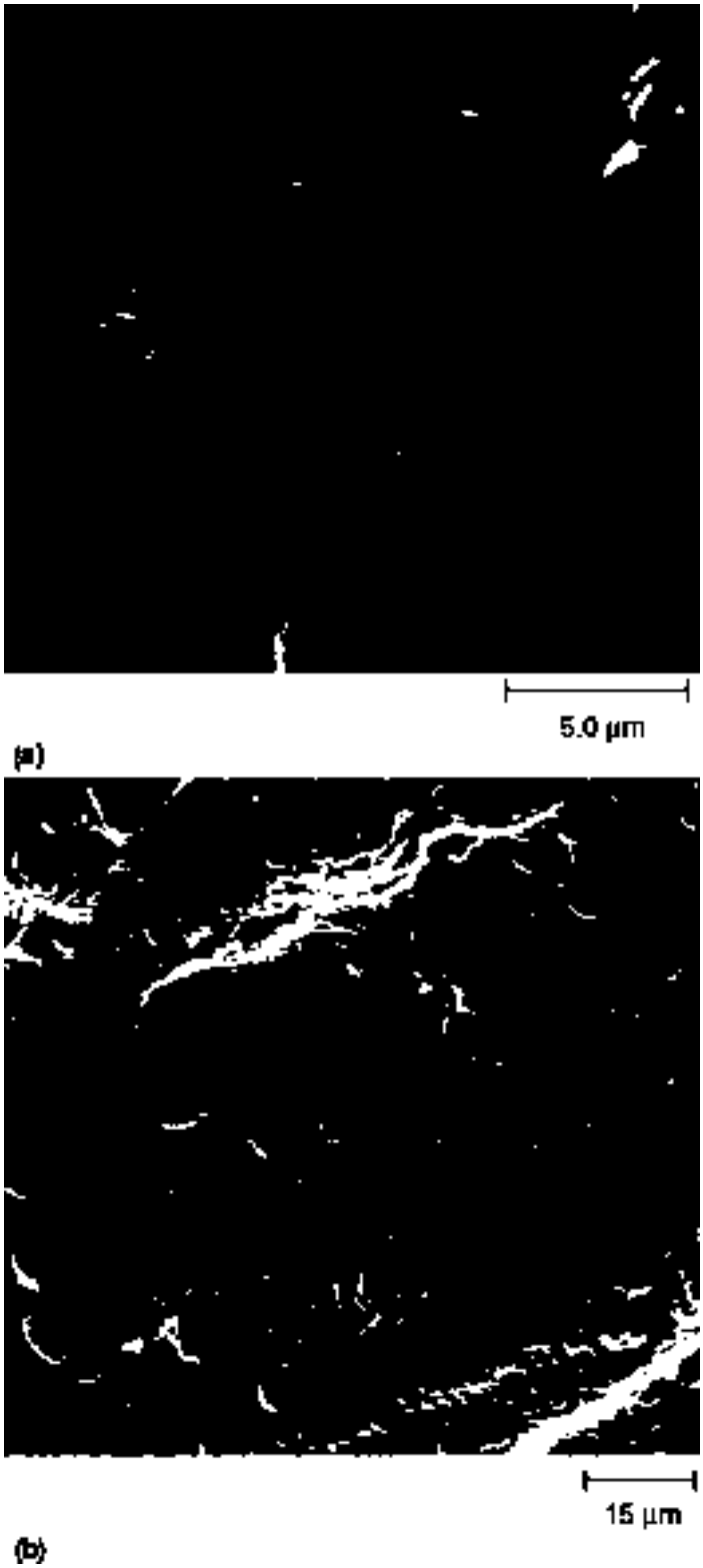


Fig. 33 Fatigue striations in 18-8 austenitic stainless steel tested in rotating bending. (a) Fine striations were located midway between origin and final overload fracture, while (b) coarse striations were located closer to the overload area. Overall direction of crack growth in these SEM views is from lower left to upper right.

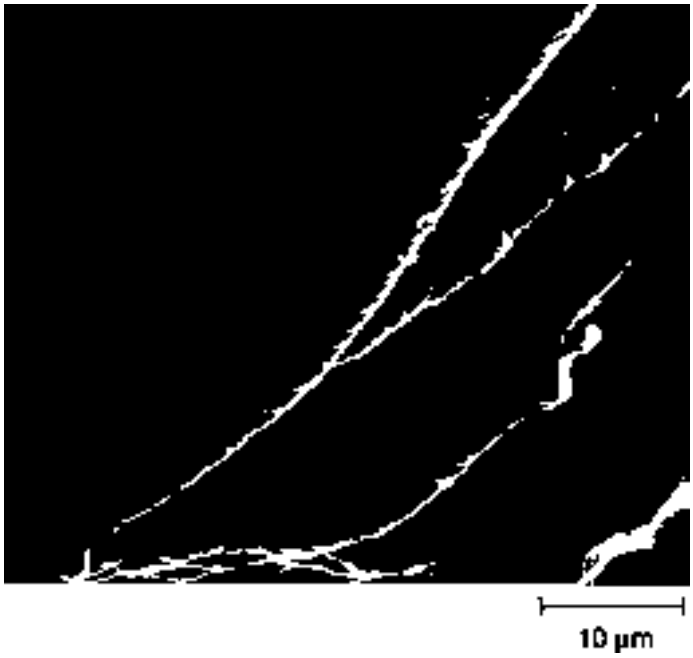


Fig. 34 SEM view of fatigue striations in aluminum forging tested under cyclic loading

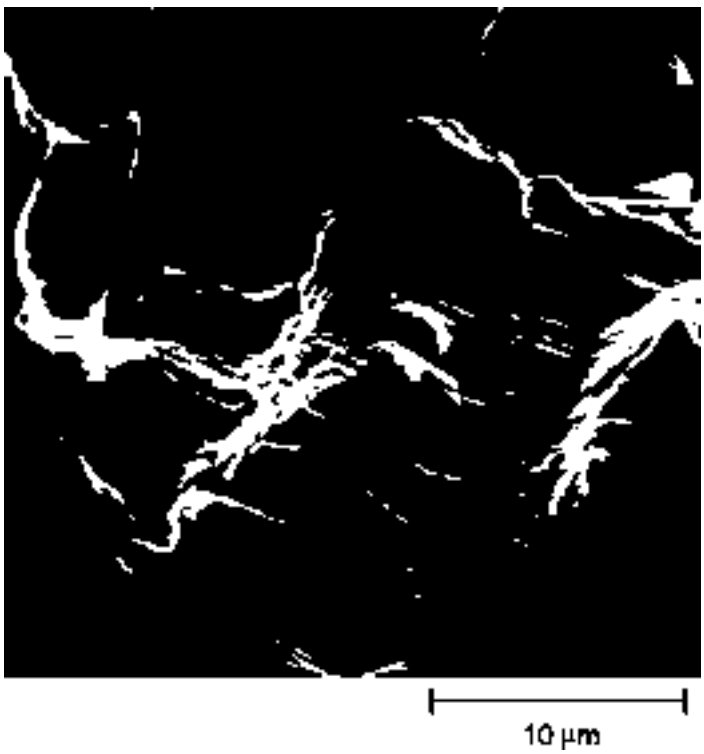


Fig. 35 Striations observed by SEM on rotating beam fatigue specimen made of cold-worked electrolytic tough pitch copper. Crack growth direction is from upper right to lower left, and specimen was tested at relatively high stress.

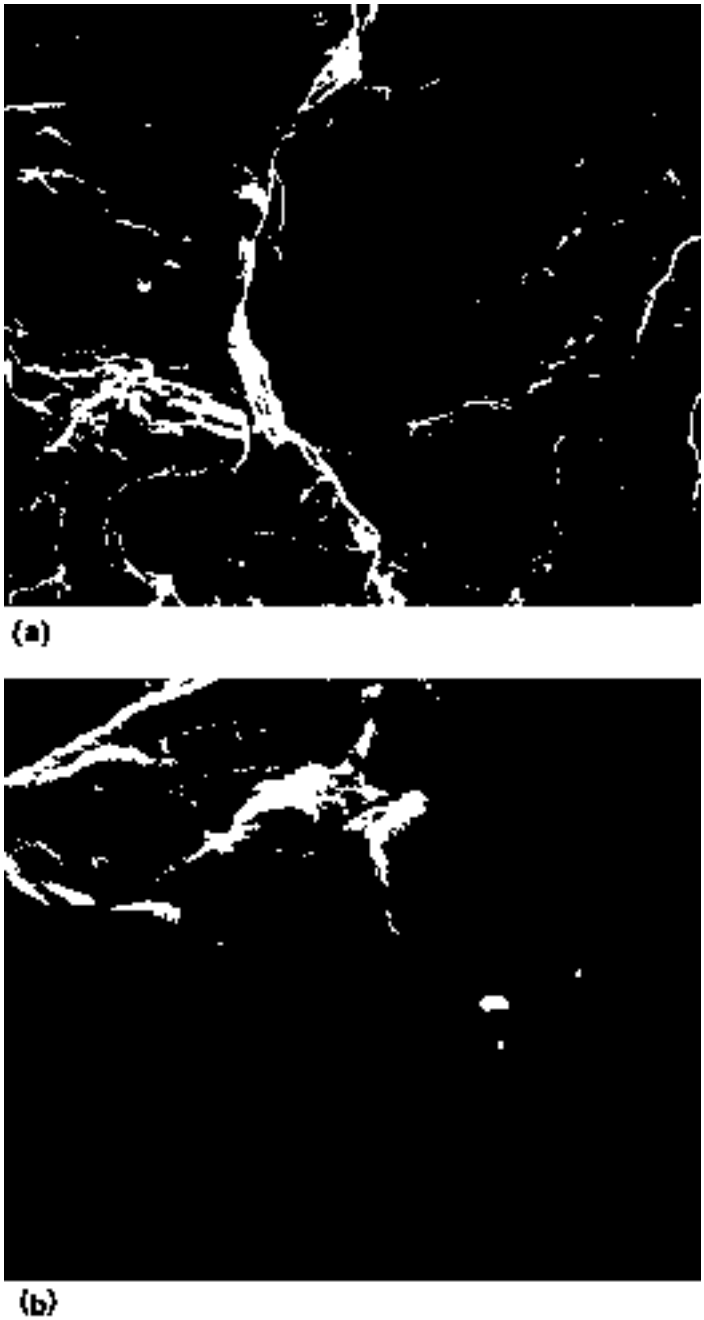


Fig. 36 (a) Coarse and (b) fine striated morphology observed on fatigue of a hot isostatically pressed titanium alloy. Source: Ref 7

Because each striation corresponds to one loading cycle, the fatigue crack growth rate can be estimated from the spacing between true fatigue striations. Even under controlled conditions, measured striation spacings within a region of macroscopically constant fatigue crack growth rate can vary severalfold. Multiple striation spacing measurements must therefore be made to obtain reliable sampling. For correlation with applied stress intensity or load, the location at which striation spacing is measured must be well established. Small loads may not produce resolvable striations. This must be considered in circumstances involving total cycle counts for variable amplitude loading.

The presence of true fatigue striations is unambiguous evidence of fatigue crack propagation. However, depending on material and loading conditions (as well as condition of the fracture surface), the absence of striations may not mean that fatigue did not occur. Not all metals form distinct striations in fatigue, and those that do may not form striations under all stress ranges. Examples of other microscopic fatigue morphologies are presented subsequently, followed by some examples of striationlike features that have been mistaken for true striations.

Ill-Defined Striations and Other Characteristic Appearances. Many common steels either exhibit ill-defined (unclear) striations under fatigue crack growth or have characteristic fatigue features without discernable striations. Examples of characteristic fatigue features in steels are presented in Fig. 37 and 38. See the article “Fracture Appearance and Mechanisms of Deformation and Fracture” for more details.

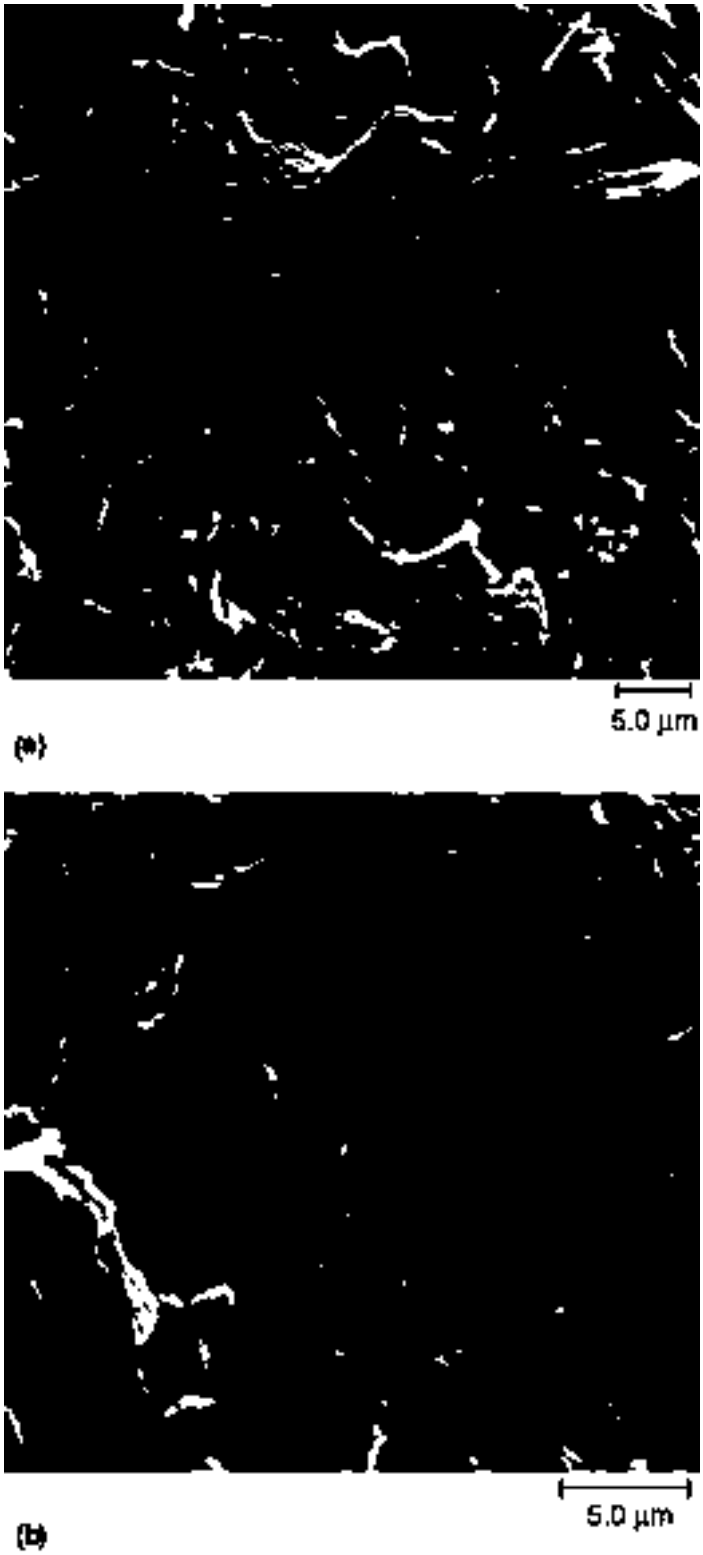


Fig. 37 Two views at different magnifications showing fatigue appearance characteristic of many steels. Striations are not resolved or are ill-defined. Quenched-and-tempered medium-carbon alloy steel tested in rotating bending, imaged using a SEM

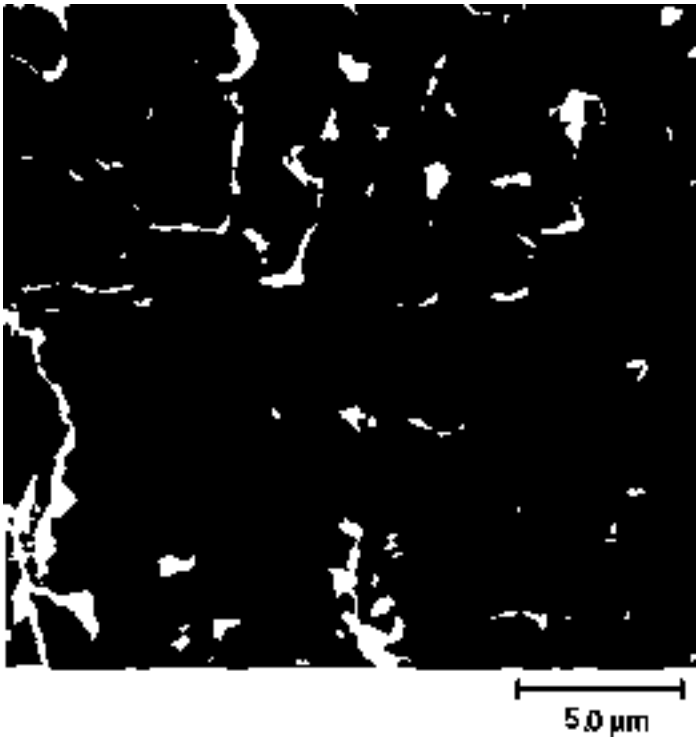


Fig. 38 SEM view of fatigue fracture surface of annealed medium-carbon alloy steel tested in rotating bending. No distinct fatigue striations could be resolved. Crack growth direction from right to left

Crystallographic Fatigue. Fatigue fracture of metals with low stacking fault energy, which favors planar slip, can occur over extended areas along crystallographic planes, producing a faceted appearance. For example, a cobalt-base alloy used in medical implants exhibits crystallographic fracture in laboratory fatigue testing at all levels of applied load (Ref 23) (Fig. 39). No striations were found. Many nickel-base superalloys can also fatigue with this faceted appearance over large regions.

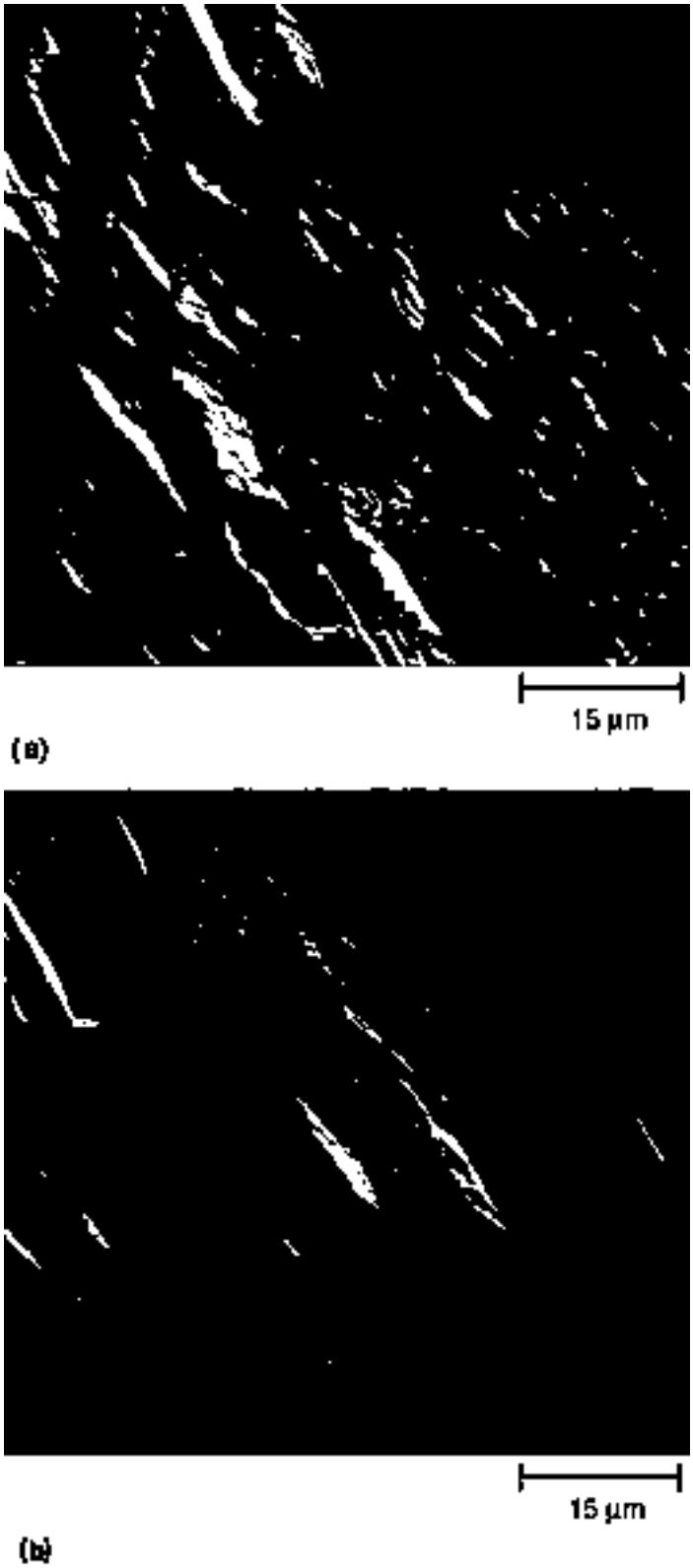


Fig. 39 Crystallographic fatigue in fracture mechanics specimen of cast Co-Cr-Mo-C medical implant alloy. SEM views located (a) 0.025 mm (0.001 in.) from machined notch tip, and (b) 7.6 mm (0.3 in.) from notch tip. Fatigue striations were not resolvable at any location, and the entire fatigue fracture surface displayed similar crystallographic features. Courtesy of D.O. Cox, Cox and Company

Metals such as austenitic stainless steel (Fig. 40) and aluminum (Fig. 41) fracture crystallographically under conditions of very low-stress high-cycle fatigue (i.e., propagation at low ΔK near the fatigue threshold). Flat facets can also appear at near-threshold fatigue crack growth rates in materials such as carbon steel (Fig. 42).

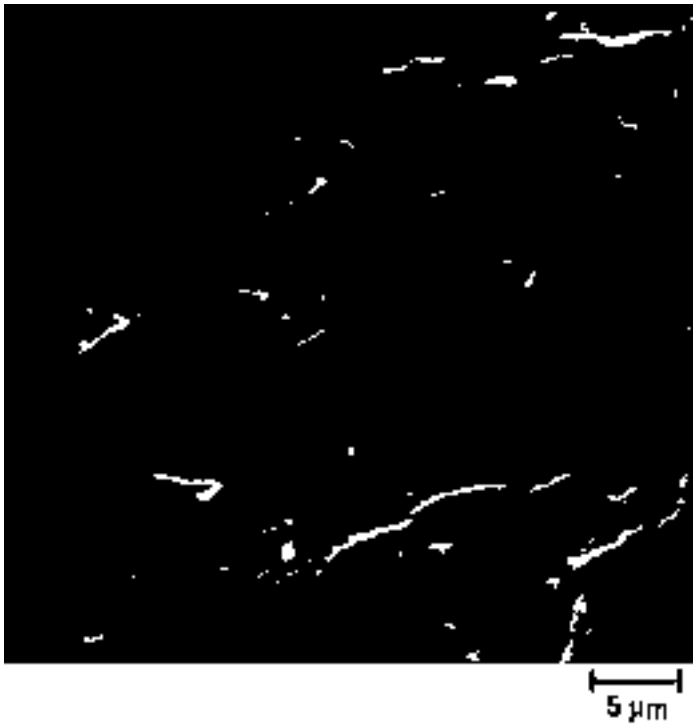


Fig. 40 Crystallographic fatigue of 18-8 austenitic stainless steel near fracture origin in rotating beam specimen. Global crack propagation direction from lower left to upper right in this SEM view

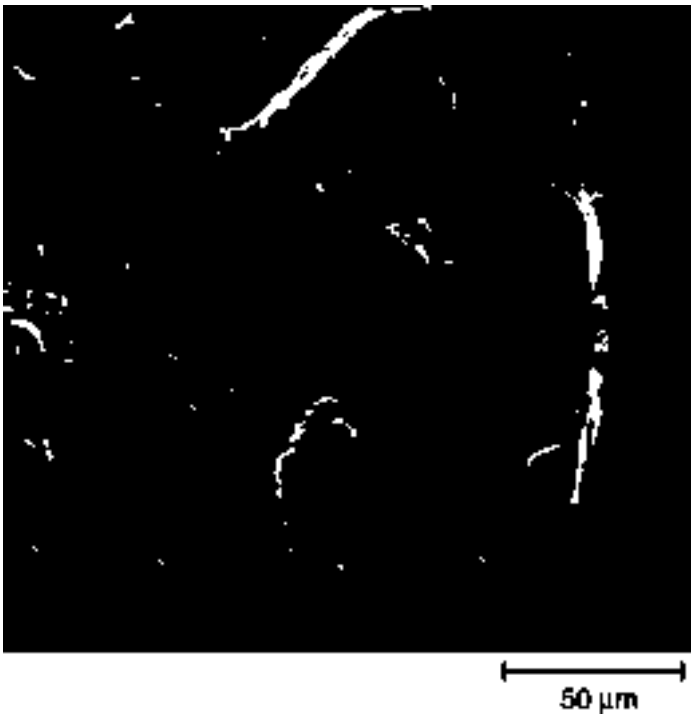


Fig. 41 Crystallographic fatigue of 6000-series aluminum extrusion near fracture origin in rotating beam specimen. Global crack propagation direction from bottom to top in this SEM view

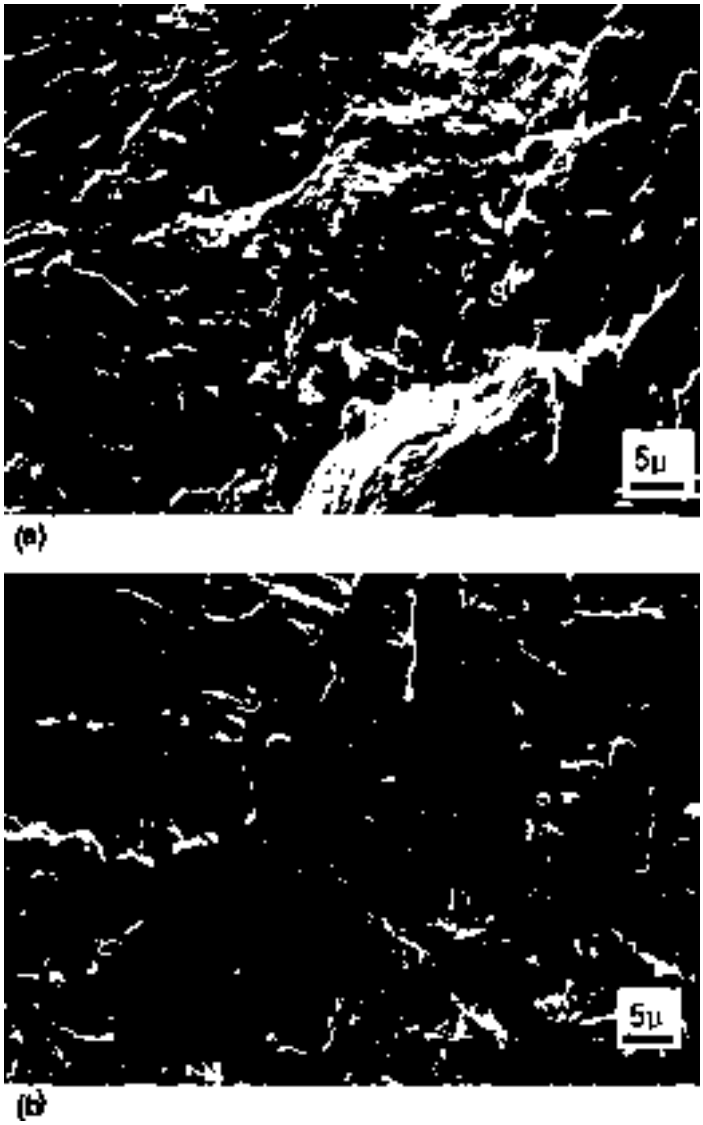


Fig. 42 Fully pearlitic steel fatigue fracture surfaces. Crack growth direction is from left to right in both images. (a) Intermediate crack growth rate (~0.1 μm/cycle), and (b) low crack growth rate (~0.001 μm/cycle). No fatigue striations were resolved by SEM at any crack growth rate. Unmarked facets are more prevalent at low (near-threshold) growth rates.

Intergranular Fatigue. Some metals can exhibit fatigue crack propagation along grain-boundaries or grain-boundary regions. For example, mixed intergranular/transgranular fracture of a nickel-copper alloy has been reported (Ref 24) (Fig. 43). Copper can also exhibit intergranular fracture in fatigue (Ref 25) (Fig. 44). Some alloys that exhibit transgranular fracture at intermediate fatigue crack growth rates, with more-or-less distinct striations, display regions of intergranular fracture at low (near-threshold) growth rates (e.g., quenched-and-tempered steel (Ref 26).

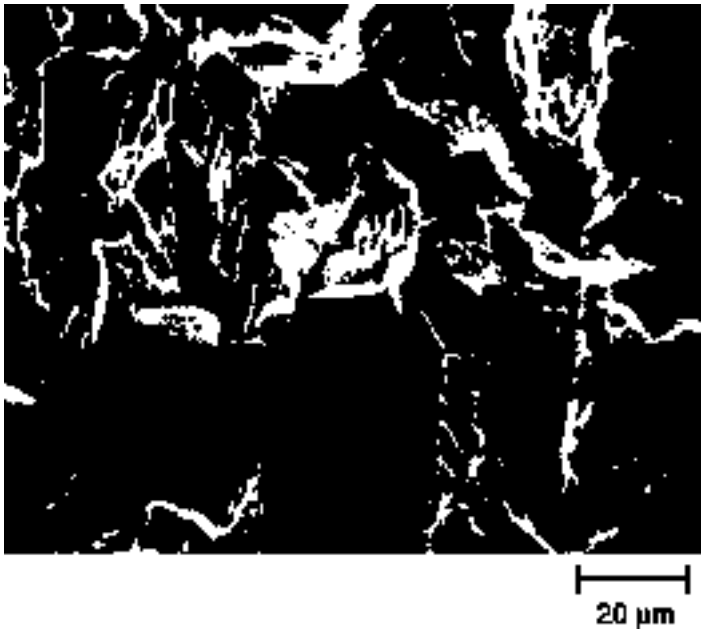


Fig. 43 SEM view of laboratory fatigue fracture of a 70-30 nickel-copper alloy showing mixed intergranular and transgranular morphology. Source: Ref 24

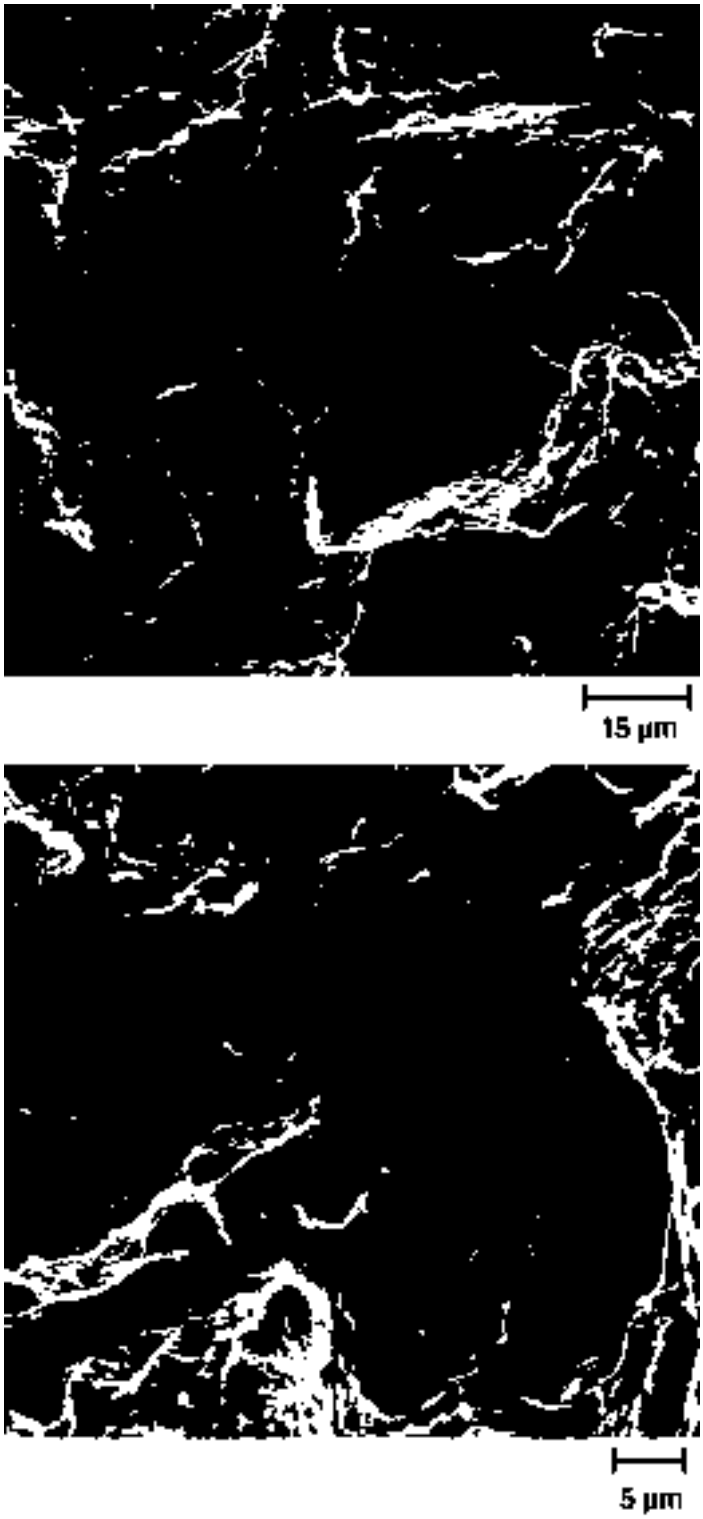


Fig. 44 SEM views of intergranular facets within fatigue crack propagation area of cold-worked electrolytic tough pitch copper tested in rotating bending at moderately low stress

Near-eutectoid steel with a narrow, continuous network of ferrite outlining prior austenite grain boundaries can exhibit intermittent discontinuous areas of fatigue fracture along the ferrite network (Ref 27). Fatigue fracture within these ferrite layers is reportedly flat and featureless.

Fatigue crack propagation in most structural alloys is typically transgranular. However, in some alloys environmental factors can contribute to intergranular fatigue. For example, corrosion of alloy steel fatigue tested in humid air can occur intergranularly.

High Growth Rates. As a fatigue crack approaches critical size, the growth rate generally increases. Monotonic overload features often appear on the progressive fracture surface. For example, isolated cleavage facets are present on pearlitic steel fatigue fracture surfaces formed at high growth rates (Fig. 45). Depending on alloy and condition, microvoids and intergranular fracture can also appear in high-growth-rate fatigue.

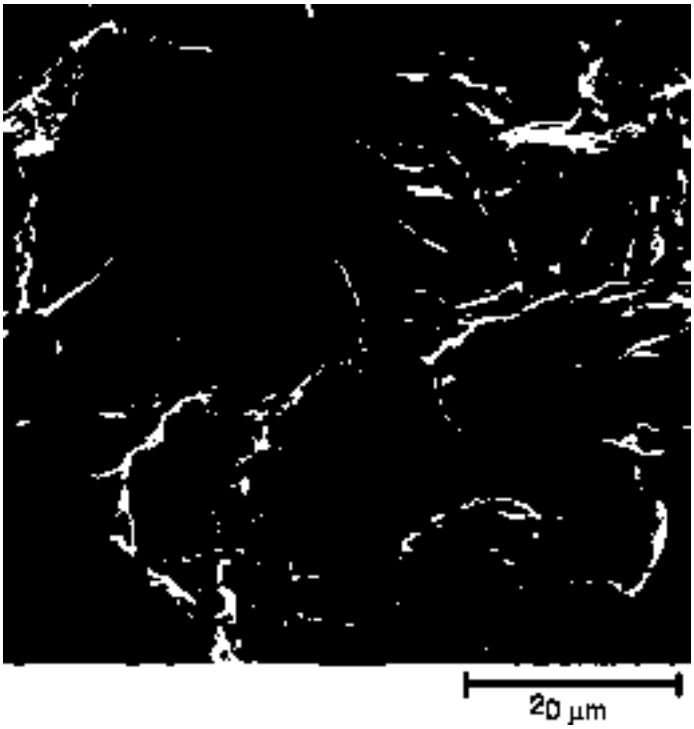


Fig. 45 Isolated cleavage facet within progressive high-growth-rate fatigue fracture of fully pearlitic steel, as viewed in the SEM

Microscopic Features Mistaken for Fatigue. Inexperienced or uninformed analysts too frequently interpret any set of roughly parallel markings observed in a SEM during examination of a fracture surface as fatigue striations. Features such as postfracture mechanical damage often exhibit fine, parallel marks or scratches (Fig. 46). Overload fracture of relatively tough and ductile materials can display fine, local markings produced by slip processes associated with monotonic loading. Such slip traces, known as serpentine glide, are illustrated in Fig. 47. Underlying microstructure can influence fracture paths. For example, fracture of pearlitic steel in fatigue or overload can contain patches of parallel marks formed by the influence of underlying pearlite lamellae.

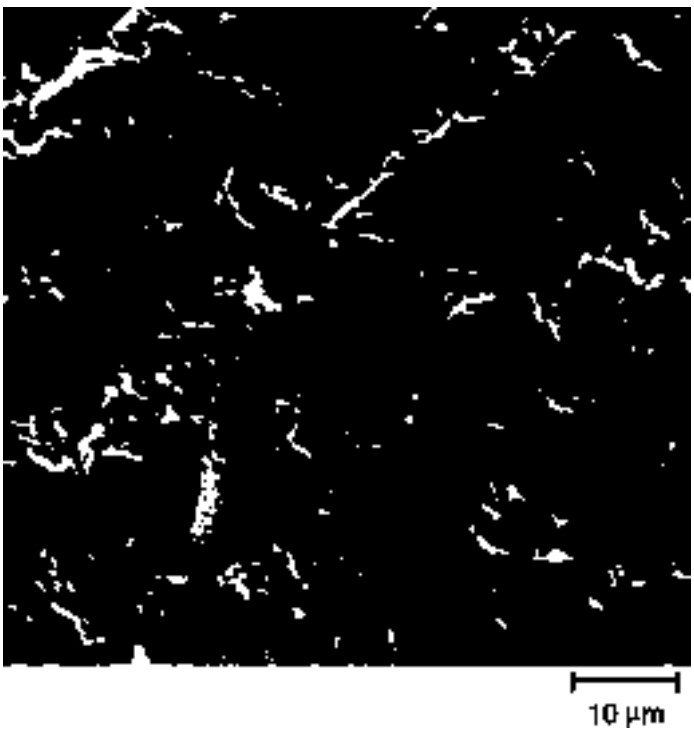


Fig. 46 Postfracture mechanical damage can exhibit fine, parallel markings, as illustrated in this SEM view of a steel sample.

Fig. 47 SEM view of serpentine glide on elevated-temperature (650 °C, or 1200 °F) laboratory tension test of nickel-molybdenum alloy. Although similar in appearance, these markings are not fatigue striations.

As noted previously, near-threshold fatigue cracking of austenitic stainless steel can be crystallographic. Transgranular chloride stress-corrosion cracking of austenitic stainless steel exhibits very similar features (Ref 28). Distinction between these fracture modes by fractography is difficult or impossible, and proper identification requires knowledge of service conditions.

References cited in this section

7. G.W. Powell, S. Cheng, and C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys*, Battelle Press, 1992

16. *Fractography and Atlas of Fractographs*, Vol 9, *Metals Handbook*, (8th ed.), American Society for Metals, 1974

22. D.O. Cox, R.A. Lund, and I. Roman, Fractography for Failure Analysis, *Pressure Vessel and Piping Technology—1985—A Decade of Progress*, C. Sundararajan, Ed., The American Society of Mechanical Engineers, 1985, p 1003–1012

23. D.O. Cox, "The Fatigue and Fracture Behavior of a Low Stacking Fault Energy Co-Cr-Mo-C Casting Alloy Used for Prosthetic Devices," Ph.D. dissertation, UCLA, 1977

24. R.A. Lund and G.J. Fowler, Intergranular Fracture in Engineering Materials, *International Symposium for Testing and Failure Analysis*, ISTFA, 1984, p 311–321

25. P.K. Liaw, T.R. Leax, R.S. Williams, and M.G. Peck, Near-Threshold Fatigue Crack Growth Behavior in Copper, *Metall. Trans. A*, Vol 13, Sept 1982, p 1607–1618

26. R.O. Ritchie, Near-Threshold Fatigue Crack Propagation in Ultra-High Strength Steel: Influence of Load Ratio and Cyclic Strength, *J. Eng. Mater. Technol. (Trans. ASME)*, Vol 99, July 1977, p 195–204

27. G.J. Fowler, "Fatigue Crack Initiation and Propagation in Pearlitic Rail Steels," Ph.D. dissertation, UCLA, 1976

28. T. Magnin and J. Lepinoux, Metallurgical Aspects of the Brittle S.C.C. in Austenitic Stainless Steels, *Parkins Symposium on Fundamental Aspects of Stress Corrosion Cracking*, S.M. Bruemmer et al., Ed., Minerals, Metals and Materials Society, 1992, p 323–339

Russell A. Lund, Consultant; Shahram Sheybany, Pacific Metallurgical Company

Fatigue of Polymers and Composites

Exemplar SEM fractographs of polymers are provided in Ref 9, and fractography of fatigue in engineering plastics is included in Ref 29. Macroscopic features of fatigue of structural polymers can parallel those of metals in many circumstances: relatively flat fracture, beach marks, and ratchet marks.

Because most polymers are not electrically conductive, they must be coated with a thin conductive layer, such as carbon or a metal, to allow examination using a standard SEM. The coating is applied by a method such as vacuum evaporation or sputtering. Coating must be performed carefully to avoid damaging the fracture surface (e.g., by excessive heating) or introducing artifacts associated with the coating process (Fig. 48). Scanning electron microscope fractographs of polymer fatigue are presented in Fig. 49, 50, 51.

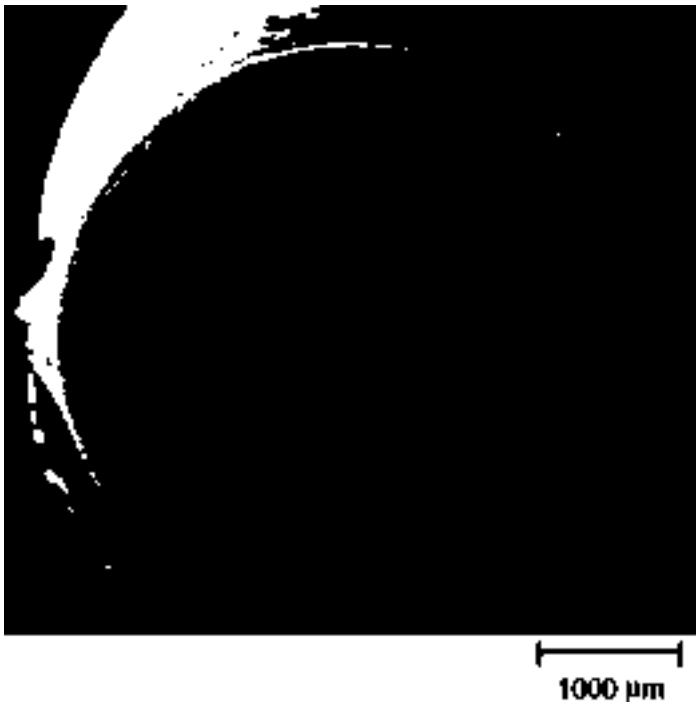


Fig. 48 Artifacts generated by improper platinum sputter coating of a 4.6 mm (0.18 in.) diameter polycarbonate rotating beam fatigue specimen. This SEM view shows a pattern in the coating reminiscent of "mud-cracking."

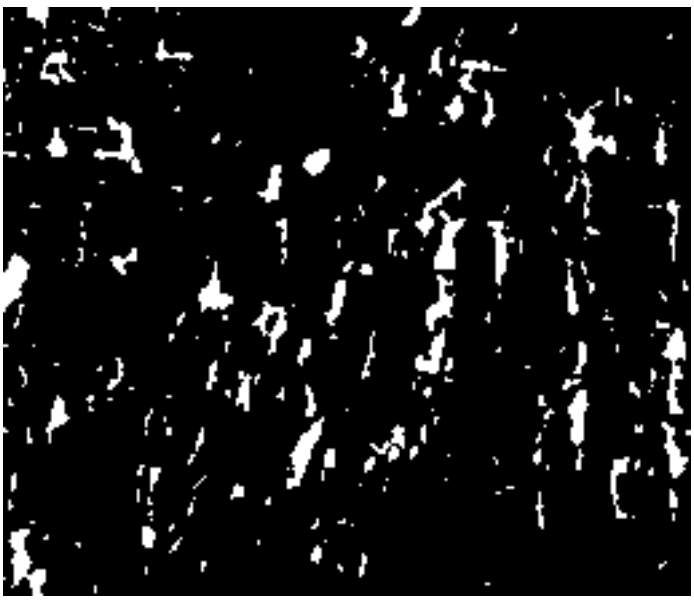


Fig. 49 SEM view of fatigue striations in medium-density polyethylene, laboratory tested at 0.5 Hz with maximum stress 30% of the yield strength. Crack growth is upward in this view. Original magnification 200×. Source: Ref 4

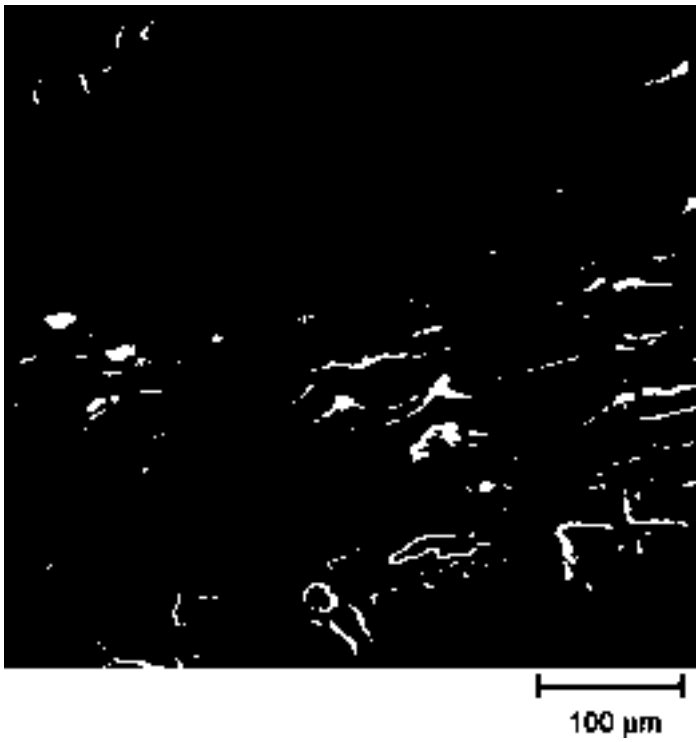


Fig. 50 Features observed on fatigue area of polymethyl methacrylate rotating beam specimen. Sample was sputter coated with platinum for SEM examination.

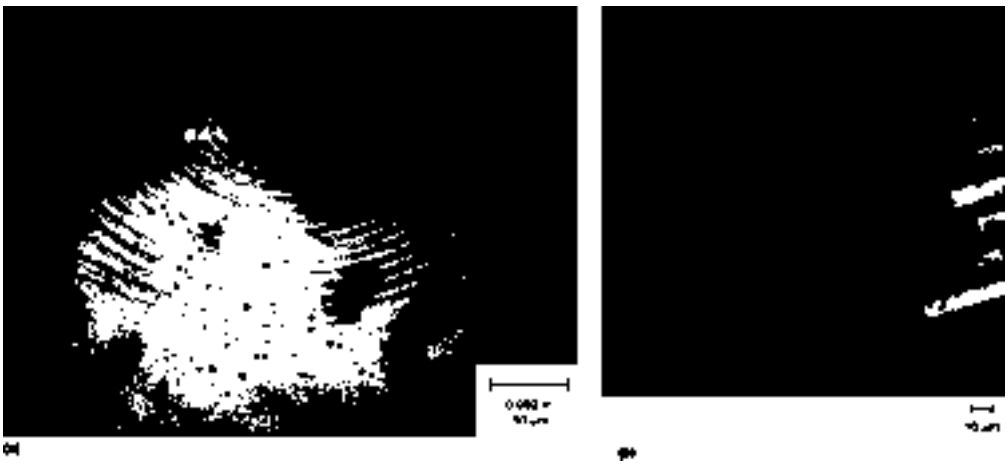


Fig. 51 Features observed on fatigue area of polycarbonate rotating beam specimen. (a) Optical view at base of notch. (b) Higher magnification electron fractograph. Sample was sputter coated with platinum for SEM examination.

The properties of polymers result in fatigue behavior different from that ordinarily encountered in metals. Except at elevated temperature or in corrosive environments, fatigue of structural metals depends on number (and magnitude) of loading cycles rather than time under load. However, because of time and rate sensitivity of many polymers at near-ambient temperatures, fatigue depends not only on the number of cycles at a given stress or stress-intensity level but also on frequency and time history of loading (Ref 30). Heat resulting from mechanical hysteresis during cyclic loading of plastics can cause thermal failure.

Additionally, depending on load level and time history, fatigue cracks in polymers may not propagate steadily (with an increment of growth for each fatigue cycle) but may grow in bursts or spurts (Ref 30). These spurts or discontinuous growth bands are associated with a large number of cycles (Ref 29). They can produce microscopic markings that appear very much like striations but which do not correspond to single load cycles. Consequently, without companion laboratory fatigue crack growth rate data and careful fractographic evaluation, estimation of service history from the spacing between fracture markings can be problematic (Ref 29). Some

polymers, such as polycarbonate and polymethyl methacrylate, can exhibit either true striations or discontinuous growth bands, depending on load levels and loading history.

A brief discussion of fractography of fibrous organic-matrix structural composites is provided in Ref 31. According to that article, fractography of fatigue failures in composite materials can be more difficult than for other materials, such as metals. There may be little macroscopic difference between interlaminar fracture features formed by fatigue and those formed in overload. Composite materials may lack visual indications of the initiation site. Although microscopic fatigue striations may form, areas exhibiting striations are usually isolated and limited in extent.

References cited in this section

4. *Fractography,* Vol 12, *Metals Handbook,* (9th ed.), ASM International, 1987

9. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage,* Prentice-Hall, 1981

29. R.W. Hertzberg and J.A. Manson, *Fatigue of Engineering Plastics,* Academic Press, 1980

30. G.C. Pulos and W.G. Knauss, Nonsteady Crack and Craze Behavior in PMMA Under Cyclical Loading, *Int. J. Fract.,* Vol 93, 1998, p 145–207

31. P.L. Stumpff, Fractography, *Composites,* Vol 21, *ASM Handbook,* ASM International, 2001, p 977–987

Fatigue Fracture Appearances

Russell A. Lund, Consultant; Shahram Sheybany, Pacific Metallurgical Company

References

1. R. Stephens, A. Fatemi, R. Stephens, and H. Fuchs, *Metal Fatigue in Engineering,* 2nd ed., John Wiley & Sons, 2001

2. J.A. Collins, *Failure of Materials in Mechanical Design; Analysis, Prediction and Prevention,* 2nd ed., John Wiley & Sons, 1993

3. S. Suresh, *Fatigue of Materials,* Cambridge University Press, 1991

4. *Fractography,* Vol 12, *Metals Handbook,* (9th ed.), ASM International, 1987

5. S. Bhattacharyya, V.E. Johnson, S. Agarwal, and M.A.H. Howes, Failure Analysis of Metallic Materials by Scanning Electron Microscopy, *IITRI Fracture Handbook,* IIT Research Institute, 1979

6. L. Engel and H. Klingele, *An Atlas of Metal Damage,* Prentice-Hall, 1981

7. G.W. Powell, S. Cheng, and C.E. Mobley, Jr., *A Fractography Atlas of Casting Alloys,* Battelle Press, 1992

8. J. Aleszka, Y. Kim, J. Scott, and A. Kumar, “Fracture Characteristics of Structural Steels: Reference Manual,” U.S. Army Construction Engineering Research Laboratory Technical Report M-258, 1979

9. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage,* Prentice-Hall, 1981

10. G. Parrish, *The Influence of Microstructure on the Properties of Case-Carburized Components,* American Society for Metals, 1980

11. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999

12. L.E. Alban, *Systematic Analysis of Gear Failures,* American Society for Metals, 1985

13. F.B. Stulen and W.C. Schulte, Fatigue Failure Analysis and Prevention, *Source Book in Failure Analysis,* American Society for Metals, 1974

14. J. Schijve, Fatigue Crack Growth Under Variable-Amplitude Loading, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 110–133

15. Fatigue Failures, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook* (9th ed.) American Society for Metals, 1986, p 102–135

16. *Fractography and Atlas of Fractographs,* Vol 9, *Metals Handbook,* (8th ed.), American Society for Metals, 1974

17. K.S. Ravichandran, Effect of Crack Shape on Crack Growth, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 159–167

18. D. McIntyre, Fractographic Analysis of Fatigue Failures, *J. Eng. Mater. Technol. (Trans. ASME),* Vol 97, July 1975, p 194–205

19. P.S. Pao and R.P. Wei, Corrosion-Fatigue Failures, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook* (9th ed.), American Society for Metals, 1986, p 252–262

20. L. Orr, Practical Analysis of Fractures in Glass Windows, *Materials Research and Standards,* Jan 1972, p 21–23

21. P.K. So, Fractography, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 805–810

22. D.O. Cox, R.A. Lund, and I. Roman, Fractography for Failure Analysis, *Pressure Vessel and Piping Technology—1985—A Decade of Progress,* C. Sundararajan, Ed., The American Society of Mechanical Engineers, 1985, p 1003–1012

23. D.O. Cox, "The Fatigue and Fracture Behavior of a Low Stacking Fault Energy Co-Cr-Mo-C Casting Alloy Used for Prosthetic Devices," Ph.D. dissertation, UCLA, 1977

24. R.A. Lund and G.J. Fowler, Intergranular Fracture in Engineering Materials, *International Symposium for Testing and Failure Analysis,* ISTFA, 1984, p 311–321

25. P.K. Liaw, T.R. Leax, R.S. Williams, and M.G. Peck, Near-Threshold Fatigue Crack Growth Behavior in Copper, *Metall. Trans. A,* Vol 13, Sept 1982, p 1607–1618

26. R.O. Ritchie, Near-Threshold Fatigue Crack Propagation in Ultra-High Strength Steel: Influence of Load Ratio and Cyclic Strength, *J. Eng. Mater. Technol. (Trans. ASME),* Vol 99, July 1977, p 195–204

27. G.J. Fowler, "Fatigue Crack Initiation and Propagation in Pearlitic Rail Steels," Ph.D. dissertation, UCLA, 1976

28. T. Magnin and J. Lepinoux, Metallurgical Aspects of the Brittle S.C.C. in Austenitic Stainless Steels, *Parkins Symposium on Fundamental Aspects of Stress Corrosion Cracking*, S.M. Bruemmer et al., Ed., Minerals, Metals and Materials Society, 1992, p 323–339

29. R.W. Hertzberg and J.A. Manson, *Fatigue of Engineering Plastics*, Academic Press, 1980

30. G.C. Pulos and W.G. Knauss, Nonsteady Crack and Craze Behavior in PMMA Under Cyclical Loading, *Int. J. Fract.*, Vol 93, 1998, p 145–207

31. P.L. Stumpff, Fractography, *Composites*, Vol 21, *ASM Handbook*, ASM International, 2001, p 977–987

Intergranular Fracture

S. Lampman, ASM International

Introduction

INTERGRANULAR FRACTURE is the decohesion that may occur along a weakened grain boundary. Typically, the grain boundaries in polycrystalline materials are stronger than individual grains in a properly processed material below its creep-regime temperature. The grain boundaries are disruptions between the crystal lattice of individual grains, and this disruption provides a source of strengthening by pinning the movement of dislocations (so long as the temperature is not elevated into the creep range for significant dislocation climb). Thus, a finer-grain alloy imparts more grain-boundary regions for improved strength. Moreover, because a greater number of arbitrarily aligned grains are achieved when grain size is reduced, the stressed material has more opportunity to allow slip and thus improve ductility. Thus, grain-size refinement provides a means for an increase of both strength and ductility (toughness).

However, the grain boundaries are also a region with many faults, dislocations, and voids. This relative atomic disarray of the grain boundaries, as compared to the more regular atomic arrangement of the grain interiors, provides an easy path for diffusion-related alterations. Grain boundaries are thus the preferential region for congregation and segregation of impurities. Weakening or embrittlement can also occur by preferential phase precipitation or absorption of environmental species in the grain boundaries. In addition, the grain-boundary regions are also weakened by high temperature, whereby the predominate fracture path becomes intergranular (although transgranular fractures have been observed in high-purity metals at rather high temperatures). Such observations in the early 20th century (Ref 1, 2, 3) led to the concept of the equicohesive temperature, for which the grains and grain boundaries exhibit equal strength. Above the equicohesive temperature, coarse-grain specimens exhibit greater strength than fine-grain specimens because of the lower grain-boundary surface area.

Intergranular (IG) fracture is easily distinguishable from cleavage or microvoid coalescence (Fig. 1). The fracture path of IG cracking is generally along the highest-angle grain boundary present either between grains or a particle/matrix boundary coincident with the grain boundary (Ref 4). When the grain structure is equiaxed, the fracture surface typically reveals a faceted “rock candy” fracture surface (Fig. 2a), which may be visible at low magnification (hand lens) if the grain size is coarse. In some cases, however, the intergranular features may not be as obvious. For example, intergranular fractures may have a relatively smooth, featureless fracture surface along elongated grains in heavily worked material, such as severely extruded alloys, or along the parting lines of some forgings (Fig. 2b). Grain facets may also be masked, such as by oxides, as is common of IG creep fracture surfaces.

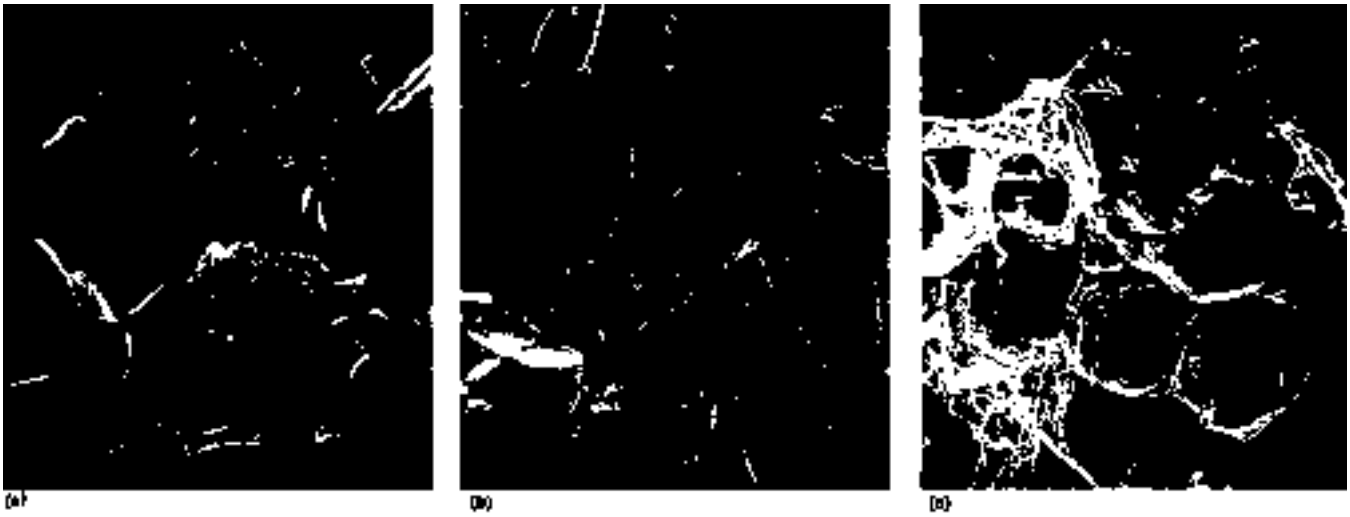


Fig. 1 SEM images of (a) IG fracture in ion-nitrided layer of ductile iron (ASTM 80-55-06), (b) transgranular fracture by cleavage in ductile iron (ASTM 80-55-06), and (c) ductile fracture with equiaxed dimples from microvoid coalescence around graphite nodules in a ductile iron (ASTM 65-40-10). Picture widths are approximately 0.2 mm (0.008 in.) from original magnifications of 500×. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

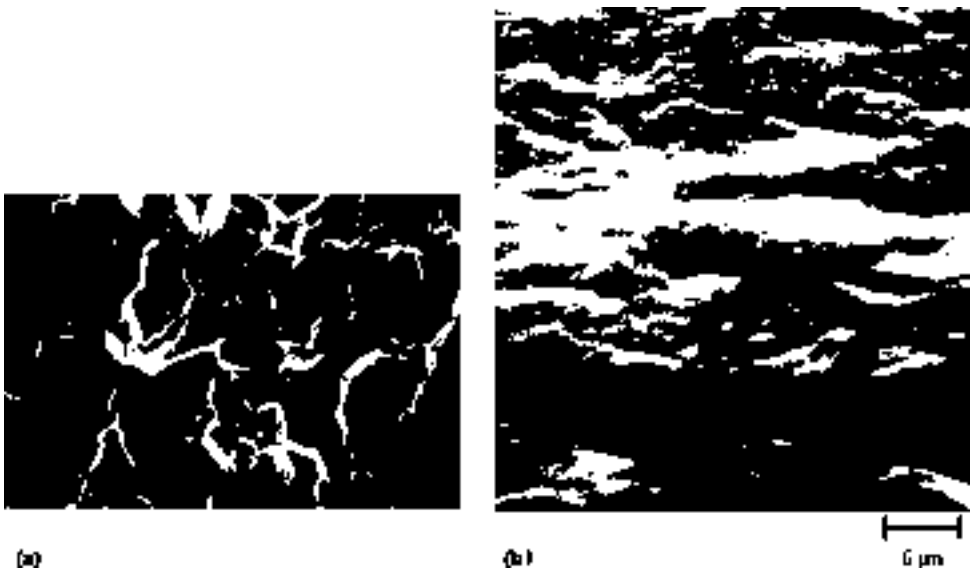


Fig. 2 SEM images of intergranular fracture with different grain morphologies. (a) Rock candy appearance from atmospheric stress-corrosion cracking of a high-strength aluminum alloy with equiaxed grains. 130×. (b) Intergranular fracture along the part line of an aluminum forging

This article briefly reviews the factors that influence the occurrence of IG fractures. Because the appearance of IG fractures is often very similar, the principal focus is placed on the various metallurgical or environmental factors that cause a weakening of the grain boundaries. In general, IG fractures are classified and termed as either:

- Dimpled IG fracture
- Intergranular brittle fracture
- Intergranular fatigue fracture

Dimpled IG fracture refers to the cases when microvoid coalescence occurs in (or perhaps adjacent to) grain boundaries from conditions such as creep cavitation or void nucleation at precipitates in the grain boundaries. The term *IG brittle fracture* is used when the grain surface does not have dimples indicative of microvoid coalescence.

In some cases, the term *decohesive rupture* has been used for IG fracture with dimples on the fracture surface. This use is perhaps historical in nature, based on the loss of grain-boundary cohesiveness at elevated temperatures. However, the term *dimpled IG fracture* is preferred, because it clearly conveys an image of the fracture surface (i.e., the fracture path and the topography of the fracture surface), and because it is described in the literature in very specific terms (e.g., Ref 5 and 6). Moreover, there is no reason for limiting the application of the term *decohesive rupture* only to separation at the grain boundaries, because it could be used for the case of decohesive separation along an active slip plane as well (Ref 7). When the term *decohesive rupture* is used in relation to IG fracture, it appears to be a general term referring to *any* macroscopic brittle fracture that does not occur by dimple rupture, cleavage, or fatigue (Ref 8). This broad use may not convey distinctions between specific features or mechanisms, and so this article attempts to use the term *decohesive separation* only in reference to specific mechanisms. The terms *dimpled IG fracture* and *IG brittle fracture*, as defined previously, are used to distinguish between ductile and brittle cracking mechanisms in the grain boundaries. This distinction, in some cases, may be helpful for appropriate evaluation of cause and corrective action.

References cited in this section

1. W. Rosenhain and D. Ewen, The Intercrystalline Cohesion of Metals, *J. Inst. Met.*, Vol 10, 1913, p 119–149
2. Z. Jeffries, The Amorphous Metal Hypothesis and Equicohesive Temperatures, *J. Am. Inst. Met.*, Vol 11 (No. 3), Dec 1917, p 300–324
3. Z. Jeffries, Effects of Temperature, Deformation, and Grain Size on the Mechanical Properties of Metals, *Trans. AIME*, Vol 60, 1919, p 474–576
4. C.L. Briant and S.K. Banerji, Intergranular Fracture in Steel: The Role of Grain-Boundary Composition, *Int. Met. Rev.*, No. 4., 1978, p 164–196
5. P.G. Shewmon, Grain Boundary Cracking, *Metall. Mater. Trans. B*, Vol 29, June 1998, p 509–518
6. S.P. Lynch, Mechanisms of Intergranular Fracture, *Mater. Sci. Forum*, Vol 46, 1989, p 1–24
7. G. Powell, private communication, 2002
8. V. Kerlins, Modes of Fracture, *Fractography*, Vol 12, *ASM Handbook*, ASM Interantional, 1987, p 112

Intergranular Fracture

S. Lampman, ASM International

Mechanisms of IG Fracture

On the atomic scale, crack growth occurs by any one or a combination of the following (Ref 6):

- Tensile separation of atoms (decohesion)
- Shear movement of atoms (dislocation egress or insertion)
- Removal or addition of atoms by dissolution or diffusion

All of these processes can occur preferentially along the grain boundary by various phenomena, such as (Ref 6):

- Segregation of embrittling elements to the grain boundaries
- More rapid diffusion of elements along grain boundaries than along grain interiors
- More rapid nucleation and growth of precipitates in grain boundaries than in grain interiors
- Greater adsorption of environmental species in the grain-boundary regions

In addition, another complex mechanism is quench cracking, which almost always occurs intergranularly. Quench cracking may be aggravated by the various mechanisms of grain-boundary weakening and grain size, but it also is heavily influenced by volumetric expansion during transformation hardening and the temperature extremes of quenching. Quench cracking is not addressed in this article.

Although it is generally easy to recognize IG fracture, identification of the cause of fracture is much more complex. The underlying mechanisms of IG fracture are much more varied than transgranular fractures. The mechanisms of transgranular fracture are fatigue, brittle cracking by cleavage, or ductile fracture by microvoid coalescence. In contrast, IG fracture is not a unique fracture process; it is just a preferential crack path where brittle, ductile, or fatigue cracking may occur due to the weakening of grain boundaries by various time-temperature factors, such as species transport or diffusion or thermal activation.

Moreover, synergistic effects of environmental conditions and sustained stress may be involved in IG cracking. The conditions under which a slowly growing crack may follow an IG path before final overload fracture can be influenced by various time- or temperature-dependent mechanisms, such as fatigue, stress-corrosion cracking (SCC), liquid-metal embrittlement, solid-metal embrittlement, hydrogen embrittlement, and creep. In addition, transgranular cracks also can, at some point, follow a grain-boundary path for short distances under certain circumstances. For example, during the process of cleavage (elastic) fracture, the fracture crosses a grain on a particular crystal plane, but it can progress further along grain boundaries that are weakened by segregates. (For more information on the mechanical and microstructural that influence crack path, see Table 3 in the article “Fracture Appearances and Mechanisms of Deformation and Fracture.”)

Reference cited in this section

6. S.P. Lynch, Mechanisms of Intergranular Fracture, *Mater. Sci. Forum,* Vol 46, 1989, p 1–24

Intergranular Fracture

S. Lampman, ASM International

Intergranular Brittle Cracking

Causes of IG brittle fracture include:

- Brittle second-phase particles and/or films in grain boundaries
- Fracture where no film is visible and, due to impurity, atom segregation at the grain boundary
- Environmentally induced embrittlement where there is neither a grain-boundary precipitate or solute segregation

Grain-boundary segregation of elements (such as oxygen, sulfur, phosphorus, selenium, arsenic, tin, antimony, and tellurium) is known to produce IG brittle fractures. Studies of the effects of such impurities in pure iron have been greatly aided by the development of Auger electron spectroscopy. In the case of brittle grain-boundary films, it is not necessary for the film to cover the grain boundaries completely; discontinuous films are sufficient. Some common examples of IG embrittlement by films or segregants include:

- Grain-boundary carbide films in steels
- Iron nitride grain-boundary films in nitrided steels
- Temper embrittlement of alloy steels by segregation of phosphorus, antimony, arsenic, or tin
- Grain-boundary carbide precipitation in austenitic stainless steels (sensitization)
- Embrittlement of molybdenum by oxygen, nitrogen, or carbon
- Embrittlement of copper by antimony

Grain-boundary strengthening is characteristic of IG fractures caused by embrittlement. Intergranular brittle fracture can usually be easily recognized, but determining the primary cause of the fracture may be difficult. Fractographic examinations can readily identify the presence of large fractions of second-phase particles on

grain boundaries. Unfortunately, the segregation of a layer a few atoms thick of some element or compound that produces IG fracture often cannot be detected by fractography.

Dimpled IG Fracture

Dimpled IG fracture results in low macroscopic ductility (i.e., the grains separate rather than deform) and a fracture surface that reveals microscopic dimples at higher magnifications (typically on the order of 1000 to 5000×). For example, a fracture surface from stress rupture of a nickel-base alloy is shown in Figure 3 at two levels of magnification. The higher-magnification image reveals a dimpled topology on the grain facets. Another example is shown in Fig. 4, which shows frequent and uniform void nucleation aided by the formation of methane bubbles at the grain boundaries. In addition, dimpled IG fracture can occur by void nucleation at precipitates in the grain boundaries. Dimpled IG fracture is aided by impurities that adsorb strongly on the grain boundary surface. It can also appear in the stress-relief cracking of chromium-molybdenum steels (Ref 5).

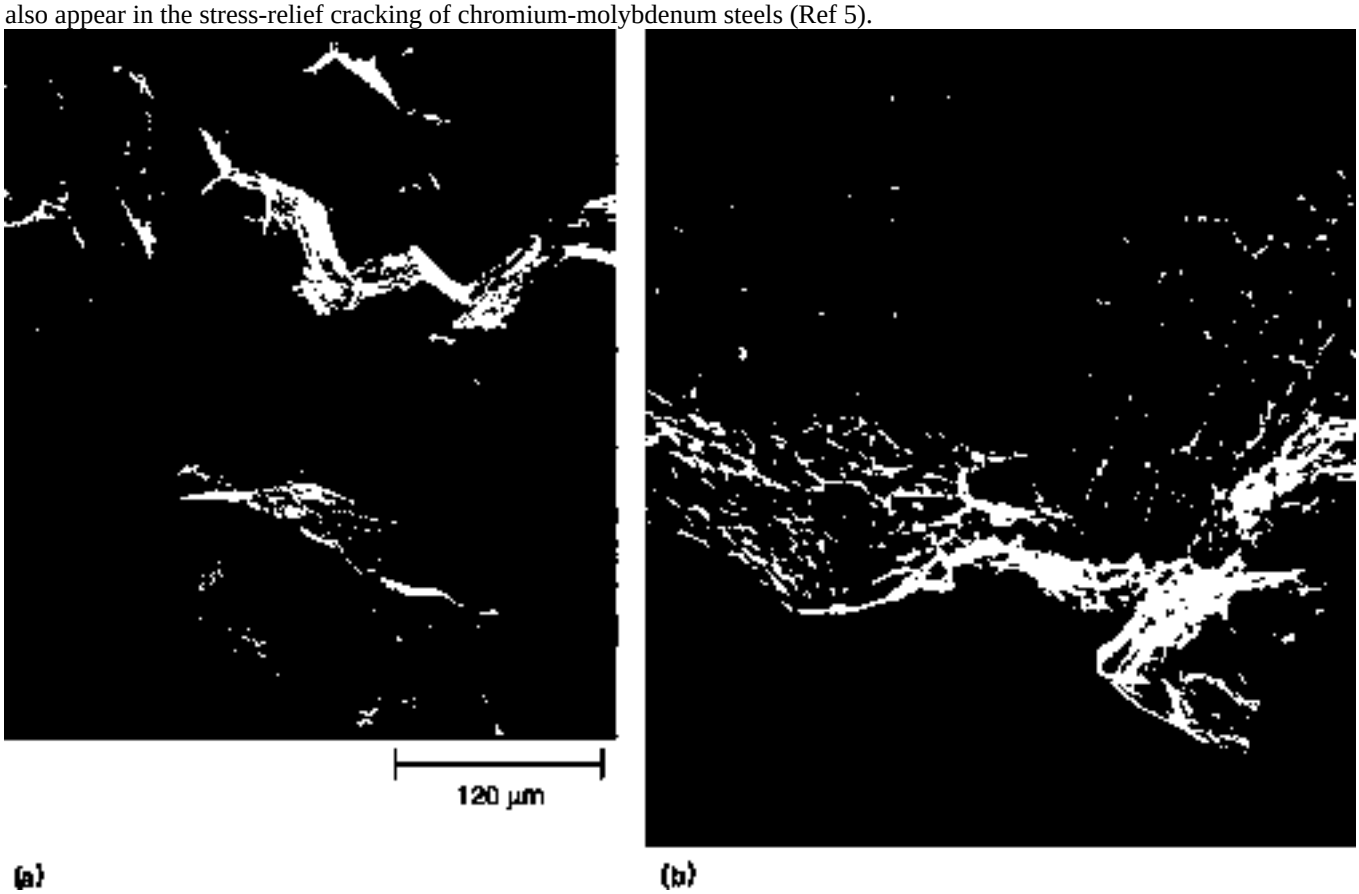


Fig. 3 SEM image of fracture surface of nickel-base alloy (Inconel 751, annealed and aged) after stress rupture (730 °C, or 1350 °F; 380 MPa, or 55 ksi; 125 h). (a) Low-magnification view, with picture width shown at approximately 0.35 mm (0.0138 in.) from original magnification of 250×. (b) High-magnification view, with picture width shown at approximately 0.1 mm (0.004 in.) from original magnification of 1000×. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

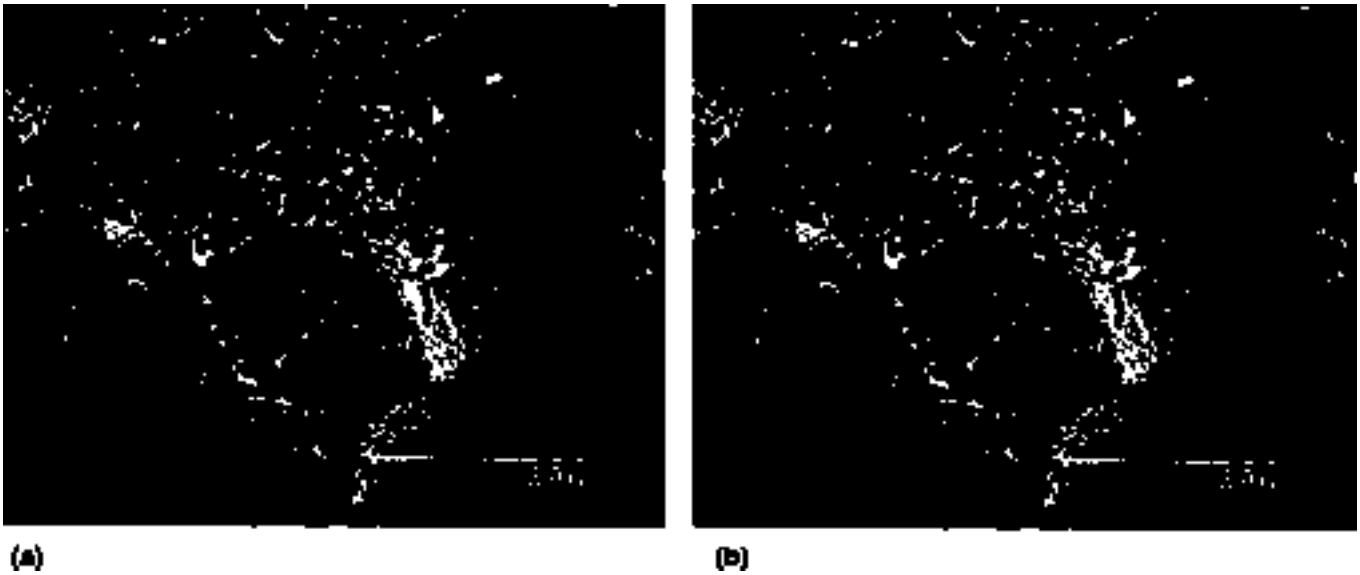


Fig. 4 Dimpled grain-boundary fracture in a small wedge-opening fracture sample, which aided formation of methane bubbles on the grains of 2.25 Cr-1.0 Mo steel exposed to high-pressure (21 MPa, or 3 ksi) hydrogen at 475 °C (887 °F). This is below the temperature where hydrogen attack would occur in service without an applied stress-intensity factor crack. (a) Low magnification. (b) High magnification. Source: L.-C. Chen and P. Shewmon, *Metall. Mater. Trans. A*, Vol 26, p 2317–2327

An example is shown in the scanning electron microscopy (SEM) image of Fig. 5 (Ref 9), which is reportedly from a high-purity aluminium-copper pH alloy with a coarse grain structure. In this example (with the coarse grain size), ductility is limited, and the yield strength in the local region of the grain boundary becomes lower than the matrix; thus, fracture tends to develop first within the grain-boundary zone by microvoid coalescence. The two levels of magnification provide a useful combination of images, one demonstrating the IG fracture path and the other revealing the microscopic mechanism of microvoid coalescence.

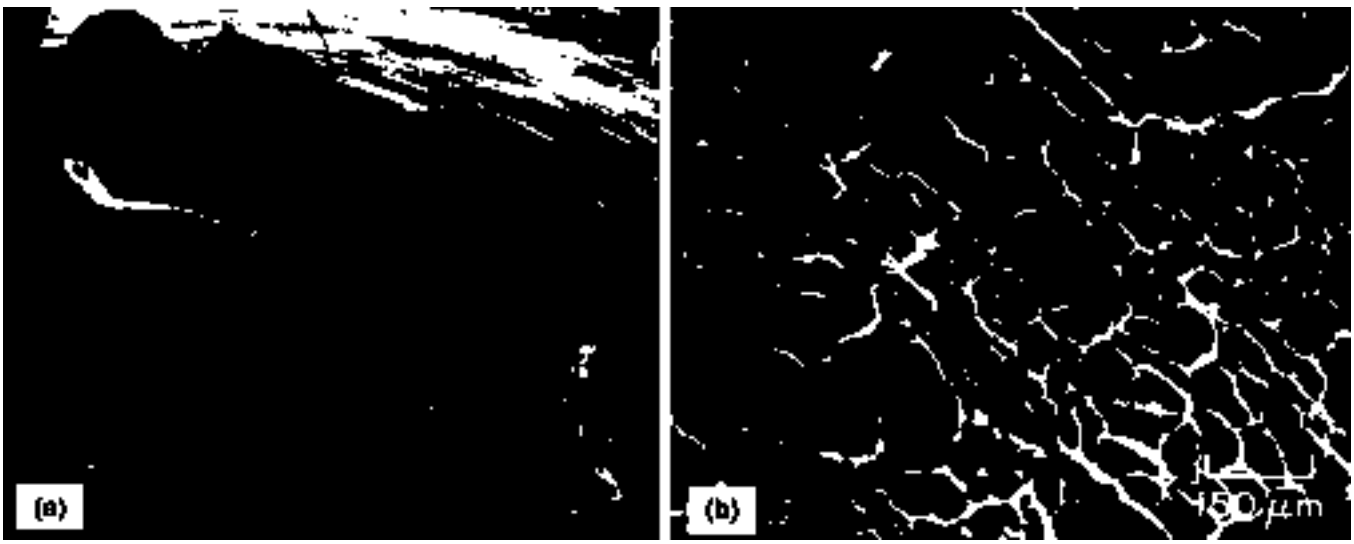


Fig. 5 Tensile test fracture surface of a high-purity, coarse-grained Al-4.2 Cu alloy with (a) IG facets at low magnification (10×) and (b) uniform dimples on one facet at higher magnification

Dimpled IG fracture is also observed in other precipitation-hardening alloys (such Al-Mg-Zn alloys) with large precipitates on grain boundaries and wide precipitation-free zones (PFZs). This fracture process was found to be unaffected by temperature for some aluminum alloys (Ref 10). Microscopically, the grain-boundary fracture may also reveal microvoid coalescence within the PFZ; that is, the alloy-depleted, denuded zone adjacent to the grain boundaries. The presence of the second phases may deplete the matrix adjacent to the grain boundary of solute content, lowering its strength. Nucleation of voids at the grain boundaries has also been

reported to occur in the absence of PFZs at the boundaries when a large fraction of particles (especially large ones) are present at the grain boundaries (Ref 6).

In many instances, SEM fractography provides a means to identify the fracture path as IG, but it may yield little other information. However, important clues on the underlying cause of IG separation may be revealed by fractography at assorted magnifications, because dimples on the IG fracture may be apparent at higher magnification. This may suggest mechanisms of microvoid coalescence, as noted. However, voids left by pulled-out second phases that initially failed by cleavage on the grain boundary might be mistaken at times for microvoid dimples. Additional important information can be obtained by chemical analysis in the SEM and by microstructural examination. Use of the SEM for microstructural examination in addition to optical light examination may be appropriate depending on the scale of microconstituents present.

References cited in this section

5. P.G. Shewmon, Grain Boundary Cracking, *Metall. Mater. Trans. B,* Vol 29, June 1998, p 509–518

6. S.P. Lynch, Mechanisms of Intergranular Fracture, *Mater. Sci. Forum,* Vol 46, 1989, p 1–24

9. *Fractography and Atlas of Fractographs,* Vol 9, *Metals Handbook,* 8th edition, 1974, p 112

10. S. Kuramoto, G. Itoh, and M. Kanno, Intergranular Fracture in Some Precipitation-Hardened Aluminum Alloys at Low Temperatures, *Metall. Mater. Trans. A,* Vol 27 (No. 10), Oct 1996, p 3081–3088

Intergranular Fracture

S. Lampman, ASM International

Intergranular Fatigue

Most fatigue and corrosion fatigue cracks are transgranular, although IG fatigue is not particularly uncommon in nonbenign environments or with improperly processed materials. For example, IG fatigue failure in austenitic stainless steels usually occurs only in the presence of grain-boundary embrittlement. Intergranular fracture has also been reported as a fracture mechanism in low-cycle fatigue, such as when IG oxidation initiates IG fatigue cracks (depending on temperature and the cyclic waveform and frequency).

On the microscopic level, crack nucleation in persistent slip bands and transgranular cracking during fatigue seem to be the rule, especially in smooth-section components (the zone of persistent slip bands is unlikely to be visible in materials with a crack-like imperfection) Transgranular cracking is encouraged by strain localization, which is promoted by reducing the grain size. In small-grain-sized metals, most of the applied plastic strain is concentrated in a few persistent slip bands, some of which develop into cracks with continued cycling. Use of coarse-grained structures causes strain homogenization and, therefore, IG fatigue cracking (Ref 11). Cyclic deformation of face-centered cubic (fcc) metals, even at low plastic strain amplitudes, leads to IG cracks from the action of persistent slip bands against grain boundaries at the surface (Ref 12). An example of IG fatigue in copper is also shown in Fig. 44 of the article “Fatigue Fracture Appearances” in this Volume.

Cyclic microvoid and cyclic IG processes (Ref 13) have also been observed. The former is shown in Fig. 6 for an ultrahigh-strength D6AC steel tested at a stress-intensity factor range (ΔK) greater than 60 MPa $\sqrt{m}$ (54.6 ksi $\sqrt{in}$). The crack advance per cycle is very rapid (Fig. 6a). With transmission electron microsopy (TEM) replicas, the microscopic growth process in Fig. 6(b) was shown to be elongated shear dimples. Apparently, few of these are nucleated by second-phase particles. It should be noted that this specimen is relatively thin, and that, for the high ΔK involved, fatigue growth is under plane-stress conditions. This contributes to the shear dimples noted in Fig. 6(b). At much lower ΔK values at both low temperature (Ref 14) and room temperature (Ref 15), IG ductile fatigue has been noted.

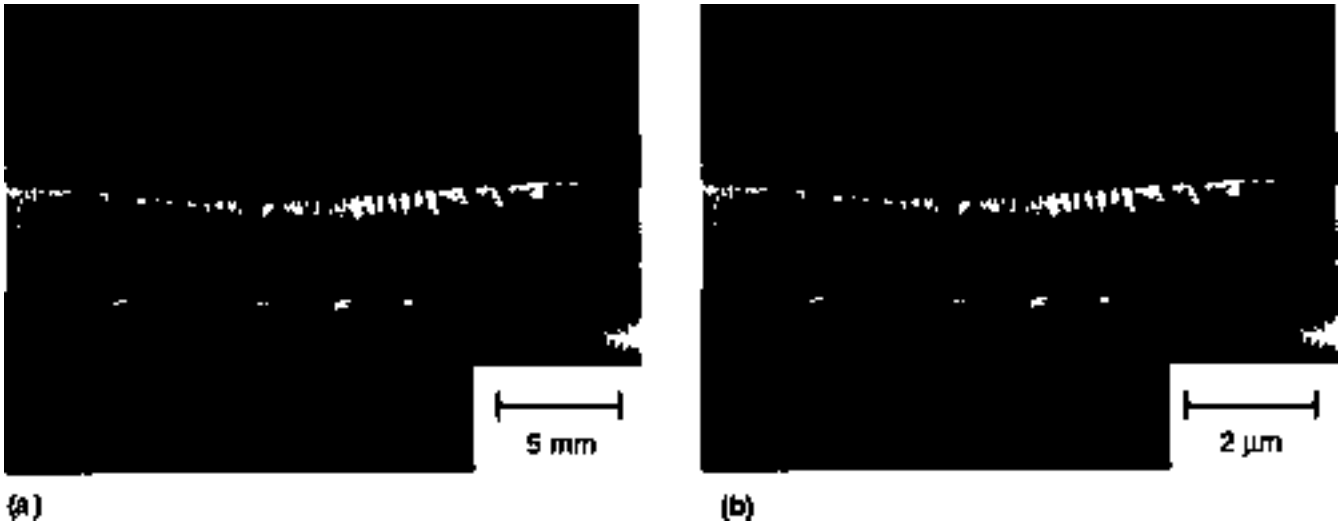


Fig. 6 High-cycle fatigue in D6AC steel at ΔK > 60 MPa $\sqrt{m}$ (54.6 ksi $\sqrt{in}$). (a) Macroscopic view of rapid fatigue growth under plane-stress conditions. (b) High-quality TEM replica showing elongated shear dimples from fatigue region

For the Fe-4Ni (at.%) in Fig. 7(a) the type of loading environment that produced this fracture surface is not obvious. It could be brittle IG cracking in a benign or aggressive environment. In fact, however, the higher magnification in Fig. 7(b) shows this to be a ductile IG fatigue process. This was determined not to be a moisture-related phenomenon because the partial pressure of water vapor at 123 K is negligible.

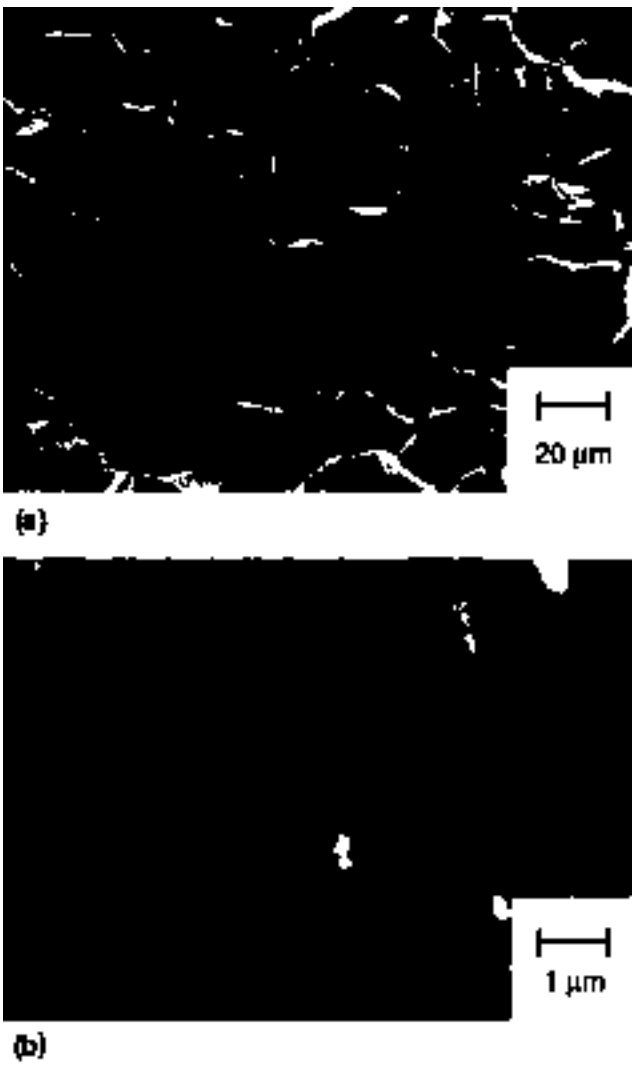


Fig. 7 Fracture surface topography of Fe-4Ni (at.%) fatigue cracked at 123 K with ΔK < 10 MPa$\sqrt{m}$ (9.1 ksi$\sqrt{in}$). (a) Overall IG appearance at low magnification. (b) Higher-magnification SEM resolution of ductile IG fatigue striations

Carburized Steels (Ref 16). Intergranular fatigue cracking can occur in the case of a carburized steel or in a material that has a high density of second-phase particles at the grain boundaries. Intergranular cracking at prior-austenite grain boundaries is an almost universal fracture mode for the high-carbon case of direct-quenched carburized steels (Ref 16). Auger electron spectroscopy shows that such IG fracture surfaces have higher concentrations of phosphorus and carbon, in the form of cementite. Thus, the brittle IG fracture that occurs in the stressed high-carbon-case microstructures of carburized steels is associated with the combined presence of segregated phosphorus and cementite at prior-austenite grain boundaries. The IG cracking may be associated with other microstructural features, such as the surface oxides generated by gas carburizing, but several studies have documented bending fatigue crack initiation by IG fracture even in the absence of surface oxidation (Ref 16). It may also be facilitated by the rather large prior austenite grain size resulting from the carburizing cycle.

Not only do the fatigue cracks initiate by IG cracking, but also the unstable crack propagates largely by IG fracture until it reaches the lower-carbon portion of the case, where ductile fracture by microvoid coalescence becomes the dominant fracture mode. In fact, sensitivity of the case microstructures to IG fracture makes possible the quantitative characterization of the size and shape of the stable fatigue crack, as shown in Fig. 8. The transition from the transgranular fracture of the stable crack to the largely IG fracture of the unstable fracture is identified by the dashed line in Fig. 8.

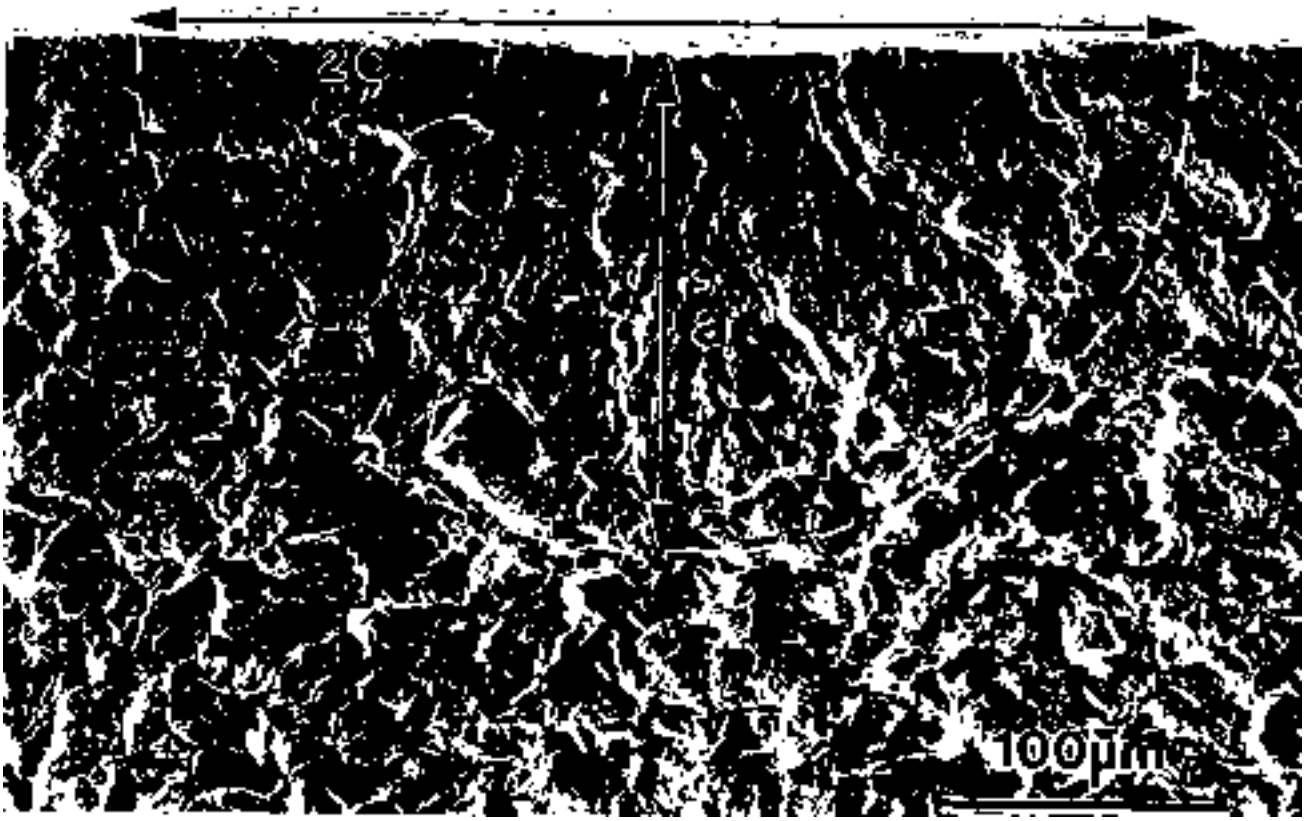


Fig. 8 Fatigue fracture in gas-carburized and modified 4320 steel showing overview of initiation, stable crack propagation, and unstable crack propagation, with extent of stable crack indicated by dashed line

References cited in this section

11. F.-L. Liang and C. Laird, Control of Intergranular Fatigue Cracking by Slip Homogeneity in Copper, Part I: Effect of Grain Size, *Mater. Sci. Eng. A,* Vol 117 (No. 1–2), Sept 1989, p 95–102

12. H.-J. Christ, On the Orientation of Cyclic-Slip-Induced Intergranular Fatigue Cracks in Face-Centered Cubic Metals, *Mater. Sci. Eng. A,* Vol 117 (No. 1–2), Sept 1989, p L25–L29

13. W.W. Gerberich, Microstructure and Fracture, *Mechanical Testing,* Vol 8, *Metals Handbook* (9th ed.), American Society for Metals, 1985, p 476

14. W.W. Gerberich and N.R. Moody, A Review of Fatigue Fracture Topology Effects on Threshold and Growth Mechanisms, “Fatigue Mechanisms,” STP 675, J.T. Fong, Ed., ASTM, 1979, p 292–346

15. W. Yu, K. Esaklul, and W.W. Gerberich, Fatigue Threshold Studies in Fe, Fe-Si and HSLA Steel, Part II: Thermally Activated Behavior of the Effective Stress Intensity at Threshold, *Metall. Trans. A*, Vol 15, 1984, p 889–900

16. G. Krauss, Bending Fatigue of Carburized Steels, *Fatigue and Fracture*, Vol 19, *ASM Handbook*, ASM International, 1996, p 683

Intergranular Fracture

S. Lampman, ASM International

Causes of IG Fracture

The fundamental causes for IG cracking are fairly limited in number, and IG fracture is often associated with improper processing of a material and/or some aggressive service environment, where the grain boundaries are embrittled and/or weakened. The general circumstances that have been known to induce IG cracking have been classified into four general categories (Ref 4):

- Presence of grain-boundary precipitates
- Thermal treatment or exposure that causes segregation of certain impurities to the grain boundaries without an observable second phase
- Stresses applied at elevated (creep-regime) temperatures
- Environmentally assisted alteration or weakening of the grain boundaries
- Additionally, a large grain size also plays a role in causing a change from TG to IG cracking and can enhance any of the above mechanisms

Environmentally assisted fracture may be fracture at elevated temperature, but it could also include interaction of the material with the environment. This would include hydrogen embrittlement, liquid-metal embrittlement, solid-metal embrittlement, oxidation or reduction potentially in the grain boundaries, radiation embrittlement, and SCC.
Because the appearance of IG fracture is relatively easy to recognize, and because the causes are fairly limited, the presence of IG fracture (especially in the region of crack initiation) is often helpful in narrowing the potential cause for failure. Except under conditions of creep stress rupture at elevated temperatures, IG fracture is not common in properly processed materials in a benign environment. Some specific situations of IG fracture include:

- High-carbon steels with a pearlitic microstructure
- Segregated phosphorus and cementite at prior-austenite grain boundaries in the high-carbon-case microstructures of carburized steels
- Stress-relief cracking
- Grain-boundary carbide films due to eutectoid divorcement in low-carbon steels
- Grain-boundary hypereutecoid cementite in carburized or hypereutectoid steels
- Iron nitride grain-boundary films in nitrided steels
- Temper embrittlement in heat treated steels due to segregation of phosphorus, antimony, arsenic, or tin
- Embrittlement of copper due to the precipitation of a high density of cuprous oxide particles at the grain boundaries
- Embrittlement of steel due to the precipitation of MnS particles at the grain boundaries as a result of overheating
- Grain-boundary carbide precipitation in stainless steels (sensitization)
- Improperly precipitation-hardened alloys, resulting in coarse grain-boundary precipitates and a denuded region (PFZ)
- Embrittlement of molybdenum by interstitials (carbon, nitrogen, oxygen)
- Embrittlement of copper by antimony
- Reduction of Cu_2O in tough pitch copper by hydrogen
- Hydrogen embrittlement by grain-boundary absorption of hydrogen
- Stress-corrosion cracking (sometimes)
- Liquid metal induced embrittlement (LMIE), for example, mercury in brass, lithium in 304 stainless steel
- Solid metal induced embrittlement (SMIE)

The following sections describe more detail of some typical embrittlement mechanisms that cause IG fracture of steels. Mechanisms of IG fracture in alloy steels are also described in Ref 17.
Intergranular Fracture of Steels due to Grain-Boundary Precipitates. This type of IG fracture is caused by the presence of well-defined second-phase precipitates at the grain boundaries. It may also be associated with precipitation of inclusions during initial solidification

or after partial solution and reprecipitation after annealing at a high temperature (where MnS precipitates, nitrides, and carbides form in the grain boundaries). These fracture mechanisms are more important in cases where temperature control is poor, as in hot forging.
The distribution of the grain-boundary phase is almost always particulate rather than continuous, although it may cover a large area of the grain-boundary region. The phase can usually be detected by careful metallography and can also be extracted or evaluated for more specific identification. In most cases where such particles cause IG fracture, small cavities form around the grain-boundary precipitates. The growth and eventual linking of these cavities cause the crack to proceed along the grain boundaries. In steels, three well-documented cases of IG fracture are described as follows (Ref 4):

- Precipitation of AlN phase on the prior-austenite grain boundaries when steel is slowly cooled after austentization at temperatures above 1300 °C (2370 °F), or slowly cooled after casting
- Thermal embrittlement of maraging steels by titanium carbonitrides
- Overheating of steels

Precipitation of AlN in aluminum-killed steels occurs along the austenite grain boundaries, and, when the steel is fractured above the ductile-brittle transition temperature, the crack path is usually along the prior-austenite grain boundaries. In as-cast specimens, ferrite films are generally observed at these grain boundaries. The incidence of such cracking has been shown to increase with increases in aluminum and nitrogen contents and with slower cooling rates after casting.
The minimum amount of aluminum nitride required to produce IG fracture is lower for alloy steels than for plain carbon steels (0.002 versus 0.004%) (Ref 18). Minimizing the nitrogen content, using the smallest possible amount of aluminum for deoxidation, and increasing the cooling rate after solidification are recommended, and it has been demonstrated that cracking can be prevented by deoxidizing with titanium or zirconium or by combined titanium and aluminum (Ref 18).
Thermal embrittlement of maraging steels occurs when they have been heated above 1095 °C (2000 °F) and are slowly cooled through, or held within, the temperature range of 980 to 815 °C (1800 to 1500 °F). The embrittlement is caused by the precipitation of TiC and/or Ti(C, N) on the austenite grain boundaries during cooling through, or holding within, the critical temperature range. The degree of embrittlement increases with time within the critical range. Increased levels of carbon and nitrogen render maraging steels more susceptible to thermal embrittlement. Auger analysis has shown that embrittlement begins with the diffusion of titanium, carbon, and nitrogen to the grain boundaries, and observation of TiC or Ti(C, N) precipitates represents an advanced stage of embrittlement.
The overheating of steels occurs when they are heated to excessively high temperatures prior to or during hot working. One of the most common grain-boundary precipitates in overheated steels is MnS. This usually requires an initial austenitization treatment above 1300 °C (2370 °F). Below this temperature, MnS does not dissolve to any appreciable extent, so that particles of MnS remain in the matrix.
Overheating may also cause incipient grain-boundary melting, a problem known as burning. Hot working after burning generally results in the tearing or rupture of the steel due to the liquid in the grain boundaries. Hot working after overheating generally does not result in cracking; if sufficient hot reduction occurs, the influence of overheating may be minor or negligible. If the degree of hot reduction is small, mechanical properties, chiefly toughness and ductility, are affected. However, it should also be noted that overheating may occur during the solution anneal of age hardening alloy systems. In this case, grain boundary porosity by melting of the eutectic constituent occurs and can result in significantly reduced ductility and toughness.
Impurity-Induced IG Fracture. The presence of certain tramp elements at the grain boundaries of an alloy is one of the major causes of IG fracture. These elements are believed to lower the cohesive strength of the boundaries and, at a given concentration (depending on yield strength, grain size, and microstructure), cause a change in the brittle fracture path from the cleavage plane to the grain boundaries.
In steels, the most common grain-boundary embrittlers are from the IV to VI groups in the periodic table (Table 1). Historically, research on this topic has been divided into two regimes for steels with very high yield strength from tempering for short times at low temperatures and lower-yield-strength steels from high-temperature tempering (at 600 to 700 °C, or 1110 to 1290 °F). In the latter, embrittlement occurs when the steel is subjected to an additional isothermal treatment near 500 °C (930 °F). This division does overlap, as described further in Ref 4, but it is used to describe impurity-induced embrittlement.

Table 1 Common grain-boundary embrittlers

Group	Elements
IVa	Si, Ge, Sn
Va	P, As, Sb
VIa	S, Se, Te

Liquid metal induced embrittlement is caused by adsorption of an embrittling material (liquid or sometimes vapor) with low solubility in the base material. Combinations of liquid metals that cause LMIE in base metals/alloys are as follows:

Base alloy	Liquid embrittlers
Aluminum	**Hg, Ga, Zn, In, Na**
Steel	**Hg, Ga, Cd, Zn, In, Li**
Copper	**Hg, In, Li, Na**
Magnesium	**Zn, In**
Titanium	**Hg, Cd**
Silver	**Hg, Ga**

References cited in this section

4. C.L. Briant and S.K. Banerji, Intergranular Fracture in Steel: The Role of Grain-Boundary Composition, *Int. Met. Rev.*, No. 4., 1978, p 164–196

17. C.J. McMahon, Jr., Mechanism of Intergranular Fracture in Alloy Steels, *Mater. Charact.*, Vol 26 (No. 4), June 1991, p 269–287

18. B.C. Woodfine and A.G. Quarrell, Effects of Al and N on the Occurrence of Intergranular Fracture in Steel Castings, *J. Iron Steel Inst.*, Vol 195, Aug 1960, p 409–414

Intergranular Fracture

S. Lampman, ASM International

Intergranular SCC and Hydrogen Embrittlement

Grain boundaries can become weakened or embrittled by a variety of environmental conditions, such as hydrogen embrittlement, SCC, and corrosion. If the process occurs by oxidation or dissolution, it is corrosion. If the cracking is exacerbated by corrosion and the simultaneous presence of an applied or residual tensile stress, then it is regarded as SCC. Other forms of environmentally induced embrittlement of the grain boundaries include neutron embrittlement, SMIE, and LMIE.

Intergranular fracture processes, whether hydrogen or stress-corrosion enhanced, are frequently observed in high-strength steels, aluminum alloys, and nickel-base alloys. The resistance to SCC or hydrogen embrittlement of single-phase materials or materials with dispersoids or precipitate phases for hardening is affected by heat treatment. In these materials, microstructure seems to play a secondary role in terms of how well hardening precipitates, dispersoids, or other second-phase interfaces “tie up” segregants. For example, for ultrapure high-strength steels, refining the grains (increasing the grain-boundary area) decreases the tramp element concentration in a given boundary. This refinement increases the threshold stress intensity in AISI 4340 steel (Ref 19).

Addition of some alloying elements may either enhance or diminish the susceptibility of a boundary to hydrogen embrittlement or SCC. For example, manganese added to AISI 4340 steel decreases resistance to hydrogen embrittlement. In a series of special heats, it was demonstrated that manganese additions reduce stress-intensity thresholds (K_{th}) (Ref 20). Increasing the manganese content also increases the frequency of IG fracture as opposed to ductile microvoid coalescence in heat treated steels, because manganese promotes segregation of phosphorus at prior-austenite grain boundaries (Ref 20).

It is well known that the combination of tramp elements and hydrogen lowers grain-boundary fracture resistance (Ref 19, 21). Tying up similar deleterious species at other interfaces may change the microstructural fracture mode. For example, 7075-T651, an Al-Mg-Zn-Cu alloy, fails by IG fracture. Evidence indicates that this may be a hydrogen embrittlement process along grain boundaries with excess magnesium at the primary boundaries (Ref 22). If this excess magnesium can be removed either by adding excessive zinc to promote matrix-hardening precipitates or by tying it up in dispersoids such as $Al_{12}Mg_2Cr$, then the SCC resistance may be improved. There is evidence that additions of dispersoid-forming elements, such as chromium, manganese, zirconium, titanium, and vanadium, tend to reduce susceptibility to slow crack growth, which may be due to changing grain-boundary composition through the dispersoids or through grain refinement.

Hydrogen embrittlement can be a complex phenomenon that may be difficult to identify conclusively. It can be enhanced by grain-boundary phases and may be difficult to distinguish and separate from the various grain-boundary phenomena that may occur and act synergistically, in some cases. Hydrogen also produces various forms of damage, such as nucleation and growth of grain-boundary bubbles at inclusions and precipitates.

The appearance of hydrogen embrittlement fractures is influenced by the strength of the material. As the strength level increases, the fracture path becomes more IG. Impurities also influence fracture mode. For example, at low impurity levels, hydrogen-induced cracking was shown to occur by cleavage at very high stress intensities (Ref 23). At high impurity levels (grain-boundary impurities), the fracture path was IG, and the stress intensity required for crack growth decreased. In steels, the fracture mode also can be influenced by the cooperative action of temper embrittlement and hydrogen embrittlement. A common fracture feature in hydrogen-embrittled low-strength steels is flat fracture regions (100 to 200 mm, or 4 to 8 in., in size) that are circular or elongated and centered around an inclusion or a cluster of inclusions (Ref 24). These zones are transgranular.

In austenitic AISI 304 or 304L stainless steels, hydrogen embrittlement can occur from hydrogen charging that partially transforms the austenite to ferrite or martensite (Ref 25). These phases can also be formed in 304 or 304L, and in other low-stability austenitic stainless steels, by cold working (uncharged) specimens at low temperatures. Tests of more stable austenitic grades, such as AISI 310, have produced conflicting results. The unstable 304 alloy exhibits transgranular cleavage fractures, and the stable 310 alloy exhibits ductile fractures, when tested in hydrogen gas environmental cells (Ref 26).

Hydrogen embrittlement of high-strength steels is typically characterized by IG fractures. A very small amount of hydrogen, often a few parts per million, can cause hydrogen embrittlement. As noted, hydrogen may promote cleavage fracture in some steels. The bcc metals are especially susceptible to hydrogen embrittlement, although austenitic stainless steels and other fcc metals and alloys can have their ductility reduced by hydrogen. As with ferrous alloys, hydrogen in nickel and its alloys may introduce IG, transgranular, or

quasicleavage cracking, and, although the macroscopic features appear to be brittle, on a microscopic scale there is a high degree of local plasticity, suggesting that hydrogen enhances flow at the crack tip.

For many years, environmental cracking of aluminum and its alloys was considered as SCC; however, testing in specific hydrogen environments has revealed the susceptibility of aluminum to hydrogen damage. Hydrogen damage in aluminum alloys may take the form of IG or transgranular cracking or blistering. Blistering is most often associated with the melting or heat treatment of aluminum, where reaction with water vapor produces hydrogen. Blistering due to hydrogen is frequently associated with grain-boundary precipitates or the formation of small voids. Blister formation in aluminum is different from that in ferrous alloys in that it is more common to form a multitude of near-surface voids that coalesce to produce a large blister.

Intergranular or transgranular cracking of aluminum by hydrogen embrittlement may be difficult to distinguish from mechanisms responsible for SCC. However, current understanding is based on variations of two basic theories: crack advance by anodic dissolution or hydrogen embrittlement. The controlling factors in these two SCC models are as follows:

Anodic dissolution (SCC) is characterized by:

- Grain-boundary precipitate size, spacing, and/or volume fraction
- Grain-boundary PFZ width, solute profile, or deformation mode
- Matrix precipitate size/distribution and deformation mode
- Oxide rupture and repassivation kinetics

Hydrogen embrittlement is characterized by:

- Hydrogen absorption leading to grain-boundary or transgranular decohesion
- Internal void formation via gas pressurization
- Enhanced plasticity (adsorption and absorption arguments exist)

Intergranular Stress-Corrosion Cracking (IGSCC). In IGSCC, corrosion products are frequently observed on the separated grain facets (Fig. 9). However, depending on the environment, cracks from hydrogen embrittlement may also reveal corrosion products on a fracture surface. Analysis of the cause of fracture in metal parts and components that have been exposed to corrosive environments also is often difficult because of interactions of fracture mechanisms or because fractures generated by different mechanisms have similar appearances.

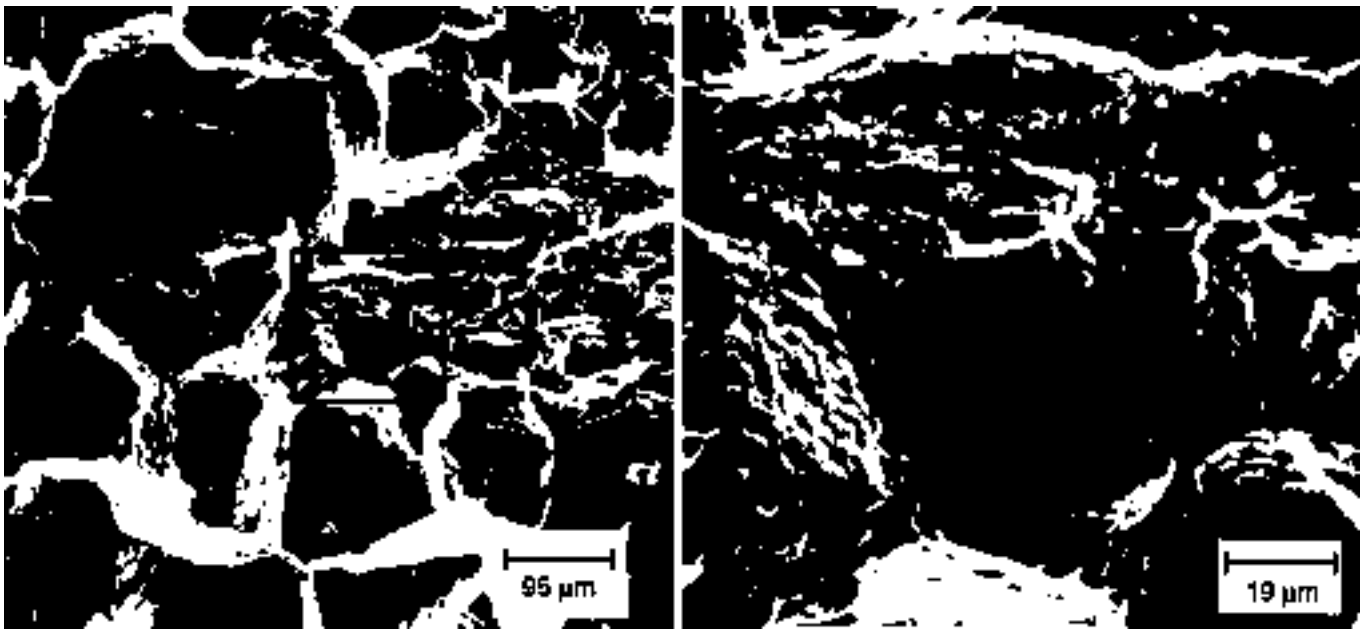


Fig. 9 Corrosion products on the grain facets from SCC of a U-700 turbine blade, presumably from combustion-gas attack that induced SCC, with IG and transgranular modes shown

Intergranular SCC is usually, but not exclusively, associated with alloys susceptible to IG corrosion (Ref 27), although IGSCC does occur in alloys not susceptible to IG corrosion, and vice versa. As with IG corrosion, IGSCC is generally related to one or more of the following conditions (Ref 27):

- Preferential penetration of a corrosion product, usually an oxide, along the grain boundary
- Presence of second-phase particles on grain boundaries
- Alloy depletion in the matrix adjacent to grain boundaries as a result of precipitation

- Impurity or alloy segregation in grain boundaries

The effects of material chemistry and microstructure on IGSCC can generally be divided into the following two categories: grain-boundary segregation and grain-boundary precipitation. Grain-boundary segregation of impurities such as phosphorus, sulfur, carbon, and silicon can be concentrated (perhaps up to 50%) within a 1 to 2 nm (10 to 20 Å) thick region of the grain boundary. These impurities can alter the corrosion and mechanical properties of the grain boundary and can therefore cause cracking by anodic dissolution and perhaps mechanical fracture. The extent of their effect depends on the electrochemical potential, the presence of other stress-corrosion processes such as chromium depletion, and their concentration in the grain boundary.

The IGSCC of austenitic stainless steel (Ref 28) is primarily dependent on the nature of the chromium-depleted zone, which is generally explained by the depletion of a passive film-forming element along a continuous path through the material. The stress-corrosion susceptibility and crack growth rate of austenitic stainless steel can be described by the degree of sensitization (DOS), as measured by corrosion tests such as the Strauss or electrochemical potentiokinetic reactivation tests. Quantitative comparisons between susceptibility, as measured by the presence or absence of IGSCC in a SCC test or the crack growth rate and the DOS or chromium depletion parameters, have been successful in cases in which sufficient data have been available, but these correlations are limited to specific alloys, environments, and stress conditions. Semiquantitative comparisons in which IGSCC is predicted for DOS values greater than some critical value are less costly to obtain and have been used more extensively.

Intergranular SCC in nickel-base alloys (Ref 28) is more complex than in austenitic stainless steel. Carbon solubility is considerably lower, and chromium diffusion is faster in nickel-base alloys than in austenitic stainless steels; therefore, the nose in the carbide precipitation or sensitization curves is shifted to shorter times and higher temperatures. Significant heat-to-heat variations in the chromium depletion of alloy 600 have been observed for a given heat treatment. This variation is attributed to the sensitivity of alloy 600 to carbon concentration and to the variability in mill-anneal heat treatment conditions.

Some significant differences between the IGSCC of nickel-base alloys and that of austenitic stainless steels include the positive effect of a semicontinuous distribution of carbides at the grain boundary in single-phase material and the galvanic couple between the γ' and γ phases in precipitation-hardened alloys. The beneficial effect of a semicontinuous distribution of carbides is thought to be related to dislocation generation from the carbides that results in a crack blunting or stress relaxation at the tip of a crack. The detrimental effect of the γ' phase is thought to result from a galvanic couple that polarizes the γ phase anodically relative to the γ' phase.

Intergranular SCC of aluminum alloys (Ref 28) in aqueous environments, including sodium chloride (NaCl) solutions can be a complex process that varies with alloy composition. However, grain-boundary precipitation has been identified as a contributing factor in the IGSCC of aluminum alloys; galvanic effects between the precipitates and matrix are considered to be important. In some cases, the precipitates are anodic, and in others, they are cathodic to the matrix. Peak-aged materials that give maximum hardnesses and microstructures in which slip bands are produced on straining are considered to be vulnerable to SCC. Overaged structures are considered to be less susceptible. The mechanism of crack growth is thought to be a combination of local anodic dissolution and hydrogen embrittlement. Whether crack extension is by anodic dissolution or by hydrogen embrittlement remains an open question; however, in aqueous solutions, the primary source of hydrogen is anodic dissolution.

Grain-boundary enrichment of magnesium without precipitation of a magnesium-rich phase has recently been identified as a factor in the IGSCC of aluminum alloys (Ref 22, 29). This enrichment is associated with increased corrosion activity at the grain boundary and the possible formation of magnesium hydride (MgH) or enhanced hydrogen entry along the grain boundaries. Crack advance is thought to occur by discontinuous crack jumps (possibly nucleated by hydrides) and crack arrest after approximately 150 to 350 nm (1500 to 3500 Å) of advance, followed by repetition of this process.

References cited in this section

19. C.J. McMahon, Jr., Intergranular Fracture in Steels, *Mater. Sci. Eng.*, Vol 25, 1976, p 233–239

20. M. Nakamura and E.-I. Furubayashi, Crack Propagation of High Strength Steels in a Gaseous Hydrogen Atmosphere, *Metall. Trans. A*, Vol 14, 1983, p 717–726

21. W.C. Johnson and D.F. Stein, A Study of Grain Boundary Segregants in Thermally Embrittled Maraging Steel, *Metall. Trans.*, Vol 5, 1974, p 549–554

22. E.C. Pow, W.W. Gerberich, and L.E. Toth, Analysis of Stress Corrosion Crack-Tip Surface Chemistries in 7075 Type Aluminum Alloys, *Scr. Metall.*, Vol 15, 1981, p 55

23. C.L. Briant et al., Embrittlement of a 5% Nickel High Strength Steel by Impurities and Their Effects on Hydrogen-Induced Cracking, *Metall. Trans. A*, Vol 9, May 1978, p 625–633

24. B.D. Craig, A Fracture Topographical Feature Characteristic of Hydrogen Embrittlement, *Corrosion*, Vol 37, Sept 1981, p 530–532

25. M.L. Holzworth and M.R. Louthan, Jr., Hydrogen-Induced Phase Transformations in Type 304L Stainless Steels, *Corrosion*, Vol 24, April 1968, p 110–124

26. N. Narita et al., Hydrogen-Related Phase Transformations in Austenitic Stainless Steels, *Metall. Trans. A,* Vol 13, Aug 1982, p 1355–1365

27. E.E. Stansbury and R.A. Buchanan, *Fundamentals of Electrochemical Corrosion,* ASM International, 2000, p 404

28. R.H. Jones and R.E. Ricker, Stress-Corrosion Cracking, *Corrosion,* Vol 13, *Metals Handbook* (9th ed.), ASM International, 1987, p 145–163

29. N.J.H. Holroyd and G.M. Scamans, *Scr. Metall.,* Vol 19, 1985, p 915

S. Lampman, ASM International

Example

Intergranular cracking can be a complex process where synergistic effects or multiple mechanisms may be involved, such as in the following example. In this case, a lowering of the operating temperature was proposed to reduce failures in a steam generator.

Example 1: Intergranular Fracture of Steam Generator Tubes (Ref 30). Power plants use steam generators, which have hot reactor coolant water flowing through the generators. One power plant experienced a steam generator tube rupture in the early 1990s. Metallurgical evaluation of the ruptured tube determined the failure mechanism was intergranular attack (IGA)/IGSCC, which occurred as a result of deposits formed on the tubes. This deposit, along with the consequential heat flux, created an aggressive environment, and, as a result, a long, deep crack initiated under the deposit under normal operating conditions.

The unit used two steam generators, which are vertical U-tube and shell heat exchangers, approximately 21 m (68 ft) high with a steam drum diameter of 6 m (20 ft). Hot reactor coolant from the reactor vessel enters the steam generators through a 107 cm (42 in.) inlet nozzle in the primary head. The reactor coolant flows through the U-tubes, where it gives up heat to the secondary coolant at the shell side, to the outlet side of the primary head, where the flow splits and leaves through two outlet nozzles. Saturated steam produced in the steam generators is passed to the turbine. Each steam generator contains 11,012 Inconel alloy 600 (nickel-base alloy) tubes, which are 19 mm ($\frac{3}{4}$ in.) outside diameter (OD) and have a nominal wall thickness of 1.0 mm (0.042 in.) and an average length of 18 m (57.75 ft). Prior to the implementation of the temperature-reduction program, the original operating temperature of the reactor coolant was 328 °C (621 °F).

To determine the mechanism of failure, a tube removal effort was conducted following the tube rupture event. A portion of the burst tube, shown in Fig. 10(a), along with seven other tubes, was removed for metallurgical analysis. Scanning electron microscope fractography of the burst tube showed evidence of IGSCC initiating at the OD, as shown in Fig. 10(b). Optical metallography revealed IGA under ridgelike deposits, shown in Fig. 10(c). Deposit analysis indicated a slightly alkaline to a very alkaline (caustic) environment. Environmental impurities consisting of sulfate, sulfide, sodium, lead, and copper were also detected. The levels of these impurities were not sufficient to indicate an acute introduction of these species into the steam generator secondary environment. However, crack oxide analysis indicated sulfate levels in excess of expected values. The literature indicated that sulfate and reduced sulfides in a caustic environment can increase the rate of IGA attack of Inconel 600 tube material. Therefore, the presence of sulfates and reduced sulfur on the crack surface contributed to the degree of IGA and IGSCC in the alkaline-to-caustic environment.

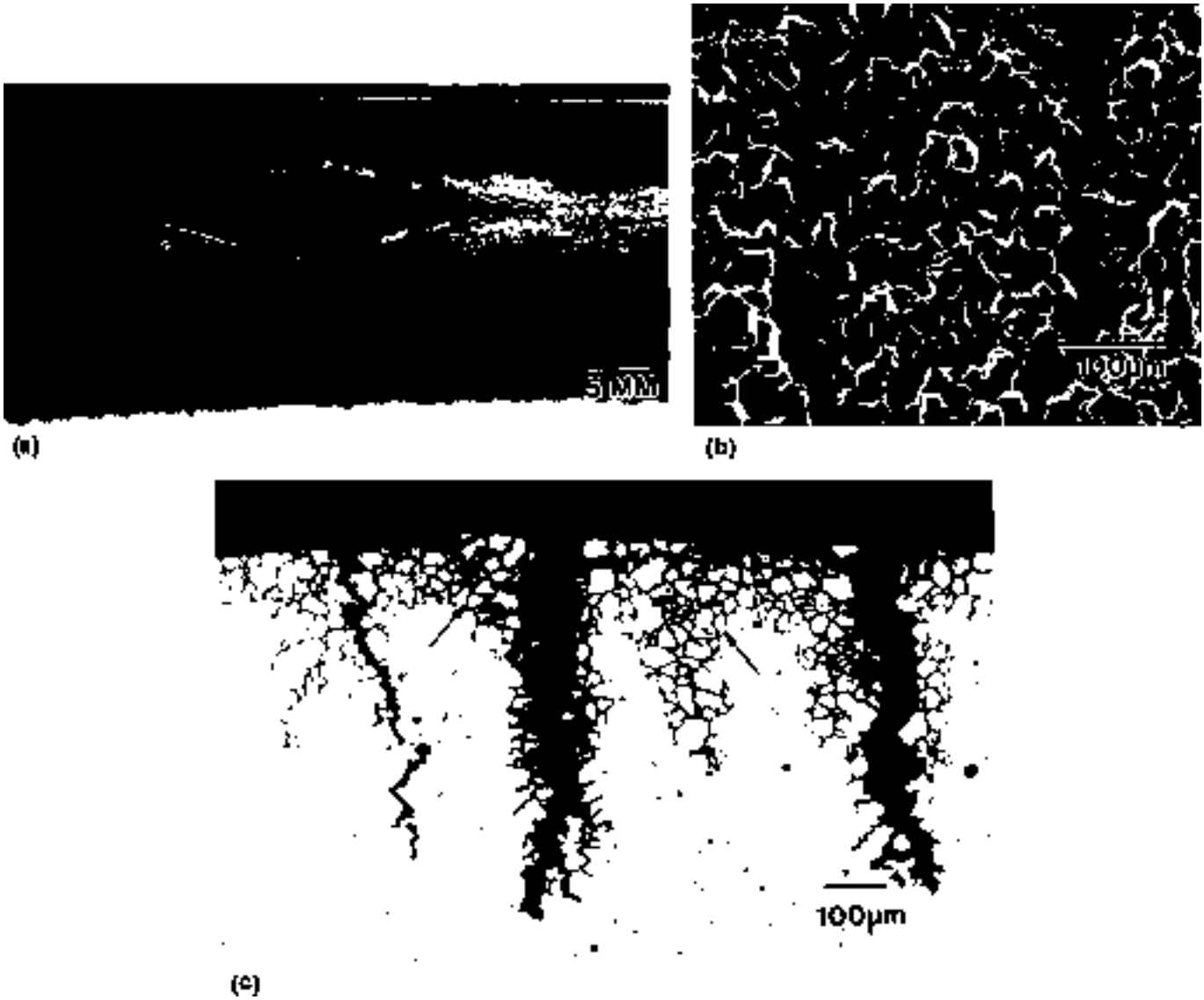


Fig. 10 Photographs of a ruptured Inconel 600 steam generator tube. (a) Tube rupture. (b) SEM fractograph showing the IG fracture surface. (c) Micrograph showing the IG attack that extended from the OD surface

It was concluded that the deposits formed at locations that experienced steam blanketing or dryout at the higher levels of the steam generators. Laboratory analysis determined these deposits contained higher-than-normal chemical contamination concentrations, suggesting that the deposits act as a "host" for chemical impurities. As a result, severe IGSCC and IGA occurred preferentially under these deposits.

Several remedial measures were developed to mitigate the effects of the conditions that contribute to IGA/IGSCC. These include steam generator water-chemistry controls, chemical cleaning, and reduction of the primary reactor coolant system (RCS) temperature (T_{hot} reduction). Of these, a reduction of the primary RCS temperature was determined to have the greatest impact. Based on industry literature, the reduction in primary RCS temperatures was modeled by the plant as an effective means of reducing the rate of attack for corrosion-related degradation of the Inconel 600 tubes. This was concluded because SCC is a thermally activated process, and temperature effects are evaluated in terms of activation energy.

The RCS temperature-reduction program was implemented for all three units at the plant. Primary temperature was reduced from 328 to 321 °C (621 to 611 °F). The decision to reduce the temperature to 321 °C (611 °F) was also based on minimizing the effect of reduced temperature on generating capacity, while maximizing the benefit in reducing tubing degradation. In addition, reducing the primary side temperature also increased the hot side recirculation ratio and reduced the dryout region in the steam generators where deposits form.

Reference cited in this section

30. D. Benac, Elevated-Temperature Failures, Materials Engineering Institute, Lesson 7, *Principles of Failure Analysis*, ASM International, to be published

Acknowledgments

Portions of this article are adapted from:

- G. Krauss, Bending Fatigue of Carburized Steels, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 680–690
- W. Gerberich, Microstructure and Fracture, *Mechanical Testing,* Vol 8, *Metals Handbook,* 9th edition, American Society for Metals, 1985, p 476–491
- R.H. Jones and R.E Ricker, Stress-Corrosion Cracking, *Corrosion,* Vol 13, *Metals Handbook,* 9th edition, ASM International, 1987, p 145–163

The author would also like to thank William Becker and Gordon Powell for their review of this article.

Intergranular Fracture

S. Lampman, ASM International

References

1. W. Rosenhain and D. Ewen, The Intercrystalline Cohesion of Metals, *J. Inst. Met.,* Vol 10, 1913, p 119–149

2. Z. Jeffries, The Amorphous Metal Hypothesis and Equicohesive Temperatures, *J. Am. Inst. Met.,* Vol 11 (No. 3), Dec 1917, p 300–324

3. Z. Jeffries, Effects of Temperature, Deformation, and Grain Size on the Mechanical Properties of Metals, *Trans. AIME,* Vol 60, 1919, p 474–576

4. C.L. Briant and S.K. Banerji, Intergranular Fracture in Steel: The Role of Grain-Boundary Composition, *Int. Met. Rev.,* No. 4., 1978, p 164–196

5. P.G. Shewmon, Grain Boundary Cracking, *Metall. Mater. Trans. B,* Vol 29, June 1998, p 509–518

6. S.P. Lynch, Mechanisms of Intergranular Fracture, *Mater. Sci. Forum,* Vol 46, 1989, p 1–24

7. G. Powell, private communication, 2002

8. V. Kerlins, Modes of Fracture, *Fractography,* Vol 12, *ASM Handbook,* ASM Interantional, 1987, p 112

9. *Fractography and Atlas of Fractographs,* Vol 9, *Metals Handbook,* 8th edition, 1974, p 112

10. S. Kuramoto, G. Itoh, and M. Kanno, Intergranular Fracture in Some Precipitation-Hardened Aluminum Alloys at Low Temperatures, *Metall. Mater. Trans. A,* Vol 27 (No. 10), Oct 1996, p 3081–3088

11. F.-L. Liang and C. Laird, Control of Intergranular Fatigue Cracking by Slip Homogeneity in Copper, Part I: Effect of Grain Size, *Mater. Sci. Eng. A,* Vol 117 (No. 1–2), Sept 1989, p 95–102

12. H.-J. Christ, On the Orientation of Cyclic-Slip-Induced Intergranular Fatigue Cracks in Face-Centered Cubic Metals, *Mater. Sci. Eng. A,* Vol 117 (No. 1–2), Sept 1989, p L25–L29

13. W.W. Gerberich, Microstructure and Fracture, *Mechanical Testing,* Vol 8, *Metals Handbook* (9th ed.), American Society for Metals, 1985, p 476

14. W.W. Gerberich and N.R. Moody, A Review of Fatigue Fracture Topology Effects on Threshold and Growth Mechanisms, "Fatigue Mechanisms," STP 675, J.T. Fong, Ed., ASTM, 1979, p 292–346

15. W. Yu, K. Esaklul, and W.W. Gerberich, Fatigue Threshold Studies in Fe, Fe-Si and HSLA Steel, Part II: Thermally Activated Behavior of the Effective Stress Intensity at Threshold, *Metall. Trans. A,* Vol 15, 1984, p 889–900

16. G. Krauss, Bending Fatigue of Carburized Steels, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 683

17. C.J. McMahon, Jr., Mechanism of Intergranular Fracture in Alloy Steels, *Mater. Charact.,* Vol 26 (No. 4), June 1991, p 269–287

18. B.C. Woodfine and A.G. Quarrell, Effects of Al and N on the Occurrence of Intergranular Fracture in Steel Castings, *J. Iron Steel Inst.,* Vol 195, Aug 1960, p 409–414

19. C.J. McMahon, Jr., Intergranular Fracture in Steels, *Mater. Sci. Eng.,* Vol 25, 1976, p 233–239

20. M. Nakamura and E.-I. Furubayashi, Crack Propagation of High Strength Steels in a Gaseous Hydrogen Atmosphere, *Metall. Trans. A,* Vol 14, 1983, p 717–726

21. W.C. Johnson and D.F. Stein, A Study of Grain Boundary Segregants in Thermally Embrittled Maraging Steel, *Metall. Trans.,* Vol 5, 1974, p 549–554

22. E.C. Pow, W.W. Gerberich, and L.E. Toth, Analysis of Stress Corrosion Crack-Tip Surface Chemistries in 7075 Type Aluminum Alloys, *Scr. Metall.,* Vol 15, 1981, p 55

23. C.L. Briant et al., Embrittlement of a 5% Nickel High Strength Steel by Impurities and Their Effects on Hydrogen-Induced Cracking, *Metall. Trans. A,* Vol 9, May 1978, p 625–633

24. B.D. Craig, A Fracture Topographical Feature Characteristic of Hydrogen Embrittlement, *Corrosion,* Vol 37, Sept 1981, p 530–532

25. M.L. Holzworth and M.R. Louthan, Jr., Hydrogen-Induced Phase Transformations in Type 304L Stainless Steels, *Corrosion,* Vol 24, April 1968, p 110–124

26. N. Narita et al., Hydrogen-Related Phase Transformations in Austenitic Stainless Steels, *Metall. Trans. A,* Vol 13, Aug 1982, p 1355–1365

27. E.E. Stansbury and R.A. Buchanan, *Fundamentals of Electrochemical Corrosion,* ASM International, 2000, p 404

28. R.H. Jones and R.E. Ricker, Stress-Corrosion Cracking, *Corrosion,* Vol 13, *Metals Handbook* (9th ed.), ASM International, 1987, p 145–163

29. N.J.H. Holroyd and G.M. Scamans, *Scr. Metall.*, Vol 19, 1985, p 915

30. D. Benac, Elevated-Temperature Failures, Materials Engineering Institute, Lesson 7, *Principles of Failure Analysis*, ASM International, to be published

Fracture of Plastics

Introduction

POLYMERS are long-chain molecules (or high-molecular-weight macromolecules) consisting of a series of small repeating molecular units (monomers) that are commonly made from compounds of carbon but are also made from inorganic chemicals, such as silicates and silicones. These repeating monomer units exhibit strong covalent bonds that form chains, which are also frequently linked by strong *primary bonds* (i.e., ionic, covalent, and metallic) and/or weaker *secondary bonds* (i.e., hydrogen and van der Waals). Polymers may also exhibit limited crystallinity, which occurs when the polymer chains arrange themselves into an orderly structure. In general, simple polymers (with little or no side branching) crystallize very easily. Crystallization is inhibited in heavily cross-linked (thermoset) polymers and in polymers containing bulky side groups.

There are three categories of polymers: thermoplastics, thermosetting plastics, and elastomers. Thermoplastics have linear chain configurations where chains are joined by weak secondary bonds. The secondary bonds holding adjacent macromolecules together are a direct result of the formation of molecular dipoles along the length of the polymer chain. Secondary bonds are much weaker than primary bonds, as indicated in Table 1, which accounts for the low melting temperatures, low stiffness, and low strength exhibited by many thermoplastic polymers. Thermoplastics can be melted by heating and will resolidify when cooled.

Table 1 Bond energies for various materials

Bond type	**Material**	Bond energy	
		kJ/mol	kcal/mol
Ionic	NaCl	640	**153**
	MgO	1000	**239**
Covalent	Si	450	**108**
	C (diamond)	713	**170**
Metallic	Hg	68	**16**
	Al	324	**77**
	Fe	406	**97**
	W	849	**203**
van der Waals	Ar	7.8	**1.8**
	Cl_2	31	**7.4**
Hydrogen	NH_3	35	**8.4**
	H_2O	**51**	**12.2**

Source: Ref 1, 2

In thermosetting polymers, covalent cross links or strong hydrogen bonds occur between polymer chains, resulting in three-dimensional networks of cross-linked molecular chains. Phenol formaldehyde is a good example. When these types of bonds are present, an increase in temperature does not lead to plastic deformation. Thermosetting is an irreversible chemical change that occurs during processing. Materials such as vulcanized rubbers, epoxies, and phenol formaldehyde change chemically during processing and will not melt on reheating. Rather, they will remain strong until they break down chemically via charring or burning. Thermoset polymers are seldom crystalline, because cross linking inhibits the mobility of individual chains such that crystalline arrangements do not occur.

Elastomers differ from thermoplastics and thermosetting polymers in that they are capable of rubbery behavior and much recoverable deformation (often in excess of 200%). Structurally, these materials consist of networks of heavily coiled and heavily cross-linked polymer chains. These materials are typically stronger than thermoplastics because of the presence of heavy cross links, which inhibits the sliding of polymer chains past each other. Elastomers are typified by natural rubber and by a series of synthetic polymers exhibiting similar mechanical behavior. Polyisoprene, for example, is a synthetic rubber that exhibits the same structure as natural rubber but lacks the impurities.

The physical state of a polymer under a given condition is primarily governed by chain flexibility, which in turn is determined by its backbone and the size and shape of the side groups and their spatial organization. Polymers can act as a fiber, plastic, or elastomer (Fig. 1), depending on the relative strength of intermolecular bonds and molecular structure or geometry. Noncrystalline polymers with weak intermolecular forces are usually elastomers or rubbers at temperatures above the glass transition temperature, T_g. In

contrast, polymers with strong hydrogen bonds and the possibility of high crystallinity can be made into fibers. Polymers with moderate intermolecular forces are plastic at temperatures below T_g. Some polymers, such as nylon, can function both as a fiber and as a plastic. Other polymers, such as isotactic polypropylene, lack hydrogen bonds, but because of their good structural geometry, they can serve both as a plastic and as a fiber.

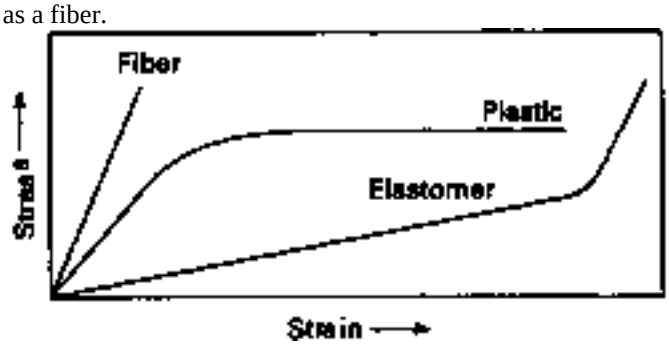


Fig. 1 Typical stress-strain curve for a fiber, a plastic, and an elastomer

Polymers are increasingly used in load-bearing structural applications. This article briefly reviews the mechanical behavior and fracture characteristics that discriminate structural polymers from metals. One of the most important differences between polymers and other materials is the statistical nature of their molecular chains. A distribution of chain lengths is a consequence of the random nature of most polymerization reactions. Thus, polymers are generally characterized by a molecular-weight distribution and the associated averages. This is discussed in more detail in the article "Characterization of Plastics in Failure Analysis" in this Volume.

References cited in this section

1. W.D. Callister, *Materials Science and Engineering—An Introduction,* 4th ed., John Wiley & Sons, 1997, p 21

2. M.L. Weaver and M.E. Stevenson, Introduction to the Mechanical Behavior of Nonmetallic Materials, *Mechanical Testing and Evaluation.* Vol 8, *ASM Handbook,* ASM International, 2000, p 15

Deformation and Fracture (Adapted from Ref 3)

Under circumstances in which metals may be regarded as purely elastic, polymers may deform nonelastically; that is, in a viscoelastic fashion. Their stress-strain response is nonlinear with time-dependent (creep) deformation; consequently, their fracture behavior is strongly influenced by temperature and the rate and overall time of load application. In addition, the considerable stress relaxation displayed by many polymers must be considered when assessing service loads in failure analysis. Another consequence of the viscoelasticity of polymers becomes obvious when fatigue loading is involved, particularly at high frequency. Hysteretic viscoelastic effects cause energy dissipation as heat, which can accumulate due to the low thermal conductivity of polymers, thus influencing failure mechanisms.

The generalized uniaxial tensile creep behavior of plastics under constant load, isothermal temperature, and a given environment can be illustrated as ductile creep behavior (Fig. 2) or as brittle creep behavior (Fig. 3). At very low stress levels, both types of plastics exhibit similar first- and second-stage creep deformation. The onset of creep rupture may not occur within the service life of the product (let alone the test). As the stress level increases, first- and second-stage creep deformation rates remain relatively the same, but the time of failure is, of course, considerably reduced. In addition, third-stage creep deformation characteristics now differ considerably. The ductile plastic exhibits typical ductile yielding or irreversible plastic deformation prior to fracture. The brittle plastic, on the other hand, exhibits no observable gross plastic deformation, only abrupt failure.

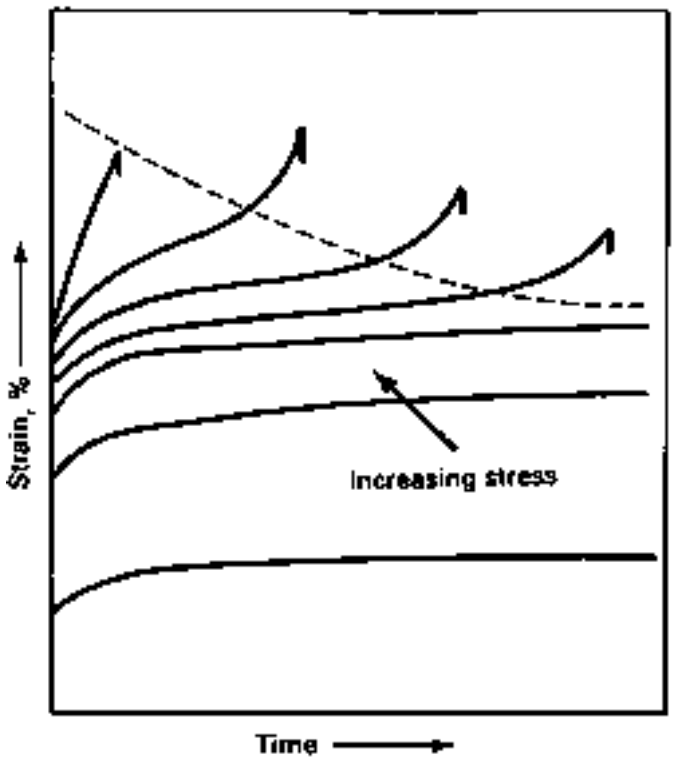


Fig. 2 Typical creep and creep rupture behavior of ductile polymers

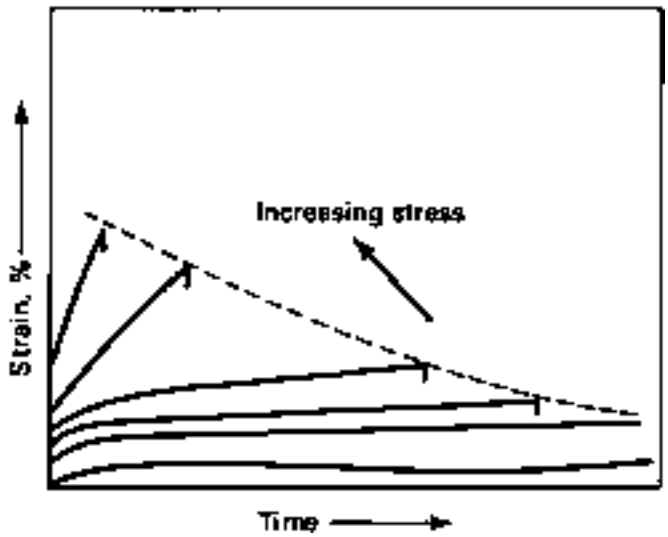


Fig. 3 Typical creep and creep rupture behavior of nonductile, or brittle, polymers

Cavitation and microvoid formation in polymers differs from that in metals. Cavities form in the necked region of metals after the onset of necking, because the geometry changes induce local triaxial stress components from the tensile load. The cavities grow and, through microvoid coalescence, result in a crack that soon causes fracture. Such cavitation processes in metals typically occur after the onset of necking (or from debonding at inclusions before necking), and fracture typically occurs soon thereafter. In contrast, cavitation processes in polymers can dominate the plastic deformation. These voids do not coalesce into a crack, but instead become stabilized by fibrils containing oriented polymeric material. The resulting region consisting of voids and fibrils is known as a craze (Fig. 4). The long-chain nature of polymers is responsible for crazing. Without the long chains, it would not be possible to form the fibrils that span the craze and prevent the conversion of the craze into a crack. Although crazing is usually associated with the deformation of amorphous polymers, it has also been observed in semicrystalline polymers and in thermosetting resins.

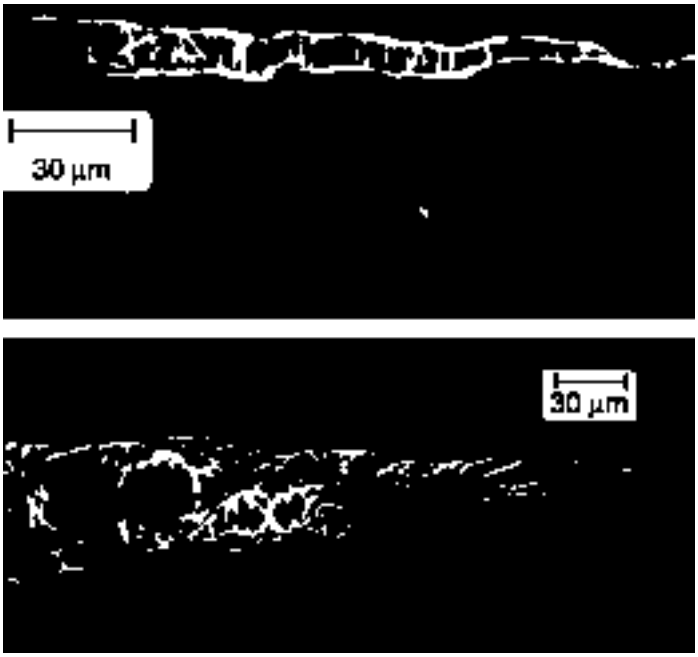


Fig. 4 Crazing fibrils in linear polyethylene (density, 0.964 g/cm^3)

In polymers, macroscopic yielding and fracture may not always be appropriate criteria for long-time duration material failure, because cavitation and localized damage can occur prior to necking (which often occurs at yielding rather than at maximum load for a metal). The designer must keep in mind that by their very nature, polymers cannot be fabricated at their maximum packing density. Molecular scale voids, known collectively as free volume, exist within the polymeric macromolecular structure. Applied stress serves to straighten the polymer chains and perhaps to redistribute the free volume so that sizable microvoids are formed. Depending on the nature of the polymer, these microvoids continue to grow into localized areas of stress whitening, stress crazing, stress cracking, or brittle failure. Therefore, design limitation and product failure for some plastics may be associated with stress crazing, stress cracking, or stress whitening. This is shown in Fig. 5 by a set of time-dependent stress curves for unplasticized polyvinyl chloride (PVC). The upper line indicates ductile failure due to necking. The region around the dotted line denotes stress-whitening failure. The dashed line, representing craze initiation, is curved in a shape similar to the ductile failure curve, but it is apparent that craze initiation occurs at stress levels substantially below those of ductile failure. To prevent failure initiation, the strain level must be below 1%. This means that if the product has a lifetime of 10^8 s, or 3.2 years, the applied stress cannot exceed 17 MPa (2.5 ksi).

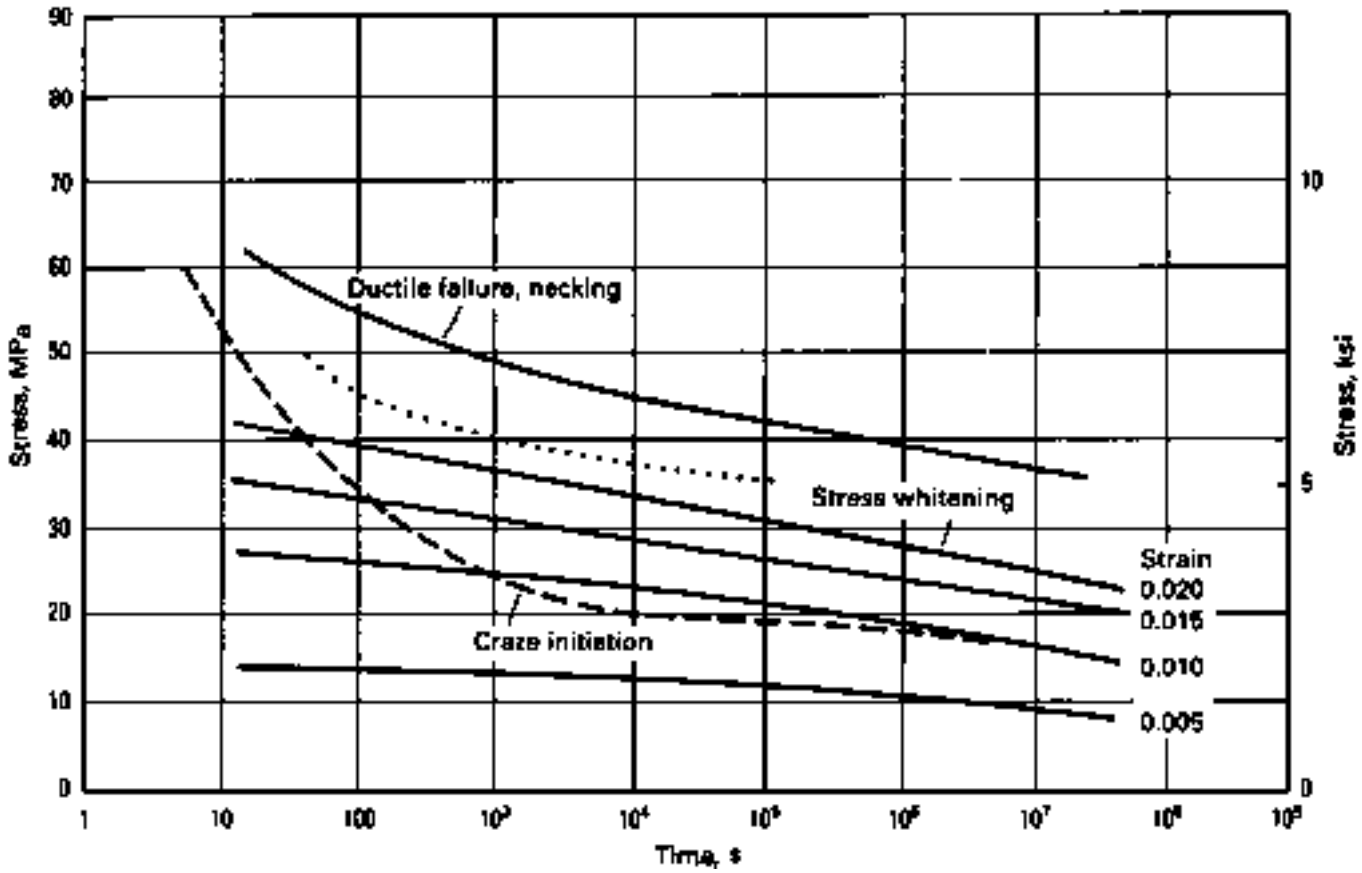


Fig. 5 Isometric tensile creep curves for unplasticized polyvinyl chloride at 20 °C (68 °F), 50% relative humidity

The deformation and fracture behavior is generally a complex phenomenon that varies with material composition, microstructure, stress environment, time, and temperature (Ref 4). Such behavior for polymers, like metals, depends on external loading conditions, component geometry, and imperfection geometry. One way of displaying the effect of temperature on creep is by temperature-dependent isochronous creep moduli (typically for 10^3 h). In Fig. 6, the creep moduli for eight unfilled plastics are shown for three temperatures. Fillers, reinforcements, plasticizers, and other adducts can significantly influence creep data. Consider the temperature effect on the isochronous creep plot for polycarbonate (PC) (Fig. 7). As expected, the overall creep effect increases with increasing temperature. The crazing boundary is also influenced by temperature (Fig. 7).

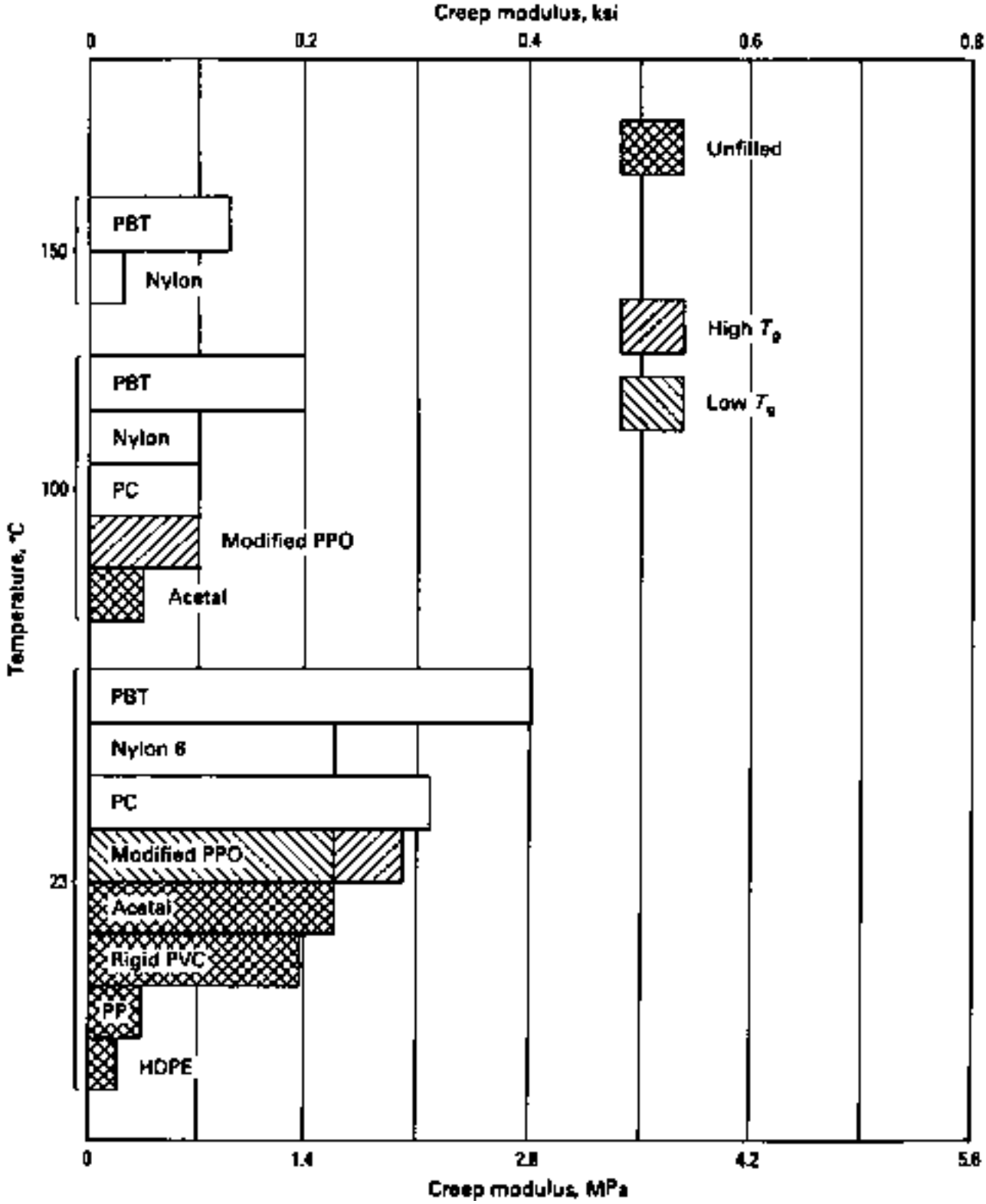


Fig. 6 A 1000 h creep modulus of several polymers as a function of temperature. PBT, polybutylene terephthalate; PC, polycarbonate; PPO, polyphenylene oxide; PVC, polyvinyl chloride; PP, polypropylene; HDPE, high-density polyethylene; T_g, glass transition temperature

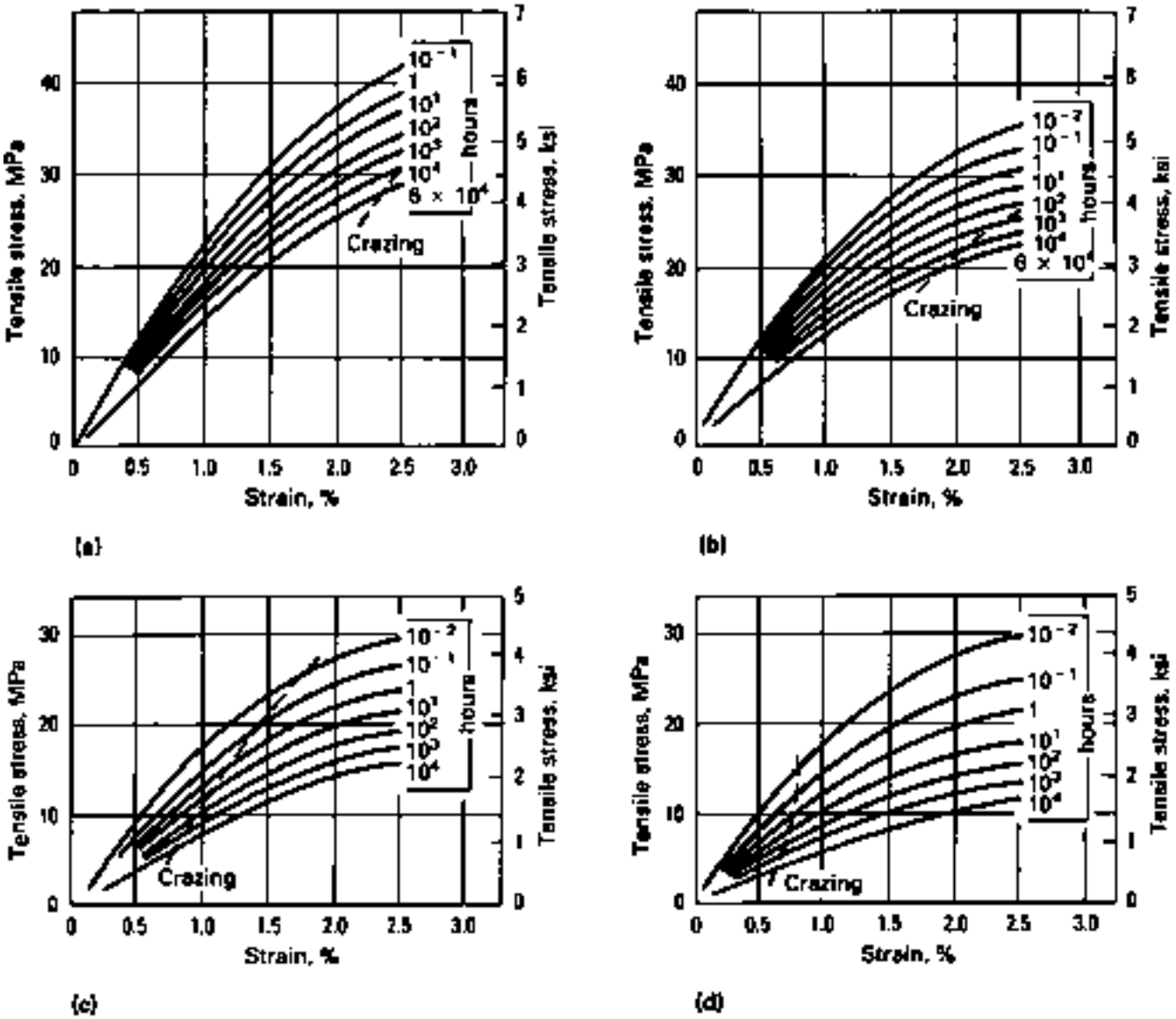


Fig. 7 Isochronous plot of polycarbonate stress-strain behavior as a function of temperature. Note that the crazing locus decreases in strain value with increasing temperature. (a) 23 °C (73.5 °F). Relative humidity, 50%. (b) 40 °C (104 °F). (c) 80 °C (176 °F). (d) 100 °C (212 °F). Courtesy Mobay Chemical Company

Stress cracking implies the localized failure that occurs when localized stresses produce excessive localized strain. This localized failure results in the formation of microcracks that spread rapidly throughout the local area. Brittle materials are more prone to stress cracking than to stress whitening.

Stress whitening is a generic term describing many different microscopic phenomena that produce a cloudy, foggy, or whitened appearance in transparent or translucent polymers in stress. The cloudy appearance is the result of a localized change in polymer refractive index. Thus, transmitted light is scattered. Microvoid clusters of dimension equal to or greater than the wavelength of light are thought to be the primary cause of stress whitening. The microvoids can be caused by the delamination of fillers or fibers, or they can be localized failure around occlusions, such as rubber particles or other impact modifiers. Although stress whitening results in visually apparent changes, the load-bearing capabilities of a specimen may not be substantially reduced during stress whitening.

Shear bands (Fig. 8, Ref 5) are also microscopic localized deformation zones that propagate ideally along shear planes. Like crazes, shear-deformation bands, or slip lines, are traditionally thought to be the mechanism of irreversible tensile deformation in ductile amorphous polymers. Almost invariably, a compressive-stress state will cause shear deformation in polymers. Under monotonic tensile loading, PC is reported to deform by shear banding.

Fig. 8 Section from a polystyrene sample that was deformed past its compressive yield. The section is viewed between cross polars, showing shear bands. 50×

Stress Crazing. As previously noted, a craze is a microcrack that is spanned by plastic microfibrils, typically oriented in the direction of applied stress. The width of a craze is of the order of 1 to 2 μm, and it may grow to several millimeters in length, depending on its interaction with other heterogeneities. Being dilatational, crazes grow normal to the applied tensile component of the stress field. Craze fibrils are load-bearing elements whose strength and density depend to some extent on the molecular weight of the polymer. Because the fibrils are load-bearing, the microcrack usually does not open substantially before parallel microcracks form.

Fiber-forming plastics such as nylon, polyethylene (PE), and polypropylene (PP) readily stress craze, but crazing also occurs in non-fiber-forming plastics such as PC and polymethyl methacrylate (PMMA). Crazes occur in type I (normally brittle in tension) amorphous polymers, such as polystyrene (PS) and PMMA, under monotonic tensile loading. Crazes, however, have been observed in semicrystalline polymers, such as PP, and in type II (normally ductile in tension) amorphous polymers, such as PC (Ref 6), under tensile-fatigue loading. Higher-molecular-weight polymers develop fewer, longer, and stronger (i.e., stable) crazes. Although crazing is commonly associated with brittleness, higher-molecular-weight polymers have higher resistance to fracture. Implications of this phenomenon have been exploited commercially in the development of toughened polymers.

Because crazes have a different refractive index and because they reflect light, they are easily visible, especially when viewed at the correct angle with the aid of a directed light source, such as a fiber-optic illuminator. When viewed with this type of light source, the craze appears to have a silvery appearance much like that of a very fine crack. The formation of crazes is sensitive to many variables, such as polymer molecular weight, loading conditions, temperature, pressure, and the presence of solvents. Crazes will initiate in a plastic when a critical limit is reached in stress, strain, dilation, or distortion strain energy. Higher molecular weight plastics generally have greater resistance to crazing, while more crystalline plastics have lower resistance to crazing. This is because of the fewer number of tie molecules that hold it together.

The effect that solvents have on crazing and, therefore, on polymer deformation is another contrast to metals. When solvents initiate crazing in polymers, they generally lower the stress at which crazes form. This solvent-crazing phenomenon can drastically affect failure characteristics. Particular, large effects are observed in polymers that normally deform by shear yielding but can be induced to craze in the presence of solvents. When an active solvent causes the crazing stress to fall below the shear yielding stress, a polymer that normally exhibits ductile failures under tensile loading can show brittle failures precipitated by the crazing. Such solvent effects can be observed in failures of PC and polysulfone (Ref 7).

Solvent effects can be a source of failure in metals (such as in stress-corrosion cracking), but solvent effects in polymers are more numerous and more likely to be precursors to failure. Solvent-induced failures are a serious engineering problem (Ref 8) that remains a concern. Example 1 in this article is just one case of this, and more information on environmental stress cracking (ESC) of plastics is described in the article “Effect of Environment on the Performance of Plastics” in this Volume. Nonsolvents that possess specific physiochemical affinity leading to wetting of the polymer are also known to cause premature failure (Ref 9). Environmentally enhanced failure generally involves crazing or cracking as the underlying mechanism.

Degradation by Heat and Light (Ref 5). Another serious consideration in failure analysis of polymers is the fact that their mechanical properties may severely deteriorate by exposure to heat and/or light. Thus, structural components serving at elevated temperatures or in outdoor applications may undergo degradation if the polymers employed are not properly stabilized. Mechanical stress is known to enhance degradative effects. Degradation usually involves molecular-weight reduction by one of several mechanisms (Ref 10). Surface embrittlement and consequent microcracking occur. This promotes crack initiation and possibly assists crack propagation. Figure 9 illustrates surface microcracking induced in a polyoxymethylene specimen exposed to ultraviolet (UV) light in the laboratory

for 1000 h. It should be noted, however, that some polymers are intrinsically more resistant to degradation than others. The following common polymers are ranked with respect to their relative resistance to the deterioration of mechanical properties due to photo-degradative effects (Ref 5):

Polymer	Relative resistance
Polymethyl methacrylate	**n**
Polyacrylonitrile	**n**
Polyoxymethylene	**m**
Polyethylene	**m**
Polyvinyl chloride	**n**
Polystyrene	**w**
Nylon 6	**m**
Polypropylene	**vs**
Polycarbonate	**s**
Polyurethane	**m**
Polysulfone	**s**
Polyphenylene oxide	**s**

n, not significant deterioration; m, moderate; w, weak; s, severe; vs, very severe

More information on polymer degradation by heat and light is in the article “Effect of Environment on the Performance of Plastics” in this Volume.

Fig. 9 Surface-microcracking network developed on polyoxymethylene due to ultraviolet exposure. 200×

Example 1: Solvent-Induced Cracking Failure of PC Ophthalmic Lenses (Ref 11). A chemical laboratory accident occurred when a solvent splashed on metal-framed safety glasses with PC ophthalmic lenses. The metal-framed glasses had been substituted for plastic-framed safety glasses.

Investigation. Metal-framed PC ophthalmic lenses appeared to have shattered from acetone solvent-induced cracking. The failed lenses exhibited primary and secondary cracks (Fig. 10), which were associated with solvent swelling and crazing. All previous solvent splashes on plastic-framed ophthalmic safety glasses did not produce any fractures.

Fig. 10 Failed polycarbonate lenses exhibited primary and secondary cracking associated with solvent swelling and cracking

The lenses were specified to be unfilled PC, which exhibits glass-like clarity. The metal frames were a custom designer series, chosen exclusively for style. The frames gripped approximately two-thirds of the periphery of the lenses. The failed ophthalmic lenses were verified by Fourier-transform infrared (FT-IR) analyses to be the specified PC material. The source of the splashed liquid was found to contain primarily acetone.

Discussion. Polycarbonate plastics are often used in applications that require glass-like clarity with extreme toughness. One such application is safety glasses. Under most circumstances, when PC material is stressed in the absence of a solvent environment, no shattering will occur. The converse of this event is also true; that is, shattering does not occur in the absence of stress with solvent exposure. However, when PC material is stressed in the presence of a solvent environment, solvent-induced or environmentally induced cracking can occur, resulting in a catastrophic brittle-like failure. This phenomenon also occurs with other plastics and other solvents.

Conclusion. The most probable cause of failure is acetone solvent-induced cracking. Solvent-induced cracking of the metal-framed ophthalmic lenses (but not the plastic-framed ophthalmic lenses) occurred because the metal frames were exerting an uneven stress distribution on the polycarbonate lenses. The failure could have been prevented by using the correct type of safety glasses with plastic frames instead of metal-framed lenses. However, it was also emphasized that PC ophthalmic lenses are usually not designed or designated for chemical splash protection.

Recommendation. Polycarbonate ophthalmic lenses should not be exposed to service environments with solvents, such as acetone. In addition, marking pens, adhesives, or soaps, which contain undesirable solvents, should never be used on the ophthalmic lenses. Excessive stress can originate from working stress or from residual stress and may result in shattering in the presence of a solvent.

References cited in this section

3. Deformation and Viscoelasticity, *Engineered Materials Handbook Desk Edition,* ASM International, 1995, p 393–409

4. I.M. Ward, *Mechanical Properties of Solid Polymers,* John Wiley & Sons, 1982

5. A. Moet, Failure Analysis of Polymers, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 758–765

6. A. Moet, Fatigue Failure, *Failure of Plastics,* W. Brostow and R.D. Coneliussen, Ed., Hanser Publishers, 1986

7. C.B. Bucknall, Toughened Plastics, *Appl. Sci.,* 1977

8. R.W. Hertzberg and J.A. Manson, *Fatigue of Engineering Plastics,* Academic Press, 1980

9. K. Matsushige, S.V. Radcliff, and E. Baer, *J. Mater. Sci.,* Vol 10, 1975, p 833

10. W. Schnabel, *Polymer Degradation,* Hanser International, 1981

11. Edward C. Lochanski, Solvent-Induced Cracking Failure of Polycarbonate Ophthalmic Lenses, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 493

Crack Propagation (Adapted from Ref 5)

In most cases, failure of load-bearing structural components involves fracture—that is, complete or partial separation of a critical member of the component under service loading that renders the component nonfunctional. The role of failure analysis, therefore, is to reconstruct the sequence of processes leading to failure, aiming at determining the primary cause. This effort requires knowledge of the various mechanisms by which the material responds to a loading environment similar to that under which a failure occurred. Frequently, laboratory experiments must be conducted to establish such cause and effect relationships. Again, failure mechanisms observed in a particular field failure should bear specific similarity to that observed in the laboratory to render comparison valid.

Fracture of load-bearing engineering components is generally a macroscopic phenomenon resulting from a series of microscopic and submicroscopic processes. Although molecular theories of polymer fracture are a subject of continuing research, practical fracture analysis is achievable from examining deformation events within the resolution of optical and scanning electron microscopy (SEM). Irreversible deformation mechanisms in polymers may fall into two basic categories: dilatational, such as crazes, voids, and microcracks; or nondilatational, such as shear bands. Commonly, a mixture of both mechanisms may be encountered.

In terms of their fracture behavior, polymers are generally classified as brittle or ductile, as discussed further in the next section in this article, “Fractography.” Brittle polymers are those that are known to fracture at relatively low elongations in tension (2 to 4%). These include PS, PMMA, and rigid (unplasticized) PVC. Crazing is the dominant mechanism of failure in such polymers. Highly cross-linked polymers, such as epoxies and unsaturated polyesters, are also brittle, yet their fracture involves a microcracking mechanism. On the other hand, semicrystalline polymers, such as PE and nylons, and some amorphous polymers, such as PC and polyethylene terephthalate (PET), exhibit considerable post-yield plastic deformation and are thus classified as ductile. In spite of this classification, however, relatively low loads applied over long periods of time are reported to cause failure to occur even in the most ductile polymers. Indeed, early investigations on PE, a ductile polymer, have shown that the failure mechanism switches from ductile to brittle at longer times and lower stress levels (Fig. 11, Ref 12).

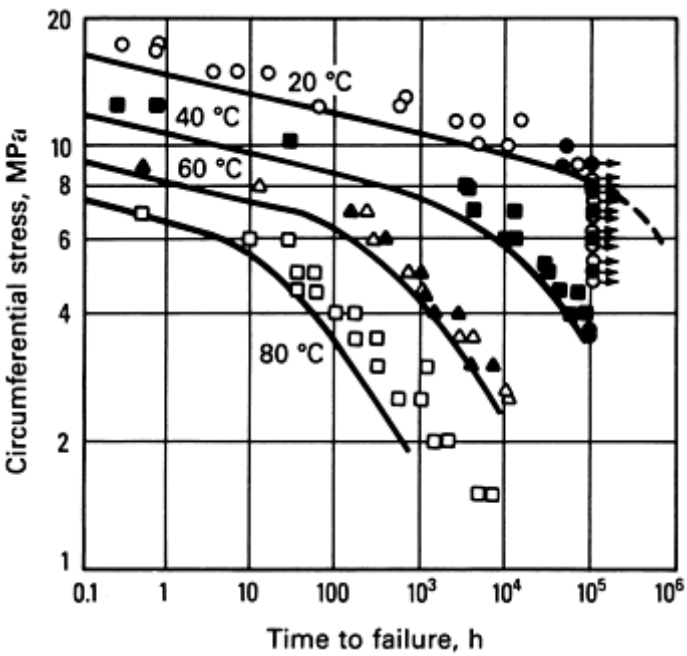


Fig. 11 Time-to-failure of high-density polyethylene pipes at different stresses and temperatures. Source: Ref 12

The crack-propagation behavior of polymers resembles that in metals, as a propagating crack in a polymer is usually preceded by a zone of transformed material (i.e., a plastic zone). This may involve any of the known irreversible deformation mechanisms. In most cases, a zone of crazes precedes crack propagation, as in PP (Fig. 12) or in PS (Fig. 13). It is important to note that PP is a semicrystalline polymer that usually deforms by yielding and recrystallizing under monotonic loading. The transformation at the crack tip may appear as a yielded zone, as in thin PC sheet material (Fig. 14). A pair of shear bands are also reported to evolve at the tip of propagating cracks in PC (Ref 13). Recently, it was observed that when the thickness of the PC sheet is 6 mm (0.25 in.) or more,

brittle microcracking ahead of the propagating crack becomes the dominant mechanism of fracture. Thus, one polymer displays more than one fracture mechanism. In all cases, the crack appears to propagate through a craze at its very tip.

Fig. 12 A thinned section of fatigue-cracked polypropylene specimen. Crazes are visible surrounding and preceding the crack. 8×

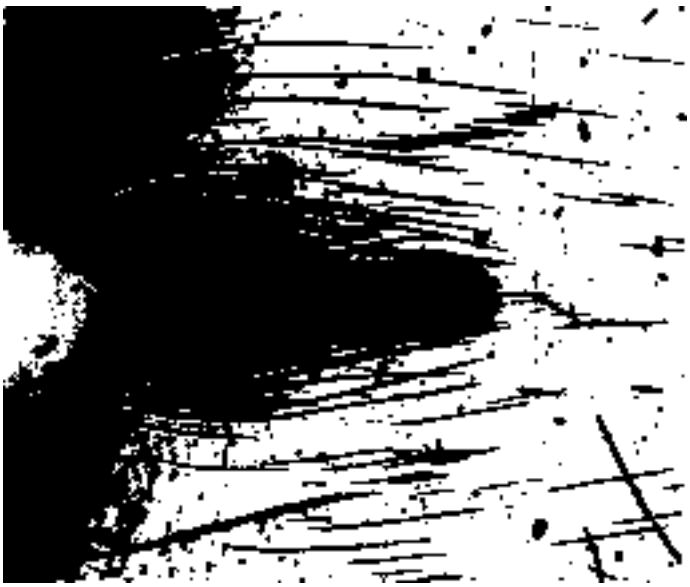


Fig. 13 Fatigue-crack initiation in polystyrene from a V-notch. Note crazes surrounding and preceding the crack. 37×

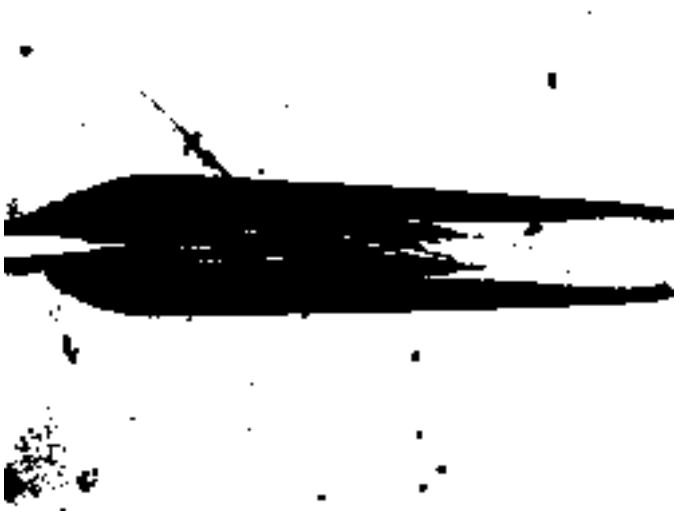


Fig. 14 Transmitted-light micrograph showing a yielded zone surrounding and preceding a fatigue crack in 0.25 mm (0.01 in.) thick polycarbonate sheet. 20×

The magnitude of transformation preceding the crack in a particular polymer is influenced by the loading history and consequently determines the resistance to crack propagation (Ref 14, 15, 16). Investigations show that damage evolution ahead of the crack tip accounts for the discrepancies in fracture toughness of polymers (Ref 17). These results suggest that examination of the material above and below the crack-propagation plane should be considered in failure analysis in addition to the commonly accepted fracture surface studies. As was briefly shown in preceding paragraphs, vital information about the resistance of the material to crack propagation and its loading history can easily be decoded from a thinned section normal to the crack-propagation plane.

References cited in this section

5. A. Moet, Failure Analysis of Polymers, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 758–765

12. H.H. Kauch, *Polymer Fracture,* Springer-Verlag, 1978

13. M.T. Takemori and R.P. Kambour, *J. Mater. Sci.,* Vol 16, 1981, p 1108

14. F.W. Billmeyer, Jr., *Text Book of Polymer Science,* John Wiley & Sons, 1984

15. J.G. Williams, *Fracture Mechanics of Polymers,* John Wiley & Sons, 1984

16. A. Chudnovsky, A. Moet, R.J. Bankert, and M.T. Takemori, *J. Appl. Phys.,* Vol 54, 1983, p 5562

17. N. Haddaoui, A. Chudnovsky, and A. Moet, *Polym. Mater. Sci. Eng.,* Vol 49, 1983, p 117

Fractography (Adapted from Ref 18)

Fractography often reveals important clues about the cause of fracture and therefore plays an important role in the choice of subsequent testing or analyses to determine the cause of failure (Ref 19). Fractography involves the examination and interpretation of fracture surfaces, although the examination is not necessarily confined to the fracture surface alone. In most cases the fracture surface examination is intended to reveal the location of the fracture origin, providing valuable information about the local service environment as well as the state of stress responsible for the crack initiation and growth that eventually led to fracture. The subsequent stress analysis of the failed part can be considerably simplified because attention can be focused at the location of crack initiation (Ref 20).

The stereo zoom optical microscope is the instrument of choice for the preliminary fracture surface examination at moderate magnifications. A failed plastic part can be thoroughly inspected to assess the extent of cracking as well as the presence of other surface microcracks not revealed during the visual inspection. In addition, the origin of the fracture can frequently be identified under the optical microscope. Because virtually no specimen preparation is involved, the specimen is unaltered. This is sometimes an important consideration in situations where the specimen being examined must be preserved without contamination for subsequent analyses.

The scanning electron microscope (SEM) is frequently used to study fracture surfaces. Its superior depth of field is particularly useful in examining rough fracture surfaces at higher magnifications. However, because most polymers are electrically nonconductive, the specimen must be coated with a thin, conductive layer such as gold or carbon, which has a tendency to alter the surface appearance. If discoloration is evident on the failed part due to either environmental degradation or chemical attack, the specimen should be examined and documented first in the optical microscope. Because SEM is also more time consuming than optical microscopy, specimens should be judiciously selected by preliminary examination under the optical microscope before a more detailed examination is conducted with the SEM. Additional information on polymer microscopy techniques and specimen preparation methods is available in Ref 21.

Modes of Fracture

When fracture occurs, a fracture surface is produced on a plane normal to the maximum principal tensile stress where the local stress exceeds the local strength of the material. Crack extension takes place along the path of least resistance to fracture. The characteristic fracture surface appearance depends on the complex interactions of the prevailing conditions of stress, materials properties, environment, and time. The distinction between the ductile and brittle fracture modes is generally made on the basis of macroscopic appearance.

Ductile fractures involve gross plastic deformation that is commonly described as yielding, tearing, plastic flow, necking, and shear—all of which involve shape changes or distortions of varying extent. Examples of commercial polymers that normally exhibit ductile fracture behavior are acrylonitrile-butadiene-styrene (ABS), polyethylene (PE), polypropylene (PP), high-impact polystyrene (HIPS), polyamide (PA), polybutylene, polycarbonate (PC), polyethylene terephthalate (PET), and polyvinyl chloride (PVC).

The most obvious example of a ductile fracture can be seen in the standard tensile specimen of a ductile polymer tested under ordinary conditions of temperature and strain rate. In a tensile test, a specimen containing no stress concentration is stretched at a constant rate until fracture. The stress in the specimen is essentially uniform uniaxial tension until onset of yield and then again after chains are oriented. Figure 15 shows the tensile load-displacement curve of a ductile polymer. Typically, the initial portion of the load-displacement record is nearly a straight line. This stress-strain response is termed *linear elastic behavior* because the stress is proportional to the strain, and on unloading, the specimen will return to its original dimensions.

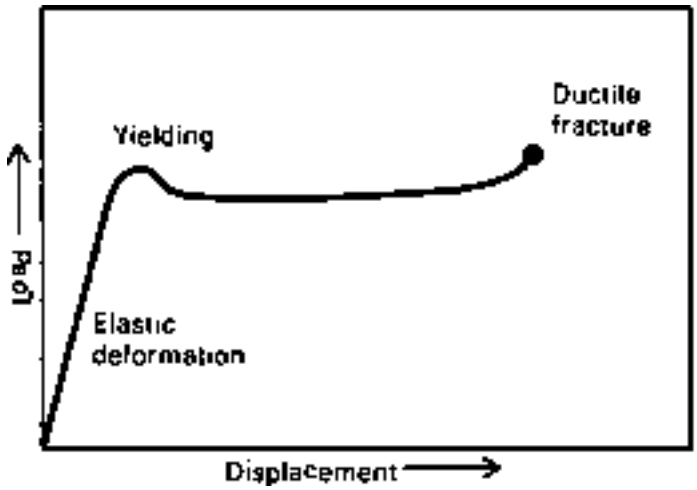


Fig. 15 Typical load-displacement curve for a ductile polymer tested in uniaxial tension

As the deformation increases, the stress is no longer linearly proportional to the strain. The slope of the curve begins to decrease, and a knee is eventually formed. For most engineering plastics, this typically occurs at a strain of a few percent. The rate of stress increase with deformation is substantially lower than the initial slope, which is a measure of the elastic modulus of the material. This sudden change in slope signifies the onset of net section plastic deformation or yielding. Even if the stress is removed, the specimen will not return to its original dimensions, in which case the specimen is permanently deformed. This suggests that ductile fractures occur with net section yielding after the applied stress has exceeded the yield strength of the material.

In practice, ductile fractures rarely become the subjects of failure analysis. The maximum stress normally expected in a well-designed component, or the maximum allowable design stress, should not exceed a small fraction of the yield strength of the material. Exceptions are those products that are intended to be broken in use, such as the tamper-evident rings found on the plastic closures of beverage bottles. When an unexpected ductile fracture occurs, it is usually found that the part was either grossly underdesigned or overloaded, so that the maximum stress it encountered not only exceeded the allowable design stress but also surpassed even the yield and tensile strengths of the material. Problems of underdesign or overloading are usually diagnosed early in the development of a product, such as during prototype testing. Ductile fractures can also result from other causes, such as excessive creep deformation or inadvertent exposure to elevated temperatures, which can substantially reduce yield strength.

Brittle fractures occur on a macroscopic level with little or no gross plastic deformation. Because shape changes or distortions are absent, the fracture faces can be precisely matched, although this practice is not recommended if the fracture surfaces are to be subsequently examined with a microscope. Inadvertent contact can produce artifacts that complicate the fracture surface examination. Examples of commercial plastics that normally fracture in a brittle manner are PS, styrene-acrylonitrile (SAN), PMMA, and thermosetting resins such as epoxy and polyester. A common example of a material that fractures in a truly brittle manner is inorganic glass.

The typical tensile stress-strain plot for a brittle plastic is essentially a straight line from the origin to the point of fracture, which coincidentally occurs at about the same strain as yielding in a ductile polymer. Except for a slight decrease in slope immediately before fracture, the deformation may be considered linearly elastic. Hooke's law for a homogeneous, isotropic material can be applied for most practical engineering analyses of stress and strain in a brittle plastic part.

When a ductile polymer fractures in a brittle manner, the stress analysis of the part may be similarly conducted assuming linear elastic behavior. This is permissible because brittle fractures in a normally ductile polymer also occur at small strains before the onset of gross yielding. The stresses involved in producing the fracture are below the yield strength of the material and therefore lie within the linear elastic portion of the stress-strain plot.

Although no gross plastic yielding is evident in brittle fractures, plastic deformation is nevertheless involved in polymer fractures. Except for those rare cases in which polymer molecular motions are totally inhibited, such as at very low temperatures, brittle

fractures of even brittle polymers are accompanied by plastic flow processes, which occur on a much more localized level than gross plastic yielding.

One form of plastic deformation that frequently leads to brittle fractures is crazing, sometimes called craze yielding (Ref 7). In many polymers, crack initiation is preceded by craze formation. In transparent polymers, the easily observed crazes appear as cracklike structures that are macroscopically indistinguishable from cracks. Geometrically, a craze is a planar defect similar to a crack, with its areal dimensions much larger than its thickness. A craze differs from a crack in that it contains an interpenetrating network of voids among highly drawn polymer fibrils bridging the craze faces.

Crazing begins with microvoid formation under the action of the hydrostatic tension component of the stress tensor. It is particularly sensitive to biaxial tensile stresses that frequently occur at sites of stress concentration with material constraints. After crazing has initiated, the voids increase in size and elongate along the direction of the maximum principal tensile stress. The polymer bulk material among the voids also undergoes gradual elongation to form thin fibrils. The strain in the craze fibrils depends on the amount of craze thickening but has been estimated to be about 50 to 100% in a well-developed craze section.

As the craze thickens, its growth in the lateral dimensions occurs by additional void nucleation at its leading edge. New craze matter is generated at the craze tip as a result of the triaxial tension there.

The deformation of a craze frequently leads to the initiation of a true crack. As the craze faces separate, the craze fibrils increase in length, while their diameters contract. The craze fibrils are not unlike microtensile specimens undergoing uniaxial extension. When the longitudinal strain in the fibrils exceeds the maximum extensibility of the molecular network, they rupture and form a crack. Figure 16 illustrates craze formation and its subsequent development into a crack.

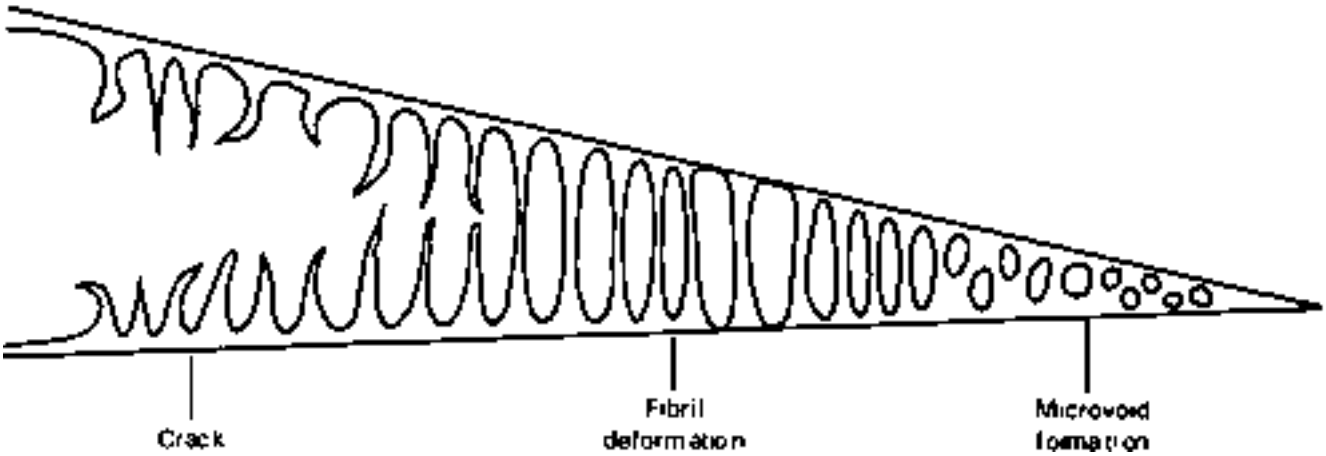


Fig. 16 Crack formation from a craze

Crack initiation is then followed by crack growth in which more craze fibrils undergo extensive plastic deformation until they rupture. At the same time, craze growth continues, and new craze matter is generated at the craze tip. When the crack reaches a certain critical size, crack extension occurs in an uncontrollable manner, resulting in an unstable, brittle fracture. In terms of linear elastic fracture mechanics, the criterion of brittle fracture in the tensile crack opening mode is described by the stress intensity factor, K_I, at the crack tip, reaching a critical value, K_{Ic}, for the plane-strain fracture toughness, which is considered to be a material property (Ref 22). Linear elastic fracture mechanics properties per ASTM D5045 are appropriate for highly cross-linked thermosets or glassy thermoplastics incapable of significant plastic deformation (e.g., polystyrene.) Unlike shear yielding, which occurs at constant volume, craze yielding is a cavitation process that is accompanied by an increase in volume. Depending on the stage of craze development, the void volume in a craze has been calculated from refractive index measurements to be roughly 40 to 60%. Crazing, frequently described as the precursor of brittle fractures, is therefore favored by the presence of hydrostatic tensile stresses.

The influence of the state of stress on the mode of fracture is best illustrated by the deformation of the craze itself. The formation and growth of a craze, which frequently lead to brittle fractures, are promoted by a stress condition known as plane strain, in which triaxial tensile stresses exist. The plane-strain condition often results from the nonuniform stress distribution near stress risers, such as cracks and other defects. Because of elastic constraints, lateral deformations near the crack tip or other stress concentrations are restricted, resulting in the development of lateral tensile stresses. The resultant hydrostatic tension tends to produce cavitation or void formation in the material, eventually leading to craze formation and brittle fractures.

On the other hand, the fibrils within the craze deform under a state of uniaxial tension in which lateral contraction is unconstrained, resulting in extensive shear band formation before rupture. A large amount of mechanical strain energy per unit volume is dissipated through craze yielding, which contributes to the toughness of the material. For this reason, even brittle polymers, such as PS and PMMA, are more impact resistant than ordinary window glass, in which brittle fractures occur without plastic deformation.

Despite the extensive plastic deformation that occurs within the craze, the total strain energy absorbed is small because the plastic deformation takes place in a relatively small volume of material occupied by the craze. Therefore, only a small amount of strain energy is absorbed in a plane-strain brittle fracture. In contrast, a large amount of energy is absorbed in a ductile fracture, in which the plastic deformation takes place over a much larger volume of material.

Certain polymers that normally deform in a ductile manner in the standard tension test frequently sustain brittle fractures when a sharp notch or crack is introduced into the specimen. Some examples of notch-sensitive polymers are PC, PVC, PP, PET, PA, ABS, and HIPS. The transition from a ductile to a brittle fracture mode when a deep notch or crack is present is partly due to the change from a uniaxial to a triaxial tensile stress state.

Evidence of craze yielding can frequently be observed on the brittle fracture surface. As a result of craze fibril rupture, the fracture surface is lined with a thin layer of highly oriented polymer fibrils whose index of refraction differs from that of the underlying bulk polymer. As a result, colorful interference fringes similar to those seen on an oily, wet pavement can frequently be observed at the fracture origin on a brittle fracture surface.

Fracture Surface Features

The fracture origin is the point at which a crack is first nucleated. It usually coincides with the location of the maximum principal tensile stress or the minimum material resistance to fracture. Strictly speaking, because neither a uniform distribution of stress nor material homogeneity exists throughout an engineered component, certain points within a part are expected to be more probable sites of crack initiation due to either higher-than-average stresses or lower-than-normal crack resistance.

Local stress variation can result from a variety of factors, including loading configuration, part design, and discontinuities or inhomogeneities such as voids, cracks, inclusions, and other imperfections. The presence of holes, sharp corners, or sudden changes in wall thickness all contribute to stress concentration.

Other imperfections that can raise local stresses are introduced during the manufacture or the service life of the part. Some examples are shrinkage voids (Fig. 17) and contaminant inclusions resulting from poor molding practices. Improper techniques during postmolding operations, such as machining, decorating, bonding, or mechanical fastening, can produce partial fractures or cracks that escape detection and serve as sites of crack initiation.

Fig. 17 Shrinkage void on field fracture surface of polycarbonate. 12×

The location of the fracture origin can also reveal points of material weakness. Material defects within a molded part include weak weld lines, inclusions of debris, poorly fused cross-linked gel particles, and other inhomogeneities such as agglomeration of pigment, filler, and reinforcement. Because these defects are caused by poor processing, their occurrence may be more sporadic than defects in design.

Material degradation caused by either natural aging or an aggressive environment is frequently revealed by examining the failed part in the vicinity of the crack origin. One of the leading causes of brittle fracture in polymers is environmental stress cracking that results from exposure to incompatible chemicals. Examples of potential stress crack agents are mold releases, cleaners, lubricating oils and greases, plasticizers, and solvents in paints and coatings. Prolonged exposure to ultraviolet radiation in sunlight can also embrittle the surface layers, resulting in microcracking. In these cases, fracture initiation is typically found at multiple locations on the affected areas and is frequently accompanied by surface microcracking.

Mirror Zone. Brittle fractures in many polymers are preceded by craze formation and its subsequent breakdown. The initial stage of crack growth results from the rupture of fibrils at the trailing edge of a craze. As the nucleated crack increases in size, new craze matter is generated at the craze tip. Crazes are very thin, planar defects; therefore, they form a very flat and smooth fracture origin, commonly known as the mirror zone or mirror region, which actually contains craze remnants. Because of the presence of a thin layer of highly oriented polymer with a different refractive index from that of the bulk, interference color fringes are frequently observed in the mirror region when the specimen is viewed in visible light. Figure 18 shows the fracture surface area near the fracture origin of a PC specimen after Izod impact. The mirror zone appears in the lower central portion of Fig. 18 in the shape of an ellipse.

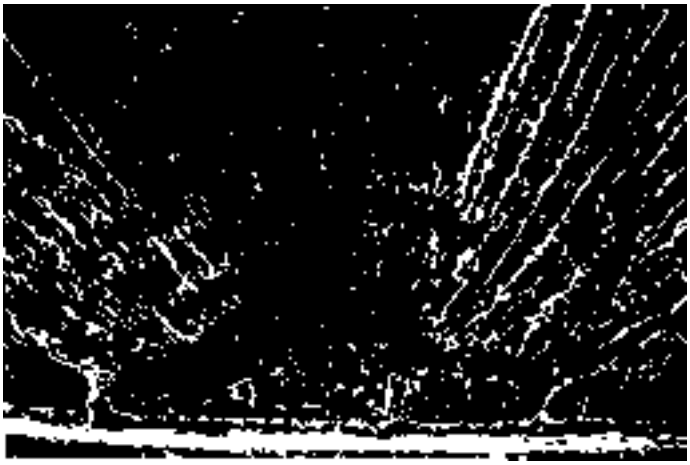


Fig. 18 Polycarbonate fracture surface showing mirror zone, mist and hackle regions, and Wallner lines. 14×

The crack initiation stage of fracture is sometimes referred to as nucleation, which is followed by slow, subcritical crack growth. As the newly formed crack increases in size, a critical size is eventually reached at the point when cracking becomes unstable overload fracture at a very rapid rate. The boundary of the mirror region marks the transition of crack velocity from a slow, stable extension to sudden acceleration, which leads to catastrophic fracture. In terms of fracture mechanics, the condition of crack instability can be described by:

$$K_{IC} = Ys_f\sqrt{a_c}$$

where K_{Ic} is the plane-strain fracture toughness dependent on temperature and environment, Y is the geometric factor dependent on loading and specimen and crack geometry, σ_f is the stress at fracture, and a_c is the critical crack size.

In other words, brittle fracture is said to occur when the stress intensity factor, K_I, exceeds the fracture toughness, K_{Ic}, of the material. Because the stress intensity factor depends on both the prevailing stress and crack geometry, the condition for brittle fracture can be satisfied by different combinations of stress and crack sizes. When the stress is high, the crack size at the moment of instability is expected to be smaller than when the stress is low. This suggests that the magnitude of the stress at fracture can be estimated by measuring the size of the slow-growth region.

According to this condition of instability, the mirror region size should be inversely proportional to the square of the fracture stress. In practice, however, this is of limited utility unless the other conditions of fracture, such as temperature and environment and their effects on fracture toughness, are also precisely known.

Mist Region. Surrounding the mirror zone is an area known as the mist region, commonly found in glass. It is also observed in some glassy polymers, such as PS, PMMA, and PC. This region is typically a flat, smooth area that is essentially featureless, except for a slight change in surface texture resembling a fine mist. Figure 19 shows the mist region immediately adjacent to the mirror zone at the fracture origin of a PC specimen after Izod impact. In polymers, mist regions are not necessarily confined to the vicinity of the fracture origin but can be observed elsewhere on the fracture surface.

Fig. 19 Fracture initiation region of polycarbonate specimen after Izod impact showing mirror zone and mist region. 27×

Hackle Region. Unlike mirror and mist regions, which are relatively smooth surface features, hackle regions are particularly rough surfaces. They are easily recognized by the outward divergent lines pointing along the crack propagation direction (Fig. 18, 19).

Hackle regions are associated with a more violent stage of fracture in which a large amount of strain energy is absorbed through both plastic deformation and the generation of new fracture surface areas. The hackle region shown in Fig. 19 is slightly magnified in Fig. 20 to illustrate ductile shear yielding.

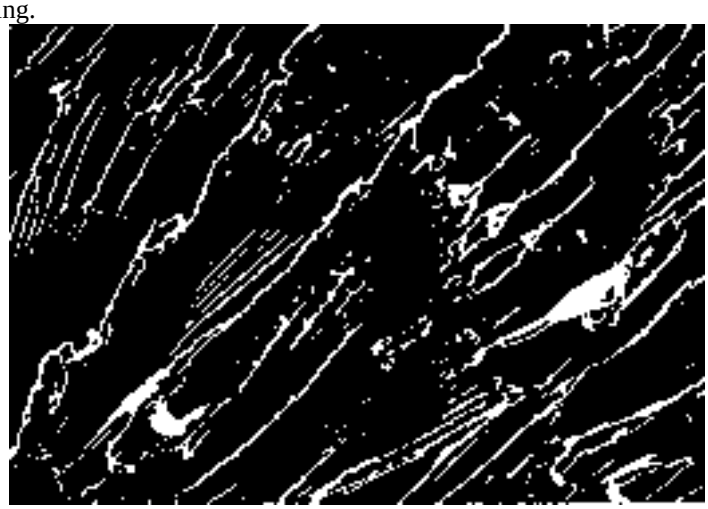


Fig. 20 Hackle region from Fig. 19 showing ductile shear yielding and crack front branching. 65×

Hackle regions tend to appear in areas where the stress field is changing rapidly (either in direction or magnitude) or when the stress state changes from one of plane strain to plane stress. For example, they are frequently observed in the region of a specimen subjected to bending on the original compression side of the specimen but changes to tension as the crack approaches from the tension side. Figure 21 shows the hackle region in the last stage of fracture of a PC specimen after Izod impact. The stress state rapidly changes from plane strain to plane stress as the crack approaches the final ligament of the specimen.

Fig. 21 Hackle region in final ligament of polycarbonate specimen after Izod impact. 14×

As previously stated, cracking is a stress-relief mechanism. Elastic strain energy is released during crack extension, providing the crack driving force. As the driving force increases, the crack is driven to higher velocities. As the velocity of propagation approaches the limiting velocity in the material, the rapidly moving crack tends to branch into two or more cracks, thus increasing the rate of energy dissipation by creating additional fracture surface areas.

For a material incapable of plastic deformation, such as glass or a polymer at a very low temperature with totally inoperative flow processes, crack branching is the only mechanism for increasing the rate of energy dissipation. For this reason, an overpowering blow to a sheet of glass or a polymer at a very low temperature will produce a shattering, brittle fracture.

Another example involves the dicing fracture of tempered glass. Without any external stresses, a sheet of tempered glass is already under a state of high residual stress corresponding to a high level of stored elastic strain energy. If a crack develops in a sheet of tempered glass, a large amount of strain energy is suddenly released. To dissipate this energy, the material responds with profuse crack branching, which creates many small glass fragments that have many more new surfaces than if only a single crack is formed.

The rough appearance of hackle regions can be partially explained by the following description of a similar mechanism, but in this case the crack branching occurs on a much finer scale. As the crack driving force increases and the crack velocity is sufficiently high, a single crack front begins to split up into many smaller crack fronts. These continue to propagate and diverge outward alongside each other, but at slightly different crack angles (Fig. 20, 22). Because this results in a series of smaller crack fronts propagating on slightly different crack planes or elevations, their side boundaries may overlap or even undercut each other. Sharp slivers are frequently observed on glass fracture surfaces when two crack fronts whose side boundaries have previously undercut each other reemerge onto the same crack plane.

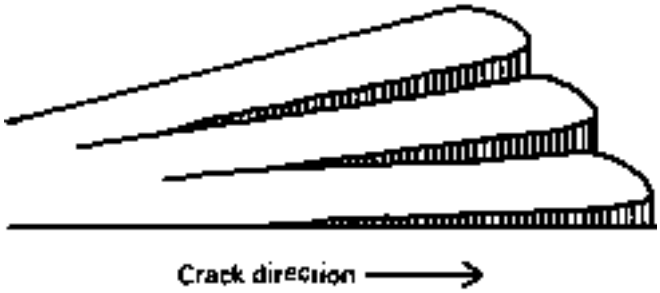


Fig. 22 Formation of hackle lines from crack front branching

In polymeric materials, ductile plastic deformation is evident in the hackle regions of a brittle fracture surface (Fig. 19 to 21). The ductile tearing mode of deformation is typically observed in the side boundaries of adjacent cracks propagating on slightly different crack planes (Fig. 20). While the cracks are propagating on planes normal to the tensile stresses, the material at the crack boundaries is subjected to shear stresses. This results in shear yielding, a form of plastic deformation that occurs with no volume change. A large amount of strain energy is dissipated by this process. Energy absorption by plastic deformation has been estimated to be greater by several orders of magnitude than through the creation of new surfaces alone. The rough appearance of the hackle regions is due to both the occurrence of plastic flow on the fracture surface and the presence of non-coplanar crack surfaces.

Because hackle regions are prominent features on a fracture surface, the crack propagation direction can be easily identified. The divergent nature of the hackle lines is advantageous in locating the crack origin. If the area of the fracture surface under examination is remote from its origin, some back tracing is necessary. Otherwise, hackle lines diverge from the fracture origin (Fig. 18, 19).

Wallner lines are sometimes observed as faint ridged pattern on otherwise smooth fracture surfaces. They resemble fatigue striations with periodic spacing, but are formed when stress waves reflected from the specimen boundaries interact with a propagating crack front. Very subtle changes in the fracture surface texture result when the stress waves produce a slight perturbation of the stress field ahead of the crack front. Wallner lines are shown near the central and upper portions of Fig. 18 and are interspersed among the hackle lines in Fig. 19 to 21.

Because Wallner lines are formed when reflected stress waves intersect a propagating crack, they are not true crack-front markings. In practice, however, stress wave velocities are so much higher than crack velocities that Wallner lines may be considered snapshots of the crack front during its propagation. Typically, they are curved markings similar to crack front markings, with the fracture initiation site located on their concave side. As such, they may be useful in locating the fracture origin.

Rib Markings. A true crack front marking is produced when a moving crack is stopped or arrested. These markings are commonly called crack arrest lines or rib markings because of their resemblance to curved rib bones. Figure 23 shows the fracture surface area in a PE natural gas pipe where the crack has stopped.

Fig. 23 Brittle fracture surface of a polyethylene gas pipe showing rib marking at crack arrest. 14.5×

Fatigue striations are also true crack arrest markings because they are formed during the repetition of crack extension and arrest. Figure 24 shows the fatigue fracture surface of a PC plumbing fixture subjected to the effect of a water hammer. In a laboratory specimen subjected to a well-controlled, uniform cyclic load, the fatigue striations are also well defined and nearly regularly spaced (Ref 8). This is generally not true of field fractures, in which the specimens are subjected to a more random loading history. The distance between crack arrests is generally more irregularly spaced because it is entirely determined by the conditions affecting the progression of the crack.

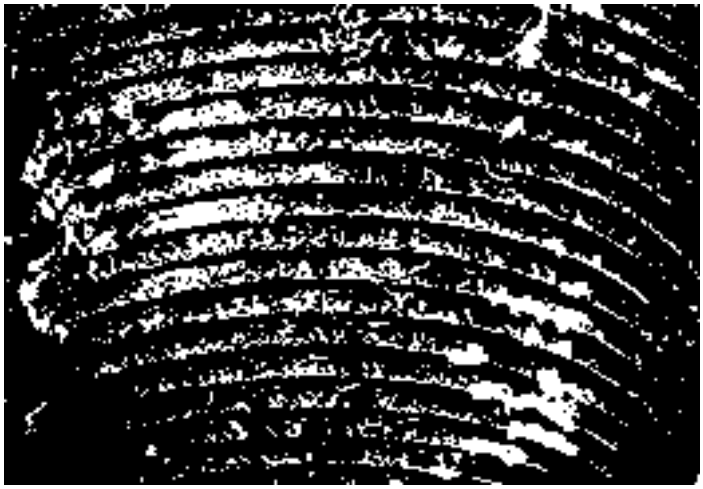


Fig. 24 Fatigue striations on the fracture surface of a polycarbonate plumbing fixture after field failure. 32×

Crack velocity changes also produce rib markings that are not as prominent as when the crack is totally stopped. Therefore, in this case, the term *crack arrest marking* is not entirely accurate even though it is more descriptive of the formation mechanism than the term *rib marking.* Figure 25 shows the fracture surface of the PE gas pipe in the vicinity of the origin near the inner pipe wall, where semicircular rib marks are seen expanding outward from the origin.

Fig. 25 Rib markings near the origin of polyethylene gas pipe fracture. 14×

Because both fatigue striations and rib markings are true crack front markings, the crack propagation direction at any point on the crack front can be determined by drawing the outward direction normal to the crack front. The fracture origin will be located by tracing back along the crack direction on the concave side of the curved markings.

Parabolic markings, which resemble the shape of a parabola, can also appear on the fracture surface of a plastic. Although often microscopic, they can be large enough to be observed with the unaided eye in certain plastics, such as PVC and ABS. Four parabolic markings are shown on the ABS fracture surface in Fig. 26. These markings are often useful surface features for determining the crack propagation direction when more prominent features are lacking.

Fig. 26 Parabolic markings on acrylonitrile-butadiene-styrene fracture surface. 12.5×

Parabolic markings are formed when a primary or main crack front intersects another crack that has initiated at a small distance just ahead of the main crack. Secondary crack origins arise at sites of local stress concentration because of material inhomogeneities and the rapidly rising stress field ahead of the primary crack. The intersection of the two crack fronts, frequently on slightly different crack elevations, produces a parabolic marking that diverges in the propagation direction of the main crack. The secondary crack origin can be found near the focus of the parabola. The primary crack origin, however, will be located on the convex side of the parabolic marking. The crack origin shown in Fig. 26 is near the lower central portion of the micrograph.

References cited in this section

7. C.B. Bucknall, Toughened Plastics, *Appl. Sci.*, 1977

8. R.W. Hertzberg and J.A. Manson, *Fatigue of Engineering Plastics*, Academic Press, 1980

18. P. So, Fractography, *Engineering Plastics*, Vol 2, *Engineered Materials Handbook*, ASM International, 1988, p 805

19. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage*, Prentice-Hall, 1981

20. J.G. Williams, *Stress Analysis of Polymers*, John Wiley & Sons, 1973

21. L.C. Sawyer and D.T. Grubb, *Polymer Microscopy*, Chapman & Hall, 1987

22. J.G. Williams, *Fracture Mechanics of Polymers*, John Wiley & Sons, 1984

Fracture of Plastics

Case Studies (Ref 5)

Several cases of field failure in various polymers are considered to illustrate the applicability of available analytical tools in conjunction with an understanding of failure mechanisms.

Example 2: Failure of an Irrigation Pipe. Figure 27 shows an optical micrograph of the fracture surface of an irrigation pipe made of medium-density PE that failed in service (Ref 23). This pipe was subjected to severe cyclic-bending strain of the order of 6% while under tensile stress of approximately 6.9 MPa (1000 psi) and a hoop stress of approximately 6.2 MPa (900 psi). These conditions of operation were far more stringent than those encountered in most applications of PE pipes. A subsurface imperfection in the pipe wall (dark diamond-shaped spot, Fig. 27) acted as a crack starter. Contrary to the dominant belief that pipe failure initiates from surface defects, this example indicates that a critical size flaw within the pipe wall can also initiate failure. This crack starter (flaw) was located closer to the outside wall, where compressive residual stresses may be dominant.

Fig. 27 Reflected-light optical micrograph of the fracture surface of medium-density polyethylene pipe. Arrows indicate the direction of crack propagation.

Concentric circular striations originate from the crack starter and grow simultaneously in radial and circumferential directions. It is well established that such striations represent crack-arrest lines, where the distance between two striations (a band) is due to a crack excursion. Thus, evolution of the band width reflects the nature of crack propagation. The band width measured from larger micrographs is plotted as a function of crack length in Fig. 28. A smooth transition is observed at a point, *T*, as catastrophic failure (pipe separation) is approached. This transition is indicative of considerable increase in crack speed and coincides with a transition in band geometry from circular to elliptical. Plausibly, this occurred when the crack-tip stress field interacted with the inside wall of the pipe. It should also be noted that maximum residual tensile stress dominates close to the inside wall.

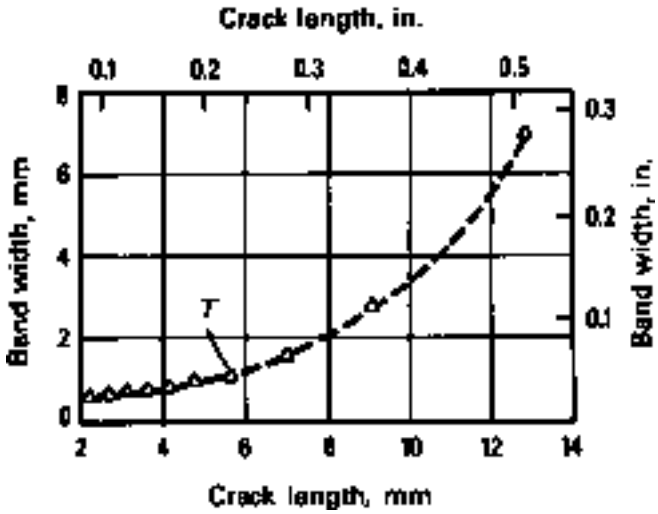


Fig. 28 Fracture band width as a function of crack length for the polyethylene pipe shown in Fig. 27

A similar transition has been noted to occur when the sufficient thermodynamic condition for crack instability is fulfilled. The major axis of the ellipse increases faster than the minor axis until no more striations are observed and ultimate failure results in large-scale yielding (about 50%) of the pipe wall (not shown in Fig. 28).

Whenever possible, similarity criteria should be established between the fracture behavior of a component in service and that observed in the laboratory. In this case, the band width appears to be a suitable candidate (Ref 24). This agreement indicates that discontinuous crack-growth band width, when available, can be employed as a similarity criterion to establish correspondence in loading history.

Example 3: Failure of a PVC Water-Filter Housing. Figure 29 shows an injection-molded PVC water-filter housing that fractured in service. An initial fissure (arrow *F*) is believed to have started first due to residual stresses developed during injection molding. Failure seems to have occurred due to fatigue crack propagation, as indicated by the presence of discontinuous crack-growth bands and their evolution. Although a tensile-stress component normal to the fracture surface was the dominant cause of failure, considerable triaxial stress seems evident in the early stages of fracture, as indicated by the successively smaller fissures to the left of the crack starter. As would be expected in PVC, catastrophic failure occurred by brittle failure as opposed to large-scale yielding, as in the PE pipe discussed in Example 2. This is evident from the relatively smooth appearance of the fracture surface beyond the last fatigue band in Fig. 29.

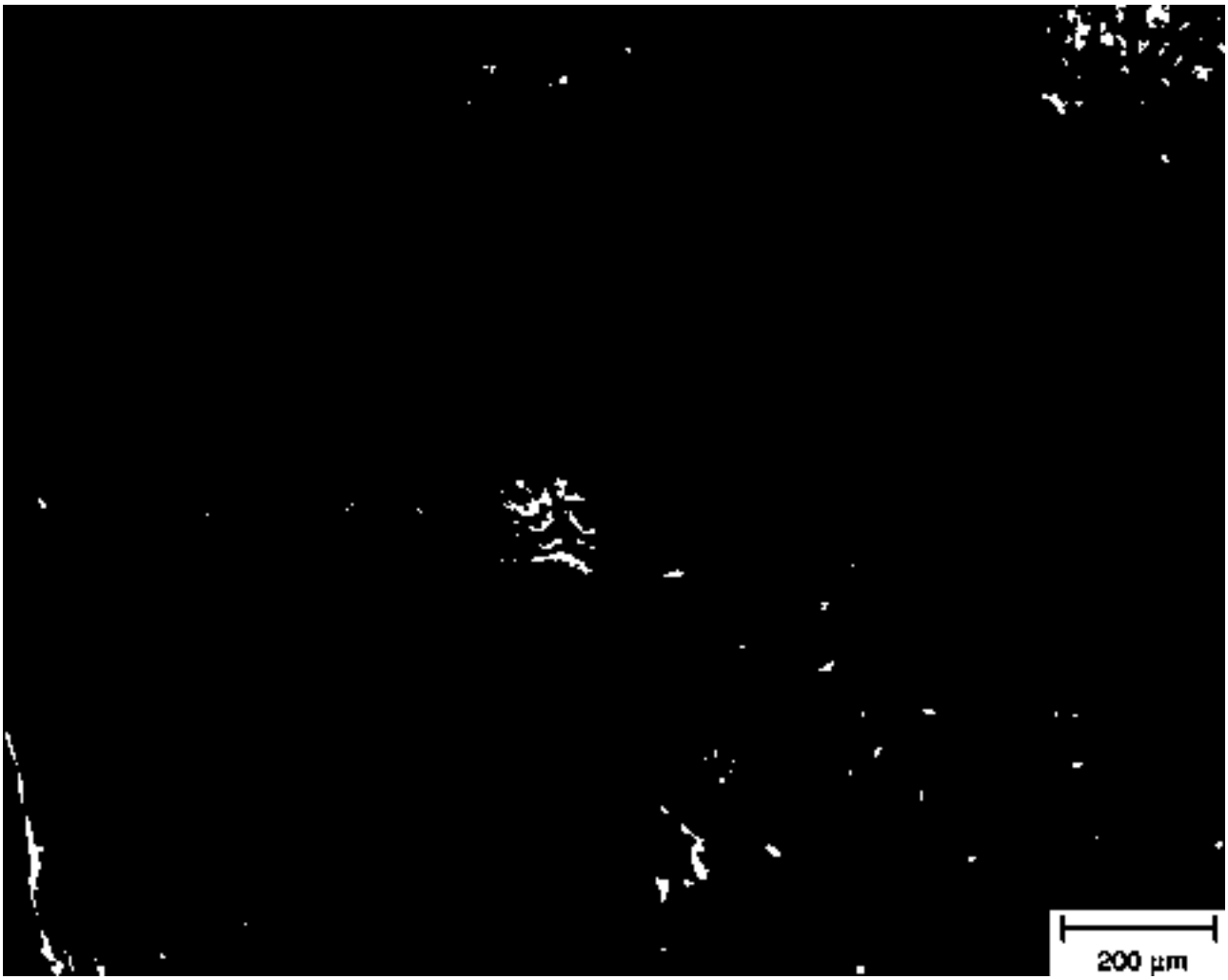


Fig. 29 Fracture in a polyvinyl chloride water filter. The fracture surface of the fatigue crack started from a fissure (arrow F). The lower dark zone is an artifact due to sectioning of the filter wall. 75×

References cited in this section

5. A. Moet, Failure Analysis of Polymers, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 758–765

23. K. Sehanobish, A. Moet, A. Chudnovsky, and P.P. Petro, *J. Mater. Sci. Lett.,* Vol 4, 1985, p 890

24. J.R. White and J.W. Teh, *Polymer,* Vol 20, 1979, p 764

References

1. W.D. Callister, *Materials Science and Engineering—An Introduction,* 4th ed., John Wiley & Sons, 1997, p 21

2. M.L. Weaver and M.E. Stevenson, Introduction to the Mechanical Behavior of Nonmetallic Materials, *Mechanical Testing and Evaluation.* Vol 8, *ASM Handbook,* ASM International, 2000, p 15

3. Deformation and Viscoelasticity, *Engineered Materials Handbook Desk Edition,* ASM International, 1995, p 393–409

4. I.M. Ward, *Mechanical Properties of Solid Polymers,* John Wiley & Sons, 1982

5. A. Moet, Failure Analysis of Polymers, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 758–765

6. A. Moet, Fatigue Failure, *Failure of Plastics,* W. Brostow and R.D. Coneliussen, Ed., Hanser Publishers, 1986

7. C.B. Bucknall, Toughened Plastics, *Appl. Sci.,* 1977

8. R.W. Hertzberg and J.A. Manson, *Fatigue of Engineering Plastics,* Academic Press, 1980

9. K. Matsushige, S.V. Radcliff, and E. Baer, *J. Mater. Sci.,* Vol 10, 1975, p 833

10. W. Schnabel, *Polymer Degradation,* Hanser International, 1981

11. Edward C. Lochanski, Solvent-Induced Cracking Failure of Polycarbonate Ophthalmic Lenses, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 493

12. H.H. Kauch, *Polymer Fracture,* Springer-Verlag, 1978

13. M.T. Takemori and R.P. Kambour, *J. Mater. Sci.,* Vol 16, 1981, p 1108

14. F.W. Billmeyer, Jr., *Text Book of Polymer Science,* John Wiley & Sons, 1984

15. J.G. Williams, *Fracture Mechanics of Polymers,* John Wiley & Sons, 1984

16. A. Chudnovsky, A. Moet, R.J. Bankert, and M.T. Takemori, *J. Appl. Phys.,* Vol 54, 1983, p 5562

17. N. Haddaoui, A. Chudnovsky, and A. Moet, *Polym. Mater. Sci. Eng.,* Vol 49, 1983, p 117

18. P. So, Fractography, *Engineering Plastics,* Vol 2, *Engineered Materials Handbook,* ASM International, 1988, p 805

19. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage,* Prentice-Hall, 1981

20. J.G. Williams, *Stress Analysis of Polymers,* John Wiley & Sons, 1973

21. L.C. Sawyer and D.T. Grubb, *Polymer Microscopy,* Chapman & Hall, 1987

22. J.G. Williams, *Fracture Mechanics of Polymers,* John Wiley & Sons, 1984

23. K. Sehanobish, A. Moet, A. Chudnovsky, and P.P. Petro, *J. Mater. Sci. Lett.,* Vol 4, 1985, p 890

24. J.R. White and J.W. Teh, *Polymer,* Vol 20, 1979, p 764

Selected References

- M. Ezrin, *Plastics Failure Guide: Causes and Prevention,* Hanser/Gardner Publications, Cincinnati, 1996
- J. Moalli, *Plastics Failure Analysis and Prevention,* Society of Plastics Engineers, 2001

Fracture Modes and Appearances in Ceramics

James R. Varner, New York State College of Ceramics, Alfred University

Introduction

FRACTOGRAPHY is the “means and methods for characterizing a fractured specimen or component” (Ref 1). This includes the examination of fracture-exposed surfaces and the interpretation of the fracture markings as well as the examination and interpretation of crack patterns. This article deals with the former of these two parts of fractography. Techniques of fractography are presented, fracture markings are explained using glass and ceramic examples, fracture modes in ceramics are discussed, and examples of fracture origins are given. The approach is qualitative; readers who wish to learn more about the quantitative side of fractography of ceramics should consult the references given at the end of the article. The reference list also includes sources that will help readers explore various aspects of fractography of ceramics in more depth.

Reference cited in this section

1. “Standard Practice for Interpreting Glass Fracture Surface Features,” C 1256, ASTM

James R. Varner, New York State College of Ceramics, Alfred University

Techniques of Fractography

How does one go about examining fracture markings or crack patterns? The techniques of fractography provide the answer to that question. Observations of fracture surfaces may begin with the unaided eye and may include various types of microscopy: light microscopy, scanning electron microscopy (SEM), transmission electron microscopy, atomic force microscopy, and other specialized methods. The bulk of the observations by far are done with low-magnification optical microscopy, often using a stereographic microscope, and with higher-magnification SEM. The primary goal of most fractographic observations is to locate and identify the fracture origin. The term “fracture origin” denotes both the location where fracture began and the discontinuity (crack, pore, large grain, machining damage, etc.) that concentrated the stress at that location. As a general rule, low-magnification observation is most effective for seeing the fracture markings that allow determination of the location of the origin, while high-magnification observation is most effective for identification of the origin. As an example, Fig. 1 shows a low-magnification image, taken with an optical microscope, of the fracture surface of a glass plate. The location of the origin is shown by the fracture markings, especially the Wallner lines (see the section “Wallner Lines” for further discussion) that envelop the origin. In Fig. 1, the curved arcs (the Wallner lines) indicate that the origin is toward the top of the image, along the edge of the fracture surface. There seems to be extensive surface damage along this edge, but it is not seen clearly. The higher-magnification SEM image in Fig. 2, taken at the right side of the origin region, reveals details of the contact damage on the surface. This image leaves no doubt that there is a complex system of surface cracks created by mechanical contact. The depth of focus of the SEM is particularly advantageous for revealing the complex morphologies of fracture origins.

Fig. 1 Fracture surface of a glass plate broken in bending. The fracture origin is at the top edge of the fracture surface, to the left of the center of the image. Optical microscope; reflected light; picture width ~1 mm

Fig. 2 High-magnification view of the glass plate shown in Fig. 1. There is a complex system of cracks created by mechanical contact of the glass surface with a sharp, hard object. SEM; picture width ~100 µm

Another way to locate the fracture origin, especially in strength tests, is to make use of the overall crack pattern. This can be facilitated in bend tests by applying tape to the compression side of the specimen before breaking it. The tape holds the pieces together without influencing the strength results. For example, Fig. 3 shows four disks of MgF_2, broken in flexure (the ring-on-ring test, in which the disk is placed on a ring and loaded to failure by pressing on the other surface with a smaller, concentric ring; equibiaxial tension is produced within the smaller ring, on the opposite surface), which failed at different stresses. The branching pattern is used to determine where the origin is located. Also, the branching is more extensive when the failure stress is greater.

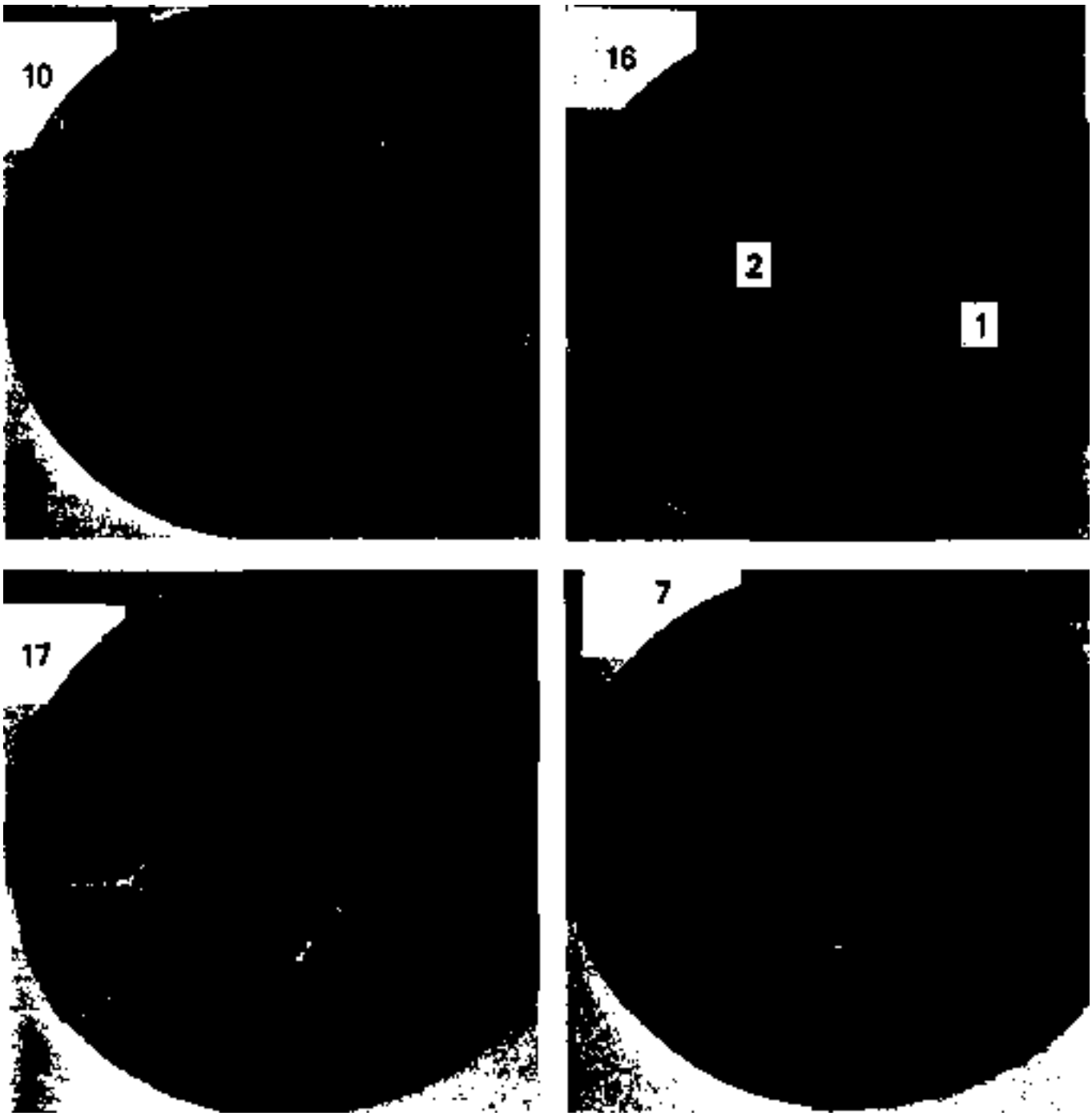


Fig. 3 Crack patterns of four specimens of MgF_2 broken in biaxial flexure. The more extensive branching is associated with higher stress at failure. Arrows indicate locations of fracture origins. Disks are ~50 mm (2 in.) in diameter. Source: Ref 2

When examining the origin, both fracture surfaces should be looked at, because the key visual clue, such as an inclusion, may be on only one of the surfaces. An example of this is shown in Fig. 4. Figure 4 shows both halves of a silicon nitride bar that was broken in bending. The tensile surfaces are touching. At this low magnification, the location of the origin is clear on both halves. However, on the one half, it appears that the origin is a circular (spherical) hole, while on the other half, it is clear that the origin really is a solid inclusion. Another reason for looking at both halves is that postfracture damage often happens right at the origin, causing part or all of the origin to be missing on one or both pieces. Looking at both makes this kind of postfracture damage more obvious.

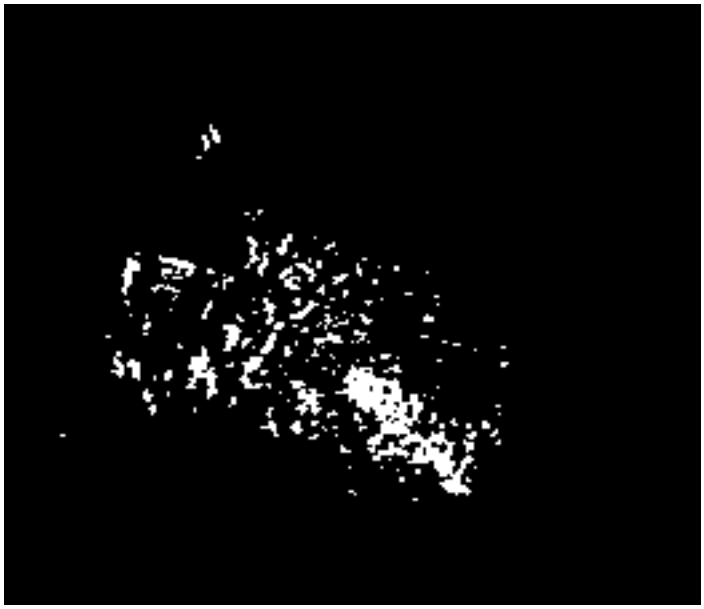


Fig. 4 Both halves of a silicon nitride bar broken in bending. The tensile surfaces are in contact with each other. The fracture origin appears to be a hole (pore) in the bottom piece, but the origin is really an inclusion, as seen on the top piece. Optical microscope; reflected light; picture width ~4 mm

Adjustment of the illumination when using an optical microscope is critical to seeing fracture markings. Figures 5 and 6 show the fracture surface of a silicon nitride rod that was broken in uniaxial tension. In Fig. 5, the light was nearly normal to the fracture surface, while in Fig. 6 the light was coming from the side, nearly parallel with the fracture surface. The fracture markings and the location of the fracture origin are much more obvious in Fig. 6 than in Fig. 5. As a general rule, oblique illumination, such as used to take the image shown in Fig. 6, should be used with ceramic specimens. However, one must always be willing to adjust the illumination in order to provide the best view of the fracture markings. Fracture origins, in particular, are often best seen with vertical illumination, because origin cracks are often fairly flat and highly reflective, especially when transgranular fracture is present (see the section “Transgranular Fracture” for a further discussion).

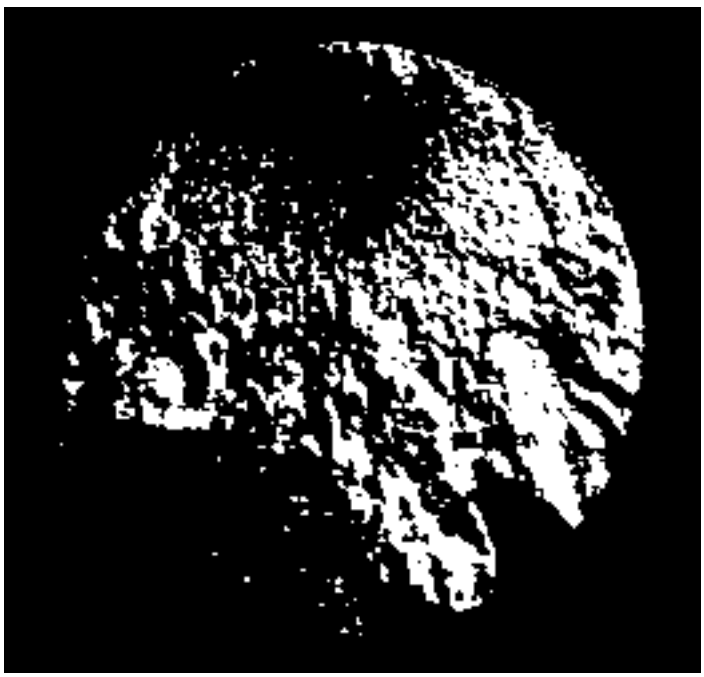


Fig. 5 Silicon nitride rod broken in uniaxial tension. Fracture origin is at the top of the image. Optical microscope; reflected light (direct illumination); picture width ~5 mm

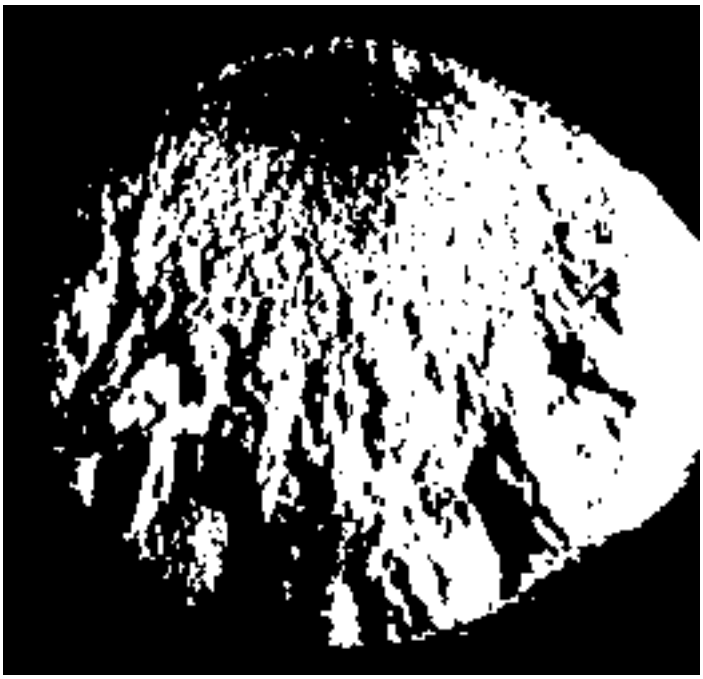


Fig. 6 Silicon nitride rod broken in uniaxial tension. The same rod shown in Fig. 5. Fracture origin is at the top of the image. Optical microscope; reflected light (oblique illumination); picture width ~5 mm

In the SEM, imaging techniques need to be matched to the specimen for best results, and stereographic pairs are extremely helpful in revealing the surface topography. With most SEMs, the fracture surfaces of ceramic specimens need to have a thin coating with a conductive material, usually a metal, to get the best images. Thin metal coatings (on the order of 10 nm, or 100 Å), produced by sputtering or evaporation, also help in getting high-quality images in optical microscopes. The reason is that the coatings eliminate subsurface scattered light, thus making it easier to see markings on the fracture surface.

Replicas can be used to advantage in fractography. Replicas are another way to eliminate effects of subsurface scattered light. They also provide a means for examining fracture surfaces of large pieces without having to cut them down to size. Cutting takes time, may contaminate the fracture surface, valuable information may be lost, and there may be prohibitions against doing anything destructive to the fracture pieces. It is beyond the scope of this article to discuss replicating procedures. However, Fig. 7 shows an image taken of a cellulose acetate replica of a fracture surface of an electrical porcelain insulator. Fracture markings are readily seen, especially in the glaze but even in the body. The wake hackle lines (see the section “Gull Wings and Wake Hackle”) at many of the bubbles in the glaze are especially good indicators of fracture direction. In this case, the fracture started in the body, or at the glaze-body interface, and moved into the glaze.

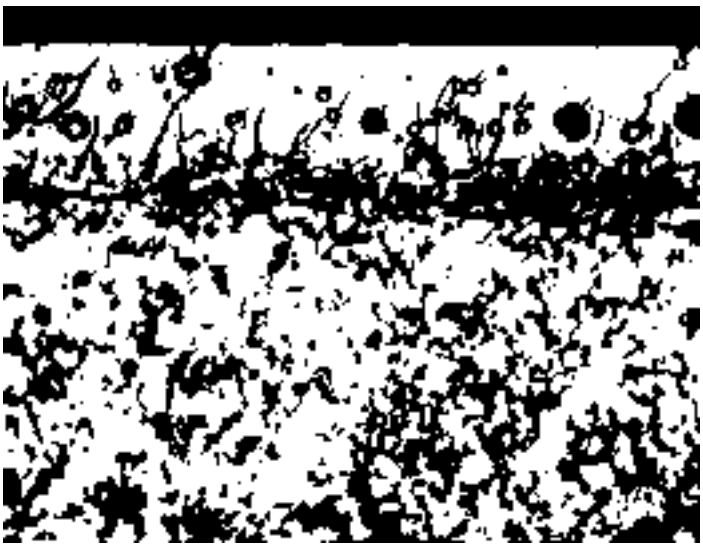


Fig. 7 Cellulose acetate replica of the fracture surface of a glazed electrical porcelain insulator. The fracture markings in the glaze, in particular, clearly indicate that the fracture started in the porcelain, not in the glaze (fracture moved from the porcelain into the glaze). Optical microscope; transmitted light; picture width ~1 mm

Regardless of the specific techniques that are used, any fractography should be well documented with notes, sketches, and photographs. Sketches are especially important, because they help the observer organize his or her thinking. They also provide a way to indicate locations of photographs, help in matching pieces for later reconstructions, and serve as reminders for subsequent observations of the fracture pieces.

Reference cited in this section

2. R.W. Rice, Ceramic Fracture Features, Observations, Mechanisms, and Uses, *Fractography of Ceramic and Metal Failures*, STP 827, J.J. Mecholsky, Jr. and S.R. Powell, Jr., Ed., ASTM, 1984, p 5–103

Fracture Modes and Appearances in Ceramics

James R. Varner, New York State College of Ceramics, Alfred University

Fracture Markings

A fracture tends to propagate as a spreading front, much like a wave front, that is, as a line. Markings are formed when something happens to disturb the propagation of this front, for example, if the fracture stops or if the local stress field changes direction. Typical markings on the fracture surfaces of ceramics are illustrated and explained in this section. The markings are shown and discussed using glass examples first. The reason is that typical silicate glasses are homogeneous and do not have the complications introduced by the microstructure of polycrystalline ceramics (grains, grain boundaries, second phases, pores, etc.). Therefore, it is much easier to see fracture markings on glass fracture surfaces. Examples of the markings on fracture surfaces of polycrystalline ceramics are then shown, so that the reader can see how the ceramic microstructure influences the appearance of the markings.

Brittle materials, such as ceramics, fail in tension. Even in those situations where the applied stresses are not tensile, resolved tensile stresses act on discontinuities in the material (cracks, pores, and so forth). The absence of significant plastic deformation (at normal ambient temperatures) means that the only way for the material to relieve the stress concentration is through fracture. As a fracture moves through a glass or a ceramic, its propagation direction and velocity are affected by the stresses it encounters. Central to understanding fracture markings and their interpretation in brittle materials is the so-called law of normal tension, which states that the fracture propagates normal to the direction of the local principal tension. In isotropic materials such as glasses, the law of normal tension alone governs crack propagation. In single crystals and in polycrystalline materials at the microscopic level, strong cleavage tendencies also play a role, particularly in the direction of propagation.

Mirror, Mist, and Velocity Hackle. Figure 8 shows a portion of the fracture surface of a glass rod broken in bending. The fracture origin is not seen in detail, but it is on the edge of the fracture surface, at the center of the smooth area. The fracture surface close to the origin is very smooth, and, when the illumination is just right, it reflects light similar to a mirror. Hence, this region is called the fracture mirror. A smooth surface such as this indicates that the fracture was moving relatively slowly as a single spreading front.

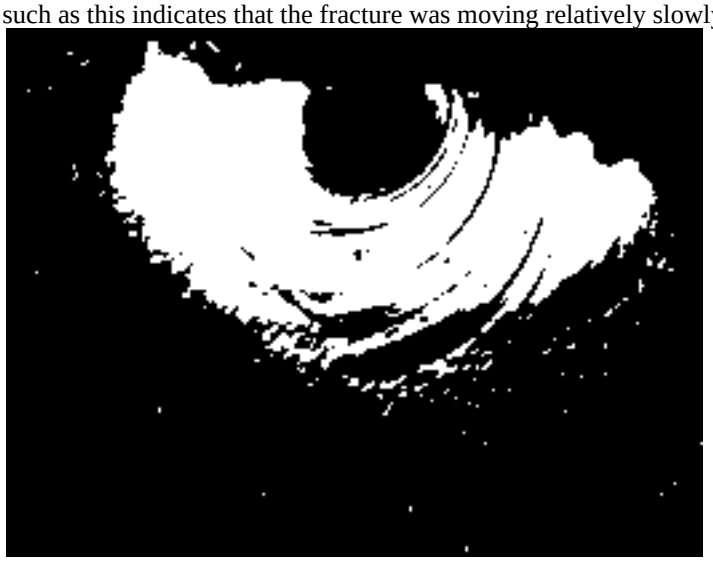


Fig. 8 Fracture surface of a glass rod broken in bending. The fracture origin is at the top center of the image. Wallner lines are seen in the fracture mirror, which is bordered by mist and velocity hackle. Optical microscope; reflected light; picture width ~2 mm

In this example, typical of a strength test, the fracture was accelerating as it moved away from the origin. There is a considerable amount of stored elastic energy in the specimen at the time of failure, and much of this energy is used to create the fracture surface. However, when the speed of the spreading crack approaches the terminal fracture velocity of the material, the single crack front begins to break up on a microscopic scale. This is seen on the fracture surface as a roughening of the surface, giving it a misty appearance. This microscopic breaking up of the crack front allows an increased dissipation of the stored energy. This is often not sufficient, and the single crack branches (splits) into two cracks. The fracture at this point has reached its terminal velocity (about one-half to two-thirds the velocity of the transverse stress wave in the material). Branching is shown on the fracture surface by the presence of large, daggerlike marks called velocity hackle. In Fig. 8, the mist and velocity hackle form the boundary of the fracture mirror. Velocity hackle is parallel to the direction of crack propagation. Depending on the amount of stored elastic energy at the time of failure, the fracture may branch many times. This is the reason for the formation of the many fragments that occur in high-stress failures of glasses and ceramics.

Figure 9 shows the fracture surface of a porcelain insulator that was broken in bending. A comparison with the fracture surface of the glass rod shown in Fig. 8 shows similarities but also some important differences. The mirror of this ceramic piece is flat and relatively smooth, as with the glass rod, but there is a small-scale roughness in the mirror region that is absent in the mirror of the glass rod. This is the influence of the microstructure. Similarly, the mist and velocity hackle are rougher and less distinct on the ceramic fracture surface. In fact, in many cases, one is hard pressed to define a region of mist hackle on ceramic fracture surfaces.

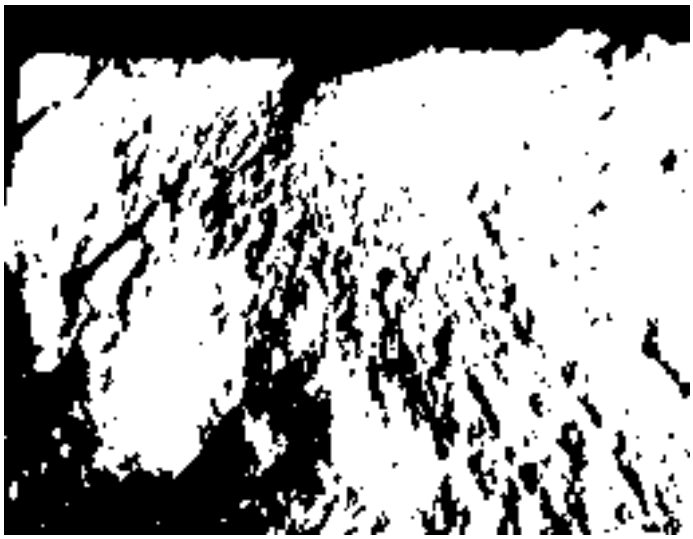


Fig. 9 Fracture surface of an electrical porcelain insulator. The fracture origin is at the top center of the image. The relatively smooth fracture mirror is bordered by mist and velocity hackle. Optical microscope; reflected light; picture width ~4 mm

Wallner Lines. Wallner lines (named for the scientist who first explained their formation) are lines on the fracture surface that are caused by disturbances of the fracture front by sonic waves. As shown in Fig. 10, Wallner lines resemble waves on the surface of a still pond when a stone is dropped in. Wallner lines are also seen in Fig. 1 and 8. The blow that broke the glass in Fig. 10 also created strong sonic waves that intersected the advancing crack front, causing momentary deviations in the direction of the local principal tension. The crack front followed these deviations, tilting out of its original plane momentarily and creating the wavy lines seen on this surface. It is important to note that Wallner lines do not give the shape of the fracture front, unless the sonic wave encounters the entire front at the same time, as, for example, when the fracture was moving slowly. However, the fracture origin is on the concave side of the Wallner lines, and this fact can be used to determine the direction of propagation and the location of the origin.

Fig. 10 Fracture surface of a piece of glass broken by striking it with a hammer. Origin is at the lower left; the wavelike lines are Wallner lines. Optical microscope; reflected light; picture width ~3 mm. Source: Ref 3

Wallner lines are often very subtle lines; that is, they are very small deviations (less than 1 μm, or 0.04 mil) from the plane of fracture. For that reason, they are much more difficult to see on ceramic fracture surfaces. Microstructure again plays a major role, in that Wallner lines are more likely to be seen on fracture surfaces of materials with small grain sizes. The issue is not one of resolution of the microscope but rather how strongly the microstructure affects fracture propagation. Wallner lines can be seen on glass fracture surfaces (and on some ceramic fracture surfaces) with the unaided eye, that is, at very low resolution. More intense (higher-energy) sonic waves are also more likely to create readily visible Wallner lines in ceramics. Figure 11 shows a ceramic fracture surface that has obvious Wallner lines as well as mist and velocity hackle. The material, electrical porcelain, has fine grains and also has a substantial amount of glassy phase. The shape of the Wallner lines and the fact that the mist and velocity hackle appear only along one edge indicate a stress gradient when failure occurred. It might be concluded that the piece failed in bending (see the section "Bending" for a more complete discussion of bending failure). However, this piece was a hollow cylinder that broke spontaneously while being cut with a diamond saw. Therefore, the stress gradient was present as a result of processing. In fact, the insulator was cooled too rapidly, resulting in compressive stress on the outside surface (at the top in Fig. 11) and tensile stress on the inside surface (at the bottom in Fig. 11).

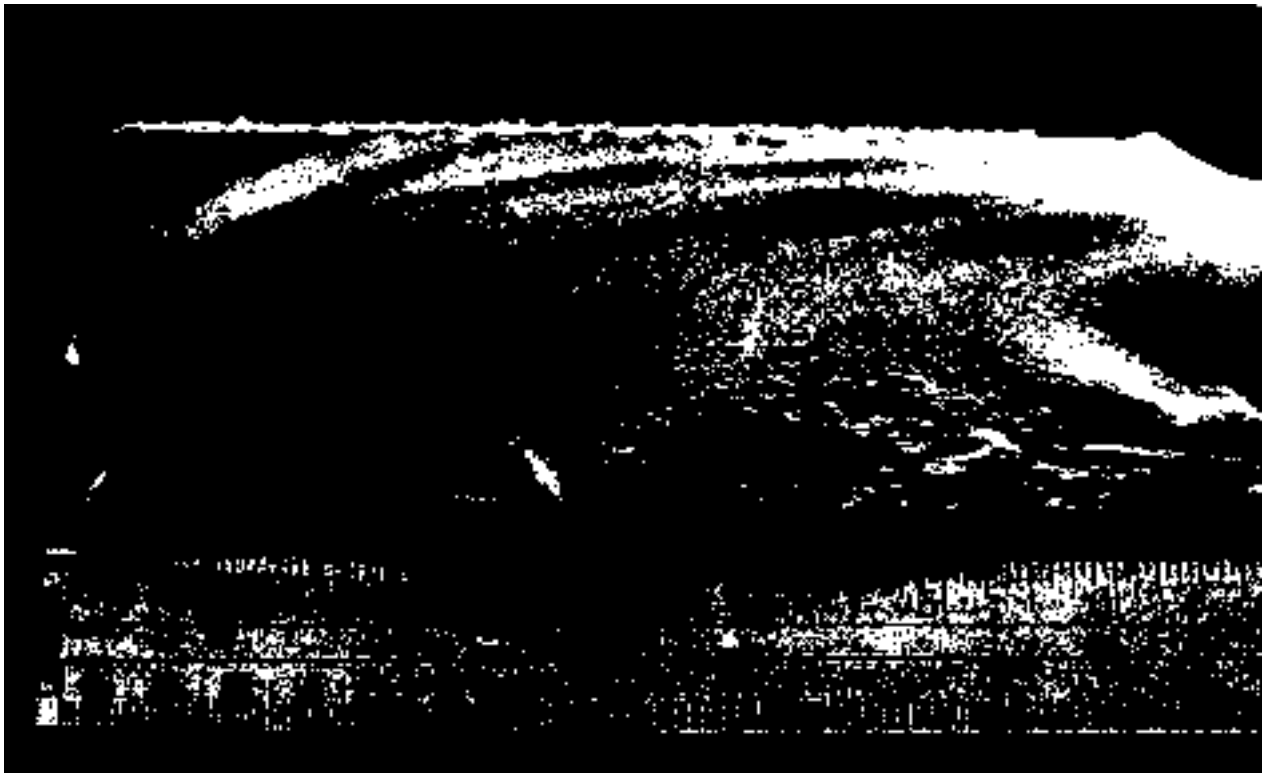


Fig. 11 Fracture surface of a porcelain insulator, which broke during cutting with a diamond saw. Fracture moved from right to left. Mist and velocity hackle and Wallner lines are readily seen in this fine-grained material. Camera image; picture width ~20 mm. Source: Ref 3

Gull Wings and Wake Hackle. This set of fracture markings is actually a combination of two other markings: Wallner lines and twist hackle. The appearance of gull wings and wake hackle is shown in Fig. 12. Part of the fracture surface of a glass capillary is shown; the dark circle is the hole of the capillary, and the fracture was moving from top to bottom, as can be seen from the intersecting Wallner lines. The mark at the bottom of the hole is twist hackle (here, called wake hackle), and the two strong Wallner lines running away from the wake hackle are the gull wings. These markings are associated with inclusions in materials—pores, bubbles, and solid particles—and are useful in determining the direction of propagation. As seen here, the gull wings and wake hackle are on the side of the inclusion that is farthest away from the fracture origin. The wake hackle (sometimes called a “tail”), therefore, points back toward the origin.

Fig. 12 Fracture surface of a glass capillary broken in bending. Fracture propagated from top to bottom. Gull wings and wake hackle formed when the front moved around

the hole of the capillary (the black circle in the picture). Optical microscope; reflected light; picture width ~4 mm. Source: Ref 4

Gull wings and wake hackle form when the fracture front encounters an inclusion. Even in the case of a pore or bubble, the front must move around the inclusion; that is, it does not simply pass through as though nothing was there. Instead, the front moves around the inclusion, effectively splitting into two fronts. When these two fronts meet on the opposite side of the inclusion, they nearly always are no longer traveling in exactly the same plane. Therefore, they overlap slightly, and the material separating them breaks through. This is the same mechanism that is described for the formation of twist hackle. The sudden breaking of the material between the two fronts generates a sonic pulse that interacts with the moving fracture, creating the gull wings.

Intragranular porosity is often seen in polycrystalline ceramics. When the fracture is transgranular (see the section "Transgranular Fracture" for further discussion), gull wings and wake hackle are generated at many of these pores. This can be the primary way in which the direction of fracture propagation is determined, thus aiding in locating and identifying the origin of failure.

References cited in this section

3. V.D. Fréchette, *Failure Analysis of Brittle Materials,* Vol 28, *Advances in Ceramics,* The American Ceramic Society, 1990

4. V.D. Fréchette, Fractology of Glass, *Introduction to Glass Science,* L.D. Pye, H. Stevens, and W. LaCourse, Ed., Plenum Press, 1972

Fracture Modes and Appearances in Ceramics

James R. Varner, New York State College of Ceramics, Alfred University

Fracture Modes

The term "fracture modes" is used in two ways in this section. First, it refers to how the fracture moves through the material on a microscopic scale, either through or around grains. Second, it refers to the way in which the piece is stressed and the fracture response to the stress, specifically, whether the piece broke in response to uniaxial tension, bending, or thermal shock. (Nominal external torsion and compression loading are excluded from this discussion, because the internal local state of stress that causes fracture in both cases is still tension.)

Transgranular Fracture. Fracture through grains is transgranular fracture. Examples are shown in Fig. 13 and 14. Transgranular fracture is favored when grain boundaries are strong in comparison with the strength of grains or when cleavage planes of neighboring grains match up with (or nearly match up with) each other (in other words, when misorientations between cleavage planes are small). A high percentage of transgranular fracture tends to produce relatively smooth fracture surfaces on a macro-scale. Fracture through individual grains is smooth on a microscale, as seen in Fig. 13 and Fig. 14, for example. Fracture through grains often leads to easily observed fracture markings—Wallner lines, twist hackle, and gull wings and wake hackle at pores—on the grain fracture surfaces, such as shown on the fracture surface of the large grain in Fig. 13. The large grain in Fig. 14 also shows some of these fracture markings. For this reason, transgranular fracture can prove to be very useful in making determinations about the direction of crack propagation. One needs to keep in mind that the direction of crack propagation through a single grain may differ from the overall fracture direction through the piece. Nonetheless, observation of fracture direction in several grains is usually enough to establish the overall direction of propagation in that region of the fracture surface.

Fig. 13 Fracture surface of a piece of polycrystalline alumina. Wallner lines, twist hackle, and gull wings are seen on the fracture surface of the large grain in the middle of the image. Fracture moved through this grain (transgranular fracture). Fracture around grains (intergranular fracture) is seen at the top of the image. SEM; picture width ~50 µm

Fig. 14 Fracture surface of a piece of polycrystalline alumina. The large grain in the center broke by transgranular fracture (note the fracture markings on its fracture surface). Most of the rest of the fracture shown in this image is intergranular. SEM; picture width ~50 µm

Intergranular Fracture. Fracture around grains is intergranular fracture. Examples are shown in both Fig. 13 and 14. Intergranular fracture is favored when grain boundaries are weak in comparison with the strength of grains or when cleavage planes of neighboring grains do not match up (in other words, when misorientations between cleavage planes are large). See the previous comment about cleavage plane versus fracture plane. A high percentage of intergranular fracture tends to produce relatively rough fracture surfaces on a macroscale. The rough surface of the region of intergranular fracture above the large grain in Fig. 14 illustrates this quite well. Intergranular fracture obscures fracture markings and provides no opportunity for having fracture markings within grains. As a result, it is usually much more difficult to discern fracture direction, locally or overall, on fracture surfaces that were created largely through intergranular fracture. Unfortunately, from the standpoint of performing fractography, polycrystalline ceramics with higher values of fracture toughness often have predominately intergranular fracture. Therefore, the benefits of higher fracture toughness are often accompanied with increased difficulties in finding and identifying fracture origins. For the same reason, those tests for measuring fracture toughness that rely on fractography may not work well with high-toughness ceramics.

Uniaxial Tension. Failure in pure uniaxial tension is rare in ceramics, except in carefully designed and carried-out laboratory tests on rods or fibers. More complex stress states cause most failures of pieces during manufacturing or in service, even though concentrated local tension is the ultimate cause of failure. Figure 6 shows an example of failure in pure uniaxial tension. Several characteristics of failure in uniaxial tension are seen in this example. Failure tends to occur in a plane (macroscale), namely the plane that is normal to the direction of the applied uniaxial tension (law of normal tension). Deviation out of this plane occurs only when branching occurs, that is, when failure occurs at high stress.

Fracture mirrors are symmetrical, and, if terminal velocity was reached, mist and velocity hackle occur uniformly along the mirror boundary. Figure 15 shows another silicon nitride rod that was broken in uniaxial tension. The origin is internal (note the mist/velocity hackle around the edge of the fracture surface). This is often the case in specimens broken in uniaxial tension, because the entire volume of the test section is under uniform tensile stress.

Fig. 15 Silicon nitride rod broken in uniaxial tension. Fracture origin is just to the left of the center of the rod. Optical microscope; reflected light; picture width ~5 mm

Bending. Failure of ceramics in bending (flexure) is very common. Most strength testing of ceramics is done using a bend test—three-point or four-point flexure (uniaxial stress state, used on bars or rods) or ball-on-ring or ring-on-ring flexure (biaxial stress state, used on disks or plates). Bending is a common type of loading of pieces in service.

Bending produces tension on one surface and compression on the other opposite surface of the piece. By its very nature, therefore, there is a gradient in stress through the thickness of the piece. Failure starts at, or very close to, the surface in tension, but the propagating crack moves into lower tension and toward compression. This causes the fracture surface to have some characteristics that are different from fracture surfaces produced in pure uniaxial tension. Figure 16 shows a silicon nitride rod broken in bending that has these characteristics. The fracture started out in a plane that was normal to the direction of tension, but it eventually turned sharply, only to make another turn as it approached the compression side of the rod. This produces a lip or curl on the compression side that, in fact, is called the cantilever curl. Mist and velocity hackle do not occur uniformly along the mirror boundary if failure occurred at a relatively low stress, such as in the example shown in Fig. 16. The absence of mist and velocity hackle in the middle part of the mirror indicates that the fracture did not reach terminal velocity along the middle part of the spreading fracture front. The reason is that the fracture was moving into lower tension. When the fracture origin is not at the location of maximum tension in bending failures, the mirror has an asymmetrical shape, because the fracture responds to the stress gradient. This is illustrated in the glass fracture surface shown in Fig. 8. These different characteristics of pieces broken in pure uniaxial tension and in bending are useful in failure analysis, of course. One can easily tell which of these two modes of failure created a particular fracture surface. For example, compare Fig. 15 (rod broken in uniaxial tension) with Fig. 16 (rod broken in bending).

Fig. 16 Silicon nitride rod broken in bending. Fracture origin is at the top of the image. The horizontal line near the bottom of the image is the cantilever curl, typical of bending failure as the fracture approaches the compression side. Optical microscope; reflected light; picture width ~5 mm

Thermal Shock. Sudden changes in temperature produce thermal gradients in pieces, and the thermal gradients, in turn, produce stress gradients in these pieces. A ceramic piece that is at a high temperature experiences surface tension, with balancing interior compression, when subjected to a downshock. Conversely, a piece at room temperature experiences surface compression, with balancing interior tension, when subjected to an upshock. Failure is more likely in downshock, because the surface often has contact damage in addition to inherent flaws. In thermal shock, the level of stress depends on a variety of factors: the magnitude of the temperature change, the thermal conductivity of the material, the shape and thickness of the piece, the rate of heat transfer between the piece and the quenching medium, and the coefficient of thermal expansion of the material.

The fracture markings are no different in thermal-shock failure than in failures produced by mechanical loading. However, there are some characteristics of crack patterns, especially in plates, that set thermal-shock failure apart from failures due to uniaxial tension or bending. Edge cracks in plates broken by thermal-shock failure are at right angles to both the edges of the plates and to the plate surfaces. Edge cracks in plates broken by bending may be at right angles to the edges but are nearly always at something other than right angles to the plate surfaces. Thermal-shock edge cracks in plates usually “meander” across the plate, and this includes cracks that have branched. The branching and the meandering are often remarkably symmetrical. Figure 17 shows a glass plate that failed in thermal shock, with the fracture origin being on one edge. The meandering crack pattern shown in this example is also seen in polycrystalline ceramic plates that have broken in thermal shock, when the origin is on one edge. Edges are particularly susceptible to contact damage, and thermal gradients and resulting stresses are more pronounced at edges.

Fig. 17 Glass plate broken in thermal shock. The fracture origin is at the lower edge of the plate. Note how the crack is normal to this edge. The "meandering" shape of the crack is typical of thermal-shock failure of plates. Camera image; picture width ~20 mm

James R. Varner, New York State College of Ceramics, Alfred University

Fracture Origins

The key to understanding strength and failure of test specimens and manufactured parts is identifying the cause of fracture, that is, the fracture origin. Cracks, pores, bubbles, inclusions, and grain boundaries are all discontinuities that serve as stress concentrators in glasses and/or ceramics and can, therefore, be fracture origins. This section presents examples of fracture origins grouped together according to their mechanism of formation: indentation or impact, machining, and processing. More examples can be found in Ref 2, 3, and 5, 6, 7, 8.

Indentation and Impact Sites. Surfaces of brittle materials are particularly susceptible to contact damage caused by hard objects, either during indentation or impact (or machining, covered in the section "Machining Flaws"). Figure 18 shows an indentation site caused by pressing a Vickers diamond onto the surface of a piece of glass. The cracks running from the corners of the indentation are normal to the surface and extend some distance into the glass. These are the cracks that become fracture origins when stress is applied to the piece. There are other cracks on the surfaces of the indentation, and the bright areas outside of the indentation are reflections from subsurface lateral cracks that have not made it up to the surface. When they do, they cause material removal. The indentation origin shown in Fig. 19 shows two radial cracks that together form a semicircular region, a so-called half-penny crack. Such a crack, or combination of cracks, is particularly effective in concentrating applied stress. Therefore, these radial cracks at indentation sites are often fracture origins.

Fig. 18 Vickers indentation site on a glass surface. The dark lines are radial cracks that are normal to the surface; the bright areas are subsurface lateral cracks that are nearly parallel to the surface. Optical microscope; reflected light (differential interference contrast)

Fig. 19 Vickers indentation origin in a glass plate. The indentation is seen in cross section at the center of the image. Two radial cracks together form a semicircular region, a so-called half-penny crack, that became the fracture origin. Optical microscope; transmitted light; picture width ~300 µm

Similar damage is produced when sharp particles impact glass or ceramic surfaces, as illustrated by the example in Fig. 20. The lateral cracks reached the surface, causing material to spall off. Complex cracking and loss of material occurred at the center of the impact. Subsurface cracking is not visible in this SEM photograph. However, Fig. 21 shows a similar impact site after the specimen was broken in a strength test. The specimen was tilted in the SEM to allow the original surface (at the left) and the fracture surface (at the right) to be seen simultaneously. The relationship between the impact site and the fracture origin is clear. The size of the origin flaw, as measured along the surface, is about the same as the size of the spalled-off region. This is typical, but there are many cases where the spalled-off region is much larger than the origin flaw, and vice versa. Note the twist hackle along part of the origin flaw.

Fig. 20 Impact site on a glass surface made by a 100 µm (4 mil) particle of SiC. SEM; picture width ~200 µm. Source: Ref 6

Fig. 21 Impact fracture origin in glass caused by impact damage from a 100 µm (4 mil) SiC particle. Specimen was tilted in the SEM to reveal original surface (left) and fracture surface (right). SEM; picture width ~300 µm. Source: Ref 9

Another impact origin is shown in Fig. 22 to make the point that impact origins often have a more complicated morphology than indentation origins. Here, the origin looks like a truncated half-penny. In fact, the origin crack is intersected by a second crack (the narrow, bright feature at the impact site), and the origin flaw itself continues behind the fracture surface seen in the photograph. Nonetheless, this other crack influenced how the fracture started; that is, it affected the strength.

Fig. 22 Impact fracture origin in glass caused by impact damage from a 100 µm (4 mil) SiC particle. Specimen was tilted in the SEM. Original surface is at left, fracture surface at right. SEM; picture width ~200 µm. Source: Ref 9

Blunt objects, such as spheres, also can produce contact damage, as shown in Fig. 23. The elliptical lines on the left side of the photograph are contact cracks on the original surface of the specimen. (They are actually circular but appear elliptical in this picture, because of the tilting of the specimen in the SEM.) These contact cracks run normal to the surface for a short distance before flaring out into cone-shaped cracks (so-called Hertzian cones, named in honor of Hertz, the scientist who described the stress field associated with blunt-object contact). This is what happens when a “BB” hits a window, for example. Portions of two of these cones are seen in Fig. 23. Notice the twist hackle and Wallner lines on the surface of the cone flaring out from the larger surface crack. Figure 24 presents a clearer picture of a Hertzian cone as a fracture origin. A single contact crack can be seen on the original surface, flaring out into the subsurface cone. The twist hackle seen at two locations makes it evident that the main fracture started from the base of the cone.

Fig. 23 Hertzian impact site in glass. Specimen was tilted in the SEM to reveal original surface (left) and fracture surface (right). SEM; picture width ~200 µm. Source: Ref 9

Fig. 24 Hertzian fracture origin in glass. Original surface is at left, fracture surface at right. Main fracture started from base of Hertzian cone. SEM; picture width ~300 µm. Source: Ref 9

Machining Flaws. When glasses and ceramics are machined or when they are cut using a diamond saw, extensive subsurface damage is produced. The crack systems are similar to those seen at indentation or impact sites but are still more complex, because many particles are involved instead of just one. Two examples of machining flaws are presented here. More examples can be found in Ref 5. The first is a machining flaw produced in glass when a diamond saw was used to cut a groove in the glass (Fig. 25). It appears that a series of subsurface cracks linked up to produce the large flaw that caused failure of this piece when external loading was applied. Note the similarity with the damage in the glass plate shown in Fig. 1 and 2. The second example shows a machining flaw in hot-pressed silicon nitride that resulted from diamond grinding of the surface (Fig. 26). Machining grooves on the original surface are evident in the upper part of the picture. Notice that this flaw is very long compared to its depth, and that it is aligned with the machining grooves. Both are features commonly observed when machining flaws are the origins of fracture.

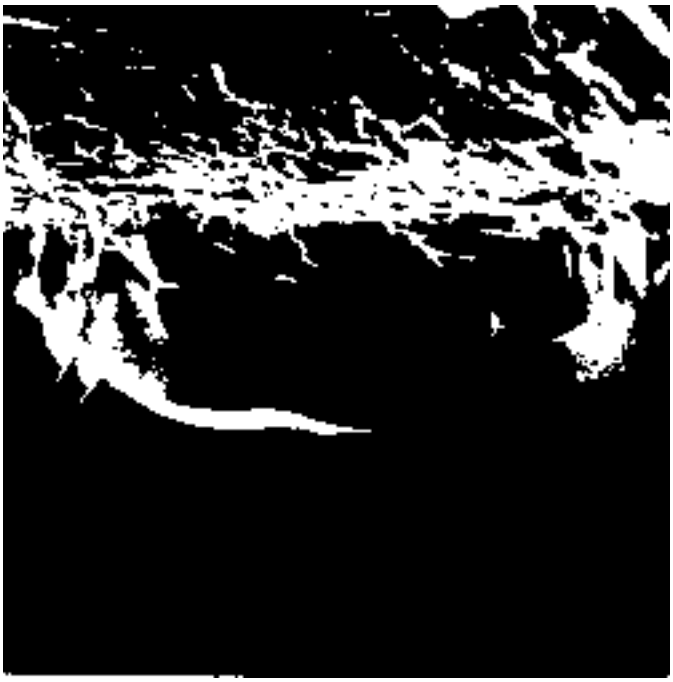


Fig. 25 Machining flaw as fracture origin in glass. Rough surface is the bottom of a groove cut by a diamond saw. SEM; picture width ~200 µm. Source: Ref 9

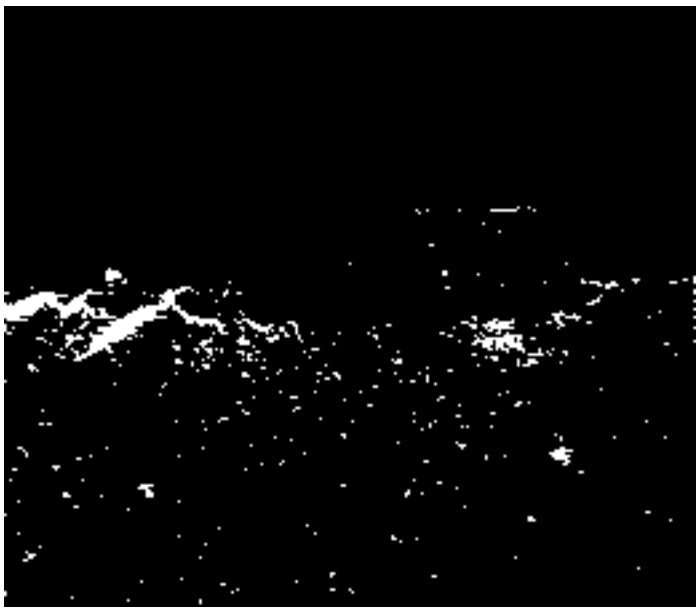


Fig. 26 Fracture surface of silicon nitride with machining flaw as origin. Specimen was tilted in the SEM showing the machined surface at the top and the fracture surface at the bottom. Machining flaw is aligned with grooves on the original surface. SEM; picture width ~1 mm. Source: Ref 7

Processing Defects. As used here, processing defects include all discontinuities in a material that can serve as fracture origins and that can be traced back to some step in manufacturing, except for those defects produced by contact damage. Examples include bubbles and crystalline inclusions in glasses, and pores, large-grained regions, and microcracks in polycrystalline ceramics. Although pores are not as effective in concentrating stresses as cracks, they are a frequent cause of failure in polycrystalline ceramics. The example in Fig. 27 shows a pore with an unusual shape that happened to be right on the tensile surface of this alumina specimen when it was loaded in bending. Apparently, something organic was in the green material (possibly a hair or a segment of poorly mixed binder), which burned out during sintering, leaving this large pore. Although the material overall is fine grained with only very small pores, this large pore is a processing defect that significantly reduced the strength of this particular piece.

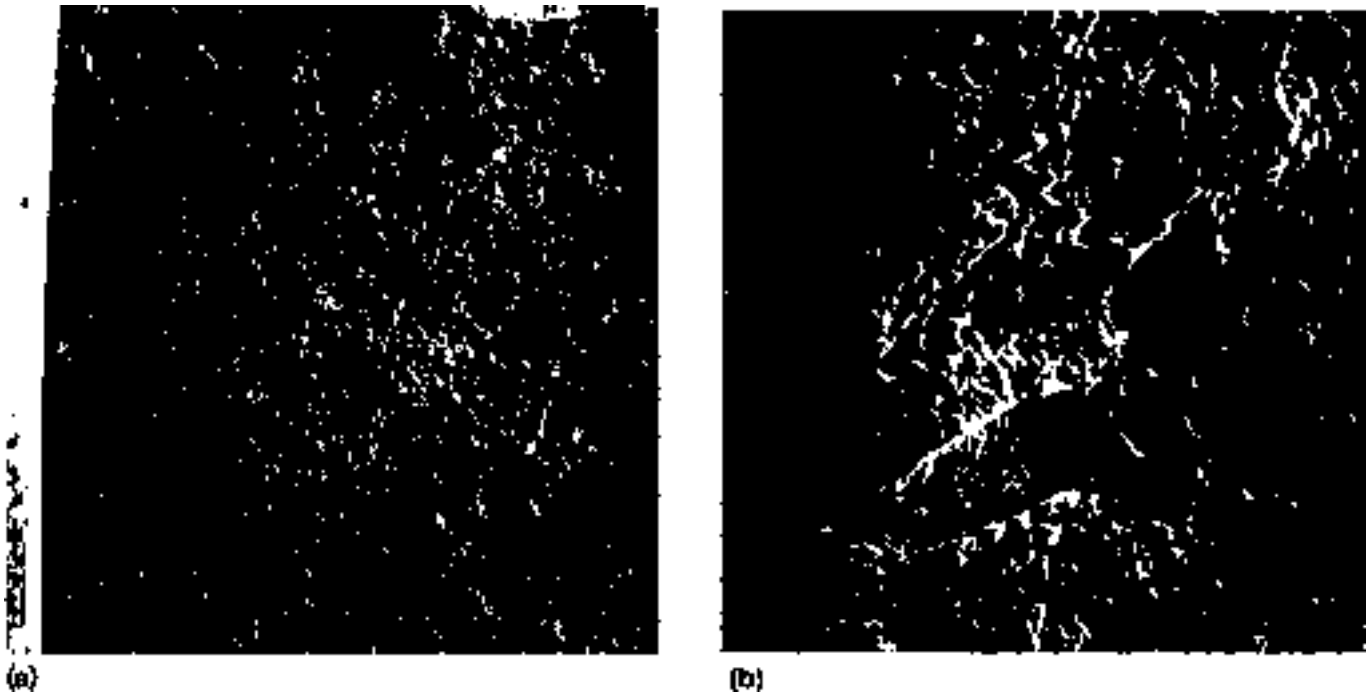


Fig. 27 Lower- and higher-magnification views of fracture surface of fine-grained alumina with a surface pore as fracture origin. SEM; picture widths (a) ~2 mm. (b) ~200 µm. Source: Ref 9

Figure 28 shows an example of a large-grained region as fracture origin. In the high-magnification view of Fig. 28, the origin is seen to be a region where the grains are larger than the average and where densification is very incomplete. Notice the transgranular fracture surrounding the large-grained area. It is common to see transgranular fracture close to the origin, where the fracture was traveling slowly.

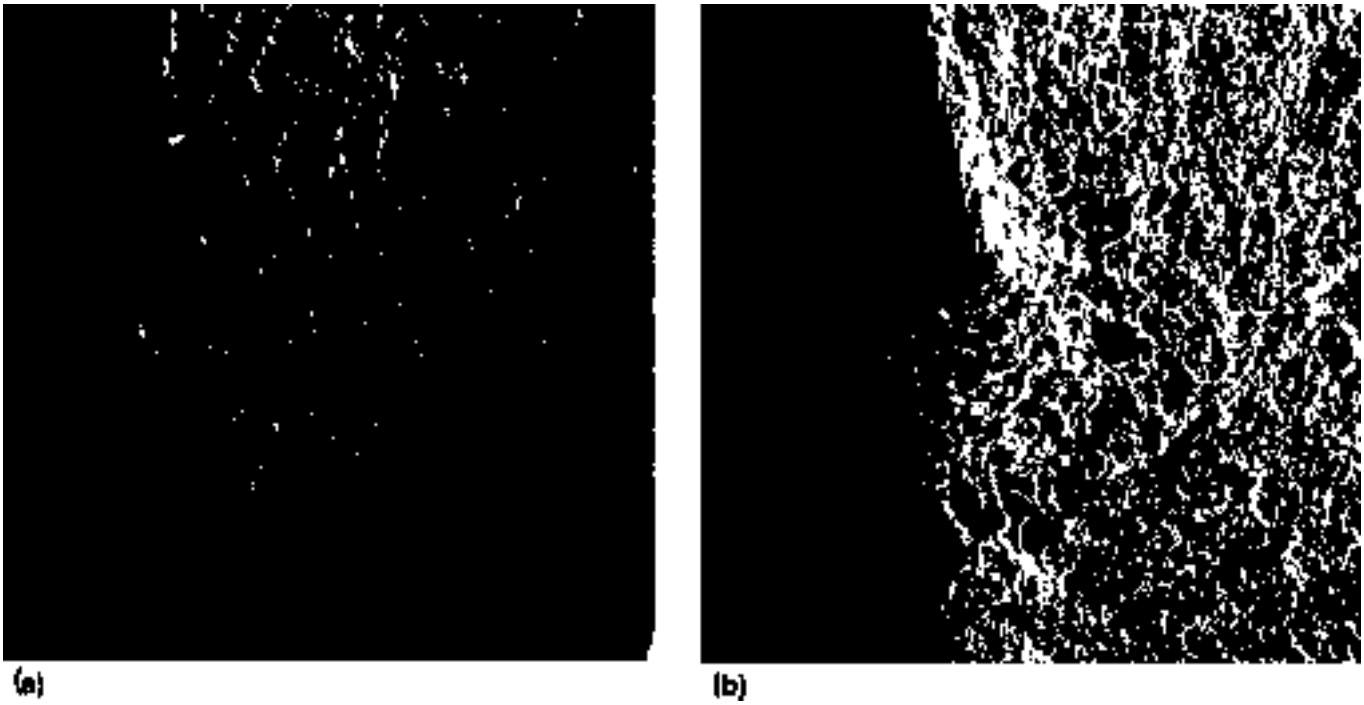


Fig. 28 Fracture surface of fine-grained alumina shown at low and high magnifications. Specimen was tilted in the SEM, allowing original surface (gray area at left) and fracture surface to be seen simultaneously. SEM; picture widths (a) ~2 mm, (b) ~300 µm. Source: Ref 9

Two other examples of processing flaws that became fracture origins (as well as one example of machining damage as a fracture origin) are shown in Fig. 29, taken from Ref 8. These are just two examples of the many types of processing flaws that are found in polycrystalline ceramics. Reference 5 is an excellent source for detailed information on fracture origins in ceramic materials.

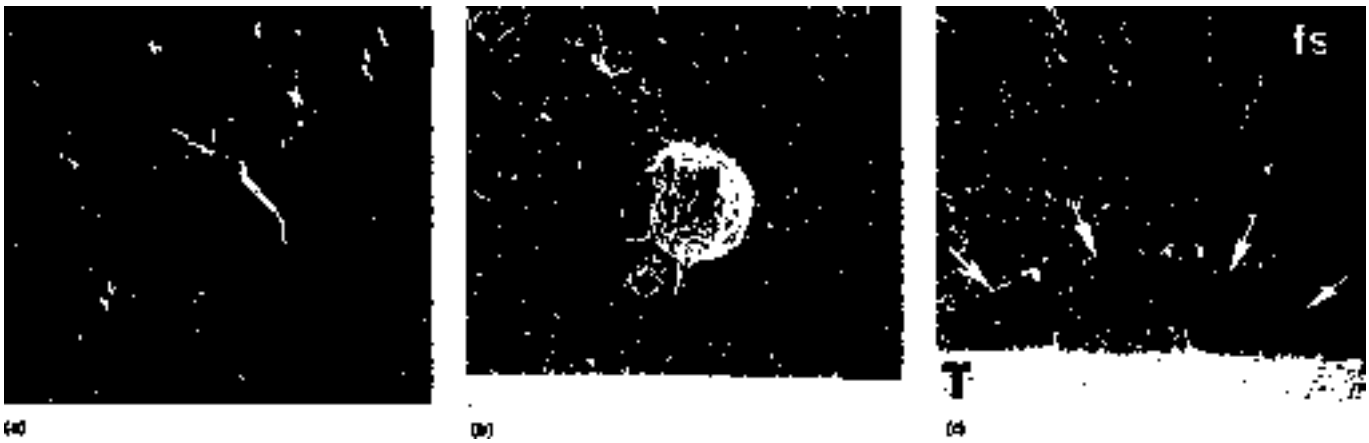


Fig. 29 Examples of strength-limiting defects in ceramics. (a) Silicon inclusion in reaction-bonded silicon nitride. (b) Powder agglomerate in sintered silicon carbide. (c) Machining damage in hot pressed silicon nitride. SEM; picture widths (a) ~150 µm. (b) ~300 µm, (c) ~150 µm. Source: Ref 8

Glazed ceramics are interesting because origins can be in the glaze, at the glaze-body interface, or in the body. It is, therefore, especially critical to locate and identify fracture origins in order to understand causes of failure. Figure 30 shows an example of an origin in the glaze. In this case, there was a large grain of unreacted quartz located between two large bubbles. There is cracking around such quartz particles, owing to stresses that develop because of thermal expansion differences between the quartz and the glaze. The large bubbles are also defects, but it was the quartz particle that caused fracture.

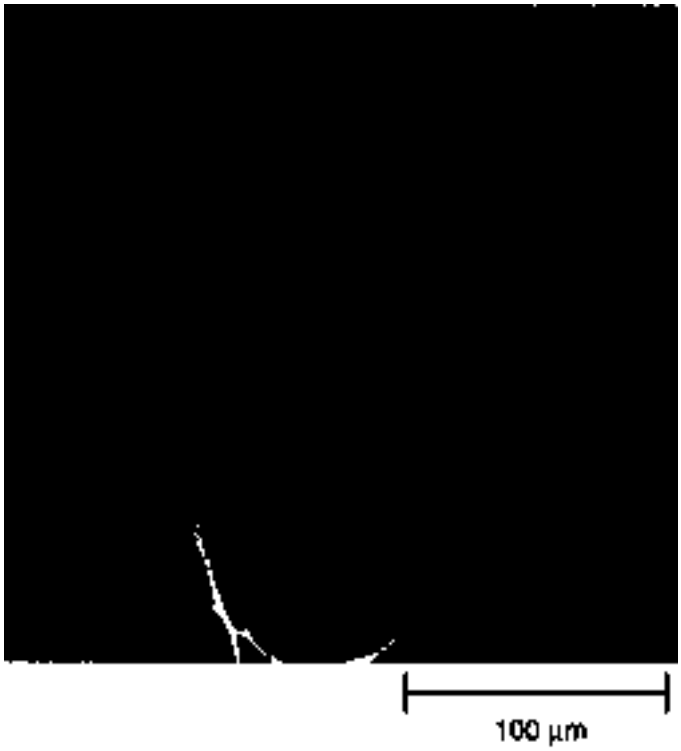


Fig. 30 Fracture surface of glazed porcelain. Fracture origin is a quartz grain between the two large bubbles in the glaze. SEM; picture width ~250 µm. Source: Ref 9

Fracture origins at the glaze-body interface are also very common. The glaze should be in compression and the body in tension. Reactions at the glaze-body interface may lead to local values of tension much higher than expected for the bulk of the body. Therefore, defects at the glaze-body interface are subjected to the combined effect of applied and residual stresses, making them frequent failure origins.

References cited in this section

2. R.W. Rice, Ceramic Fracture Features, Observations, Mechanisms, and Uses, *Fractography of Ceramic and Metal Failures,* STP 827, J.J. Mecholsky, Jr. and S.R. Powell, Jr., Ed., ASTM, 1984, p 5–103

3. V.D. Fréchette, *Failure Analysis of Brittle Materials,* Vol 28, *Advances in Ceramics,* The American Ceramic Society, 1990

5. "Standard Practice for Fractography and Characterization of Fracture Origins in Advanced Ceramics," C 1322, ASTM

6. J.R. Varner and H.J. Oel, Surface Defects: Their Origin, Characterization and Effects on Strength, *J. Non-Cryst. Solids,* Vol 19, 1975, p 321–333

7. E.S. Alfaro, J.V. Guiheen, and J.R. Varner, Insights Provided by Fractography in Strength Testing of Machined Si_3N_4 and Indented Al_2O_3, *Fractography of Glasses and Ceramics II,* Vol 17, *Ceramic Transactions,* V.D. Fréchette and J.R. Varner, Ed., The American Ceramic Society, 1991, p 485–508

8. G.D. Quinn, Strength and Proof Testing, *Ceramics and Glasses,* Vol 4, *Engineered Materials Handbook,* ASM International, 1991, p 585–598

9. J.R. Varner, Descriptive Fractography, *Ceramics and Glasses,* Vol 4, *Engineered Materials Handbook,* ASM International, 1991, p 635–644

Acknowledgments

The author thanks the following people for supplying some of the images used in this article: Roy Rice, George Quinn, Ward Votava, Christrophe Wagner, and Trevor Wilantewicz.

Fracture Modes and Appearances in Ceramics

James R. Varner, New York State College of Ceramics, Alfred University

References

1. "Standard Practice for Interpreting Glass Fracture Surface Features," C 1256, ASTM

2. R.W. Rice, Ceramic Fracture Features, Observations, Mechanisms, and Uses, *Fractography of Ceramic and Metal Failures,* STP 827, J.J. Mecholsky, Jr. and S.R. Powell, Jr., Ed., ASTM, 1984, p 5–103

3. V.D. Fréchette, *Failure Analysis of Brittle Materials,* Vol 28, *Advances in Ceramics,* The American Ceramic Society, 1990

4. V.D. Fréchette, Fractology of Glass, *Introduction to Glass Science,* L.D. Pye, H. Stevens, and W. LaCourse, Ed., Plenum Press, 1972

5. "Standard Practice for Fractography and Characterization of Fracture Origins in Advanced Ceramics," C 1322, ASTM

6. J.R. Varner and H.J. Oel, Surface Defects: Their Origin, Characterization and Effects on Strength, *J. Non-Cryst. Solids,* Vol 19, 1975, p 321–333

7. E.S. Alfaro, J.V. Guiheen, and J.R. Varner, Insights Provided by Fractography in Strength Testing of Machined Si_3N_4 and Indented Al_2O_3, *Fractography of Glasses and Ceramics II,* Vol 17, *Ceramic Transactions,* V.D. Fréchette and J.R. Varner, Ed., The American Ceramic Society, 1991, p 485–508

8. G.D. Quinn, Strength and Proof Testing, *Ceramics and Glasses,* Vol 4, *Engineered Materials Handbook,* ASM International, 1991, p 585–598

9. J.R. Varner, Descriptive Fractography, *Ceramics and Glasses,* Vol 4, *Engineered Materials Handbook,* ASM International, 1991, p 635–644

Fracture Modes and Appearances in Ceramics

James R. Varner, New York State College of Ceramics, Alfred University

Selected References

- V.D. Fréchette and J.R. Varner, Ed., *Fractography of Glasses and Ceramics,* Vol 22, *Advances in Ceramics.* The American Ceramic Society, 1988
- V.D. Fréchette and J.R. Varner, Ed., *Fractography of Glasses and Ceramics II,* Vol 17, *Ceramic Transactions,* The American Ceramic Society, 1991
- H.P. Kirchner and R.M. Gruver, Fracture Mirrors in Polycrystalline Ceramics and Glass, *Fracture Mechanics of Ceramics,* Vol 1, Plenum Press, 1974, p 309–322
- J.J. Mecholsky, Fracture Surface Analysis of Optical Fibers, *Ceramics and Glasses,* Vol 4, *Engineered Materials Handbook,* ASM International, 1991, p 663–668
- J.J. Mecholsky, S.W. Freiman, and R.W. Rice, Fracture Surface Analysis of Ceramics, *J. Mater. Sci.,* Vol 11, 1976, p 1310–1319
- J.J. Mecholsky, S.W. Freiman, and R.W. Rice, Fractographic Analysis of Ceramics, *Fractography in Failure Analysis,* STP 645, B.M. Strauss and W.H. Cullen, Jr., Ed., American Society for Testing and Materials, 1978, p 363–379
- T.A. Michalske, Quantitative Fracture Surface Analysis, *Ceramics and Glasses,* Vol 4, *Engineered Materials Handbook,* ASM International, 1991, p 652–662
- L. Orr, Practical Applications of Fractures in Glass Windows, *Mater. Res. Stand.,* Vol 12, 1972, p 21–23, 47
- J.B. Quinn, Extrapolation of Fracture Mirror and Crack-Branch Sizes to Large Dimensions in Biaxial Strength Tests of Glass, *J. Am. Ceram. Soc.,* Vol 82 (No. 8), 1999, p 2126–2132
- R.W. Rice, Fracture Topography of Ceramics, *Surfaces and Interfaces of Glass and Ceramics,* V.D. Fréchette, W.C. LaCourse, and V.L. Burdick, Ed., Plenum Press, 1974, p 439–472
- R.W. Rice and J.J. Mecholsky, The Nature of Strength Controlling Machining Flaws in Ceramics, *The Science of Ceramic Machining and Surface Finishing II,* B.J. Hockey and R.W. Rice, Ed., National Bureau of Standards, 1979, p 351–378
- J.R. Varner, V.D. Fréchette, and G.D. Quinn, Ed., *Fractography of Glasses and Ceramics III,* Vol 64, *Ceramic Transactions,* The American Ceramic Society, 1996
- J.R. Varner and G.D. Quinn, Ed., *Fractography of Glasses and Ceramics IV,* Vol 122, *Ceramic Transactions,* The American Ceramic Society, 2001

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Introduction

OVERLOAD FAILURES, from the perspective of materials failure analysts, refer to the ductile or brittle fracture of a material when stresses exceed the load-bearing capacity of the material from either excessive applied stress or degradation of the load-bearing capacity of the material from damage, embrittlement, or other factors. However, the definition of overload failure is not uniformly agreed upon. Many engineers limit the definition of overload failures to those in which the applied stresses were higher than anticipated in design. This popular misconception incorrectly presumes underdesign as the cause of overload failure, because it does not include failures where embrittlement or other causes result in fracture under normal loading.

The purpose of engineering failure analysis is to identify the root cause of component or system failure and to prevent similar occurrences. Underdesign is just one potential cause of an overload failure, and the time-consuming and potentially expensive process of redesign may indeed be necessary for prevention of further failures in some cases. However, from the standpoint of practical failure analysis, a wide variety of manufacturing and material characteristics can act singly or combine synergistically to reduce the strength,

ductility, and toughness of metallic materials. These factors must also be considered in the evaluation and prevention of overload failures. Therefore, throughout this article, the term overload failure is intended to imply fracture due to stresses exceeding the capacity of the material, whether or not the stress was higher than that anticipated by design.

The purpose of this article is to identify and illustrate the types of overload failures, which (for the purpose of discussion) are categorized as follows:

- Overload failure due to insufficient material strength and underdesign
- Overload failure due to stress concentration and material defects
- Overload failure due to material alteration

These categories are neither mutually exclusive nor collectively exhaustive, but are a useful means for pinpointing those areas of the design and materials selection that may require corrective scrutiny. These factors associated with overload failures are discussed in this article, and, where appropriate, preventive steps for reducing the likelihood of overload fractures are included. In particular, overload failure due to material alteration is very complex. No engineering component is immune from changes in material properties from mechanical strain, gross damage, and change in material microstructure or embrittlement. Complex engineering failures also may result from simultaneous multiple alterations.

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Fracture Modes and Mechanisms

When dealing with a fracture (not all failures are fractures), the failure analyst is called upon to identify the origin of the fracture and the mechanism(s) of crack propagation. In general, the three general “modes” or types of metal fracture include failures from overload, fatigue, and creep. These three modes of fracture are briefly summarized in Table 1, which includes concise summaries of contributing factors and the visual and microscopic aspects of each. In general, the growth or propagation of cracks during an overload fracture is more rapid than for the more progressive forms of cracking due to fatigue or creep. Hence, overload is described as “instantaneous” in Table 1 to distinguish it from the more progressive form of cracking due to fatigue or creep. These basic modes of fracture may also occur in combination, as when a fatigue or creep crack grows over time and ultimately leads to a rapid fracture through overload failure by ductile and/or brittle cracking.

Table 1 Fracture mode identification chart

Method	**Instantaneous failure mode**[(a)]		Progressive fracture mode	
	Ductile overload	**Brittle overload**	**Fatigue**	Creep
Visual, 1 to 50× (fracture surface)	• Necking or distortion in direction consistent with applied loads • Dull, fibrous fracture • Shear lips	• Little or no distortion • Flat fracture • Bright or coarse texture, crystalline, grainy • Rays or chevrons point to origin	• Flat progressive zone with beach marks • Overload zone consistent with applied loading direction • Ratchet marks where origins join	**• Multiple brittle appearing fissures** **• External surface and internal fissures contain reaction scale coatings** **• Fracture after limited dimensional change**
Scanning electron microscopy, 20–10,000× (fracture surface)	• Microvoids (dimples) elongated in direction of loading • Single crack with no	• Cleavage or intergranular fracture • Origin area may contain an	• Progressive zone: worn appearance, flat, may show striations at magnifications above 500×	**• Multiple intergranular fissures covered with reaction scale** **• Grain faces may show**

	branching • Surface slip band emergence	imperfection or stress concentrator	• Overload zone: may be either ductile or brittle	**porosity**
Metallographic inspection, 50–1000× (cross section)	• Grain distortion and flow near fracture • Irregular, transgranular fracture	• Little distortion evident • Intergranular or transgranular • May relate to notches at surface or brittle phases internally	• Progressive zone: usually transgranular with little apparent distortion • Overload zone: may be either ductile or brittle	**• Microstructural change typical of overheating** **• Multiple intergranular cracks** **• Voids formed on grain boundaries or wedge shaped cracks at grain triple points** **• Reaction scales or internal precipitation** **• Some cold flow in last stages of failure**
Contributing factors	**• Load exceeded the strength of the part** **• Check for proper alloy and processing by hardness check or destructive testing, chemical analysis** **• Loading direction may show failure was secondary** **• Short-term, high-temperature, high-stress rupture has ductile appearance (see creep)**	**• Load exceeded the dynamic strength of the part** **• Check for proper alloy and processing as well as proper toughness, grain size** **• Loading direction may show failure was secondary or impact induced** **• Low temperatures**	**• Cyclic stress exceeded the endurance limit of the material** **• Check for proper strength, surface finish, assembly, and operation** **• Prior damage by mechanical or corrosion modes may have initiated cracking** **• Alignment, vibration, balance** **• High cycle low stress: large fatigue zone; low cycle high stress: small fatigue zone**	**• Mild overheating and/or mild overstressing at elevated temperature** **• Unstable microstructures and small grain size increase creep rates** **• Ruptures occur after long exposure times** **• Verify proper alloy**

(a) Failure at the time of load application without prior weakening

Of course, the modes of fracture (overload, fatigue, and creep) cannot be considered independently of the underlying mechanisms of crack initiation and crack propagation. The occurrence and appearance of different fracture modes obviously depend on the mechanism(s) of crack initiation and propagation. For brittle and ductile cracking, distinguishing characteristics at different scales of observation (1× to >10,000×) are briefly summarized in Table 2. The following sections also briefly review some mechanistic aspects of ductile and brittle crack propagation, including discussion of mixed-mode cracking, which may also occur when an overload failure is caused by a combination of ductile and brittle cracking mechanisms. This section only describes the general aspects of fracture mechanisms, as more details are described in other articles in this Volume.

Table 2 Distinguishing characteristics of brittle versus ductile behavior depending on the scale of observation

Scale of observation	**Brittle**	Ductile
Structural engineer	Applied stress at failure is less than the yield stress	**Applied stress at failure is greater than the yield stress**
By eye (1×)	No necking, shiny facets, crystalline, granular	**Necked, fibrous, woody**
Macroscale (<50×)	“Low” RA or ductility	**Medium to high RA**
Microscale, scanning electron microscopy (100–10,000×)	Brittle microprocess, cleavage or intergranular	**Ductile microprocess, microvoid coalescence**
Transmission electron microscopy	**May have a large level of local**	**High amount of plasticity globally**

(>10,000×)	**plasticity**	

RA, reduction of area

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Ductile Overload Failures

Ductile overload failures are simply those that exhibit significant visible macroscopic plastic deformation. Applied stress and energy from an external source is required to propagate the fracture, rather than residual stresses alone. Ductile cracks blunt and stop when the stresses are dynamically reduced and insufficient energy to create further distortion remains. This ability to deform and absorb energy at the tip of a crack permits ductile materials to fatigue crack or crack in a stable (i.e., noncritical) overload manner over a longer period of time. Well-formed and easily identified striations are often present in fatigue fractures of ductile materials.

Due to necking or other visible signs of deformation evidenced by ductile materials prior to failure, ductile overload can often be less surprising and catastrophic than brittle failures. Ductile overload thus can be a more forgiving fracture mode than brittle fracture. Materials that exhibit yielding prior to fracture also can redistribute their loading to the adjacent support structure, avoiding system failure.

Due to their very nature, ductile overload failures often occur during manufacture or relatively early in service (for example, collapse of a roof due to snow loading). Sometimes this suggests underdesign (for example, by miscalculation of the service stresses; misuse or abuse in service; improper material selection, specification, or processing; poor manufacturing; or other reasons). Unfortunately, in some situations, ductile overload failures are routinely assumed to be design mistakes, a practice that ignores all of the other contributing factors to the failure, and may result in unnecessary or ill-advised corrective actions. Perfunctory failure analysis or incomplete root-cause investigation can be very misleading or even dangerous.

Ductile fracture occurs from the nucleation, growth, and coalescence of microvoids during deformation. Second-phase particles or inclusions typically provide the nucleation sites for initiation of the microvoids, which then grow and coalesce as deformation progresses further. Ductile fracture by void growth and coalescence can occur by two modes during tensile loading. Mode I is plane-strain fracture, where void growth and fibrous tearing occur along a crack plane that is essentially normal to the axis of the tensile load. The fracture appearance typically has a dull and fibrous appearance, as in the classic “cup and-cone” feature of ductile fracture in the simple cross section of a tensile testing specimen (for example, see Fig. 166 in the “Atlas of Fractographs” in *Fractography*, Volume 12) of the *ASM Handbook.*

The other fracture mode is plane-stress fracture (mode II fracture), which manifests itself as the “shear lips” in the cup-and-cone fracture specimens from tensile tests. During tensile loading, the fracture originates near the specimen center, where hydrostatic stresses develop during the onset of necking and where microvoids develop and grow. Multiple cracks join and spread outward along the plane normal to loading axis, as representative of mode I (plane-strain) crack propagation. When cracks reach a region near the outer surface, the mode of fracture changes to mode II (plane-stress) condition, where shear strain becomes the operative mode of deformation (as maximum stresses occur along the shear plane in the basic principles of continuum mechanics). Thus, even though the overall applied stress is still a tensile load, deformation makes a transition to the shear plane in the outer regions of the specimen. In this mode II region, voids grow in sheets at an oblique angle (~45° to the plane of the major fracture or loading axis) to the crack plane under the influence of shear strains. This type of shear-band tearing results in the classic shear lips of a ductile cup-and-cone tensile fracture specimen. Shear-band formation also commonly occurs in deformation processing, where friction and/or geometric conditions produce inhomogeneous deformation, leading to local shear bands. Localization of deformation in these shear bands also results in localized temperature increases that produce local softening.

Ductile fractures often progress as single cracks, without many separated pieces or substantial crack branching at the fracture location. They are characteristically flat and perpendicular to the maximum tensile stress for mode I (plane-strain) conditions, or at about a 45° angle to the maximum stress (although this angle may vary depending on material condition and loading condition). The crack profiles adjacent to the fracture are consistent with tearing. Ductile overload failures of engineering components typically exhibit the characteristics of dimple rupture, as discussed further.

The second mode of ductile fracture, shear-band formation, is a phenomenon associated more with deformation processes during metalworking, where localized and inhomogeneous deformation can potentially lead to crack formation. This concern of inhomogeneous deformation and shear-band formation in metalworking is not discussed in this article. However, shear-band formation may occur in dimple-rupture failures under certain circumstances. In some ductile materials, shear-band fractures may also show evidence of microvoid coalescence (MVC), typically associated with

dimple rupture. In pure metals and alloys refined to remove all impurities (where there is no ready supply of sites for microvoids to nucleate), the material may fail by adiabatic shear-band formation, resulting in a somewhat flat fracture surface devoid of dimples.

Transgranular dimple rupture due to MVC is the most prevalent fracture mechanism of ductile overload failures. During deformation, cavities (or microvoids) nucleate at small particles (such as inclusions and second-phase particles) when the matrix material separates from the particle by interfacial decohesion or when the inclusion or second-phase fractures. The level of extensional growth of the microvoids is dependent on the level of plastic strain the material can undergo prior to fracture.

The microscale fractographic features consist of ruptured dimples, and both diffuse and local necking may be macroscopically visible depending on the component geometry. Mechanical conditions and metallurgical features can influence the appearance of MVC. Examples are shown for a ferritic steel (Fig. 1a), ductile iron (Fig. 1b), and cast titanium alloy (Fig. 1c). The dimple features are evident in equiaxed form for tensile loading (Fig. 1a and b). Parabolic shear dimples occur from torsional loading (Fig. 1c) or from mode II (in-plane) shear. However, metallic materials can behave in a ductile manner, but not exhibit the classical dimple-rupture morphology, as shown in Fig. 1(d) for a cast aluminum alloy. The cast aluminum alloy fracture surface should show dimples in the aluminum matrix (original dendrites), but can be difficult to see because fracture surface morphology is impacted by cleavage in second-phase (silicon) particles and other second phases.

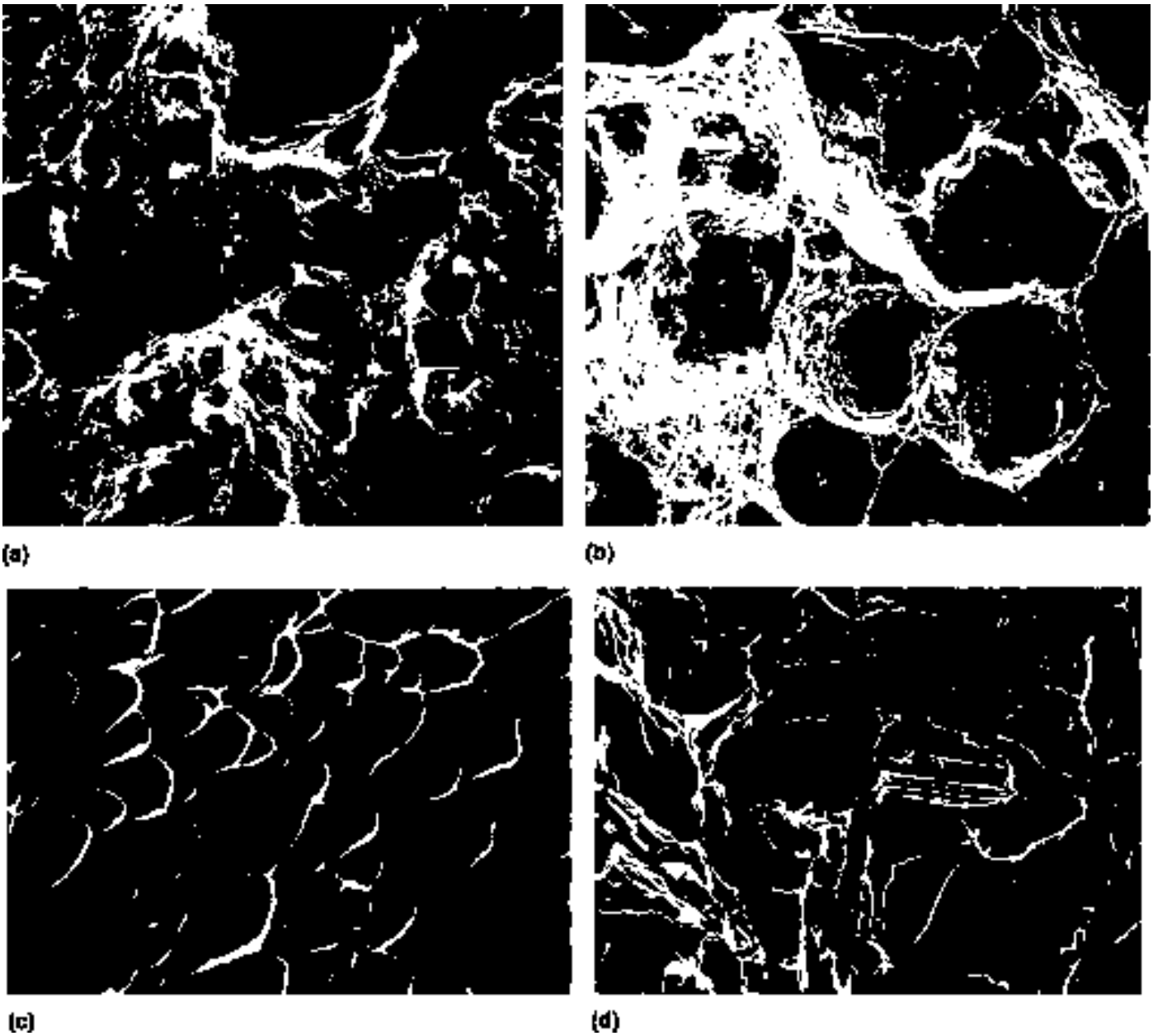


Fig. 1 SEM images of dimple-rupture fractures. (a) Fracture of low-alloy medium-carbon steel bolt (SAE grade 5). 1750×. (b) Equiaxed tensile dimples originating around the graphite nodules of ASTM 60-45-10 ductile iron. 350×. (c) Parabolic shear dimples in cast

Ti-6Al-4V from torsional loading. 1400×. (d) Dimple rupture in cast aluminum alloys. 593×. (a), (b), and (c) courtesy of Mohan Chaudhari, Columbus Metallurgical Services, Inc.

Exceedingly pure materials, particularly unalloyed metals with few inclusions, can exhibit extremely high reduction in area, or necking, at the fracture site. In most engineering materials, impurities and second-phase particles are always present. As would be expected, an increased volume fraction of these constituents will reduce the true fracture strain and the reduction in area (Ref 1). The type, distribution, and size of these constituents are also mechanistically important and are discussed in greater depth later in this article.

Simple overload fractures are not uncommon in bulk deformation manufacturing processes. Even when not resulting in fracture, severe room-temperature forming processes will necessarily result in unavoidable MVC, concentrated in regions of highest plastic deformation. Forming-induced microvoids can act as subsequent fracture initiation sites, which may be of particular importance with regard to fatigue crack propagation. Fatigue cracking also requires a level of ductile behavior; otherwise, once a crack is initiated it would result in instant unstable crack growth, fracturing in a macroscopic brittle manner. Unlike ductile overload, however, fatigue most often initiates and propagates at overall stresses below the yield stress.

Dimple-rupture failures in components of simple cross section often evidence a classic cup-and-cone fracture typical of ductile fracture of tension-test specimens. As previously noted, the fracture originates near the center of the section, where hydrostatic stresses develop and multiple cracks join and spread outward with a fibrous fracture surface perpendicular to the tensile loading axis (mode I, plane strain). When cracks reach a region near the outer surface, the fracture path makes a transition to the shear plane (mode II, plane stress), where the fracture surface is approximately at 45° to the plane of the major fracture. This is the familiar “shear-lip” region. The final stages of fracture thus occur principally along the shear plane, although the angle of this “shear” region changes continually with distance from the central fibrous region and is not always at 45° to the tension load axis.

Examples and illustrate uncomplicated types of ductile overload failures.

Example 1: Fracture of High-Strength Screws During Installation. One lot of size M5 × 0.8 mm, class 8.8 metric screws with a proprietary head design were failing during application. The screws were reportedly failing at the normal installation torque.

Investigation. Typical failed fasteners are shown in Fig. 2(a). Substantial necking accompanied all of the fractures. The fractographic features were consistent with dimple rupture, and the macroscopic necking is shown metallographically in Fig. 2(b). The chemical composition of the fasteners was not verified, but proof load and tension testing of exemplar screws from the same lot revealed satisfactory results. The level of necking on the tension-tested screws was analogous to those that failed during installation.

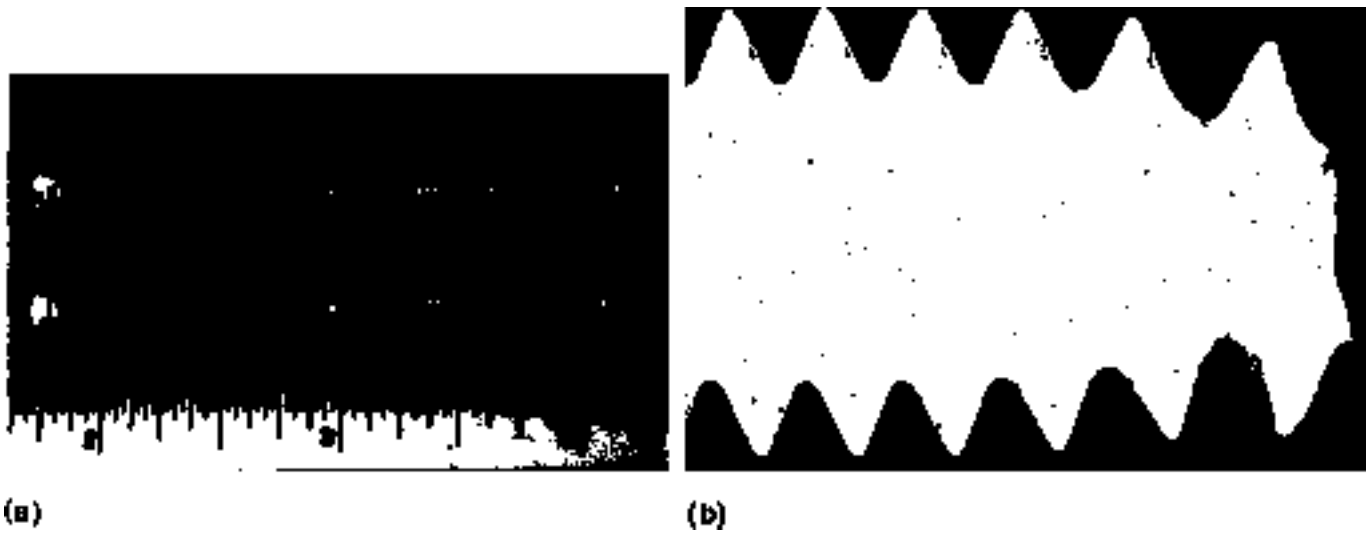


Fig. 2 Fracturing of high-strength screws (example 1). (a) Two grade 8 high-strength steel fasteners that failed on installation. (b) Necking and stretching between adjacent thread crests are evident on a cross section. Unetched. 8.9×

Metallographic study revealed a relatively uniform martensitic microstructure. Minor laps were apparent at the thread crests, but were acceptable according to the applicable specifications. No gross defects or evidence of brittle behavior were identified.

Conclusions. The analytical results indicated that the fasteners failed via ductile overload in the absence of gross defects or embrittlement. It was subsequently determined that a nonapproved lubricant was added to the questioned fastener lot during installation.

In fastener design where tension preload is controlled by torque measurement, the general relationship $T = KDP$ is often used, where T is torque, K is a torque coefficient including a friction factor, D is the fastener diameter, and P is the tensile load. Assuming a predetermined torque for bare steel screws, the tension preload can be more than twice as high on lubricated fasteners. In this case, the reduced friction level resulted in a tension preload sufficient to fracture the screws.

Example 2: Forming Cracks on Stainless Steel Wire. Cold-drawn type 303 stainless steel wire sections, 6.4 mm (0.25 in.) in diameter, failed during a forming operation. All of the wires failed at a gradual 90° bend.

Investigation. High-magnification visual examination disclosed many fine ruptures accompanying a typical fracture (Fig. 3a). These ruptures occurred at the outside of the bend. The rupture interior surfaces (Fig. 3b) were entirely ductile in nature. No defects from manufacture of the wire were evident. The fracture and crack surfaces exhibited dimples (Fig. 3c). The dimples appeared to initiate at nonmetallic inclusions inherent in this free-machining steel. Microstructural examination indicated there was no alteration in the fracture and crack regions. Substantial distortion of the grains was apparent at the ruptures.

Fig. 3 Forming cracks on stainless steel wire (example 2). (a) The fracture, which occurred during bending shows many parallel fissures. 5.3×. (b) A typical fissure on the wire surface. Scanning electron micrograph. 71×. (c) The interior of the fissures and the fracture surface exhibit dimple rupture morphology. Scanning electron micrograph. 1187×

Conclusions. It was concluded that the wires were cracking via ductile overload. The forming stresses were sufficient to initiate surface ruptures, suggestive of having exceeded the forming limit.

Recommendations. The results suggested that the forming process should be examined. In particular, lubrication and workpiece fixturing would be of interest along with examination of successfully formed wire stock.

Reference cited in this section

1. P.F. Thomason, *Ductile Fracture of Metals*, Pergamon Press, Oxford, England, 1990, p 19–24

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Brittle Overload Failures

Brittle overload failures, in contrast to ductile overload failures, are characterized by little or no macroscopic plastic deformation. The fracture features and mechanisms on a microscopic scale may have components of ductile or brittle crack propagation, but the macroscopic process of fracture is characterized by little or no work being expended in the form of permanent (i.e., plastic) deformation. The macroscopic behavior is essentially elastic up to the point of failure. The energy of the failure is principally absorbed by the creation of new surfaces, that is, cracks. For this reason, brittle failures often contain multiple cracks and separated pieces, which are less common in ductile overload failures. Brittle overload failures are sometimes be identified as “catastrophic”

because there may be no warning signs such as warpage or distortion prior to the final fracture, and these failures may cause substantial collateral damage. Energy released during brittle fracture in these failures can be very loud and in some cases explosive.

Brittle fractures are characterized by relatively rapid crack growth to final fracture. The cracking process is sometimes referred to as being “unstable” or “critical” because the crack propagation leads quickly to final fracture. Brittle fracture initiates and propagates more readily than ductile fracture (or for so-called “subcritical” crack propagation processes such as fatigue or stress-corrosion cracking). Representative crack extension rates for brittle and ductile fracture have been measured at 1000 and 6 m/s (3000 and 20 ft/s), respectively. In addition, no supplemental or continuously applied energy may be necessary to continue the brittle crack propagation once it is initiated. The elastic strain energy may be sufficient.

Brittleness does not necessarily imply material flaws or processing mistakes, although this is very often the case. Brittle fracture in some circumstances is the expected overload failure mode. This is true for inherently brittle materials, such as extremely high-strength materials where ductility is sacrificed for maximum deformation or wear resistance. Additionally, embrittling conditions and metallurgical phenomena can impose brittle fracture characteristics on materials that would generally evince the more desirable ductile overload behavior. Brittle overload failures are typically differentiated by a primary crack that occurs from either transgranular or intergranular crack propagation. These categories indicate whether the cracks propagate through the grains or around them, respectively. The use of the infrequently encountered term “intragranular” to describe transgranular features is discouraged due to its similarity in pronunciation to intergranular.

All brittle fracture mechanisms can exhibit chevron or herringbone patterns that indicate the fracture origin and direction of rapid fracture progression. Ductile cracking by MVC cannot have a herringbone pattern, although a macroscale radial pattern and chevrons can occur from MVC. River lines on a macroscopic scale may be revealed in glassy polymers, but not on a macroscopic scale in a metallic material. With regard to fatigue, materials with poor ductility can still experience crack initiation and growth since some slight plasticity may by present. Relatively brittle metals such as hardened steel and gray cast iron do not always form microscopically identifiable striations and macroscopic beach marks in cyclic failures.

Transgranular Cleavage

Transgranular cleavage is cracking through the discrete grains. Intracrystalline is an equally descriptive term, but is not preferred terminology. Crystallographic cleavage occurs preferentially in individual grains in directions that do not readily deform, by slip processes, under strain. Therefore, specifically oriented grains tend to crack, leaving a shiny, faceted appearance easily differentiated from dull and fibrous, dimple-rupture features. Cleavage in a steel sample is shown in Fig. 4. Face-centered-cubic (fcc) metals (for example, copper, aluminum, nickel, and austenitic steels) exhibit the greatest ductility during rapid fracture and in benign environments do not normally fracture via cleavage, as described further in the section “Atomic and Crystalline Structure” in this article. However, brittle cracking of fcc metals may occur under conditions of environmentally-assisted cracking (for example, transgranular stress-corrosion cracking of austenitic stainless steels). High-nitrogen austenitic stainless may also be less ductile during rapid crack growth.

Fig. 4 A cleavage fracture in a carbon steel component is shown. Scanning electron micrograph. 593×

Cleavage initiates via microcrack nucleation at the leading edge of piled-up dislocations. The crack propagates through the grain in which it initiated. The crack then continues across the grain boundaries very rapidly as the critical crack size is exceeded. The fracture surface continues across the grain boundary, but during the cracking process it is very unlikely that the crossing is strictly continuous across the grain boundary. Cleavage cracks cannot cross a grain boundary if the body has a twist component. The crack must reinitiate, and therefore the the classic fine-scale convergence of multiple river lines develop as the crack progresses. Cleaved grains often exhibit river patterns, further confirming transgranular cleavage as the mechanism of crack propagation. Cleavage occurs in materials with a high strain-hardening rate and relatively low cleavage strength, or when a geometric constraint (i.e., large hydrostatic stresses) acts as an initiator of cleavage fracture. It also occurs in materials that are embrittled within the grains rather than at the grain boundaries. All materials are sensitive to hydrostatic stresses, which may also induce brittle fracture. Fractographically, it is very unlikely that cleavage facet size can exceed the grain size, and thus the size of the fracture facets can be a measure of the grain size if there is only one cleavage facet per grain. There is also a possibility of subgrain boundaries and therefore multiple growth directions

in a single grain for those cleavage cracks that cannot propagate across a boundary that has a twist component. In lath martensitic and bainitic microstructures, cleavage facet size correlates with packet size, not prior-austenite grain size.

The term quasi-cleavage applies when significant dimple rupture and/or tear ridges accompany the cleavage morphology. Grains oriented favorably with respect to the axis of loading may slip and exhibit ductile behavior, whereas those oriented unfavorably cannot slip and will exhibit transgranular brittle behavior. A failure analysis example illustrating brittle transgranular fracture follows.

Example 3: Cracking of a Dump Truck Transmission Housing. Both halves of a gray cast iron transmission housing from a 50 ton dump truck contained numerous cracks. No service duration or material specifications were provided.

Investigation. Visual examination revealed that the cracks all occurred adjacent to stiffening ribs and at changes in section thickness. One of the cracks is shown in Fig. 5(a). A crack that was opened to permit examination is shown in Fig. 5(b). All of the cracks appeared to have the high light reflectivity of cleavage facets with no postfracture indication of ductility or mechanical damage. None of the casting cracks contained stains, casting defects, or features suggestive of fatigue cracking. The housing halves did not contain any internal damage.

Fig. 5 Cracking in a truck transmission housing (example 3). (a) A typical crack at a stiffening rib. (b) An opened crack at a stiffening rib exhibited brittle features. (c) Longitudinal metallographic cross section showing secondary brittle cracking along the graphite flakes. 2% nital etch. 59×

Tensile specimens from both halves exhibited ultimate tensile strengths exceeding 205 MPa (30 ksi). Although testing of samples removed from a cast component may not necessarily bear relation to grade/condition determination that is based on separately poured test bars, results were suggestive of the G3000 grade designation for automotive gray cast iron. The typical fracture morphology within the matrix was transgranular cleavage. The microstructure consisted of size 5 type A graphite flakes in a matrix of approximately 73% pearlite/27% ferrite. No free carbides were present. The crack followed the profile of the graphite flakes, as shown in Fig. 5(c).

Conclusions. The transmission housing castings exhibited brittle overload cracks at highly stressed locations. Failure was due to applied stresses sufficient to fracture the castings. These were probably unexpected bending forces. The cleavage cracking occurred preferentially along the graphite flakes, which is typical for this material. No cold shuts, porosity, or other casting flaws were evident within the cracks. The source of the unexpected bending was not determined. Fracture could be transgranular (and most likely is), but the fracture path is heavily influenced by cleavage cracking in the graphite.

Intergranular Fracture

Intergranular brittle fracture occurs by separation at or adjacent to the grain boundaries. In some cases it can occur at previous grain boundaries, such as in the case of martensitic steels where fracture can occur at prior-austenite grain boundaries, which may also coincide with some of the boundaries in the martensite. Intergranular cracking has been documented in nearly all engineering metals and alloys and is caused by a wide variety of mechanical and environmental factors such as grain-boundary embrittlement and decohesive separation along the grain boundaries at elevated (creep-regime) temperatures. For service temperatures below the creep regime (i.e., about 0.4 to 0.5 the melting temperature of the alloy) and with appropriate materials selection and design, intergranular brittle fracture is often (but not always) indicative of improper material processing. Transgranular cleavage is usually the brittle mode anticipated in normal conditions for brittle materials. Intergranular brittle fracture is atypical, as the grain boundaries are usually stronger than the grains at temperatures below the creep regime.

Grain-boundary fracture can occur both with and without evident MVC on the grain surfaces. Intergranular fracture exhibiting dimple-rupture features is often referred to as decohesive rupture or more generally as “intergranular dimpled fracture.” Grain-boundary fracture exhibiting no grain surface ductile rupture is identified as intergranular brittle fracture. In failure analysis, the investigator must distinguish which type of intergranular fracture has occurred as the causes and corrective actions can be substantially different.

Decohesive rupture is an intergranular fracture mechanism that is macroscopically brittle yet microscopically ductile. Visual and low-magnification stereomicroscopic examination typically reveal the facets of the individual grains along which cracking occurred. The appearance of this morphology could be easily mistaken for pure brittle intergranular fracture if high-magnification fractography is

not performed to examine for microscopic void formation in the grain boundaries. This mechanism of MVC in the grain boundaries can also result in a greater level of secondary cracking than evidenced by ductile overload failures.

In some circumstances, dimple rupture occurs along the interface between the matrix and a relatively weak intergranular phase (for example, at high temperatures). Weakening of the grain boundary may occur at elevated temperatures and result in creep stress rupture. It may also occur adjacent to the grain boundary due to elemental segregation causing a denuded zone. As materials with this inherent or imposed tendency are stressed, normal MVC initiates at inclusions, second phases, or other crystallographic imperfections (for example, the grain boundaries). Due to the preexisting crack path provided by the grain boundaries, decohesive rupture requires less energy to effect separation than that for typical transgranular ductile rupture.

Example 4: Overload Failure of a Bronze Worm Gear. A very large diameter worm gear that had been in service in a dam for more than 60 years exhibited cracks and was removed. It was reported that the cast bronze gear was only rarely stressed during service, associated with infrequent opening and closing of gates. Due to the age of the gear and the time frame of its manufacture, no original material specifications or strength requirements could be located. Likewise, no maintenance records of possible repairs to the gear were available.

Investigation. As part of the investigation, a number of representative sections excised from the gear were submitted for metallurgical analysis. The cracks were primarily located on the gear faces and were not associated with traditional regions of high service stresses, such as the tooth roots. Some portions of the gear face exhibited light discoloration, which was suggestive of weld repairs. The cracks were evident within the base metal, at the repair weld fusion lines, and within the weld metal. Figure 6(a) shows the opened crack features in an area where a weld and remaining casting imperfection were apparent. The macroscopic fracture features differed substantially between the base metal and the welds. Some of the cracking did not occur at casting imperfections or repair welds.

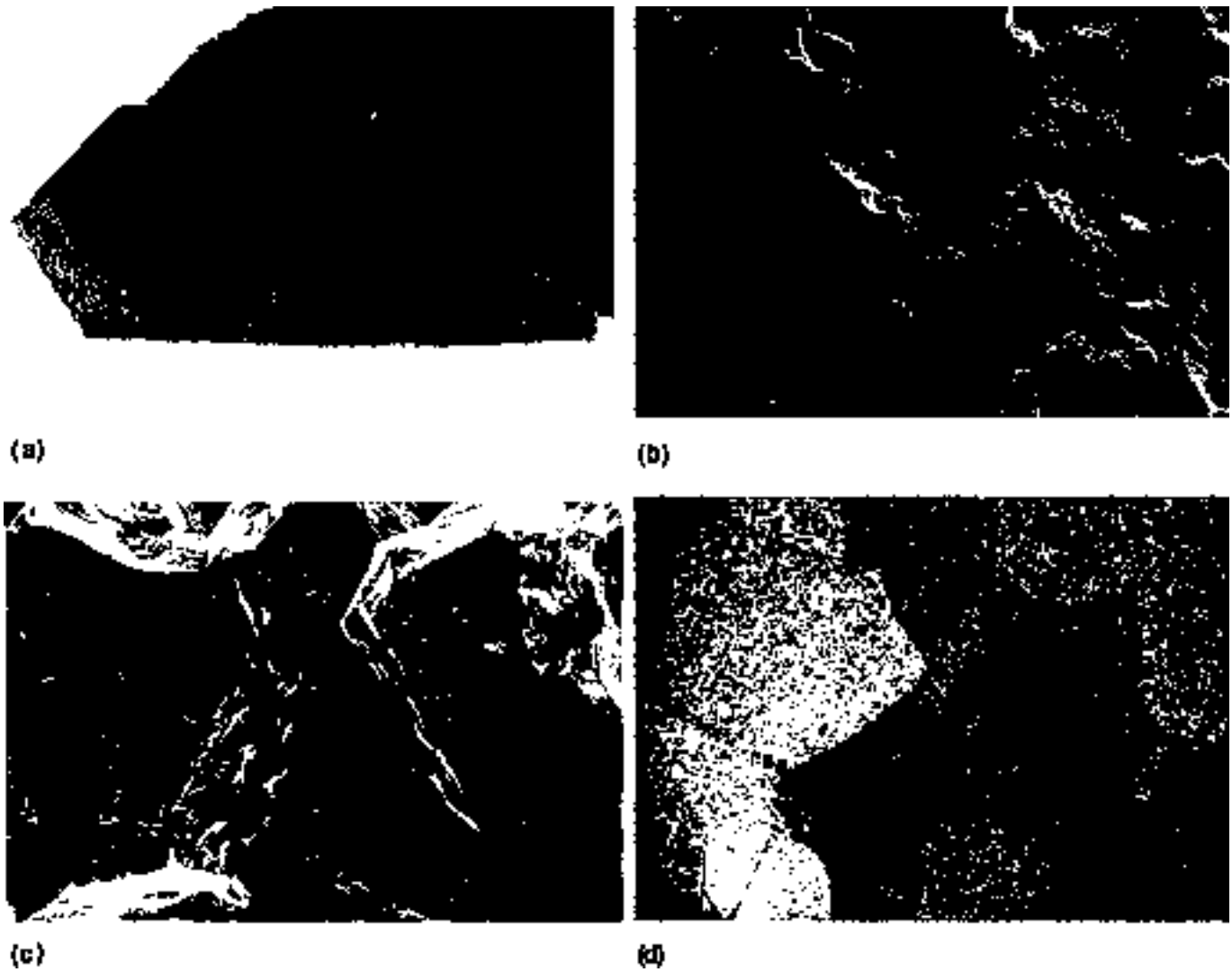


Fig. 6 Overload failure of a bronze worm gear (example 4). (a) An opened crack is shown with a repair weld, a remaining casting flaw, and cracking in the base metal. (b) Electron image of decohesive rupture in the fine-grain weld metal. Scanning electron micrograph.

119×. (c) Morphology in the large-grain base material at the same magnification as (b), showing intergranular brittle fracture features. Scanning electron micrograph. 119×. (d)

Metallographic image showing the weak grain-boundary phase in the weld. Potassium dichromate etch. 297×

Scanning electron microscopy (SEM) revealed the typical weld and base metal fracture features shown in Fig. 6(b) and (c), respectively. The weld exhibited fine features that consisted of both dimple rupture and grain-boundary dimple-rupture evidence. The base metal exhibited rather large-grain intergranular brittle cracking features with many types of inclusions.

Chemical analysis revealed a standard high-strength, manganese bronze composition, and the weld filler metal was considered compatible. Tension testing of specimens removed from gear segments remote to repairs revealed tensile strengths near 690 MPa (100 ksi), yield strengths near 517 MPa (75 ksi), and 5% elongation at fracture. Weld and base metal hardnesses were similar.

The microstructure at the transition between the base metal and weld metal is shown in Fig. 6(d). The continuous grain-boundary phase in the weld metal responsible for the decohesive rupture features is evident.

Conclusions. It was concluded from the investigation that the bronze gear cracked via mixed-mode overload, rather than by a progressive mechanism such as fatigue or stress-corrosion cracking (SSC). The cracking was not associated with regions that would be highly stressed and did not appear to be consistently correlated to casting imperfections, repair welds, or the associated heat-affected zones. The cast gear was a high-strength, low-ductility bronze. Cracking across the gear face suggested bending forces from misalignment were likely responsible for the cracking, and further review of this potential root cause could be recommended.

Intergranular brittle fracture exhibits little or no plastic deformation and fractographic features devoid of rupture evidence. This is the classic “rock-candy” appearance in brittle materials. Materials that are identified as “embrittled” are usually prone to intergranular brittle fracture, although some embrittlement mechanisms are manifested as transgranular cleavage, decohesive rupture, or dimple rupture. The term rock candy is most often applied to coarse-grain materials where this appearance is visually evident.

This fracture mechanism can be due to many conditions, including a brittle grain-boundary phase, high strain rates, or embrittlement. Visually, the macroscopic plane of the intergranular brittle fractures appears relatively flat, perpendicular to the axis of the applied stress responsible for the overloading. This mechanism is usually due to an improper condition of the materials or the alteration of the material during service. Intergranular brittle fracture is very prevalent in ceramic materials and is often present in poorly sintered powder-metallurgy constructs. It is necessary for the practicing failure analyst to understand that intergranular brittle fracture is not always due to defects and improper processing. This mechanism can result from well-controlled processes that provide benefit that can outweigh the expectation of brittle failure. One example is in the process of carburization, as included in the case history that follows.

Example 5: Valve Seat Fractures. Valve seats fractured during testing and during service. The seats were machined from grade 11L17 steel and were surface hardened by carburization.

Investigation. Fracture occurred through the relatively thin wall of the seat, between circular holes in a cylindrical barrel. Secondary cracks were evident adjacent to the fractures in every failed ligament region. The fracture surfaces visually had the faceted or “crystalline” appearance of intergranular fracture near the exterior surfaces, as shown in Fig. 7(a). No fracture origins were apparent, although corners at circular crossholes were considered the probable origins. The core fracture morphology consisted of elliptical dimples.

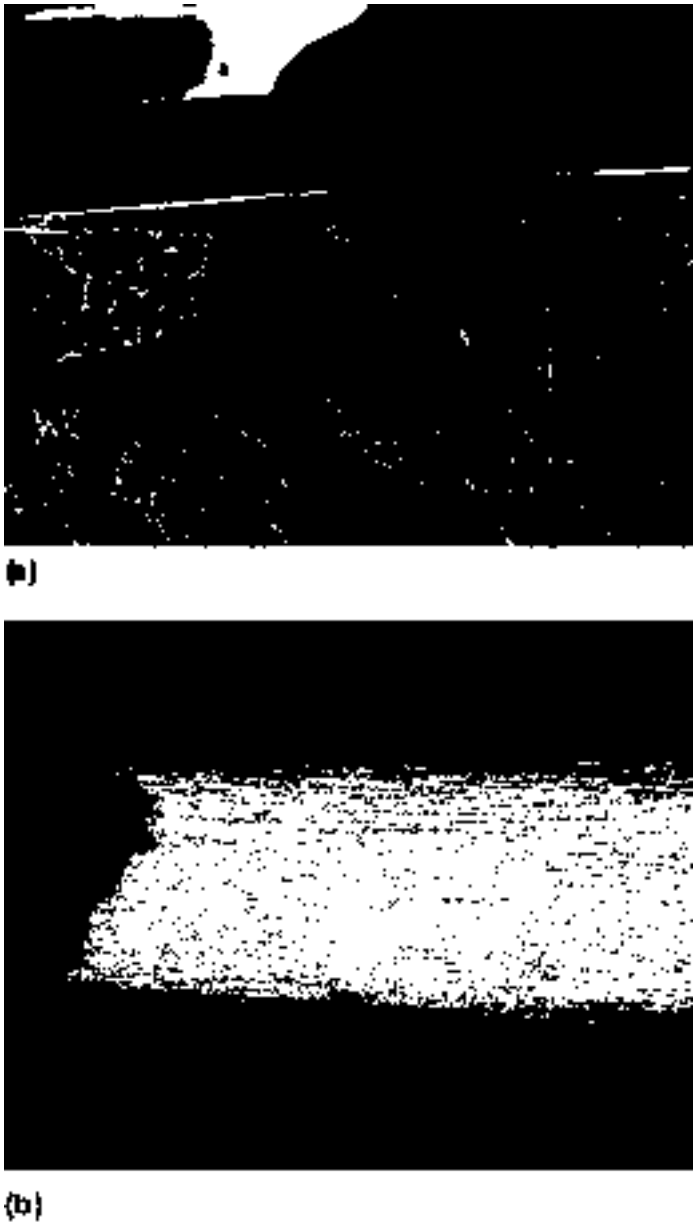


Fig. 7 Valve seat fractures (example 5). (a) View of the seat fracture surface with intergranular near-surface features from the carburizing heat treatment. Scanning electron micrograph. 59×. (b) Cross section showing the case depth and the blunting of a secondary bending crack by the ductile core microstructure. 2% nital etch. 15×

The seat surface hardness was 55 HRC, which is typical for case-hardened steel. A metallographic cross section through the fracture surface is shown in Fig. 7(b). A secondary crack has opened substantially and blunted in the ductile core. Plastic distortion of the core grain structure at the crack tip was evident. The seat wall was not through-carburized. The near-surface microstructure was tempered martensite and retained austenite, along with the large nonmetallic inclusions inherent in this material. No case-hardening microcracking was evident, and no grain-boundary carbides were apparent.

Conclusions. The fracture occurred via brittle overload, which was predominantly intergranular. The amount of bending evidence and the directionality of the core overload fracture features suggest that the applied stresses were not purely axial, as would be anticipated in this application. Resulfurized steels are typically case hardened for wear resistance, but the hardened structure has relatively poor ductility. Stress concentration occurs at the ligaments of hardened microstructure between the soft inclusions. These materials characteristically have low transverse mechanical properties, even when they are not case hardened. The level of retained austenite in the hardened case layer likely contributed to the failure.

Brett A. Miller, Stork Technimet, Inc.

Mixed-Mode Cracking

As shown in several of the previous examples, it is not unusual to observe fractographic evidence of multiple mechanisms on a single fracture surface. The failure analyst must rely on experience to interpret mixed-mechanism and mixed-mode features. The morphology at the fracture origin is of primary importance. Atypical features among uniform fracture features may not be of significance other than pinpointing microstructural inhomogeneities and discontinuities (see, for example, the section, “Material Factors,” in this article).

A blunted crack in a ductile iron impeller that failed via mixed brittle and ductile mechanism overload fracture is shown in Fig. 8. The crack stopped in this area where deformation and necking of the ductile bull's-eye ferrite is evident. Much of the remaining cracking through the lamellar pearlite was very brittle. Although the ferrite phase was not predominant in the matrix, substantial ductile character was imparted.

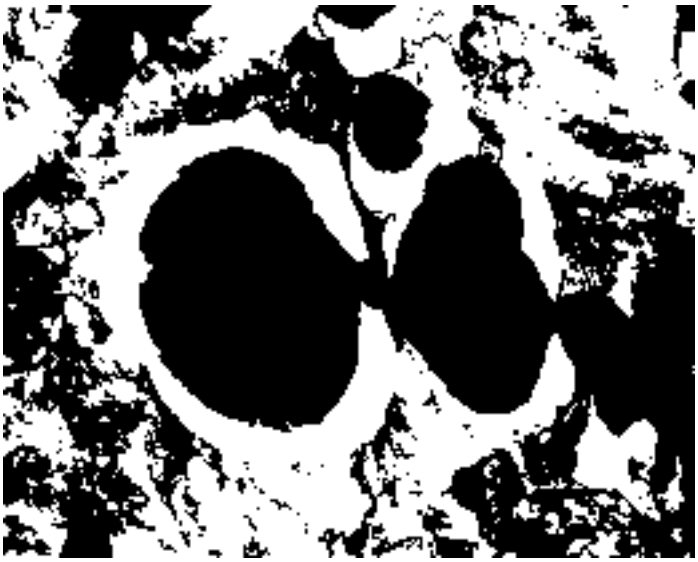


Fig. 8 Crack in a high-strength ductile iron (grade 100-70-03) impeller showing deformation in the ferrite resulting in blunting of this secondary crack. 2% nital etch. 492×

In many cases, a combination of fracture mechanisms suggests a minority phase has acted detrimentally. Figures 9(a) and (b) show fracture through globular carbides in a thin tool steel component resulting in matrix fracture. The tempered martensitic matrix was somewhat ductile, yet fracture occurred when brittle, massive carbides near the surface shattered.

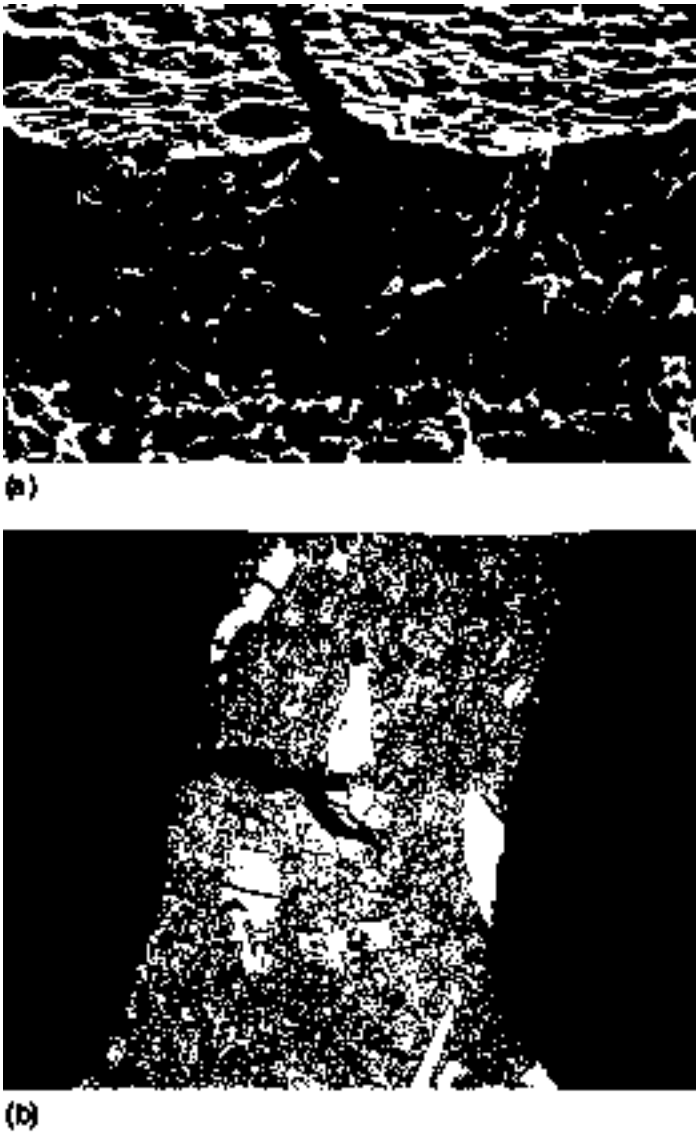


Fig. 9 Fracture in a thin medical device manufactured from type D 2 tool steel. (a) View showing a fractured massive carbide and associated matrix crack. Scanning electron micrograph. 1187× (b) Cross section through a cracked region in a similar part showing brittle fracture in the carbides and plastic deformation of the surrounding matrix. Vilella's reagent. 297×

Overload fracture by ductile or brittle crack growth is often the final result of fracture following other subcritical cracking phenomena such as fatigue and corrosion. In these cases it is necessary to identify the proximate cause that was followed by the subsequent overload behavior. The final fracture occurs at a critical point when the remaining intact ligament of material can no longer support the applied stress. Figure 10 shows the fatigue to final ductile overload transition in a relatively soft austenitic stainless steel. Some dimples due to MVC are evident in the last few striations that formed before the crack size became critical.

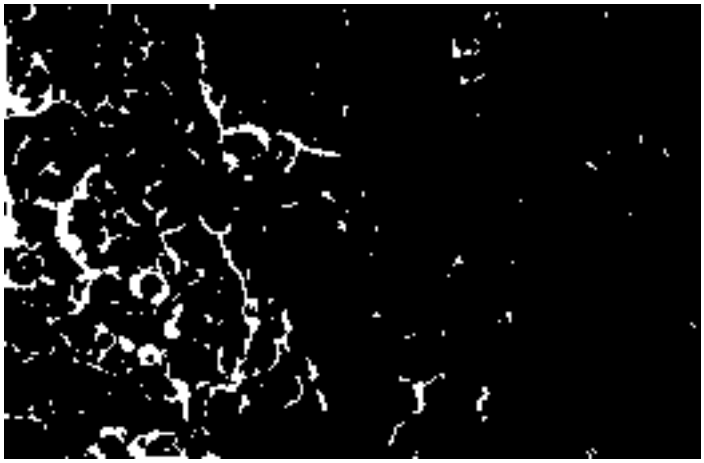


Fig. 10 Stainless steel component that was flexed in service. The transition between fatigue and dimple rupture is shown. Scanning electron micrograph. 1187×

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Material Factors

Determining the root cause of a failure requires a comprehensive understanding of the material, loading, and environmental factors that can contribute to failure. The following sections describe some of the material, mechanical, and environmental factors that may be involved. Each is described in a separate section, but it must be recognized that they are not discrete and separate factors. The mechanisms of fracture are influenced by complex and interrelated factors that may provide competing or synergistic conditions leading to fracture. This may illustrate some of the complexities or difficulties of engineering failure analysis.

This section briefly reviews some of the material factors that may be considered in the evaluation of unstable, rapid fracture. Ductile or brittle overload failures and the resultant fracture morphology are influenced by various material factors from the atomic level to the bulk material level. The inherent macroscopic ductility of crystalline materials, for example, depends on factors such as the specific crystal lattice that is present, the presence of foreign atoms in the lattice and where they are located, as well as the size of the individual crystalline regions (“grain size”). Many alloys also exhibit a combination of fracture morphologies in the region of rapid, unstable overload fracture due variations in the compositional constituents of the material. For example, Fig. 11(a) reveals a dimpled surface of ductile features typical of fcc metals such as austenitic steel, while Fig. 11(b) is a SEM image of an aluminum alloy where various fracture morphologies are revealed. This is typical of many alloy systems where rapid fracture exhibits a combination of morphologies.

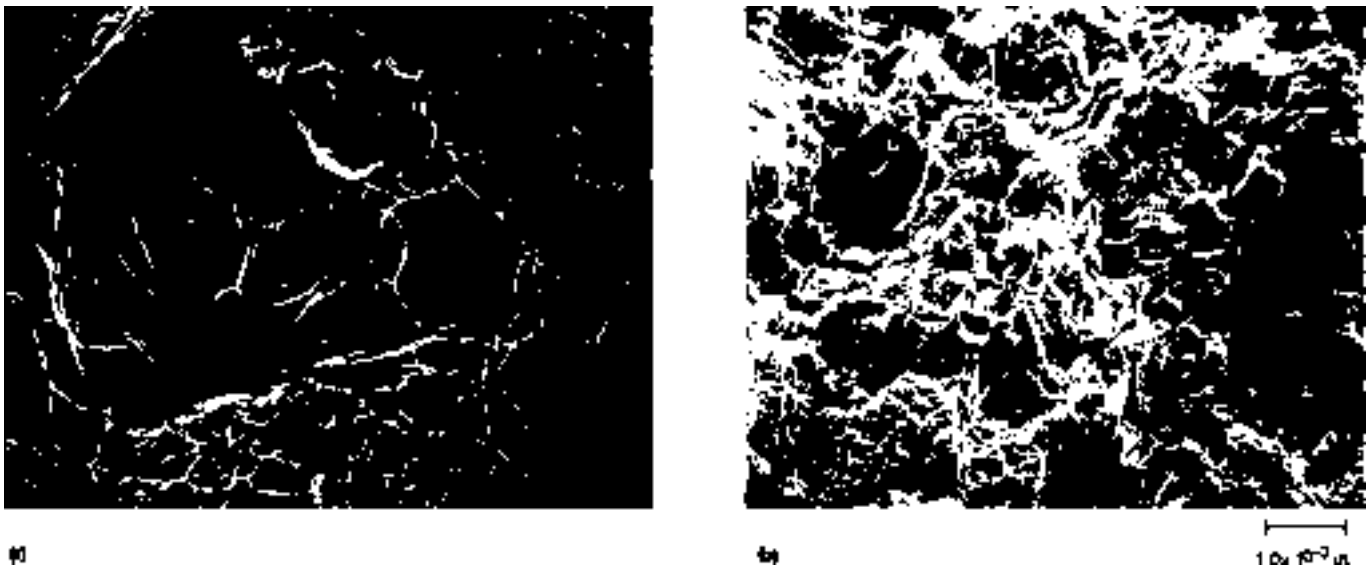


Fig. 11 Examples of fracture surfaces in face-centered cubic (fcc) metals. (a) Austenitic 316L stainless steel with variable sizes of equiaxed/tensile dimples. 690×. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services. (b) 2014-T6 aluminum alloy where unstable rapid fracture exhibits a combination of morphologies of ductile microvoid coalescence of the matrix and brittle cleavage and intergranular separation through and around second-phase particles. Courtesy of Howard Nelson (Ref 2)

Atomic and Crystalline Structure

At the atomic structure level, individual atoms exhibit differences in the number of electrons in the various electron shells. This results in different types of bonding and bond strength. Bonding may be covalent (strong), ionic (strong), metallic (intermediate), or van der Waals (weak). Metallic materials are metallically bonded and ceramic materials are predominantly ionically bonded. Carbon atoms in the backbone of polymer chain are covalently bonded and the pendant side groups are covalently bonded to the backbone. However, bonding between chains may be covalent ("cross linked") or van der Waals.

The bonds between atoms may occur in various spatial orientations. The spatial arrangement of atoms may be random ("amorphous"), or periodic in either two or three directions resulting in sheet structures (two-dimensional such as graphite) or crystalline (three-dimensional). Over still greater distances, atom arrangements may be completely amorphous, oriented (as parallel alignment of carbon backbone chains in polymers), mixtures of crystalline and noncrystalline regions, or totally crystalline but in which crystalline regions are rotated with respect to each other ("polycrystalline"). In most metals, the metallic bonds between atoms typically result in a crystalline structure, which in most engineering metals are face-centered cubic (fcc), body centered cubic (bcc), or hexagonal close-packed (hcp) structures.

The formation of crystal lattices occurs as a result of bonding between atoms. Strong bonding forces between atoms cause atoms to pack efficiently (high packing densities). These arrangements exhibit planes of high atomic density, which contain close-packed directions. This results in the three lattices (fcc, bcc, hcp) being especially common for metallic materials. The higher symmetry of the cubic lattices relative to noncubic lattices has implications regarding their inherent ductility as discussed later. In general, the extent of ductility is influenced by the strength of electronic bonding, the crystal structure, and the degree of order. The three basic atomic-level influences on ductile behavior in order of increasing influence on ductile behavior with highest ductility at the top and increasing influence toward brittle behavior at the bottom are:

Influence of electronic bonding (most to least ductile):

- Metallic bonds
- Resonating covalent bonds
- Ionic bonds
- Covalent, Van der Waals bonds
- Covalent bonds

Influence of crystal structure (most to least ductile):

- Close-packed fcc
- Body-centered cubic
- Hexagonal close-packed. Note: The relative ranking of ductility between hcp and bcc materials may vary depending on temperature and available slip systems, as discussed in the section "Ductility in Crystalline Materials" in this article)
- Rhombohedral
- Less symmetrical through triclinic

Influence of degree of order (most to least ductile):

- Random solid solution
- Ordered or intermetallic compound
- Ionic compound
- Covalent compound

Crystal structures of ceramic materials are often more complex than those of metallic materials. High packing density cubic structures do exist in some ceramic materials (rock salt and cesium chloride structures), but the presence of more than one type of atom together with ionic bonding causes differences in critical stresses for ductile (slip) and brittle (cleavage) behavior compared to the metallic materials. Qualitatively, the critical stresses for slip are so high relative to those required to propagate cleavage that these materials are considered to be inherently brittle.

The ability of polymeric materials to form crystalline solids depends in part on the complexity of the pendant side groups of atoms attached to the covalently bonded carbon atom backbone. Crystallinity is also decreased if the polymer is branched. Seldom is crystallinity complete in a polymeric material. The crystal structures that do form are less symmetrical than those of metallic materials, caused in part by the inability to tightly pack the pendant side groups. Triclinic and monoclinic lattices are common. Similar to the ceramic materials, the crystalline polymers and oriented polymers tend to not be very ductile

The stress-strain behavior depends strongly on the long-range spatial arrangement, together with the type of bonding. For example, ionically bonded ceramic materials are crystalline and typically show little plastic deformation prior to fracture. Randomly oriented polymers show considerable ductility prior to alignment of the molecular chains, but then quickly fracture. Polymeric materials containing mixtures of crystalline and noncrystalline regions show decreased ductility and increased strength and stiffness as the degree of crystallinity increases. Polymeric materials also show decreased ductility as the extent of cross linking between chains increases. Finally, polymeric materials show a considerable stiffening, increase in strength, and loss in ductility if they are loaded at temperatures below the glass transition temperature.

Ductility in Crystalline Materials

Plastic deformation occurs by the movement of dislocations, which are imperfections in the crystal structure. Plastic deformation by dislocation motion also results in dislocation-dislocation interactions that cause the dislocations to become nonmobile (“pinned”), and dislocations may also be pinned at disruptions in the lattice such as grain boundaries. These dislocation pileups exert a back stress on other mobile dislocations causing the *slow* flow stress to increase (“strain hardening”). Strength and hardness are increased, but ductility is reduced. A material has a finite level of work hardening it can withstand, and additional strain results in fracture.

Dislocations are always present in crystal materials, and their movement tends to occur in close-packed directions on close-packed planes. This combination of a slip plane and slip direction is called a slip system. Plastic deformation occurs from slip along lattice planes, permitting energy applied to the crystal structure to be dissipated as irreversible plastic deformation. This is the essence of ductile behavior. Those metals that have the greatest number of slip systems are more ductile and can deform more easily. As such, the inherent ductility of a crystalline material is in part controlled by the symmetry of the lattice and the number of equivalent packed directions and planes.

For example, in a cubic lattice, the face planes all have the same packing density, and the body diagonals all have the same packing density. If, however, the bcc lattice is distorted into a body-centered tetragonal lattice (for example, martensite), the packing density on the three mutually orthogonal face planes is no longer the same, reducing the number of equivalently packed planes. If there are more than three equivalently packed slip planes, it is impossible to orient a load such that it is simultaneously perpendicular to all planes, and therefore there will be a resolved shear stress on at least one of the slip planes.

The fcc metals (for example, copper, aluminum, nickel, and austenitic steels) have a large number of slip systems (12), which is one reason why they exhibit high ductility. The fcc metals are more closely “packed” (i.e., a shorter distance exists between atoms in the crystal cell) than are bcc metals. The bcc lattice has a slightly lower packing factor of 0.68, compared to the packing factors of 0.74 for the fcc and (ideal) hcp lattice. This partly explains why cleavage fracture does not normally occur in the matrix of fcc metals. However, this is not just a matter of slip system and cleavage system multiplicity. Two other factors control the inherent ductile-brittle behavior of crystalline materials: the critical shear stress required to initiate slip and the critical normal stress required to propagate a cleavage crack. So long as there is only one type of atom in the lattice, the shear stress for slip is low. However, the presence of foreign atoms raises this stress and, depending on the location of the foreign atoms (random or periodic), may cause a severe loss in ductility as is the case for the ordered intermetallic alloys. Thus, the critical stresses required for slip and cleavage also determine whether or not cleavage occurs.

Nonetheless, there are clearly differences between the deformation and fracture behavior of the fcc, bcc, and hcp lattice structures. The bcc metals (for example, ferritic steels, tungsten, and β brass) are not as ductile as the fcc metals and may exhibit regions of cleavage, as revealed by a SEM image of the fracture surface after unstable, rapid fracture of an annealed low-carbon steel (Fig. 12). There is also a significant difference in the variation of the yield strength with temperature in the two lattices. For bcc materials (and some hcp materials), there is a strong increase in yield strength as the temperature is decreased (see the section “Temperature Effects” in this article). The stress required for cleavage thus can be less than the yield stress at a sufficiently low temperature, and brittle cracking becomes the dominant mechanism below the so-called ductile-brittle transition temperature. In contrast, the yield strength of fcc metals does not show as much variation versus temperature, and fracture mechanisms remain ductile even at cryogenic temperatures (see the section “Temperature Effects” in this article).

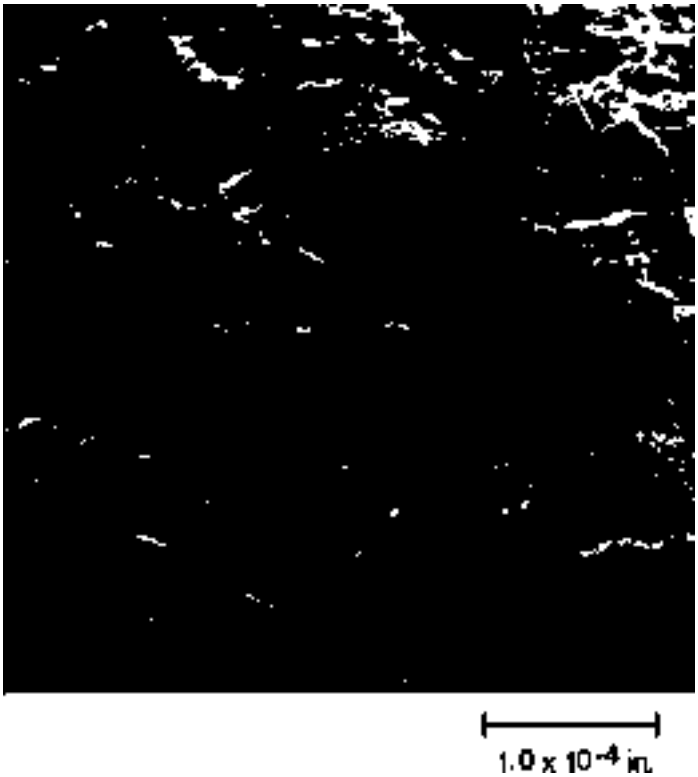


Fig. 12 Example of unstable rapid fracture in a body-centered cubic (bcc) metal (annealed low-carbon steel). Rapid fracture in this alloy occurs almost completely by microvoid coalescence, but close examination reveals a few areas of brittle cleavage. The bcc structure is not close-packed rapid fracture, and it is not unusual to see increasing amounts of brittle cleavage in alloys with higher strength or in the ductile-brittle transition for a given temperature and strain rate. Courtesy of Howard Nelson (Ref 2)

A well-accepted theoretical model to permit a randomly oriented grain in a polycrystalline material to take on any shape change, depending on the applied loads, requires that there be a minimum of five independent active slip systems. The fcc and bcc lattices have more than the required minimum number of slip planes and slip directions. The situation is more complex in the hcp lattice in that there are several different potential combinations of slip plane and slip direction, but the critical stress for slip on these systems varies with the specific material. Introductory materials classes always indicate that slip occurs on the basal plane (one plane, three directions, only two of which are independent). If that is indeed the case, there are an insufficient number of slip systems to obtain general ductile behavior. A stress applied perpendicular to the basal plane must result in brittle behavior since there is no resolved shear stress on any slip system. Additionally, since there is no slip direction out of the basal plane, the material cannot get thinner along the basal plane normal; it can only get longer and narrower. However, slip does occur on other systems, and slip on the basal plane may not be the most easily activated (as in titanium), For example, slip on the face (prism) plane in the close-packed direction (three systems, two independent) is more easily activated in titanium than is slip on the basal plane. Slip on this system provides only two more independent systems. The final result is that depending on the critical stresses to activate slip on various planes and directions, there may be an inadequate number of independent slip systems to permit an arbitrary shape change. Stated differently, if there are an inadequate number of slip systems, a single crystal may deform for some orientations, but behave in a brittle way for other orientations. Or for a given orientation, changing the loading conditions (say tension to torsion) may cause a change from ductile to brittle behavior. The conclusion can be extended to textured polycrystalline material. The material can be ductile for some types of loading, but brittle for other types depending on the specific texture that is present. Figures 13 and 14 are sample SEM images from the region of unstable, rapid fracture in hcp metals with a high *c*/*a* (crystal lattice lengths across the *c* and *a* axes) ratio (easily activated basal plane slip) and low *c*/*a* ratio (less basal plane slip), respectively.

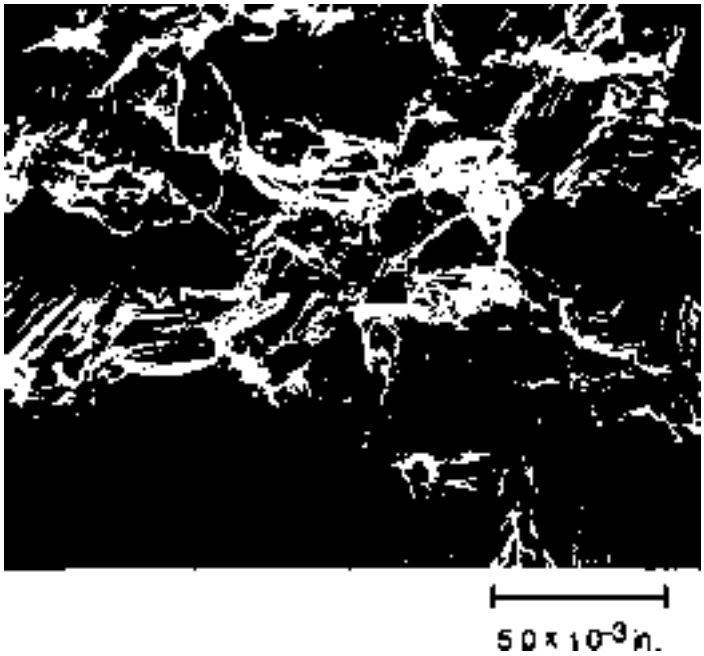


Fig. 13 Fracture surface from rapid unstable cracking of an alloy with a high *c*/*a* ratio. Rapid fracture in this hexagonal close-packed (hcp) alloy (QE 22A-T6 magnesium alloy) appears to be a very difficult process, as slip is confined essentially only to the basal planes. Cracking is more cleavagelike on the basal planes with very small tears between separated basal planes. The various orientations of the grains can be seen by the orientation of basal cracking. In some orientations, cracking is extremely difficult, and, in this case, intergranular separation occurs. Courtesy of Howard Nelson (Ref 2)

Fig. 14 Fracture surface from rapid unstable cracking of an alloy with a low *c*/*a* ratio. Rapid fracture in this hcp alloy (Ti-Al-2.5Sn) reveals a very feathery appearance from cleavagelike cracking occurring on the basal planes. Because other slip systems are also active, cracking can occur on other planes as well. Although the fracture surface appears fairly brittle, cracking is difficult in this alloy and fracture toughness is relatively high. Courtesy of Howard Nelson (Ref 2)

In general, crystalline materials can permanently deform by two transgranular shear processes: slip and twinning. Brittle fracture (cleavage) occurs due to normal separation of planes. They may also deform and fracture in the intracrystalline regions ("intergranular" deformation or fracture). Intergranular deformation becomes of engineering importance when temperatures exceed 30 or 40% of the melting point. Transgranular plastic deformation by slip produces greater intrinsic plastic strain than twinning and is the

dominant mechanism causing “ductile” behavior. As noted, plastic deformation by slip tends to occur in close-packed directions on close-packed planes in ductile deformation and fracture. No such generality can be made for prediction of cleavage planes, although there is some tendency for cleavage to occur on more widely separated planes (i.e., the basal plane in the hcp lattice and the face planes in the cubic lattice).

Effect of Grain Size. Grains are the discrete crystalline regions present in metals. A grain may exhibit a lattice orientation different than surrounding grains based on nucleation and postsolidification processing effects. The grain size is a very important factor, as a finer grain structure provides higher strength and ductility. This situation is rather unique, as most other strengthening mechanisms result in decreased toughness. Strengthening by finer grain size is achieved, in part by the increase in the amount of grain-boundary regions as the grain size is reduced. Grain boundaries are stronger than individual grains (when temperatures are below the so-called equicohesive temperature), and thus a finer grain size imparts more grain-boundary regions for improved strength. Moreover, because a greater number of arbitrarily aligned grains are achieved when grain size is reduced, the stressed material has more opportunity to allow slip and thus improve ductility. In contrast, coarser-grain material tends to exhibit generally poorer mechanical properties (lower yield strength, higher ductile-brittle transition temperature, poorer long-life fatigue properties). Thus, grain size refinement provides a means for the materials engineer to increase both strength and ductility (toughness).

Effect of Strain Rate on Ductility. As previously noted, deformation and fracture behavior can be influenced by complex material, mechanical, and environmental conditions. In addition to the material condition, ductility is also greatly affected by the rate of deformation (i.e., strain rate). Slow strain rates allow rearrangement of complicated three-dimensional dislocations and thus permit more plasticity. In contrast, the rapid application of loads and high strain rates do not permit as much lattice reorganization, resulting in a tendency toward more brittle behavior during fracture. The effect of a higher strain rate on ductility is analogous to a decrease in temperature, as shown schematically in Fig. 15. At a higher strain rate, the transition to brittle fracture of a bcc metal occurs at a higher temperature (T2 in Fig. 15). Similarly, higher strain rates lower fracture toughness, as shown in Fig. 16. This illustrates the complexity of deformation and fracture behavior, as it is affected by material factors, strain rate, and temperature.

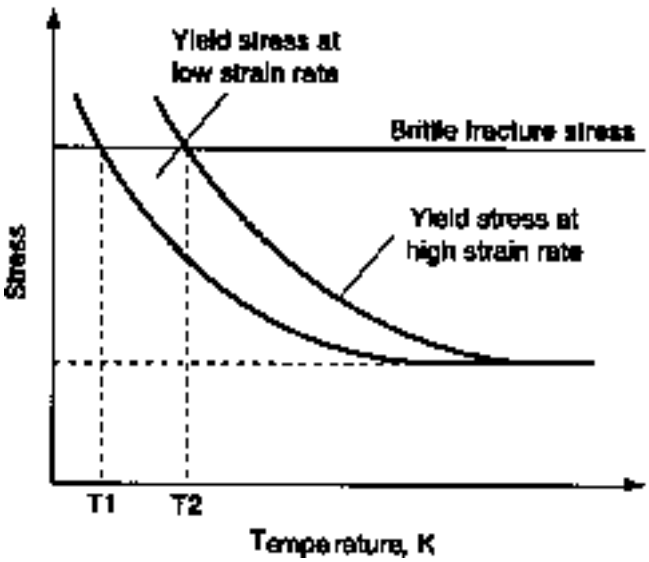


Fig. 15 Effect of strain rate on ductile-to-brittle transition temperature in body-centered cubic metals

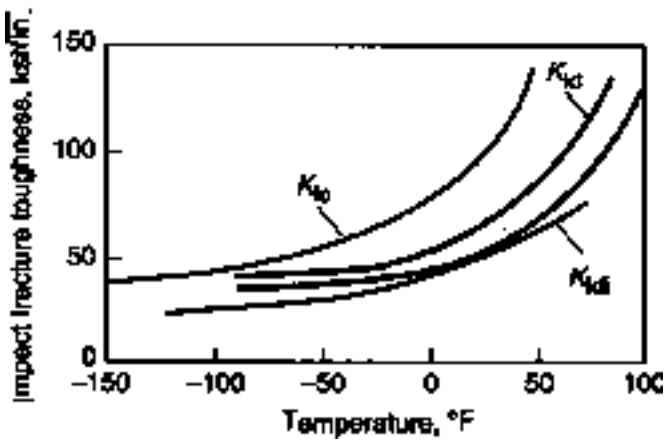


Fig. 16 Comparison of static (K_{Ic}), dynamic (K_{Id}), and dynamic-instrumented (K_{idi}) impact fracture toughness of precracked specimens of ASTM A 533 grade B steel, as a function of test temperature. The stress-intensity rate was about 1.098 × 10^4 MPa$\sqrt{m}$ · s^{-1} (10^4 ksi$\sqrt{in}$ ·

s^{-1}) for the dynamic tests and about 1.098×10^6 MPa$\sqrt{m}$ · s^{-1} (10^6 ksi$\sqrt{in}$ · s^{-1}) for the dynamic-instrumented tests.

Effect of Texture

Texture, where the crystalline grains in a metal have a preferred orientation, may result in anisotropic elastic, plastic, and fracture mechanical properties as well as physical properties (electrical conductivity, magnetic behavior, and so on). The level of anisotropy from texture is dependent on the crystal structure and elemental composition and the specific preferred orientation that is present. Polycrystalline metals consist of many individual crystalline grains, which often are largely randomly oriented throughout the material. In such a case, isotropic mechanical properties on a macroscopic scale are expected with random grain orientations. However, if grains in a polycrystalline body are not randomly oriented (i.e, they have a preferred orientation or texture), the macroscopic properties can be anisotropic. The degree of anisotropy depends on the anisotropy of the single crystal and amount of nonrandom grain orientation. In practical situations, anisotropy is created by three different mechanisms: texture, banding, and fibering. Texture is created by deformation systems in grains; banding is created by nonhomogeneous distribution of alloying elements and/or constituents. Fibering is created by the volume fraction, deformability, and distribution of inclusions in the material.

Anisotropy of mechanical and physical properties due to texture depends on whether strong textures can develop from mechanical and thermal processing. Textures tend to not develop strongly in polyphase alloys containing a large volume fraction of second phases or constituents. Thus, texture can develop more strongly in single-phase alloys used primarily for fabrication by plastic working (drawing and stretching, bending) and less strongly in hot working of a medium-carbon steel that contains a large volume fraction of pearlite. Anisotropy of the yield strength due to texture is low for materials having a cubic lattice. Anisotropy of ductility is great enough that specific heat treating procedures are used to enhance the formability of materials used for stretching and deep drawing.

The anisotropy of textured materials is greater in alloys with a noncubic lattice and can result in the anisotropy of properties of engineering importance. In some instances, anisotropy of the physical properties can also attain engineering importance as in the production of iron-silicon alloys as sheet material for transformer cores. Other examples where anisotropy can be of engineering importance include anisotropy of thermal expansion of noncubic materials and the variation in elastic moduli in single crystals. The tensile modulus of iron, for example, varies between 190 and 280 GPa (28 and 41×10^6 psi) depending on orientation. One example of yield strength variation in design is the fabrication practice used to manufacture Zircalloy tubing in the nuclear power industry. The tubing is subjected to internal pressure from fuel swelling in service, and the hoop stress is twice the longitudinal stress in internally pressurized tubing. Therefore, design can be optimized by creating an anisotropic material that has a higher yield strength in the hoop direction. Two fabrication procedures to manufacture the tubing are possible—extrusion and welding fabrication from flat sheet. The textures created by extrusion and rolling are sufficiently different and the potential anisotropy high enough that tubing is fabricated to take advantage of the anisotropy.

Texture can be developed (intentionally or unintentionally) to varying extents for different alloys by working or heat treatment. This grain texture can impart directionality of properties, especially ductility, toughness, and fatigue crack propagation rate. Also of considerable importance in commercial quality materials, and especially those having a large volume fraction of second phases and/or inclusions, is the banding of compositional segregations and grain fibering in wrought forms. Banding gradients and grain fibering in wrought metals results in nonuniform distribution of inclusions and chemical segregations. In some alloys, chemical segregation can result in banding of microconstituents and greater anisotropy. Examples include bands of pearlite in annealed or hot-rolled steels and ferrite stringers in austenitic stainless steels.

Effect of Interstitial or Substitutional Alloy Elements. In addition, undesirable tramp (impurity) element atoms and intentional alloying element atoms are typically present in engineering materials. These atoms can occupy interstitial or substitutional positions in the crystal lattices, both of which result in lattice distortion and almost always provides strengthening among many other effects. The short-range net effects of these atoms are typically to increase strength by resistance to slip. Intentional alloying to achieve these effects is referred to as solid-solution strengthening. Nonmetal atoms typically encourage brittleness more than metal atoms.

Microstructure

Long-range order in metals forms the matrix microstructure. The microstructure is dependent on the composition and processing, and many metals and alloys can exhibit a number of different structures at the same nominal composition. This is known as polymorphism. For example, some steels can contain a matrix microstructure of ferrite and pearlite, martensite, bainite, or a combination of these, based on thermal history. The strength and toughness of a material during fracture can be drastically different based on the microstructure. In the case of heat treated microstructures, tempering can reduce hardness and strength as well as either increase or decrease the toughness.

Although there are some exceptions, most metals of engineering significance are alloys rather than pure (unalloyed) metals. Solid-solution strengthening occurs with the homogeneous dispersion of solute atoms. However, uniform composition control is not usually possible with traditional cast or ingot-based processing of wrought metals, although more uniform composition control is the motivation for some powder-metal processing. Dendritically solidified metals exhibit microscopic and/or macroscopic gradients in composition concentrations that may not be completely removed by deformation processing and thermal treatment. In rolled materials, this is known as banding, which occurs in the primary direction of hot work. An overload failure where banding was of primary importance is shown in example 6.

Example 6: Failure of Brittle Lawn Mower Blades. Mower blades manufactured from grade 1566 high-manganese carbon steel failed a standard 90° bend test. The blades had been austempered and reportedly fractured in a brittle manner during testing. The austempering treatment is intended to result in a bainitic microstructure.

Investigation. The cross section of a representative blade that had been heat treated but not bend tested is shown in Fig. 17(a). Substantial microstructural banding is apparent. The near-surface zone of the blade exhibited a thin layer of total decarburization and oxide formation. The typical core microstructure contained alternating bands of martensite and bainite, as shown in Fig. 17(b).

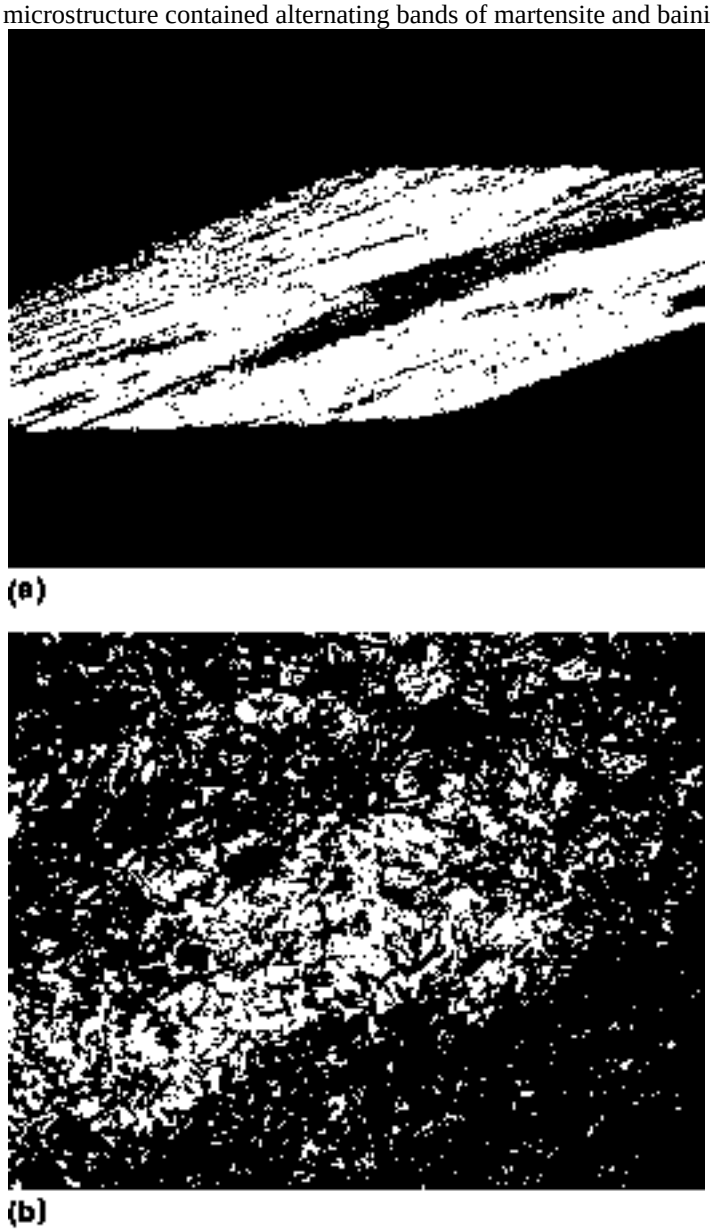


Fig. 17 Fracture of a brittle lawn mower blade (example 6). (a) Low-magnification view of the formed blade showing substantial banding. 2% nital etch. 8.9×. (b) High-magnification view of the banding showing alternating tempered martensite (light) and bainite (dark). 2% nital etch. 196×

Conclusions. The nonuniform microstructure was likely responsible for the atypical brittle behavior of the blades. Much of the microstructure consisted of martensite, which is more brittle at the same hardness level than is lower bainite. The observed structure suggests that the austempering heat treatment was performed too close to the nominal martensite start temperature (M_s). Alternate bands of slightly different alloy composition, which resulted from rolling practice and are not unusual for this alloy, resulted in actual M_s temperatures above and below the nominal M_s temperature. Since it has been shown that the M_s is generally an inverse function of alloy content, the martensitic regions were lower alloy zones.

Recommendations. It was recommended that the austempering salt bath temperature be raised 56 °C (100 °F) in the future to account for the expected localized compositional variation.

Grain Structure. As previously noted, a finer grain structure provides higher strength and ductility. The grain size of polycrystalline material plays an important role in strength levels. Qualitatively, stresses to cause plastic deformation increase with a decrease in grain size according to Hall-Petch equation:

$$\sigma_{YS} = k_1 + k_2(d)^{-1/2}$$

where k_1 and k_2 are constants and d is the grain diameter. Ductility also tends to improve with finer grain size, but the relation is more mixed for ductility than strength. Other mechanical properties (for example, fracture toughness, long-life fatigue strength) and heat

treatment behavior (for example, hardenability) directly or indirectly depend on grain size, and the dependence is strong enough that commercial mechanical/thermal processing practice is focused on keeping the grain size as small as possible.

Grain size also affects many physical and thermal properties such as magnetism and hardenability. For very specialized applications, single-crystal and directionally solidified components have been developed. These components take advantage of anisotropy, orienting the single crystal such that the strongest direction is parallel to the largest stress, as for example in turbine blades. The benefit of the single crystals or directionally solidified materials is to improve creep resistance, as creep occurs when temperatures exceed the equicohesive temperature (i.e., where grain boundaries become weaker than the crystalline grains). Single crystals or directionally solidified materials improve creep resistance by eliminating the grain boundaries or aligning them with the stress direction, respectively.

Intentional Inclusions and Included Phases. When sufficient solute is present, various phases may form that exhibit a different crystalline structure from the matrix. A supersaturated solution may develop, and the excess solute forms a discrete, pure inclusion or second-phase particle. Processing or aging can also create precipitates of compositions that are favored by diffusion kinetics. Precipitates, whether within the grains or at the grain boundaries, tend to add strength by impeding dislocation motion. This is known as “pinning.” Grain-boundary regions are also a source of pinning, and so precipitates in the grain boundaries may not add very much more pinning over that of the grain-boundary structure itself. Nonetheless, the basic intent of precipitation strengthening is to impede dislocation movement. Dispersion-strengthening processes, where very fine nonmetallic particles are mechanically added to a metal, are also done for this purpose.

Rounded precipitates and second-phase particles are often preferable from a ductility perspective. For example, graphite shape control is an essential factor on the ductility of cast irons, as shown in example 7. Rounded carbides in spheroidized steels and rounded eutectic silicon in solution-annealed aluminum-silicon alloys are other examples of preferred shapes. A channel or trough-shaped ductile morphology is often produced in metals with elongated inclusions and is described as a “woody” fracture.

In some instances, clusters of inclusions can be segregated into lamellas, which can lower strength in the short-transverse direction, depending on the orientation of the lamellas. Additionally, separation during nonuniform heating, such as welding, can cause cracking identified as lamellar tearing. This is most often seen in thick, rolled plates, where the beneficial mechanical disruption or breakup of nonmetallic inclusions does not occur to the same extent as outer regions during metalworking.

Example 7: Failure of a Ductile Iron Cone- Crusher Frame. The upper frame from a large cone crusher failed in severe service after an unspecified service duration. The ductile iron casting was identified as grade 80-55-06, signifying minimum properties of 552 MPa (80 ksi) tensile strength, 379 MPa (55 ksi) yield strength, and 6% elongation.

Investigation. The upper frame supports the crusher bowl section and can be unevenly loaded during crushing. The fracture surface of the frame section received for analysis was obliterated by postfracture corrosion. Repeated attempts at cleaning using progressively stronger chemicals revealed that no telltale fracture morphology remained. The fracture occurred at a poorly radiused thickness transition, and the fracture surface was relatively rough. No gross casting imperfections were visually apparent.

The chemical composition satisfied the typical specifications for this cast iron grade. Although samples removed from a cast component may bear little relation to grade/condition determination based on separately poured test bars, mechanical properties were obtained for a specimen removed adjacent to the fracture. Test results from the specimen were 470 MPa (68 ksi) tensile strength, 370 MPa (54 ksi) yield strength, 1.5% elongation, and 1% reduction in area. The near-surface tensile strength and elongation were far below requirements, but the measured core hardness was within the center of the acceptance range.

Microstructural evaluation indicated that the casting surface at the fracture contained corrosion damage, inclusions, and incomplete nodulization of the graphite (Fig. 18a). The etched matrix is shown in Fig. 18(b). The core microstructure consisted of pearlite and bull's-eye ferrite, typical for this material, whereas the near-surface exhibited mixed graphite morphology. Many transgranular, secondary cracks were apparent next to the fracture surface. No carbides were observed.

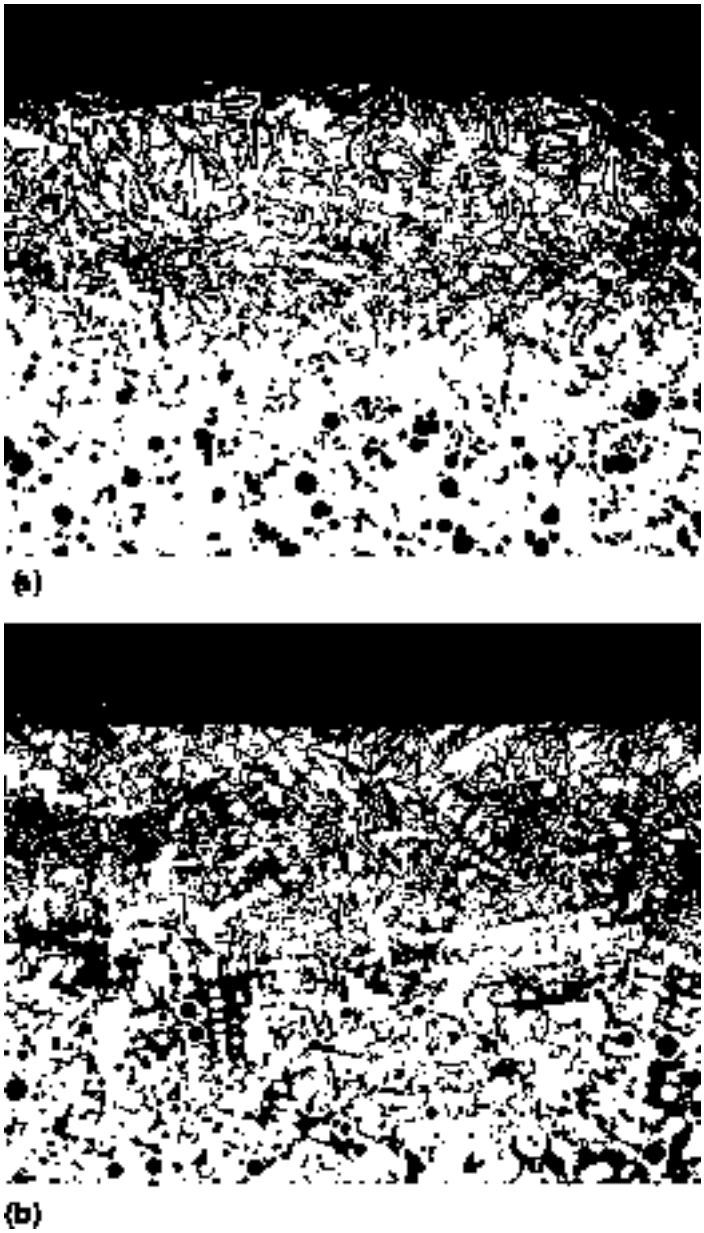


Fig. 18 Failure of a ductile iron cone-crusher frame (example 7). (a) Near-surface microstructure of the frame showing degenerate graphite morphology rather than the intended nodular structure evident in the core. Unetched. 30×. (b) The typical frame microstructure also exhibits high levels of ferrite in the undercooled surface graphite. 2% nital etch. 30×

Conclusions. The crusher frame failed via brittle overload fracture, likely due to excessive service stresses and substandard mechanical properties. Predominantly pearlitic ductile irons are very sensitive to a reduced degree of nodularity, which is exhibited as significant reduction in strength and ductility.

Recommendations. Degenerate graphite forms are not unusual at the surfaces of very large castings. Additional quality-control measures were suggested to provide better spheroidal graphite morphology at the frame surface.

Volume and Nature of Inclusions. The volume percent or number of foreign particles as well as their shape and deformability index in a material will alter the mechanical properties, as the load-bearing cross-sectional area will be affected. Large numbers of desirable inclusions can provide benefits other than strength. Normal manganese sulfide inclusions provide excellent machinability as the stringers act as chip-breakers. Naturally, these also decrease toughness and affect the level of anisotropy due to their elongated, pancake shape and parallel orientation.

Inclusions and included phases share boundaries with the matrix that may or may not be approximate continuations of the surrounding matrix lattice. These interfacial boundaries are called coherent and incoherent boundaries, respectively. The level of coherency can dictate the relative ease at which microvoids or cracks can initiate at inclusion interfaces. High levels of coherency increase the likelihood that fractures in brittle inclusions or second phases will continue as cracks in the adjacent matrix.

Impurity Inclusions. Undesirable impurity inclusions can also be inherently ductile or brittle in character and can facilitate matrix cracking, as shown in examples 8 and 9. In these cases, the properties of impurity inclusions compromised the toughness of the more

ductile matrix. Continuous intergranular phases, such as brittle carbides, or ductile phases can promote brittle overload behavior via intergranular fracture and decohesive rupture, respectively. The net effect of many brittle phases is directly proportional to their relative level of continuity. It is the individual and combined effects of atomic structure, microstructure, shape and orientation, and phases present that determine the relative ductility or brittleness of a metallic material. Processing controls all of these variables; however, service alteration is also possible in many circumstances, as discussed later in this article.

Example 8: Brittle Overload Failure of Die- Cast Zinc Snowthrower Adapters. A die-cast zinc adapter used in a snowthrower application failed catastrophically in a brittle overload manner. The mechanical properties and alloy designation were not specified. An exemplar casting from a lot known to be acceptable was provided for comparison purposes.

Investigation. The macroscopic fracture features, even in lab fracture regions, appeared brittle. Substantial shrinkage porosity was also apparent within the casting wall on the fracture surface. Examination with SEM of the fracture surface revealed transgranular brittle features and secondary cracking through platelike inclusions (Fig. 19a). A small, relative proportion of the overload fracture features appeared to be dimple rupture, and the porosity exhibited a dendritically solidified appearance.

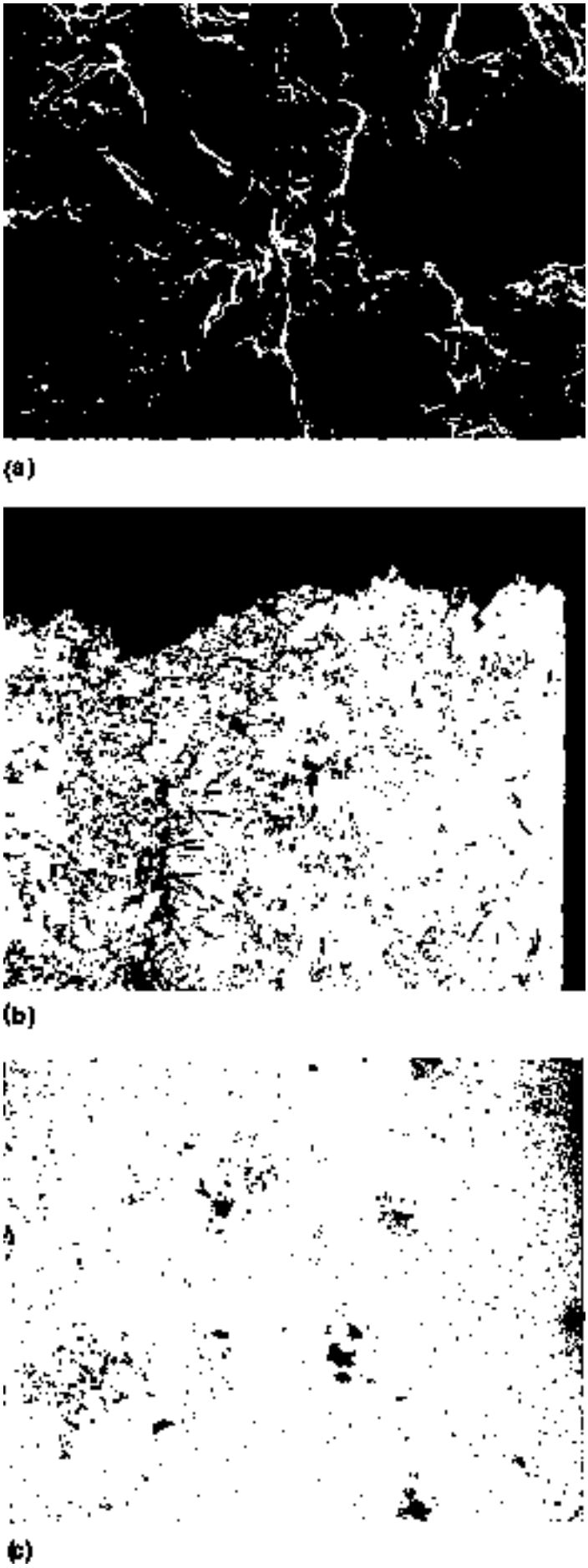


Fig. 19 Brittle overload of a die-cast zinc snowthrower adapter (example 8). (a) Fracture surface of failed zinc casting contained many large brittle constituents. Energy-dispersive

spectroscopy analysis revealed these were high-iron content inclusions. Scanning electron micrograph. 178×. (b) Microstructure of the failed casting showing massive, cracked inclusions and shrinkage. Unetched. 30×. (c) Microstructure of an exemplar zinc casting showing fewer, finer inclusions, and less shrinkage porosity. Unetched. 30×

Chemical analysis revealed a chemical composition similar to standard zinc alloy ZA-27 (UNS Z35840); however, the iron content was much higher than specified, and the copper was slightly lower than specified. The measured hardness of the failed casting in a region without porosity was 109 HB, whereas the exemplar casting was 89 HB.

Metallographic cross sections through a representative region of the fractured casting and an analogous region of the exemplar casting are shown in Fig. 19(b) and (c), respectively. Gross and fine porosity were evident in some regions of the failed casting. The microstructure contained an abundance of blocky iron-zinc platelets, as evident during SEM examination. Many of these platelets near the fracture exhibited cracking, indicating that this phase is hard and brittle. The exemplar casting did not contain a significant amount of the brittle phase, suggesting a lower iron content. Less porosity was also apparent in the exemplar die casting.

Conclusions. It was concluded that the casting failed as a result of brittle overload fracture due to excessive levels of an iron-zinc phase and gross porosity. These conditions acted synergistically to reduce the strength of the material. The composition was nonstandard, and the inherent brittleness suggests that it is unlikely that this material was an intentional proprietary alloy.

Example 9: Fracture of a Cast Steel Bracket. A cast steel bracket manufactured in accordance with ASTM A 148 grade 135/125 failed in railroad maintenance service. Ancillary property requirements included a 285 to 331 HB hardness range and minimum impact energy of 27 J (20 ft · 1bf) at -40 °C (-40 °F). The conditions at the time of failure were characterized as relatively cold.

Investigation. The bracket fractured catastrophically and exhibited no plastic deformation within the region containing several exterior fracture origins. The shiny macroscopic features were suggestive of microshrinkage and brittle overload fracture. Examination with SEM revealed primarily cleavage fracture features between dendritic shrinkage porosity and brittle fracture through inclusion troughs. Fracture had apparently occurred below the ductile-to-brittle transition temperature for this material.

The molybdenum, cobalt, and vanadium all exceeded the specification limits, and the sulfur content was near the maximum allowable. The aluminum content was relatively low, and the tensile strength appeared to be substantially lower than specified. The casting was very brittle, and shrinkage was apparent on the tensile specimen fracture surfaces. The impact strength was very low at the specified test temperature.

The typical microstructure is shown in Fig. 20(a) and 20(b). The casting exhibited a tempered-martensite microstructure with massive, eutectic type II manganese sulfide (MnS) stringer inclusions at the prior-austenite grain boundaries. These are characteristically more embrittling than the globular type I manganese sulfides. It was considered probable that grain-boundary formation of aluminum nitride contributed to the failure, but the presence of this nonmetallic compound is often difficult to verify.

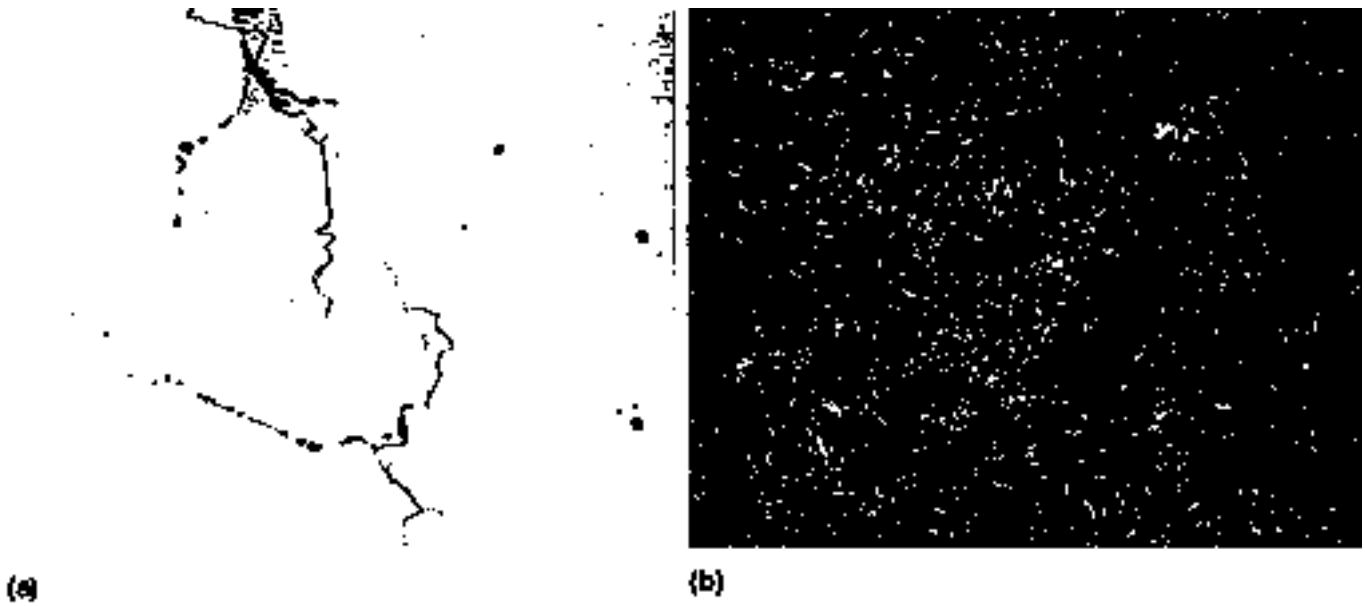


Fig. 20 Fracture of a cast steel bracket (example 9). (a) Bracket microstructure exhibiting gross type II manganese sulfide stringer inclusions. Unetched. 119×. (b) The microstructure consisted of tempered martensite, with the inclusions likely present at the prior-austenite grain boundaries. 2% nital etch. 119×

Conclusions. The bracket failed through brittle overload fracture due to a number of synergistic characteristics. The quenched-and-tempered microstructure contained solidification shrinkage, inherently poor ductility and type II MnS inclusions that are known to reduce ductility. The macro- and microscale fracture features confirmed that the casting was likely in low-temperature service at the time of failure. The composition and mechanical properties of the casting did not satisfy the design requirements.

Recommendations. It was recommended that better composition control be exerted, primarily with regard to melting, deoxidation, and nitrogen control. Better deoxidation practice would be recommended to generate the more desirable MnS inclusion morphology. Reevaluation of the casting design to minimize the shrinkage would also be beneficial.

Reference cited in this section

2. H. Nelson, “Fractography in the Failure Analysis of Engineering Structures,” Seminar, ASM Annual Meeting, 1997

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Temperature Effects

Temperature is one of the most important extrinsic variables that influence the mechanical behavior of materials. As previously noted, the temperature dependence of the mechanical properties of fcc materials is quite distinct from those of the bcc materials. This dissimilar behavior is related in part to the nature of dislocation motion within the individual crystal lattices, and fcc metals exhibit plastic deformation and ductile fracture mechanisms even as temperatures are lowered. This general difference in the temperature-dependent mechanical behavior for fcc and bcc metals is shown in the deformation behavior maps of Fig. 21(a) and (b).

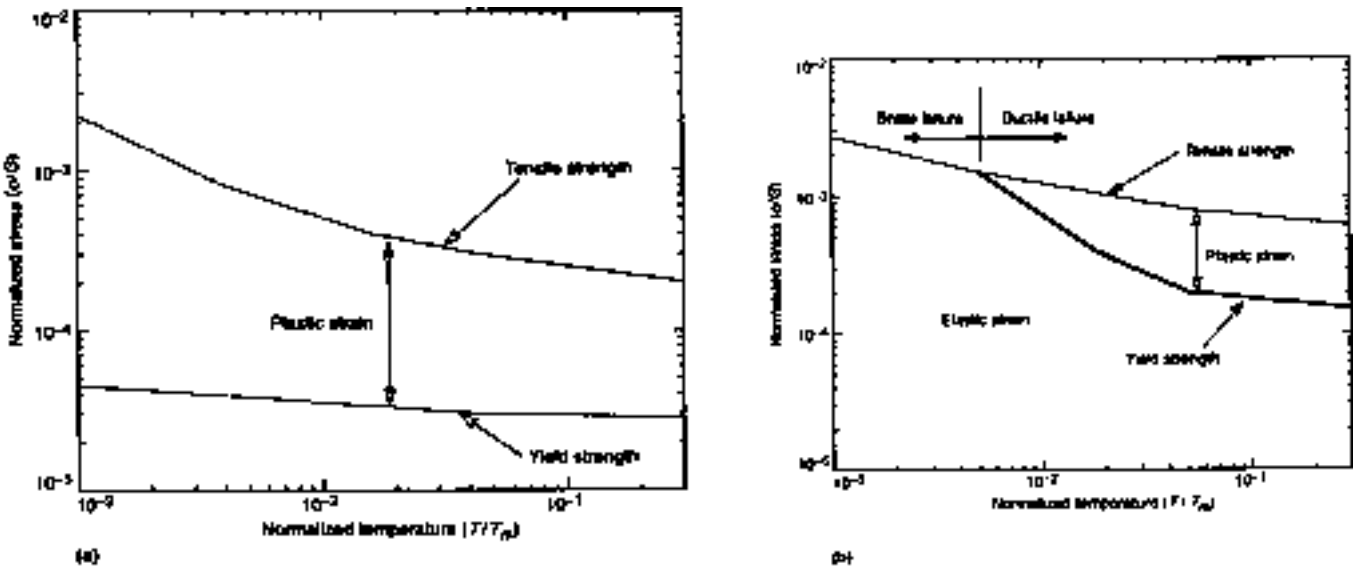


Fig. 21 Simplified deformation behavior (Ashby) maps of unalloyed annealed metals with (a) face-centered cubic crystal structure and (b) body-centered cubic crystal structure. Engineering alloys may behave somewhat differently than unalloyed metals, but these general trends are relatively consistent (see text). Source: Ref 3

Figure 21 is for annealed unalloyed metals, and it illustrates general differences. The ultimate tensile strengths of the fcc metals have stronger temperature dependence than those of bcc metals. However, the variation of yield strength with temperature of the fcc metals is not as pronounced as for the bcc metals. The characteristics may be altered with alloying, but the general trends remain relatively consistent. For example, solid-solution strengthening typically increases the yield strength and ultimate strength, while also increasing the temperature dependence in the yield strength of fcc alloys. However, the temperature dependence of ultimate strength is still greater than that of yield strength, thus allowing ductile behavior in fcc alloys.

The ductility of fcc metals is also evident in terms of toughness at low temperature, as shown in Fig. 22. This is why fcc metals such as austenitic stainless steels and aluminum alloys are used extensively in cryogenic applications. However, it is important to note that some metallurgical transformations at low temperature can alter the structure. For example, some

austenitic steels are susceptible to martensitic transformation at cryogenic temperatures (Ref 3). This can lead to low-temperature embrittlement and brittle fracture.

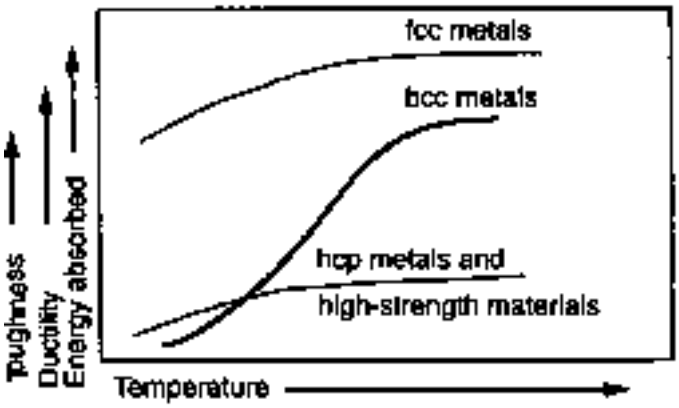


Fig. 22 Effect of temperature on toughness and ductility of face-centered cubic (fcc), body-centered cubic (bcc), and hexagonal close-packed (hcp) metals

Low Temperature

In general, lowering the temperature of a solid increases its flow strength and fracture strength. At low temperatures, the thermal activation of dislocations is minimal; therefore, more applied stress is required for deformation. This inverse proportionality is generally true and is also applicable to nonmetallic materials such as polymers and ceramics. The primary exceptions to the inverse proportionality are the result of a wide variety of microstructural alterations and other alloy specific changes. These changes are dictated by the desirable minimization of free energy in whatever manner it may be attained. Spontaneous isothermal phase changes can affect ductility as operating temperature is decreased. There are relatively few engineering alloys that undergo phase changes at cryogenic temperatures or at room temperature, although as noted, some austenitic steels are susceptible to martensitic transformation and embrittlement at low temperature. Some effects such as aging, or precipitation hardening, can occur at room temperature or lower. Low temperatures may also retard other processes. For example, the so-called “icebox” rivets often used in aerospace applications are solution-annealed aluminum rivets (typically alloys 2017 or 2024) that must be stored at low temperature prior to installation. After they are driven, the rivets naturally age to the desired moderate strength T4 temper condition.

Ductile-to-brittle transition is the well-known drop in ductility and toughness of bcc metals as temperature is reduced. The ductile-to-brittle transition temperature (DBTT) refers to the region in which the toughness drops and the fracture mechanisms and features change from ductile to brittle (Fig. 21b and 22). This transition is between dimple rupture and transgranular or intergranular brittle features. This transition is primarily evident in bcc and hcp metals, but is somewhat analogous to the glass transition temperature evinced by polymeric materials. Metals that have a high yield strength sensitivity to temperature fluctuation exhibit a DBTT. Many factors affect the transition temperature, including the grain size and strain rate. An increase in grain diameter raises the DBTT (Fig. 23), as does an increase in loading rate (Fig. 15). Some elements, such as nickel and manganese, can reduce the DBTT.

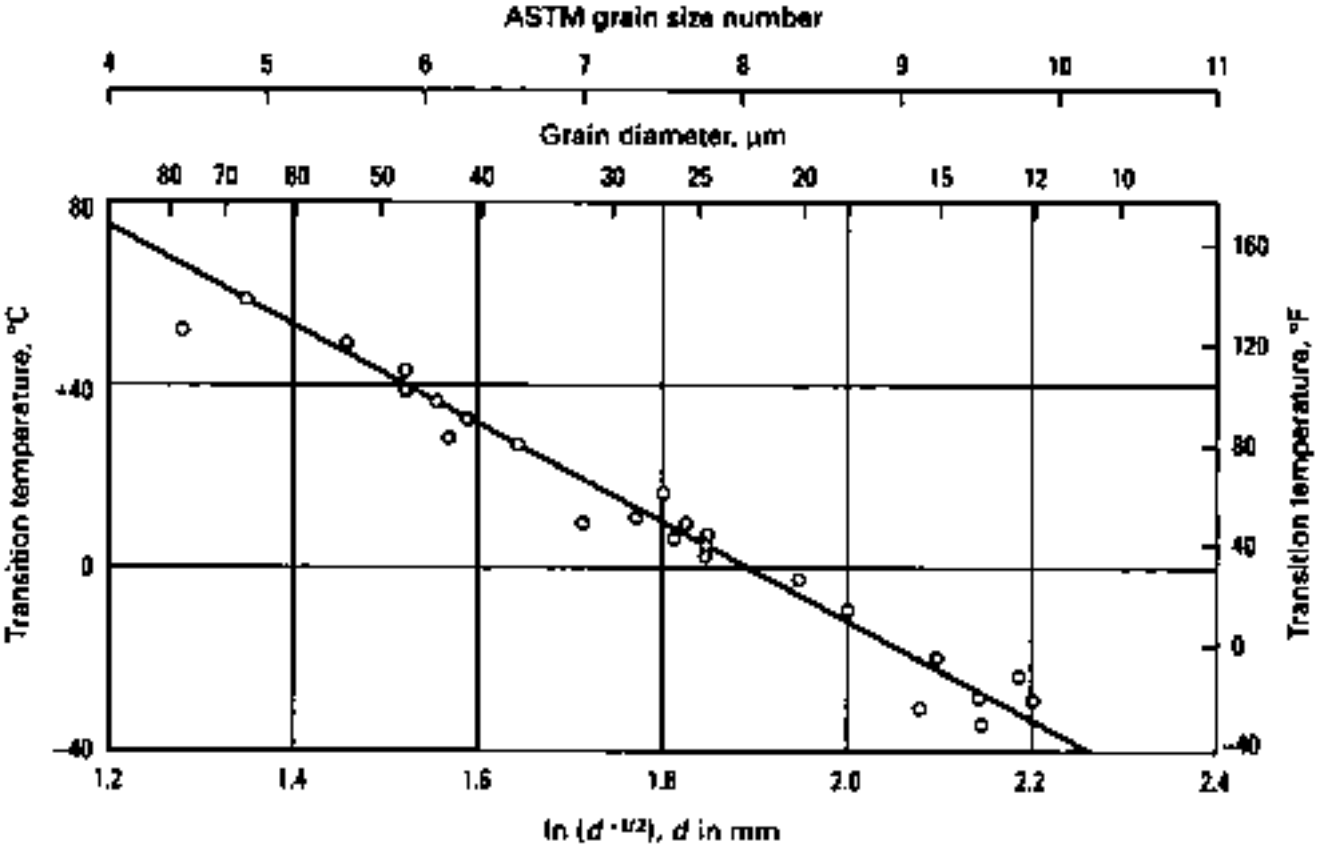


Fig. 23 Effect of grain size on the ductile-to-brittle transition temperature (DBTT) of 0.11% C mild steel. Source: Ref 4

Changes in toughness as a function of temperature are quantified by a number of different measures that have found use in materials evaluation, design, and failure analysis. In particular, dynamic fracture toughness of engineering alloys as a function of temperature are evaluated by various methods of impact testing and instrumented drop-weight testing (see the article "Impact Toughness Testing" in *Mechanical Testing and Evaluation,* Volume 8 of *ASM Handbook,* 2000). Measurement of dynamic fracture toughness is important because many structural components are subjected to high loading rates during service or must survive high loading rates during potential accident conditions. Furthermore, because dynamic fracture toughness is generally lower than static toughness, impact toughness testing can be a more conservative measure of toughness under dynamic conditions. However, measurement of dynamic fracture toughness is more complex than quasi-static (e.g., K_{ic}) toughness. Impact toughness tests, such as the Charpy test, are only a qualitative measure for materials evaluation. Results from impact toughness on laboratory specimens cannot be used for quantitative design of components because results depend on specimen size. In contrast, results from fracture-mechanic testing (e.g., K_{Ic} or K_{Id} under dynamic conditions) do scale.

Nonetheless, impact toughness testing such as the Charpy test does provide a useful way to screen or evaluate materials. Several types of criteria are used to evaluate the effect of temperature on toughness from impact testing. For example, impact testing over a spectrum of test temperatures is commonly done to determine the effects of temperature and the DBTT region of bcc metals. Other evaluation criteria include:

- *The nil ductility transition temperature (NDTT),* usually measured by drop-weight test and based on a crack/no crack criterion
- *The fracture appearance transition temperature (FATT),* a measure of the percentage of shear morphology measured on impact specimens

The upper- and lower-shelf regions and the transition zone can be identified by impact testing at a variety of temperatures and comparing the impact energy, the shear percentage, the lateral expansion, or all three. Regardless of the test procedure or the method used to define the DBTT transition, results depend on the metal and loading conditions and strain rate. Many design criteria use a standard but arbitrarily selected impact energy value, for example 20 J (15 ft · 1bf), in some cases to empirically specify the gradual change in toughness with temperature of a material when determining suitability for service.

Many catastrophic brittle failures are associated with the reduction of toughness at low temperatures. The Liberty ships of World War II are a well-known example. Of the 4694 ships considered in the final investigation, 24 sustained complete fracture of the strength deck, and 12 ships were either lost or broke in two. In this case, the need for tougher structural steel was even more critical because welded construction was used in shipbuilding instead of riveted plate. In riveted-plate construction, a running crack must reinitiate every time it runs out of a plate. Similar long brittle cracks are less likely or rare in riveted ships, which were predominant prior to welded construction (although even some riveted ships have provided historical examples of long brittle fracture due, in part, to low toughness).

Low notch toughness is a more critical factor for long brittle cracks in welded ships, as the fabricated structure provides a continuous path for brittle cracking. For example, the fracture origins in the Liberty ships were at weld discontinuities, and unstable crack propagation resulted in brittle fracture of entire welded ship structures. The existence of coarse pearlitic structure (from furnace-cooled plate) was the basis for the brittle structural failures experienced by the ships, which split decks even while at dockside (for example, see Fig. 1 in the article "Failures Related to Welding" in this Volume). The cracking was minimized by specifying normalized plate, which had fine-grain pearlite microstructure that is much more tolerant of weld discontinuities than coarse pearlite formed by furnace cooling. Low temperatures at or below the transition temperature of the steel was also a factor in these failures.

Elevated Temperature

Generally, increasing the processing or operating temperature will initially result in a decrease in mechanical strength and an increase in ductility. Hot working takes advantage of this extreme plastic behavior. The simultaneous relaxation of deformation stresses also results in the suppression of nucleated microvoids, permitting more reduction without fracture than possible during cold deformation. It should also be noted that the alloy-dependent drop in the elastic modulus with temperature limits the materials that are suitable for high-temperature service.

All metals exhibit an elevated-temperature transition from normal overload morphology to brittle intergranular fracture (with an accompanying loss in toughness and ductility). At this temperature the strength of the grains and the grain boundaries are equal; therefore, this is often identified as the equicohesive temperature (ECT). Similar to the DBTT, the ECT is affected by material characteristics and the loading conditions. The ECT has been identified as being equivalent to the recrystallization temperature. There are so many variables that affect the ECT, that quantification of an actual temperature is possible for only a specific set of conditions and may not be useful.

Heat treatment is performed to achieve desired mechanical properties by precipitation, recovery, formation of desirable matrix microstructure and included phases, grain refinement, and homogenization. Excessive time at temperature, primarily in high-temperature service, can cause undesirable diffusion of alloying elements, formation of deleterious phases, grain growth, and many other changes. All changes will affect overload behavior. At even higher temperatures, the mechanical properties of a metal are compromised as the melting point is approached.

On heating and cooling, the normal expansion and contraction stresses due to thermal gradients are complicated by the volumetric changes accompanying crystalline structure and phase changes. Metals can exhibit allotropic transformations, where different crystalline structures are evident in the material based on whether they are above or below a critical temperature. In steels, for example, normal bcc ferrite changes to fcc austenite on heating above the Ac_3 temperature. The stresses accompanying phase changes can result in brittle fracture such as hot cracking, cold cracking, and quench cracking.

References cited in this section

3. R. Reed and A. Clark, *Materials at Low Temperature,* American Society for Metals, 1983

4. N.J. Petch, The Ductile-Cleavage Transition in Alpha-Iron, *Fracture,* B.L. Averbach, Ed., Technology Press, 1959, p 54–67

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Effects of Mechanical Loading

Overload failures, by definition, are incidents in which the stress state of the component at the time of fracture is of primary concern to the failure analyst. The type and magnitude of stresses placed on a part in service must be considered, along with the physical and analytical evidence, to identify the cause(s) of the failure. Loading information is usually available in the engineering design specifications, but much can be logically inferred by analysis alone. It is not unusual for the results of a failure investigation to suggest the actual service stresses were drastically different from those anticipated during original design.

Brittle or ductile overload failures can occur simply as a result of insufficient strength for the applied forces. This can be due to either underdesign or incorrect material usage. Design faults can result from inaccurate determination of the forces that would be applied to a component in service. The type, magnitude, and application of the stress can be miscalculated, such as anticipating purely static tensile forces when cyclic impact actually occurs. The material selection process is then based on flawed assumptions, and an inappropriate material or material condition is selected. Overload failure may also be due to extreme loading that could not have been logically foreseen during design, such as intentional misuse or abuse.

In some cases, the wrong material or material condition is inadvertently used. Deviation from the correct material specification can be subtle or profound. Subtle deviation may result in reduced service life or distortion, whereas profound deviation from the specified material may result in cracking or fracture. Underdesign, poor-quality materials, improper heat treatment, and incorrect processing are often encountered in metallurgical failure analyses. Due to the differences in corrective actions typically employed, embrittlement-related failures are not typically considered underdesigned.

Overload failures due to underdesign are uncommon. This is primarily due to the current state of engineering design expertise, which has grown through analysis of past failures and applied mechanical testing. The use of design safety factors also reduces the frequency of underdesign failures. Design review is another method of preventing failure. Design reviews may also be part of the failure analysis process (for example, see the article “Design Review for Failure Analysis and Prevention” in this Volume). In some cases, materials failures analysis is also done first to identify potential materials- or manufacturing-related causes. If nothing is evident from this stage of the investigation, then the evaluation of causes may require a design review, which may include stress analysis by traditional techniques, finite-element analysis (FEA), and/or fracture mechanics. These techniques are discussed elsewhere in this Volume, but a brief description follows.

Stress Analysis and Fracture Mechanics

Stress analysis is a basic design tool and also an effective tool in failure analysis. In general, there are two types of conditions that may lead to structural failure:

- Net-section instability where the overall structural cross section can no longer support the applied load
- The critical flaw size (a_c) is exceeded by some preexisting discontinuity or when subcritical cracking mechanisms (for example, fatigue, SCC, creep) reach the critical crack size

Failures due to net-section instability typically occur when a damage process such as corrosion or wear reduces the thickness of a structural section. This type of failure can be evaluated by traditional stress analysis or FEAs, which are effective methods in evaluating the effects of loading and geometric conditions on the distribution of stress and strain in a body or structural system. This may involve traditional stress analysis of applied loads or FEA. These methods can also help identify the likely areas of stress concentration where cracking may have initiated.

However, stress analyses by traditional methods or FEA do not easily account for crack propagation from preexisting cracks or sharp discontinuities in the material. When a preexisting crack or discontinuity is present, the concentration of stresses at the crack tip becomes asymptotic (infinite) when using the conventional theory of elasticity. In this regard, fracture mechanics is a useful tool, because it is a method that quantifies stresses at a crack tip in terms of a stress-intensity parameter (K):

$$K = Ys\sqrt{a}$$

where Y is a geometric factor (typically on the order of about 1), σ is the gross stress across the fracture plane, and a is the crack length. The stress-intensity parameter K quantifies the stresses at a crack tip, and a critical stress-intensity value (K_c) thus can be defined as:

$$K_C = Ys_f\sqrt{a_c}$$

where σ_f is the fracture stress occurring with a critical crack size, a_c. The critical stress intensity, also known as fracture toughness (K_c), is the value of stress intensity (K) that results in rapid, unstable fracture. Fracture toughness (K_c) depends on both the thickness of the section and the ductility of the material. For a given material, the fracture toughness (or critical stress intensity, K_c) decreases as section thickness is increased. The value of K_c decreases with increasing section

thickness until a minimum value is reached. The toughness at this minimum, which is a inherent material property, is the plane-strain fracture toughness (K_{Ic}). Plane-strain fracture is a mode of brittle fracture without any appreciable macroscopic plastic deformation and is thus referred to as linear-elastic fracture mechanics (LEFM). The general conditions for LEFM analysis are expressed as:

$$\text{Thickness} \geq 2.5(K_{Ic}/\sigma_y)^2$$

where σ_y is the yield strength.

Linear-elastic fracture mechanics is a useful tool in failure analysis as many (and perhaps most) structural failures occur by the combined processes of crack initiation followed by subcritical crack growth mechanism (for example, fatigue, stress corrosion, creep) until a critical crack (a_c) is reached. In this regard, fracture mechanics is an effective tool for evaluating critical flaw size (a_c) that leads to rapid unstable fracture and can help answer questions during a failure analysis, such as (Ref 2):

- Where should one look for the transition from subcritical crack growth to unstable rapid fracture?
- What was the load on the component at the time of failure?
- Was the correct material used and manufacturing/processing sound?
- Was the part designed properly?
- Did the environment influence the failure?

Of course, many situations may involve thin sections and/or very ductile materials, where the conditional constraint (plane-strain) for LEFM may not apply in evaluating stress and crack size conditions for the onset of rapid (unstable) fracture. In this case, the use of elastic-plastic fracture mechanics (EPFM) is required, as the process of unstable fracture involves some plasticity. Plane-stress fracture toughness (K_c) is higher than plane-strain fracture toughness (K_{Ic}), but when thinner sections and more ductile materials are involved, net-section instability becomes a factor. Linear elastic fracture mechanics is generally valid for subcritical crack growth conditions, except in the case of creep crack growth whose plasticity is involved. More information on fracture mechanics is contained in the articles “Stress Analysis and Fracture Mechanics” and “Fracture Appearance and Mechanisms of Deformation and Fracture” in this Volume.

Stress Conditions

Stress Type. The type of stress applied to a component is of interest to the failure analyst, as it will help to explain the fracture features, fracture location, and many other attendant characteristics. Stresses are usually identified as tension, compression, or shear (i.e., the state of stress at a point is defined in terms of three normal orthogonal components and six shear components). The loading, which produces the stresses in particular regions of a component, can be pure tension, pure compression, pure bending, pure torsion, or some combination of these. The external loading conditions, the component geometry, and the presence of cracklike imperfections control the internal state of stress. Complex stress states may occur when components are not loaded uniformly and/or when geometric factors cause regions of stress concentration. An overload failure that occurred under a complicated stress state is presented in example 10.

The stress states in a component often contain a hydrostatic component, that is, the mean of the three normal stresses at any point. A compressive hydrostatic stress is often less of a concern in metals than a tensile hydrostatic stress. A tension hydrostatic stress promotes brittle fracture because the maximum shear stress, which is a function of the difference in the principal stresses, is decreased by hydrostatic tension. One observes shear lips at the surface of a cross section because plane stress exists there—that is, one of the principal stresses is zero at the surface. Triaxial tension is promoted by notches and so forth, due to the constraint exercised by the surrounding material away from the stress concentrator. Thus, plane-strain deformation (in which a hydrostatic stress is present) results in fracture that is macroscale brittle. The microscale appearance may be either ductile (i.e., MVC) or brittle (cleavage, quasi-cleavage). This state may prevent yielding by dislocation slip in the preferred directions, thereby encouraging brittle fracture during overloading. However, the effect of a hydrostatic stress component to the total stress state is usually discussed in terms of macroscopic appearances; and on a microscopic scale in a polycrystalline body, the local stresses may still cause some dislocation motion.

The formation of a hydrostatic stress state depends on the configuration of part geometry, loading, and material condition. For example, the walls of a pressurized component may be subjected to a hydrostatic stress state depending on the thickness of the wall. If the component is “thin walled” (for example, a radius/thickness ratio, r/t, greater than approximately 20), the vessel wall would be loaded in plane stress; that is, no appreciable radial stress would be present, only hoop and axial stresses would occur. If the vessel is “thick walled,” then there is a radial stress and a hydrostatic stress would be present. The only other way to develop a hydrostatic stress in a thin-walled vessel would be from an internal discontinuity adjacent to the wall surface.

When a discontinuity or cracklike imperfection is present in a material, the magnitude of the hydrostatic stress increases as the section thickness increases, and fracture is more brittle with a mode of plane-strain fracture. Thick materials can

exhibit low toughness and generally flat fracture features (although the microscopic features and mechanisms may be either MVC or cleavage). Additionally, heavy sections typically have a nonconstant microstructure through the thickness, with poorer properties at midthickness. Thin materials, or situations when cracks approach free surfaces to approximate thin materials, deform (neck) and are thus in a plane-stress state. This permits ductile fracture manifested by a shear lip at the edges of fractures and at the final fracture location under plane-stress conditions. Materials can exhibit a geometrical transition from plane-strain to plane-stress behavior analogous to a temperature transition. The various transitions evidenced by metals are not mutually exclusive occurrences.

Stress Magnitude. The magnitude of the applied and residual stresses is also an important part of an overload failure. When an overload fracture has occurred, as indicated by the physical evidence, the actual magnitude of the stresses must have been sufficient to initiate fracture; that is, the stresses applied to the material that fractured exceeded the capability of that specific material. Mechanical testing is often an integral portion of a failure analysis investigation to determine whether the stresses were excessive, whether the part or material was weak, or both. The magnitude of applied stress is usually compared to the material yield strength in distortion failures and to the material fracture strength in the cases of ductile and brittle overload fractures. The failure analyst also compares these values to what the designer anticipated for service stresses and material properties.

Stress Application. To develop a complete understanding of the forces on a failed component, the analyst must consider the application of stresses. Loading can be constant or cyclic, slow or rapid, concentrated or distributed, or any mixture of these, and thus different internal stresses occur at different locations and/or at different points in time. Failure initiates where the *local* stress exceeds the *local* strength. Materials will exhibit differing overload failure characteristics based on these stress application variables. Cyclic loading most often results in fatigue failures rather than overload failures. A delayed overload fracture in a cyclic application suggests an unusual loading event or that embrittlement of the material has occurred. Progressive fracture features in a statically loaded component may suggest a corrosion mechanism.

Example 10: Failure of a Jack Cylinder. A jack cylinder split open during simulated service testing (Fig. 24a). The intended internal test pressurization was reportedly analogous to typical service. The material and mechanical properties of the cylinder pipe were unknown.

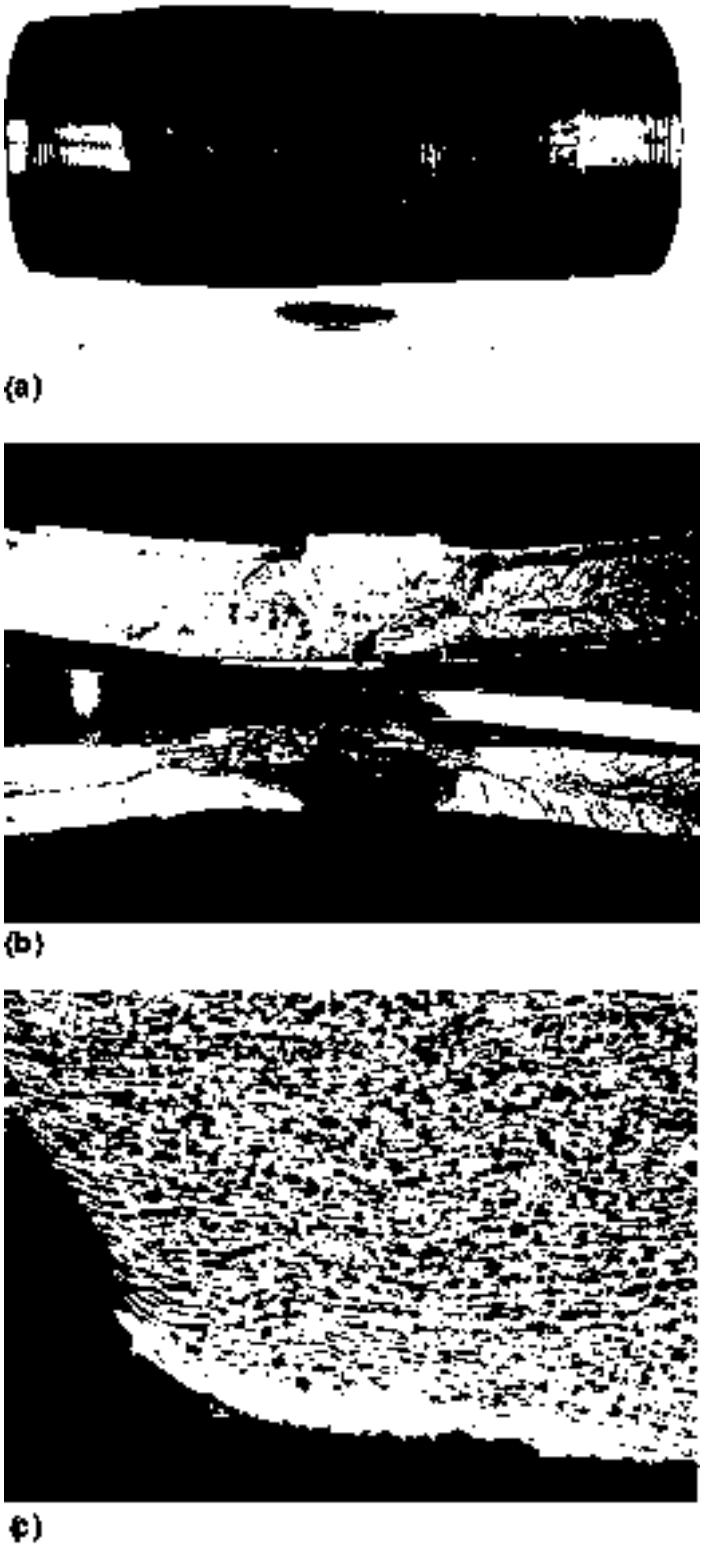


Fig. 24 Failure of a jack cylinder (example 10). (a) Cylinder that split during testing. Substantial bulging is apparent at the failure location. (b) Opened crack surfaces from the cylinder in (a). Both the inside and outside diameter surfaces exhibited a shear-lip appearance. (c) Metallographic cross section near the inside diameter surface showing decarburization and a shear lip. 2% nital etch. 119×

Investigation. The single-event crack was longitudinal and bulged noticeably in its center. The crack center, which was identified as the likely crack origin, is shown after excision in Fig. 24(b). The surfaces are relatively flat, but exhibit angled shear lips at the inside and outside diameter surfaces. Visible “chevron” markings pointed back to the origin in the center of the bulged cylinder. Fractographic examination revealed quasi-cleavage and dimple-rupture features throughout the flat-fracture region and shear-lip region, respectively.

Chemical analysis indicated that the pipe satisfied the requirements for a grade 1045 medium-carbon, plain carbon steel. Tension testing of a specimen removed from the pipe section opposite the crack revealed 779 MPa (113 ksi) tensile strength, 634 MPa (92 ksi) yield strength, and 12% elongation.

Microstructural examination confirmed the crack surface profile with somewhat flat fracture in the core with shear lips. The near-surface grains were severely distorted. The inside diameter shear lip is shown in Fig. 24(c). The microstructure away from the crack consists of cold-worked ferrite and pearlite, suggestive of the normalized condition. Surface decarburization and oxide formation were evident on the pipe surfaces.

Conclusions. The cylinder pipe burst in a mixed brittle and ductile manner due to overpressurization. The bearing strength of the pipe was likely to be slightly compromised by a low-strength layer of decarburization. The pipe was a moderately cold-worked plain carbon steel.

Recommendations. The analysis results indicated that the testing procedure should be evaluated for the possibility of inadvertent overpressurization. If no accidental burst could occur with the test configuration, it may be necessary to analyze successfully tested cylinders to identify changes in material, and perhaps heat treatment, that may have contributed to this failure.

Residual Stresses. Residual tensile stresses can be present in components, resulting from welding, casting, heat treatment, and grinding, as well as other manufacturing and assembly processes. In many cases, residual stresses can be additive to applied stresses and thereby contribute to overload failure. Hot working allows immediate stress relaxation, but cold working does not. Cold working is a strengthening mechanism for many materials, including those that may otherwise have been heat treated to achieve the necessary strength level. In wrought aluminum, for example, the “H” temper designations indicate cold-work strengthening, whereas the “T” tempers indicate heat treatment with or without cold-work strengthening. Some strain-hardened alloys (especially those of aluminum) may exhibit a spontaneous decrease in strength due to recrystallization, sometimes referred to as age softening.

Residual compressive stresses at the surface, induced by processes such as shot peening and some surface-hardening treatments such as carburization in steels, are generally considered beneficial, both in overload and fatigue resistance. However, the effects of the surface treatment may be overcome if stress concentrators, corrosion, and similar effects penetrate the compressive stress layer. Even after surface treatments, there is still a subsurface tensile stress under the compressive stress layer. Thus, the residual compressive surface stress is especially beneficial in delaying fatigue crack initiation at the surface, but has little or no influence during the stage of critical (unstable) crack propagation.

Stress Concentration and Defects. Good engineering designs, even when encompassing quality standards, cannot totally prevent the presence of inadvertent discontinuities. These deleterious discontinuities can be present as casting and forging imperfections, thermomechanical processing cracks, welding anomalies of all types, and others. These features can act as stress raisers and can impart brittle character. A failure resulting from an undetected crack is presented as example 11.

Even in enlightened engineering design, where a proper fracture mechanics approach considers the imperfections that are most assuredly present, unexpected stress concentrators can cause failure. It is not always feasible to create components without the possibility of inadvertent scratches, nicks, and gouges. All manufacturing processes contain variability that must be understood and monitored, if not controlled. For these reasons, the level of inspection and nondestructive examination applied to a component is usually a measure of its criticality and the consequences of its failure. *However, the designer should anticipate the types of imperfections likely to be present and the manufacturing process should include controls and inspections to ensure that the design allowables are not exceeded.*

Example 11: Failure of Tension Springs During Installation. Two large tension springs fractured during installation. The springs were manufactured from a grade 9254 chromium-silicon steel spring wire. The associated material specification allows wire in the cold-drawn or oil-tempered (quenched-and-tempered) condition. The specified wire tensile strength range was 1689 to 1793 MPa (245 to 260 ksi). The finished springs were to be shot peened for greater fatigue resistance.

Investigation. The springs fractured within the coils rather than at the end hooks, where fatigue fracture often occurs. The springs were painted black. Examination of the fracture locations revealed that the features were similar and both springs contained a precracked region. A representative fracture surface is shown in Fig. 25(a). No additional cracks were observed.

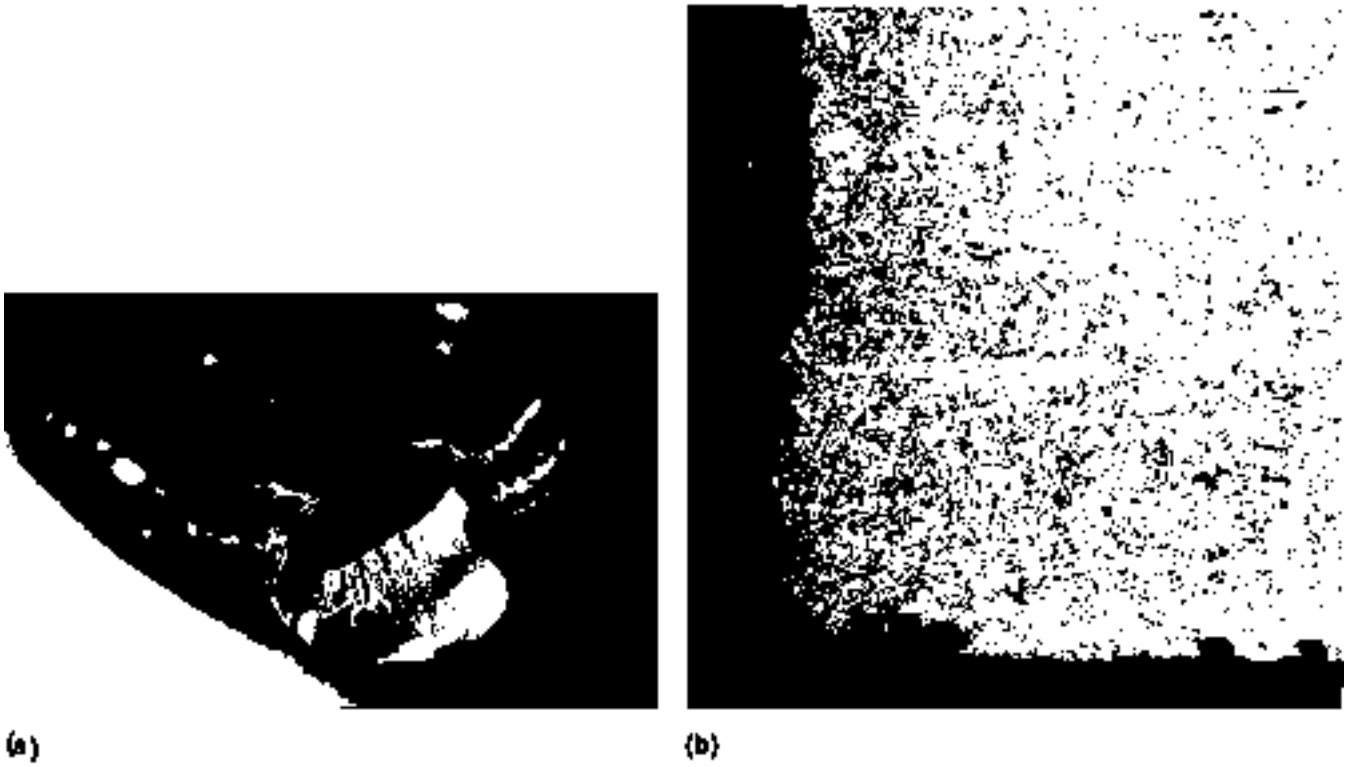


Fig. 25 Failure of tension springs (example 11). (a) Spring fracture surface showing the presence of a discolored precrack region. 3×. (b) Cross section through the precracked region of the spring revealing a thick scale (vertical surface) on the fracture surface. 2% nital etch. 148×

The chemical composition of the tested spring satisfied the specification requirements. Energy-dispersive x-ray analysis (EDS) of the discolored precracked region detected a very high level of phosphorus, suggestive of a phosphate coating applied prior to painting. This tenacious contamination could not be removed by chemical means, which were limited to minimize the potential for damage to the remaining fracture features. Examination by SEM did not reveal discernible fracture features in the precrack zone, but the remainder of the fracture consisted of dimple-rupture evidence.

A longitudinal metallographic cross section through the wire in Fig. 25(a) is shown in Fig. 25(b). The precrack region is toward the left of this micrograph. Some transgranular, secondary cracking is apparent near the surface. The near-surface microstructure consists of tempered martensite, but some partial decarburization and corrosion pits are evident on the wire surface. The surface did not appear to have been shot peened.

The microhardness of the wire averaged 591 Knoop (HK_{500}). This equates to an approximate ultimate tensile strength slightly higher than the specified range.

Conclusions. The springs failed during installation due to the presence of preexisting defects. The crack surfaces were corroded and phosphate coated, confirming that they had occurred during manufacture. Installation, which presumably entailed some axial extension, resulted in ductile overload failure at the crack sites.

Recommendations. Evaluate the manufacturing steps to identify the process(es) wherein the cracking is likely occurring. The spring-coiling process was of particular interest due to the appearance of the cracks. It was further recommended that a suitable nondestructive method such as magnetic particle inspection be implemented.

Reference cited in this section

2. H. Nelson, “Fractography in the Failure Analysis of Engineering Structures,” Seminar, ASM Annual Meeting, 1997

Brett A. Miller, Stork Technimet, Inc.

Service Damage or Alteration

Overload failures can result from service damage such as mechanical alteration of the material by dynamic loading or distortion or progressive forms of damage by corrosion, wear, and subcritical crack growth. The damage may result in a net-section instability from a reduction in section size, or the damage may result in a critical crack size that leads to overload. In addition, parts that are dynamically loaded or deflected can experience strain hardening to an extent based on the material and loading characteristics. Some materials are also sensitive to the rate of load application and may be altered by strain-rate hardening. Mechanical twinning occurs in some alloy systems. These effects can change the ductility substantially; however, fatigue cracking results more often than overload failure, as these hardening effects are often gradual.

More severe mechanical alteration can result in buckling and distortion effects. Plastic deformation and elastic instability, such as buckling, can create a site for stress concentration. Localized tensile or compressive yielding creates microvoids that can reduce the mechanical strength and the fatigue resistance. Loading history can also significantly alter the load-bearing characteristics of a component. For example, compressive yielding can result in a reduced elastic modulus upon subsequent tension testing. This phenomenon is called the Bauschinger effect (Fig. 26).

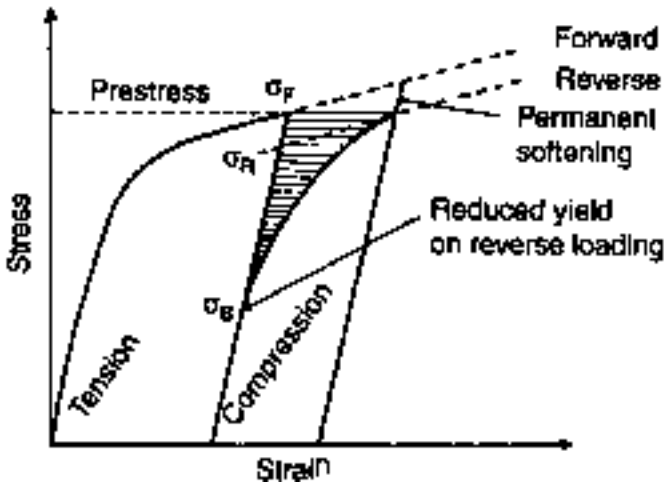


Fig. 26 Bauschinger effect. The softening effect is exaggerated for clarity. σ_F, forward stress; σ_R, reverse stress; σ_B, backward stress

Progressive forms of damage may also lead to overload failure by unstable cracking from a discontinuity or, in the absence of a critical crack size, processes that reduce the effective cross section. In static-loading applications, overload failure is the likely consequence, while cyclic loading with corrosion or wear damage is obviously more complex. In damage situations, it is imperative that the relative importance of the damage to the component be determined during the metallurgical failure analysis investigation. For instance, ductile overload failure with relatively little preexistent corrosion damage may suggest a severe loading event occurred and that the corrosion was of little consequence. When the damage is the proximate cause of a failure, such as abrasive wear or pitting corrosion, it is disingenuous to identify it as an overload failure, although overloading certainly occurred at the time of final separation. Somewhat severe mechanical alteration that occurred by an abused piece of coal mining equipment is shown in Fig. 27. Continued abrasion distortion of the near-surface material occurred after cracking. In the most severe cases of mechanical alteration during service, cracks are formed and can possibly lead to failure.

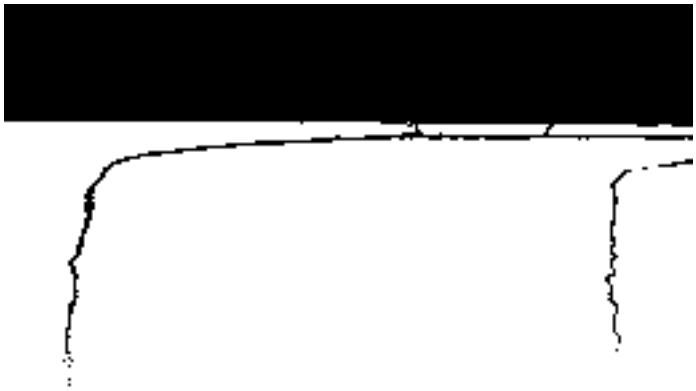


Fig. 27 Near-surface cracking present on a heavily abraded mining component. Continued abrasion after intergranular crack formation has deflected the perpendicular cracks. Unetched. 30×

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Embrittlement

Many engineering alloys are susceptible to embrittlement from specific mechanisms manifested during heat treatment, processing, storage, and service. The specific causes of embrittlement vary, depending on the metallurgical stability of the specific type or family of alloy under various manufacturing or service conditions. In general, however, embrittlement of metals is typically activated by thermal or environmental conditions, as described further in this section. Many types of embrittlement have been identified for steels, and of course nonferrous metals are also susceptible to embrittling phenomena. However, this article focuses on steels, as their varying service conditions have revealed numerous conditions for embrittlement, as briefly summarized in Table 3. Embrittlement phenomena of steels also are discussed in greater depth in the article "Embrittlement of Steels" in *Properties and Selection: Irons, Steels, and High-Performance Alloys*, Volume 1 of the *ASM Handbook* (Ref 5). Similar content is also provided in the article "Visual Examination and Light Microscopy" in *Fractography*, Volume 12 of the *ASM Handbook* (Ref 6).

Table 3 Summary of the types of embrittlement experienced by ferrous alloys

Embrittlement type	**Steels susceptible**	**Causes**	Result
Strain-age embrittlement	Low-carbon steel	Precipitation after deformation processing	**Strength increase, ductility decrease**
Quench-age embrittlement	Low-carbon steel	Quenching followed by precipitation	**Strength increase, ductility decrease**
Blue brittleness	Carbon and alloy steels	230–370 °C (450–700 °F) exposure	**Strength increase, toughness and ductility decrease**
Stress-relief embrittlement	Alloy and stainless steel	Postweld heat treatment	**Toughness decrease**
Temper embrittlement	Carbon and low-alloy steel	375–575 °C (700–1070 °F) exposure	**Increase in DBTT**
Tempered martensite embrittlement	Heat treated alloy steel	200–370 °C (400–700 °F) exposure	**Toughness decrease**
400–500 °C embrittlement	Stainless steel (>15% Cr)	400–500 °C (750–930 °F) exposure	**Strength increase, ductility decrease**
Sigma-phase embrittlement	Ferritic and austenitic stainless steel	560–980 °C (1050–1800 °F) exposure	**Toughness decrease**
Graphitization	Carbon and alloy steel	Exposure >425 °C (>800 °F)	**Toughness decrease**
Intermetallic compound	All steel	Exposure to metal that forms	**Brittleness**

embrittlement		an intermetallic	
Neutron embrittlement	All steel	Neutron irradiation	**Increase in DBTT**
Hydrogen embrittlement	Cold worked or heat treatment hardened steel	Processing, service, corrosion, etc.	**Ductility decrease, cracking**
Liquid-metal-induced embrittlement	All steel	Molten low T_M metal exposure	**Ductility decrease**
Solid-metal-induced embrittlement	**All steel**	**Near-molten low T_M metal exposure**	**Ductility decrease**

T_M, melting temperature

This article focuses primarily on the contribution of embrittlement to overload failure. The following sections describe a number of thermally and environmentally induced embrittlement effects that can alter the overload fracture behavior of metals. Most of these are embrittling phenomena, where normally ductile materials evince brittle characteristics. These phenomena are described and differentiated by their causes, effects, and remedial methods, so that failure characteristics may be directly compared during practical failure investigation. For the purposes of failure prevention, possible steps for failure avoidance or component remediation augment the diagnostic information. Naturally, when embrittlement includes crack formation this usually renders a component unusable, especially in cyclic service. Cracks may become quiescent, under null or very low loading; however, their presence remains and may be a future source of crack propagation.

Thermally Induced Embrittlement

Strain-Age Embrittlement. If low-carbon steel is deformed, its hardness and strength will increase upon aging at room or slightly elevated temperature. Strain-age embrittlement occurs when uncombined interstitial solute atoms, usually carbon and nitrogen, diffuse to dislocations and lattice discontinuities. The ordering of carbon and nitrogen atoms at the dislocations assists in the pinning of dislocations and thus causes an increase in strength and lowering of ductility.

The effect of strain-age embrittlement is manifested as a time and temperature dependent increase in strength and decrease in ductility (Ref 7). The amount of cold work is also important; about 15% reduction provides the maximum effect. Room-temperature aging is very slow. Room-temperature aging may require from a few hours to a year; generally, several months are required for maximum embrittlement. As the aging temperature increases, the time for maximum embrittlement decreases. This process may take only a few minutes at 200 °C (400 °F).

Strain-age embrittlement is of most concern in sheet materials for drawing or other cold-forming processes (although it should be noted that some components are intentionally allowed to strain age, either naturally or artificially, in circumstances when forming is completed and when the additional strength is desirable). Embrittled material can exhibit poor formability, including creation of aesthetically undesirable Lüders bands (stretcher strains) from discontinuous yielding, or rupture at normal processing stresses. The return of the yield point or the presence of Lüders strain in the stress-strain curve is evidence that strain-age embrittlement has occurred. These steels are often temper rolled to suppress the yield point, but a pronounced yield point may return due to strain-age embrittlement. In nonaluminum-killed sheet steels, a small amount of deformation, about 1%, will suppress the yield-point phenomenon for several months. If the material is not formed within this safe period, the discontinuous yielding problem will eventually return and impair formability.

Rimmed or capped sheet steels are particularly susceptible to strain-age embrittlement, although strain aging has also been encountered in plate steels and weld HAZs. Certain coating treatments, such as hot-dip galvanizing, can produce a high degree of embrittlement in areas that were cold worked the critical amount, leading to brittle fractures. This can be prevented if the material is annealed before coating.

Avoidance/Remediation. Rimmed or capped steel sheet is especially prone to strain-age embrittlement. These steel types, as well as other materials with high inclusion and interstitial element contents, should be avoided when maximum formability, such as deep drawing, is required. Many drawing-quality materials are aluminum-killed, so that the aluminum will combine with the nitrogen, and have very low carbon contents. Interstitial-free steels and those alloyed with sufficient titanium do not strain age. Some manufacturers have prevented this aging problem by prompt forming after temper rolling, or by ordering stock immediately prior to use.

Quench-Age Embrittlement. If a low-carbon steel is heated to temperatures immediately below the lower critical temperature (Ac_1) and then quenched, it becomes harder, stronger, and less ductile upon aging at room or slightly elevated temperature. As with strain aging, quench-age embrittlement is a function of time at the aging temperature until the maximum degree of embrittlement is reached. Steel may require several weeks of room-temperature aging to reach the maximum embrittled state.

This embrittlement is somewhat similar to strain aging, except the impetus for carbon diffusion is a concentration gradient rather than cold-working stresses. Quench aging involves actual precipitation of carbides (whereas strain aging involves interaction of carbon and other atoms with dislocations). The solubility of carbon in ferrite is higher at elevated temperature, resulting in carbon diffusion to relieve the stresses in the supersaturated lattice. Quench-age embrittlement results from precipitation of solute carbon at existing dislocations and from precipitation hardening because of differences

in the solid solubility of carbon in ferrite at different temperatures. Nitride precipitation can also occur, but the amount of nitrogen present is generally too low for substantial hardening.

Steels with carbon contents of 0.04 to 0.12% appear most susceptible to quench-age embrittlement; increasing the carbon content above 0.12% reduces the effect. Quenching must be from below the Ac_1 temperature yet above 560 °C (1040 °F), which is reportedly an approximate lower boundary for this condition. Embrittlement effects increase at higher temperatures within this range.

Avoidance/Remediation. The quenching process necessary to initiate this embrittlement does not have much practical, metallurgical significance. Subcritical quenching is sometimes performed to reduce the cooling time between sequential high- and low-temperature processing operations. Avoiding the rapid quench will allow the normal equalization of carbon content that occurs in slow cooling. Annealing and slow cooling of quench-age embrittled steel can restore ductility.

Blue Brittleness. When plain carbon steels and some alloy steels are heated to temperatures between 230 and 370 °C (450 and 700 °F), they experience an increase in strength and a marked decrease in ductility and impact strength. The term blue brittleness is derived from the characteristic blue coloration on steel that has been exposed to oxygen-containing atmospheres within this temperature range.

Blue brittleness is an accelerated form of strain-age embrittlement. The increase in strength and decrease in ductility are caused by carbide and/or nitride precipitation hardening within the critical-temperature range. Deformation while the steel is heated in the blue-heat range results in even higher hardness and tensile strength after cooling to room temperature. If the strain rate is increased, the blue-brittle temperature range increases.

Avoidance/Remediation. Use of susceptible steels that have been heated in the blue-brittleness range should be avoided, especially if the steels are subjected to impact loads, because the toughness of these materials will be considerably less than optimum. Tempering in the blue-heat range should be avoided for materials expected to be sensitive to blue brittleness. It can be avoided by specifying materials with low carbon and nitrogen levels or with alloying elements that are strong carbide and nitride formers.

Stress-Relief Embrittlement and Reheat Cracking. Stress-relief embrittlement results in the loss of toughness, typically within the heat-affected zone (HAZ) and/or the weld metal as a result of stress relieving of a welded structure. Stress-relief embrittlement of welds occurs during subsequent, usually subcritical, thermal treatment after welding. When severe, this can result in intergranular brittle cracking, typically referred to as postweld heat treat cracking or reheat cracking. Both phenomena have been observed only in those alloy systems that undergo precipitation hardening. These systems include low-alloy structural and pressure vessel steels, ferritic creep-resisting steels, austenitic stainless steels, and some nickel-base alloys. In addition, a tendency for creep embrittlement, notch weakening, and poor creep ductility during short-term elevated-temperature tests usually occurs in these materials.

During welding, the HAZ is exposed to high temperatures, ranging up to the melting point of the alloy. At these temperatures, existing precipitates in the base metal (in steels, carbides, and nitrides) are taken into solution, and grain coarsening occurs. During cooling, some precipitation takes place at grain boundaries or within the grains, but the majority of the precipitates remain in solution. Subsequent exposure at stress-relieving temperatures causes precipitate formation, primarily alloy carbides, within the grains and at the grain boundaries in the HAZ. This is responsible for strengthening and embrittlement. Embrittlement can be more pronounced in welds with substantial grain coarsening in the HAZ due to high heat input.

Cracking occurs when the nonuniform stresses, including both residual stresses from welding and thermal expansion stresses, exceed the local, reduced creep ductility. This often occurs at the weld toe. Rapid or nonuniform heating will increase cracking severity as less elastic recovery occurs; however, precipitate formation is still diffusion-related. In addition, the combined heating and stress effects can nucleate creep voids in the grain boundaries surrounding the strengthened grains. During stress relief, residual stresses are relieved through creep deformation. However, the strengthening of the precipitates of the grain interiors tends to concentrate creep strain at grain boundaries, leading to intergranular cracking.

Example 12 illustrates a welded component that failed from stress-relief embrittlement. Although this embrittlement is usually manifested in the HAZ or weld metal, other situations can occur. Reheat cracking is sometimes apparent upon start-up of high-temperature boilers, in areas not associated with welding. As noted, low-alloy structural steels, austenitic stainless steels, and some nickel alloys have been found to be susceptible to stress-relief embrittlement. Strong carbide-forming elements contribute to this embrittlement, vanadium, molybdenum, and boron in particular. Ferritic high-strength low-alloy (HSLA) steels can be more susceptible than quenched-and-tempered grades. Some HSLA grades should never receive a postweld heat treatment (PWHT).

Avoidance/Remediation. The possibility for stress-relief embrittlement and cracking can be reduced by specifying cleaner steels made with better deoxidation practice. In susceptible materials, service in the as-welded condition should be explored. Welding practice can also be altered somewhat to reduce residual stresses and HAZ grain coarsening by reducing heat input where possible. Steel that is embrittled, but not cracked, may be annealed to restore ductility.

Example 12: Stress-Relief Cracking of a Welded Alloy Steel Tube. A thick-walled tube that was weld fabricated for use as a pressure vessel exhibited cracks as depicted in Fig. 28. Similar cracking was apparent at the weld toes after postweld stress relief or quench-and-temper heat treatment. The cracks were not detectable by nondestructive examination after

welding, immediately prior to heat treatment. Multiple-pass arc welds secured the carbon steel flanges to the Ni-Cr-Mo-V alloy steel tubes.

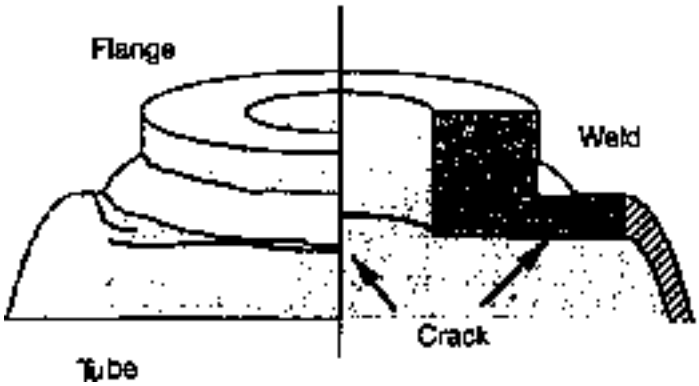


Fig. 28 Large welded tube that cracked upon postweld stress-relief heat treatment (example 12)

Investigation. Visual examination revealed that the centers of the cracks were at the weld fusion line, but the ends of the cracks were in the adjacent tube HAZ. There was no cracking in the carbon steel flanges adjacent to the circular welds. The opened crack surfaces were heavily oxidized and discolored, as an air atmosphere had been used for the postweld heat treatment. The crack features were oxidized, but appeared to be intergranular, whereas the room-temperature lab fracture features were consistent with ductile rupture. The crack was evident at the weld toe, but it did not follow the fusion line or HAZ, as is sometimes the case in hot cracks. The welding penetration was shallow and irregular. The base metal microstructure consisted of tempered martensite.

Conclusions. Evaluation of the fabrication history and the analysis data indicated that the tube failed as a result of stress-relief cracking. Very high residual stresses often result from welding thick sections of hardenable steels, even when preheating is employed. Quenched-and-tempered steels containing vanadium, as well as HSLA steels with a vanadium addition, have been shown to be susceptible to this embrittlement. Manufacturers of susceptible steels recommend use of these materials in the as-welded condition.

Temper Embrittlement. Alloy steels containing certain impurities (phosphorus, antimony, tin, and arsenic) will become embrittled during tempering in the range of 350 to 570 °C (660 to 1060 °F) or during slow cooling through this region. Embrittlement occurs because of the segregation of phosphorus, antimony, arsenic, and/or tin to the grain boundaries. Temper embrittlement is a diffusion-related process. The degree of embrittlement depends on the impurity content and the time within the critical tempering range. Embrittlement occurs fastest at about 455 to 480 °C (850 to 900 °F). Carbon steels are not susceptible to temper embrittlement. Fortunately, temper embrittlement is reversible and can be removed by retempering above the critical tempering range, followed by rapid cooling.

As noted, the impurity elements most associated with temper embrittlement are the metalloids phosphorus, arsenic, tin, and antimony. Silicon, germanium, selenium, tellurium, and bismuth are also identified as embrittling agents. The metallic element manganese is thought to both cause and facilitate temper embrittlement to an extent that the effects are difficult to differentiate. Plain carbon steels that contain less than 0.3% Mn are not considered susceptible. The alloy content also influences embrittlement. Nickel-chromium steels are particularly prone to temper embrittlement, but molybdenum additions reduce the susceptibility. Carbon, chromium, manganese, silicon, and nickel enhance embrittlement, both singly and in synergistic combination. Ferritic and pearlitic microstructures are less susceptible than the heat treated metastable martensitic and bainitic microstructures.

Prior to development of electron fractographic techniques, the degree of embrittlement was identified by macroscopic fracture examination, degradation of mechanical properties, and use of grain-boundary etchants. Temper embrittlement is manifested as a substantial increase in the DBTT, but strength and hardness are largely unaffected. It is most easily observed using toughness tests, but tensile properties will be affected by severe embrittlement. The magnitude of the DBTT shift can be dramatic (for example, see Fig. 22 in the article “Tempering of Steel” in *Heat Treating,* Volume 4 of the *ASM Handbook,* 1991) and transition temperature increases as high as 200 to 300 °C (360 to 540 °F) have been reported. Intentional doping with embrittling elements during research studies has resulted in even more severe shifts of the DBTT.

Electron fractography has added another tool for assessment of the degree of embrittlement by determination of the area fraction of intergranular facets on the fracture surface. Intergranular fracture usually results. The amount of grain-boundary fracture depends on the degree of impurity segregation to the grain boundaries, but is also influenced by the matrix hardness, the prior-austenite grain size, and the temperature. In specimens from Charpy V-notch tests, the amount of intergranular fracture will also vary as a function of the distance from the root of the V-notch.

Avoidance/Remediation. Temper embrittlement is simply avoided by tempering outside of the critical range, although this may not always be possible depending on the inherent hardenability and tempering response of the selected steel. The use of alloys with only the minimum hardenability necessary for service can employ tempering temperatures below the embrittlement range. Ferritic and pearlitic structures are also less susceptible. Embrittlement of thick materials, which can occur by slow cooling through the critical range, can be avoided by rapid cooling or quenching, or by additional tempering at a temperature above the critical range, followed by rapid cooling.

Welding research has developed equations and factors that can be used to identify the temper embrittlement susceptibility of base metals and filler metals. For example, thick-section 2Cr-1Mo steel plates used in pressure vessels for petroleum and chemical service at elevated temperatures are susceptible to temper embrittlement in service. The problem has been attributed to the presence of residual elements such as phosphorus and antimony. Bruscato (Ref 8) established an *X* factor to describe the effect of the residual elements:

$$X = \frac{10P + 5Sb + 4Sn + 1As}{100}$$

where *X* is in parts per million. Acceptable weld metal ductility can be obtained if the manganese and silicon contents and the *X* factor are low. Reduction of these elements can be achieved using basic fluxes. Another indicator, the Watanabe number, *J*, shows that welds and base metal are not susceptible to temper embrittlement if *J* is less than 200 (Ref 9):

$$J = (\mathrm{Si} + \mathrm{Mn})(\mathrm{P} + \mathrm{Sn}) \times 10^4$$

Tempered-martensite embrittlement (TME) of quenched-and-tempered HSLA steels occurs over a temperature range of approximately 200 to 370 °C (400 to 700 °F) (Ref 10). Tempered-martensite embrittlement is also referred to as 350 °C (or 500 °F) embrittlement, although the greatest alteration reportedly occurs at approximately 315 °C (600 °F). It differs from temper embrittlement in the strength of the material and the temperature exposure range. In temper embrittlement, the steel is usually tempered at a relatively high temperature, and embrittlement occurs upon slow cooling after tempering and during service at temperatures within the embrittlement range. In TME, the steel is tempered within the embrittlement range, and service exposure is usually at room temperature. Therefore, temper embrittlement is often called two-step temper embrittlement, while TME is often called one-step temper embrittlement.

Tempered-martensite embrittlement occurs in the tempering range in which ε-carbide changes to cementite. It takes place mainly in steels with a microstructure of tempered martensite, but steels with microstructures of tempered lower bainite are also susceptible to TME. Other structures, such as upper bainite and pearlite/ferrite, are not embrittled by tempering in this region. The mechanisms of TME are not completely understood, but the main explanations for TME are based on the effects of impurities and cementite precipitation on the prior-austenite grain boundaries. Early studies concluded that TME was due to precipitation of thin platelets of cementite at the grain boundaries. However, TME has also been found to occur in very-low-carbon steels, and residual impurity elements have also been shown to be essential factors in TME.

It thus appears that TME results from a combination of effects due to: (1) cementite precipitation on prior-austenite grain boundaries or at interlath boundaries and/or (2) the segregation of impurities, such as phosphorus and sulfur, at the prior-austenite grain boundaries. The predominant mechanism depends on many alloy and processing variables, but the effect of impurity segregation is more pronounced at the higher range of the embrittlement temperature region. Fracture by TME also can be either transgranular or intergranular. Transgranular TME fractures observed with tempering between 200 and 300 °C (390 and 570 °F) and are found with the formation of coarse carbides at martensite lath boundaries. Intergranular TME fractures usually, but not always, are observed from tempering at about 350 °C (660 °F) with the presence of coarse carbides at the impurity-weakened grain boundaries. These differences may be due to differences in carbon, alloy, and impurity content, as well as in strength level, test temperature, nature of the test, and grain size. Impurities appear to influence coarse-grain steels to a greater degree than fine-grain steels.

Depending on the test temperature, TME produces a change in the fracture mode from either predominantly transgranular cleavage or MVC to intergranular fracture along the prior-austenite grain boundaries. In room-temperature tests, the fracture may change from MVC to a mixture of quasi-cleavage and intergranular fracture. Tensile strength, yield strength, elongation, and reduction in area are not affected by this embrittlement, but it has an adverse effect on Charpy V-notch impact energy (Ref 11). Charpy V-notch absorbed energy plotted against the tempering temperature is the most common procedure to reveal the embrittlement. The DBTT may also increase some with tempering in this range, but the increase in DBTT is not as pronounced as in temper embrittlement.

Avoidance/Remediation. Avoiding the embrittling temperature range is necessary to prevent TME, although the range is somewhat variable. Use of steels with low impurity contents, particularly phosphorus, can prevent the impurity segregation contribution to total embrittlement. Embrittled steels can be annealed to restore maximum impact energy when necessary.

400 to 500 °C (750 to 930 °F) Embrittlement. Fine-grain high-chromium stainless steels normally possess good ductility. However, if they are held for long periods of time at temperatures of 400 to 500 °C (750 to 930 °F), they will become harder and embrittled. Time and temperature of exposure affect the level of embrittlement. This phenomenon is sometimes called 475 °C (885 °F) embrittlement, as the embrittlement is most pronounced at about 475 °C (885 °F). The

moderate temperature exposure allows diffusion of chromium to the grain boundaries for the formation of chromium-rich carbides. This segregation also results in reduced corrosion resistance in the denuded regions adjacent to the grain boundaries, analogous to sensitization in austenitic stainless steels. At least 15% Cr is considered necessary for this embrittlement. Increasing chromium contents and higher levels of silicon, niobium, and titanium appear to promote embrittlement.

Embrittlement of high-chromium ferritic stainless steels in this 400 to 500 °C (750 to 930 °F) temperature range produces iron-rich α and chromium-rich α′ ferrites. Chromium nitrides also precipitate during aging and are observed at grain boundaries, dislocations, and inclusions. The formation of α′ can occur in two ways: by nucleation and growth or by spinodal decomposition. The latter mechanism appears to be operative in higher chromium content alloys and has been the most commonly observed formation process. The reaction is reversible; heating above the embrittlement range dissolves α′. In duplex stainless steels, the embrittlement temperature range appears to be broader, with additional phases precipitating in the upper portion of the range. An example of 475 °C (885 °F) embrittlement of a duplex alloy similar in composition to type 329 stainless steel, is shown in Fig. 101 of the "Atlas of Fractographs" in *Fractography*, Volume 12 of *ASM Handbook* (1987).

Avoidance/Remediation. The occurrence of 400 to 500 °C embrittlement can be prevented in various ways. A low-chromium grade, less than 15%, could be used if critical range exposure is unavoidable. Stress relief in this range could be avoided for nonhardenable ferritic grades. Martensitic grades could be preferentially selected for tempering response that would employ a tempering temperature outside the critical range. Embrittled stainless steels can be retempered at temperatures above the critical range, at least 540 °C (1000 °F) to remove this effect (Ref 12). Tempering should be as brief as possible if it is performed at a higher temperature, as sigma phase may form.

Sigma-Phase Embrittlement. Sigma (σ) is an iron-chromium phase that forms in stainless steels with chromium contents typically between 25 and 76 wt%, although σ formation may occur in ferritic stainless steels with chromium contents below 25 wt%. Sigma forms more readily from ferrite in stainless steels than from austenite, so austenitic stainless steels are typically less susceptible to σ formation than ferritic grades. However, many austenitic grades will form σ phase. In addition, σ-phase formation is enhanced in grades of austenitic stainless steels that intentionally have a small amount of δ ferrite (about 5%) to prevent hot cracking during welding.

Some superalloys also can be embrittled by σ-phase formation. However, the presence of σ in superalloys is not necessarily damaging to properties. Various morphologies may be encountered, some of which are quite detrimental to properties. Sigma in the form of platelets or as a grain-boundary film is detrimental, but globular intragranular precipitation can improve creep properties. Considerable effort has been devoted to determining how composition influences σ-phase formation in nickel-base superalloys, and this has substantially influenced alloy development.

In stainless steels, σ generally forms with long-time exposure in the range of 565 to 980 °C (1050 to 1800 °F), although this range varies somewhat with composition and processing. This can occur during slow cooling, tempering or stress relief, and welding. Sigma formation exhibits C-curve behavior with the shortest time for formation (nose) generally occurring between about 700 and 810 °C (1290 and 1490 °F) for stainless steels. In austenitic stainless steels, σ precipitates at grain and twin boundaries at temperatures between about 595 and 900 °C (1100 and 1650 °F). Sigma precipitation occurs fastest at about 845 °C (1550 °F).

Coarse grain sizes from high solution-annealing temperatures retard σ formation, while cold working generally promotes σ formation. The influence of cold work on σ formation depends on the amount of cold work and its effect on recrystallization. If the amount of cold work is sufficient to produce recrystallization at the service temperature, σ formation is enhanced. If recrystallization does not occur, the rate of σ formation may not be affected. Small amounts of cold work that do not promote recrystallization may actually retard σ formation.

The σ phase is extremely hard and brittle and imparts a reduction in toughness with an increase in notch sensitivity. These effects are most notably observed at low temperatures following the embrittling heating event. Steels containing this phase also suffer a reduction in corrosion resistance. The creep resistance and thermal fatigue resistance are also compromised in high-temperature applications for embrittled steels. Embrittlement from σ phase is most detrimental after the steel has cooled to temperatures below 260 °C (500 °F). At higher temperatures, stainless steels containing σ phase can usually withstand normal design stresses. However, cooling to 260 °C (500 °F) or below results in essentially complete loss of toughness.

Additions of silicon, molybdenum, nickel, and manganese permit σ to form at lower chromium levels. Silicon, even in small amounts, markedly accelerates σ formation. In general, all of the elements that stabilize ferrite promote σ formation. Molybdenum has an effect similar to that of silicon; aluminum has a lesser influence. Small amounts of nickel and manganese also increase the rate of σ formation, although large amounts, which stabilize austenite, retard σ formation. Additions of tungsten, vanadium, titanium, and niobium also promote σ formation. Carbon additions retard σ formation by forming chromium carbides, thereby reducing the amount of chromium in solid solution.

Avoidance/Remediation. To a large extent, judicious alloy selection and heat treatment can prevent σ-phase embrittlement. Extraction replicas have been used to determine the amount of σ phase in the microstructure, and the amount of σ phase has been directly related to the creep rate (Ref 13). Components that are suspected of being embrittled

can be reannealed at a suitable temperature to dissolve the σ phase, along with care not to dwell in the critical embrittlement temperature range.

Graphitization is the decomposition of pearlite into ferrite and randomly dispersed graphite, which can result in embrittlement when the graphite particles form along a continuous zone through a load-carrying member. The degree of embrittlement depends on the distribution, size, and shape of the graphite. When the graphite is present in sufficient quantity and the inclusions are aligned or continuous, brittle fracture can occur along this relatively weak pure-iron/free-carbon interface. Continuous graphite formation is sometimes called chain graphitization. The severity of graphitization is frequently evaluated by bend testing.

Graphitization is a diffusion process, and the rate and severity are functions of service duration and temperature. It typically occurs in carbon or low-alloy steels subjected to moderately high temperatures (~450–600 °C, or 850–1100 °F) for extended periods of time. At higher temperatures, pearlite decomposition is dominated by the formation of spheroidal carbides. Graphitization and spheroidization are competing mechanisms with different activation energies and rates for the decomposition of pearlite (Fig. 29). Graphitization is the usual mode of pearlite decomposition at temperatures of about 550 °C (1025 °F), while the formation of spheroidal carbides predominates at higher temperatures. Although spheroidization will reduce the strength, the material will remain ductile rather than embrittled. The critical temperature is dependent on alloying and many other factors.

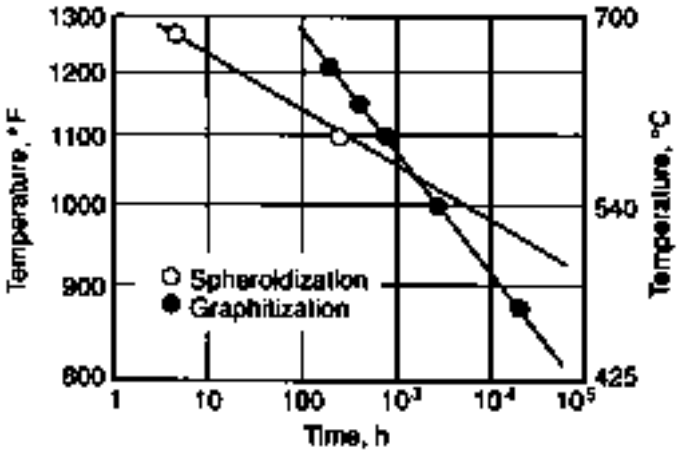


Fig. 29 Temperature-time plot of pearlite decomposition by spheroidization and graphitization. The curve for spheroidization is for conversion of one-half of the carbon in 0.15% C steel to spheroidal carbides. The curve for graphitization is for conversion of one-half of the carbon in aluminum-deoxidized 0.5% Mo cast steel to nodular graphite.

The deoxidation practice used in making the steels is the most important parameter influencing graphitization. High levels of aluminum deoxidation strongly promote graphitization for both carbon and carbon-molybdenum steels. The graphitization tendency of carbon and carbon-molybdenum steels is increased when the aluminum content exceeds about 0.025%. Steels deoxidized with silicon may also be susceptible to graphitization. Deoxidation with titanium will usually produce good resistance to graphitization. While nitrogen appears to retard graphitization, high levels of aluminum remove nitrogen and thus promote graphitization (Ref 5). Molybdenum additions (0.5%) help stabilize cementite, but do not fully offset the influence of high aluminum additions. Manganese and silicon both affect graphitization, but their influence is small at the levels used in low-alloy chromium-molybdenum steels.

Chromium appears to be the best alloy addition for stabilizing carbides. Graphitization can be resisted by steels containing more than 0.7% Cr; chromium-containing ferritic steels for high-temperature service typically contain at least 0.5% Mo as well, largely to impart elevated-temperature strength and resistance to temper embrittlement. Molybdenum is a carbide stabilizer and may limit available carbon for graphitization, and carbon-molybdenum steels exhibit greater resistance to graphitization than do carbon steels. However, use of carbon-molybdenum steel has been limited at higher temperatures because of graphitization. Chromium steels are highly resistant to graphitization and are therefore preferred when service temperatures become greater than about 425 to 455 °C (800 to 850 °F). As an example of maximum-use temperatures based on oxidation or graphitization criterion, the following are maximum-use temperatures for superheater tube materials covered in ASME boiler codes:

Material	Maximum-use temperature at surface based on oxidation/graphitization criteria	
	°C	°F
SA-106 carbon steel	400–500	**750–930**
0.5Cr–0.5Mo	550	**1020**
1.2Cr–0.5Mo	565	**1050**

2.25Cr–1Mo	580	**1075**
9Cr–1Mo	650	**1200**
Type 304H	**760**	**1400**

Graphitization of carbon and carbon-molybdenum steel piping during service at temperatures above 425 °C (800 °F) has caused numerous failures in steam power plants and refineries. Graphite formation generally occurs in a narrow region in the HAZ of a weld where the metal has been briefly heated above the lower critical temperature. If graphitization is detected in its early stages, the material can often be rehabilitated by normalizing and tempering just below the lower critical temperature. Steel that has undergone more severe graphitization cannot be salvaged in this manner; the defective region must be cut out and rewelded, or the section must be replaced. Carbon and carbon-molybdenum steels can be rendered less susceptible to graphitization by tempering just below the lower critical temperature.

Because graphitization involves prolonged exposure to moderate temperatures, it seldom occurs in boiling-surface tubing. Economizer tubing, steam piping, and other components that are exposed to temperatures from about 425 to 550 °C (800 to 1025 °F) for several thousand hours are more likely than boiler-surface tubing to be embrittled by graphitization. The HAZs adjacent to welds are among the most likely locations for graphitization to occur.

Avoidance/Remediation. Material selection for specific service temperatures is the best way to avoid graphitization embrittlement. Safe operating temperatures have been identified for susceptible materials, and many alloys that are not susceptible have been developed. In cases where graphitization is suspected, it is often possible to perform in situ metallographic examination to identify gross graphitization without component removal. Severe graphitization cannot be repaired. Slight graphitization may be removed by normalizing, followed by tempering slightly below the lower critical temperature.

Intermetallic Compound Embrittlement. Embrittlement of steel can be caused by the formation of brittle, intermetallic compounds during elevated-temperature service. This phenomenon is considered irreversible, and embrittlement by intermetallic compounds can be avoided by not utilizing susceptible material combinations at elevated temperatures.

This phenomenon has been documented in galvanized steels where a brittle, intergranular iron-zinc phase forms at the grain boundaries, resulting in embrittlement. At slightly elevated temperatures, the zinc and iron undergo spontaneous formation of intermetallic alloy particles. Iron and zinc can form a number of stable phases, all of which exhibit crystalline structures different from bcc iron. Grain-boundary formation occurs preferentially due to the relative ease of zinc atom diffusion. This is different from liquid-metal-induced embrittlement in that no melting occurs. Additionally, this is different from solid-metal-induced embrittlement by the intergranular diffusion of the foreign solute along with intermetallic compound formation.

Overaging and Precipitation Embrittlement. Many nonferrous and ferrous alloys are precipitation or age hardened, as they do not exhibit a polymorphic character for transformation hardening by quenching and tempering. Room-temperature (natural aging) and elevated-temperature (artificial aging) precipitation hardening provide strength and, in some cases, ductility. The maximum yield strength occurs at a combination of aging temperature and time. Excessive thermal exposure can lead to overaging, which typically reduces both strength and ductility. The drop in strength on overaging can be substantial. This effect is evident in high-manganese steels where carbides precipitate. Titanium alloys can also be degraded by overaging. In some cases, however, intentional overaging has been found to be beneficial when maximum strength is not required. For example, some 7*xxx* series aluminum alloys are overaged to the T7 condition for better SCC resistance, although a 10 to 15% strength decrease occurs. Additionally, optimal fracture toughness is also obtained for this temper.

Avoidance/Remediation. Precipitation from solid solution occurs during intentional precipitation treatment or during moderate-temperature service exposure. The rate of embrittlement is a function of both the exposure time and temperature, and service at embrittling temperatures should be avoided. Alloying and heat treatment in some cases can suppress precipitation.

Quench Cracking. Although quench cracking is not an alteration of a material occurring in service, it is unpredictable. These cracks can be instantaneous upon quenching, or they may be delayed. It is not unusual for some components to be crack-free, whereas seemingly identical components have cracked. Quench cracking occurs due to thermal contraction stresses coupled with the volumetric expansion accompanying the martensite transformation. The contraction from austenite to martensite or pearlite on cooling is more than 1%. The volumetric expansion of martensite is directly proportional to carbon content and microstructural factors. Thick materials and components with sharp corners and thickness transitions are most susceptible to quench cracking, as temperature gradients will form severe stress gradients. Quench cracking of water- or brine-quenched alloys steels also can be promoted by soft spots on the surface of the part. These soft (unhardened) regions may be due to the fixture used to hold the part during quenching, or from tongs, or may be caused by vapor pockets during quenching due to inadequate agitation or contamination of the quench.

Quench cracking typically initiates at the surface, particularly at geometrical changes, corners, defects, and inclusions. Quench cracks always begin at the part surface and have characteristics that are easily recognized. First, the fracture generally runs from the surface toward the center of mass in a relatively straight line with either a longitudinal or radial

orientation unless located by a change in section size. The crack is also likely to open or spread and may exhibit a shear lip at the extreme surface. The fracture surfaces of quench cracks are always intergranular.

When cracked parts are tempered, the intergranular morphology may form a thick oxide scale from tempering. A quench crack exhibiting tempering scale is shown in Fig. 30. The microstructure adjacent to the crack will not be decarburized unless a specimen with an undetected quench crack is rehardened. In quenched-and-tempered steels, proof of quench cracking is often obtained by opening the crack and looking (visually) for temper color typical for the temperature used.

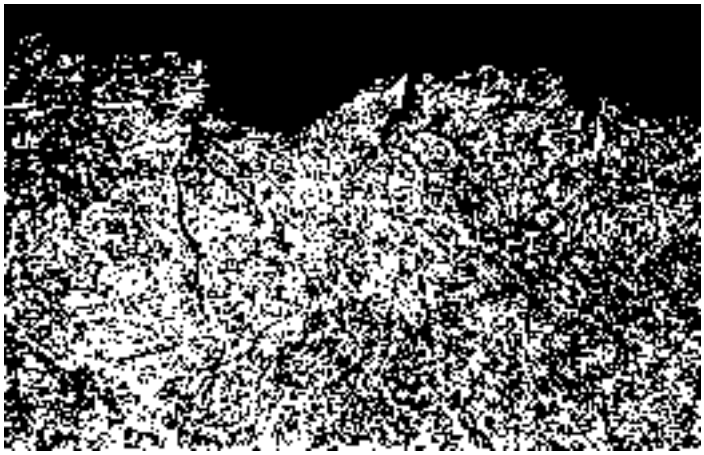


Fig. 30 Quench cracking in a hypoeutectoid steel. The intergranular cracks are filled with tempering scale. 2% nital etch. 297×

Delayed quench cracks are the result of additional transformation of retained austenite in steel. Delayed quench cracking occurs when heavily stressed retained austenite continues to transform to martensite prior to tempering heat treatment. In highly alloyed steels, delayed cracking can be avoided by tempering immediately after quenching. Many highly hardenable steels receive multiple tempering heat treatments or cryogenic treatment to minimize retained austenite. In many cases, this austenite is not considered desirable.

Avoidance/Remediation. Many factors can contribute to quench-cracking susceptibility: carbon content, hardenability, M_s temperature, part design, surface quality, furnace atmosphere, and heat treatment practice. As the carbon content increases, the M_s and martensite finish (M_f) temperatures decrease, and the volumetric expansion and transformation stresses accompanying martensite formation increase. Steels with less than 0.35% C are generally free from quench-cracking problems. The higher M_s and M_f temperatures permit some stress relief to occur during the quench, and transformation stresses are less severe. Alloy steels with ideal critical diameters of 4 or higher are more susceptible to quench cracking than lower-hardenability alloy steels.

In general, quench-crack sensitivity increases with the severity of the quench medium. Quench-cracking propensity can be reduced by avoiding excessive austenitizing temperatures, unnecessarily rapid quenchants, and quenching to an unnecessarily low temperature. Additionally, sometimes the problem can be solved by selecting a material with higher hardenability, so that quench rates can be less severe. Heated quench baths and immediate tempering are very beneficial. Salt-bath heat treating methods such as austempering and martempering may be advisable in complex components manufactured from steels with high hardenability. This often has the added benefit of reducing distortion. Components can be designed to be more quench-crack resistant. Providing generous fillets and rounded edges is beneficial.

Creep Embrittlement. Under creep conditions, embrittlement can occur and result in abnormally low rupture ductility. While creep strength and rupture strength are given considerable attention as design and failure parameters, rupture ductility is an important mechanical property when stress concentrations and localized defects such as notches are a factor in design. Rupture ductility, which varies inversely with creep and rupture strength, influences the growth of cracks or defects and thus affects notch sensitivity and toughness, as discussed in more detail in the article “Creep and Stress Rupture Failures” in this Volume.

As voids and microcracks nucleate and grow during creep deformation, the development of creep embrittlement is invariably associated with a transition from transgranular to intergranular fracture. At high stresses (low time to stress rupture), ductility is fairly high and constant with fracture characterized as transgranular. Below a threshold value of stress, ductility drops with fracture becoming increasingly intergranular. This decrease in ductility with decreasing stress for elevated-temperature (creep) regimes is generally attributed to increasingly localization of creep strain in narrow zones adjacent to the grain boundaries (Ref 14). This process is assisted by impurity effects and precipitate-free zones (PFZs) near the grain boundaries. For example, impurities such as phosphorus, sulfur, copper, arsenic, antimony, and tin have been shown to reduce rupture ductility, although rupture life increases. This behavior appears to be due to the grain-boundary segregants blocking grain-boundary diffusion, which reduces the cavity growth rate. High impurity contents

increase the density of the cavities. Substantial intergranular cracking is observed in high-impurity material and is absent in low-impurity heats (Ref 15).

Creep embrittlement of heat-resistant steels after long-term exposure at elevated temperature has been noted for many years, and one of the first indications was observed in the unexpected failure of boiler flange bolts (Ref 16). In low-alloy heat-resistant steels, embrittlement occurs as a result of small amounts of creep strain along with the precipitation of fine, intergranular precipitates during the creep process (Ref 16). Failure of creep-embrittled ferritic steels material occurs through the prior-austenite grain boundaries in a manner similar to temper embrittlement, and temper-embrittling impurities (P, Sn, Sb, As) may assist the creep embrittlement. However, creep embrittlement is irreversible, while temper embrittlement is reversible. Hence the term “reverse temper embrittlement” has been used (for example, Ref 17). In ferritic steels, creep embrittlement appears to occur in the temperature range of 425 to 595 °C (800 to 1100 °F).

In steels, creep damage often occurs along with oxidation and other changes, as illustrated by the case study in example 13. Overheating also appears to have been a factor.

Example 13: Failure of a Steel Superheater Tube. Failure occurred in a steel superheater tube in a power plant. The tube was specified as ASTM A 213 grade T 22, and the reported operating conditions were 13 MPa (1900 psi) at 482 °C (900 °F). The tube carried superheated steam and was coal fired.

Investigation. The tube failed at a bend and deformed significantly. The tube had ruptured and buckled with some secondary internal cracks apparent (Fig. 31a). The fracture surface was “thick-lipped,” and the inside diameter exhibited a tenacious, black magnetite coating.

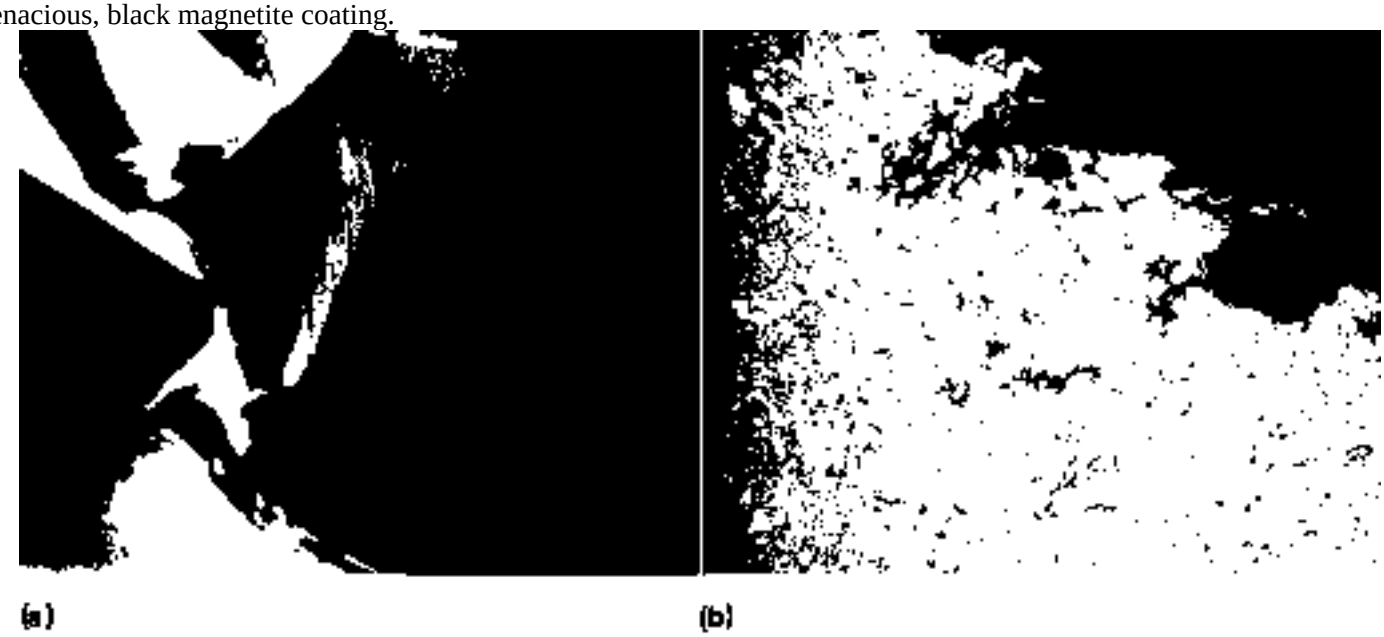


Fig. 31 Fracture of a steel superheater tube (example 13). (a) The interior of the tube that fractured showing secondary cracks and a black contamination layer. (b) Microstructure of the tube showing triple-point cracks and intrusive oxidation damage. The creep damage was throughout the entire wall thickness. 2% nital etch. 297×

Chemical analysis confirmed that the tube satisfied the specified compositional requirements, that is, a 2.25Cr-1Mo alloy steel. The strength determined on a sample remote from the fracture was higher than required, but the elongation at fracture was very low. Fractographic examination by SEM revealed generally intergranular features, but the level of oxidation damage to the features was severe. The cross section (Fig. 31b) exhibited creep voids and spheroidization of the pearlite within the ferrite matrix. Some features suggested hydrogen may have also contributed to the failure.

Conclusions. The superheater tube failed as a result of long-term overheating. Substantial creep damage reduced the strength of the tube to the point that overload failure occurred. The rupture was “thick-lipped,” which is also consistent with long-term overheating. Technical literature indicates the scaling temperature of this alloy is approximately 595 °C (1100 °F), also suggesting service in excess of the stated 480 °C (900 °F) operating temperature. There was no apparent graphitization as this alloy has been shown to resist graphitization.

Overheating and Burning. The mechanical properties of metals and alloys can be altered or degraded by overheating during hot working, heat treatment, or service. This is sometimes called burning and initiates as incipient melting, usually at the grain boundaries. Intergranular oxidation can also accompany this melting. Excessive grain growth without incipient melting can also occur, thus increasing the grain separation and assisting in the growth of cracks. Furthermore, excessive hot-working temperatures may also result in solution of inclusions that may reprecipitate, often in grain boundaries, upon subsequent cooling with a loss in fracture toughness.

In many materials, burning occurs during improperly controlled thermal processing, where heating for hot working or annealing is excessive. It is especially prevalent in hot working such as forging and swaging, where temperature control is more difficult than in simply heating the material in a furnace. Incipient melting may also occur due to poor temperature control in age-hardening alloys. A similar phenomenon is HAZ cracking in welding, as regions of the HAZ adjacent to the fusion line are nearly melted. The term burning is also sometimes used for surface damage from abusive grinding practice.

Avoidance/Remediation. Prudent manufacturing process development should obviate any overheating concerns. Material that has been overheated in this manner is not typically salvageable.

Environmentally Induced Embrittlement

Hydrogen Damage and Embrittlement. Hydrogen damage in metals occurs in many forms, including embrittlement, cracking, blistering, and grain-boundary methane formation (also called hydrogen attack) during high-temperature service. Some materials also form internal flake and fisheye defects during solidification or welding. In addition, hydride formation can occur in a number of nonferrous alloy families, such as titanium alloys, but not steels. These types of hydrogen damage are discussed in more detail in the article “Hydrogen Damage and Embrittlement” in this Volume.

Hydrogen appears to facilitate other types of embrittlement, and it can be complex to diagnose and separate from other contributing factors, such as intergranular impurities and in some cases SCC. As a single proton, hydrogen atoms are sufficiently mobile with significant diffusivity in all metals, even at low temperatures. Individual hydrogen atoms will combine into diatomic gaseous molecules, often at grain boundaries, inclusions, and lattice imperfections. Embrittling hydrogen, which can be as low as several parts per million, is supplied by an extensive list of possible sources. Hydrogen charging, or infusion of a damaging amount of hydrogen, also can occur during primary processing (for example, the steelmaking process) or a variety of fabrication and secondary operations such as acidic cleaning (pickling), electroplating, and welding. Hydrogen can also be introduced from chemical reactions in service due to gas exposure, lubricant degradation, and corrosion. In many instances, the presence of cathodic poisons can increase the quantity of hydrogen atoms absorbed. Increasing the exposure temperature also increases the level of hydrogen absorption. However, both low- and high-temperature service appear to reduce the level of hydrogen damage over room-temperature service.

Hydrogen embrittlement results in cracking or catastrophic failure during service, generally in a delayed manner under static loading. It reduces the ductility of the materials (with or without a decrease in strength depending on the material and the location and extent of hydrogen in the material) and an acceleration of subcritical crack growth rates. Cracks from hydrogen embrittlement typically exhibit little branching and are often intergranular, but caution is required when characterizing hydrogen embrittlement on the basis of fracture appearance. Other contributory factors may be involved, and in some cases (such as highly stressed conditions), transgranular fracture may be observed. These complexities are discussed in a little more detail in the article “Hydrogen Damage and Embrittlement” in this Volume.

Many ferrous and nonferrous metals are prone to hydrogen damage. Steels and precipitation-hardenable steels are particularly susceptible to hydrogen embrittlement. Increasing the hardness and/or the level of cold work increases the embrittlement susceptibility. Basically, hydrogen embrittlement requires three factors: tensile stresses (residual, applied, or both), a susceptible metal, and hydrogen above a threshold level. Many failures in steels occur at loads below the yield strength, and no plastic deformation is evident. Steels at a tensile strength level below 690 MPa (100 ksi) are often considered resistant to hydrogen cracking. These steels may, however, still be susceptible to hydrogen damage in the forms of blistering. Austenitic ferrous alloys are much less susceptible to HE than the martensitic or cold-worked ferritic grades, but fcc metals are not immune to HE. Hydrogen embrittlement can also be difficult to distinguish from SCC in some cases.

Avoidance/Remediation. Hydrogen embrittlement in steels and other metals can be avoided by removing production and service-related sources of potential embrittlement. Close control of fabrication processes that produce hydrogen is a necessity. In welding, underbead cracking and other hydrogen-assisted weld-cracking types can be reduced by better material control. Using better-quality shielding gases, special low-hydrogen electrodes, and moisture-free electrodes and fluxes is beneficial. Reduction of cooling-rate-induced hardening and joint restraint may result in a final stress state below the threshold necessary for embrittlement and cracking.

The practice of baking following electroplating has been shown to reduce the possibility of hydrogen embrittlement. Stress relief prior to plating to remove residual stresses is also beneficial. However, baking cannot remove all of the hydrogen and is by no means a panacea. Plated high-strength steel fasteners that have been properly baked, most notoriously the grade 8 bolts and higher-strength socket head cap screws, can still fail via hydrogen embrittlement. Example 14 discusses one such failure incident. This example also illustrates the difficulty of hydrogen embrittlement, as it is not always possible to *prove* hydrogen embrittlement in an exacting fashion, even though indications point to hydrogen. In many cases, the main issue may be not be proof, but rather education toward eliminating hydrogen embrittlement as a potential problem (see the article “Hydrogen Damage and Embrittlement” in this Volume).

Example 14: Hydrogen Embrittlement Failure of Cap Screws. Some socket head cap screws (SHCS) used in a naval application evinced delayed fracture in service. Standard ASTM A 574 $\frac{5}{16}$ in diam screws were zinc plated and dichromate coated. Two unused, exemplar fasteners from the same lot were submitted for comparison.

Investigation. Several cap screws exhibiting representative cracking and fracturing are shown in Fig. 32(a). All of the cracking was at the head-to-shank fillet, and no plastic deformation was apparent. One screw that cracked 360° around the circumference is shown in Fig. 32(b).

Fig. 32 Delayed fracture of cap screws (example 14). (a) Cap screws that either cracked or fractured in a delayed manner at the underside of the heads. (b) One screw that cracked 360° around the head without separating. (c) Representative electron image of the primarily intergranular fracture features. Secondary grain-boundary cracks, ductile hairlines, and micropores are apparent. Scanning electron micrograph. 1187×

The fracture and opened crack surfaces were dull and with a crystalline appearance of an intergranular fracture. Examination by SEM revealed that the peripheries of the surfaces exhibited primarily intergranular brittle and transgranular dimple-rupture features, whereas the core morphology consisted primarily of dimple-rupture features. Figure 32(c) shows some of the intergranular features, which also contain secondary grain-boundary separation, ductile hairlines, and micropores on the grain facets.

The chemical composition of a failed screw was consistent with grade 4037 alloy steel and satisfied the fastener specification. Tension testing of two exemplar screws revealed failure strengths that exceeded the minimum requirement and fracture features consisting of ductile-rupture features.

Metallographic cross sections were prepared through both fractured and cracked screws. The cracking had an intergranular aspect in many regions with some secondary grain-boundary cracks adjacent to the primary cracks. The microstructure was uniform and consisted of fine, tempered martensite, indicative of the required quenched-and-tempered condition.

Conclusions. The cap screws appear to have failed due to hydrogen embrittlement, as revealed by the delayed cracking and intergranular fracture morphology. Static brittle overload fracture occurred due to the tension preload, and prior hydrogen charging did occur in the manufacturing. The probable source of hydrogen charging is the electroplating, although postplating baking was reportedly performed.

This example illustrates the difficulty of analyzing the hydrogen embrittlement problem. The samples satisfied the composition and property requirements for the identified cap screw grade. However, antimony, phosphorus, tin, sulfur, lead, and other low-melting-point impurities, which are generally very expensive to analyze, also may be factors contributing to the embrittlement situation. These elements, like hydrogen, are more damaging when they are present in the grain boundaries, and thermal exposure may concentrate these elements in the grain boundaries. In fact, some of the extended baking times that have been recommended for hydrogen embrittlement prevention in electroplated parts may actually increase the concentration of these other low-strength elements in the grain boundaries.

The point, therefore, is to be cognizant of other potential contributory factors and to be aware of how the potential problem of hydrogen may appear be reduced. In this case, one important factor is that the cap screws were at the highest strength (1220 MPa, or 180 ksi, minimum tensile strength) for steel fasteners in wide commercial use. This increases the risk of embrittlement, and it may be necessary to examine the manufacturing processes in more detail, depending on the criticality of service and the frequency of occurrence.

Neutron embrittlement of nuclear reactor pressure vessel steels is a serious form of materials degradation that can results in an increase in the DBTT. The DBTT, or NDTT, increase can be as much as 200 °C (360 °F). In some cases, the upper-shelf energy is decreased, with an increased propensity for intergranular fracture. Most ferrous metals can be susceptible, including carbon steels, alloy steels, and precipitation-hardening stainless steels. Some austenitic stainless steels, although containing no DBTT, have been found to exhibit reductions in ductility when irradiated.

On the crystalline structure level, neutron bombardment of susceptible materials results in atomic displacement, thereby creating self-interstitials and vacancy defects, along with phase and precipitate formation. These discontinuities, singly and in clusters, also prevent plastic deformation via dislocation pinning. The extent of embrittlement is based on a large number of service variables, including exposure time, service temperature, and neutron fluence.

Research has shown that heat treatment and composition greatly affect the sensitivity of steel to neutron embrittlement. Tempered martensitic microstructures are less sensitive than ferritic or tempered bainitic microstructures, and finer grain sizes are also beneficial. Steels containing nickel as an alloying addition and/or copper and phosphorus as impurities are known to be particularly susceptible.

Avoidance/Remediation. Avoidance of neutron embrittlement is typically achieved by the selection of materials in appropriate heat treatment conditions found to be less sensitive in nuclear service. Normal annealing heat treatments have been shown to partially or fully restore embrittled materials due to the effects of recovery and, possibly, recrystallization. Some of the effects, such as precipitation of deleterious phases, may not be entirely eliminated, and affected steels may incur embrittlement at a different rate upon reintroduction to service.

Stress-corrosion cracking is manifested as transgranular or intergranular brittle cracking that can lead to final overload failure. Although it is a corrosion failure mechanism, it is included in this discussion of overload fracture due to its embrittling nature. This mechanism does not always include gross general corrosive attack, and differentiation from the other possible embrittling phenomena may be necessary during failure analysis.

The synergistic combination of a corrodent, a susceptible material, and the presence of a tensile stress can result in progressive cracking and reduction in ductility. Nearly all metallic materials have been shown to be susceptible to SCC under certain circumstances. Steels, even those of the corrosion-resistant varieties, are susceptible in many types of service. Hydrogen has been shown to accelerate stress corrosion, as in the case of hydrogen sulfide cracking. In high-strength steels, SCC has been attributed to the hydrogen generated by the corrosion process.

Avoidance/Remediation. Removal of one of the three requisite factors is typically sufficient to prevent SCC. Materials in conditions that evince little susceptibility in the intended service environment can often be selected. Alternatively, it may be possible to remove the corroding substance from the local service environment. A threshold stress is typically required for SCC, with both applied and residual stresses below this level not initiating cracking. More specific information is provided in the article “Stress-Corrosion Cracking” in this Volume.

Liquid-metal-induced embrittlement (LMIE) (Formerly liquid-metal embrittlement, or LME) occurs when contact with molten material of a lower-melting-point alloy results in a severe reduction in ductility and, very often, strength. This occurs primarily by intergranular entry of the molten metal into a stressed material, effecting progressive separation cracking or grain-boundary decohesion. Cycles of solidification and liquation of the low-melting-point alloy can also affect the extent and speed of this embrittlement, if volumetric expansion occurs during cycles of solidification or liquation.

Liquid-metal-induced embrittlement occurs for most loading conditions except for compressive loading. Both residual and applied stresses, along with a temperature dependency, contribute to this complicated embrittlement mechanism. A stress threshold value for this mechanism has been suggested by research. Parameters that have been shown to increase susceptibility to LMIE include grain size, strain rate, time under static loading, and the strength or hardness of the component. High stresses can result in instantaneous fracture, whereas low stresses may have a more delayed effect. Fracture from LMIE can occur below the elastic limit in high-strength materials, but some plastic deformation may be evident in softer materials.

Most engineering metals and alloys have been shown to be susceptible to LMIE. The presence of a molten metal is a necessary, but not by itself a sufficient, condition for LMIE. Metal couples that exhibit low intersolubilities and that do not form intermetallic compounds can be susceptible. Zinc embrittlement in steel is a notable exception to this generalization as it can participate in both LMIE and intermetallic compound embrittlement.

Metals that form surface oxides can prevent ingress of liquid metal due to the adverse effect on the necessary wettability. Initiation of LMIE attack does not apparently depend on the amount of molten metal available, and the required threshold for damage is very small. However, continued degradation requires a steady supply of molten metal. As would be expected, brazing and soldering may cause LMIE (Fig. 33). In this case, the cold-worked steel fractured due to excessive heat input during manual torch brazing. Many plating and coating metals can act as embrittling agents.

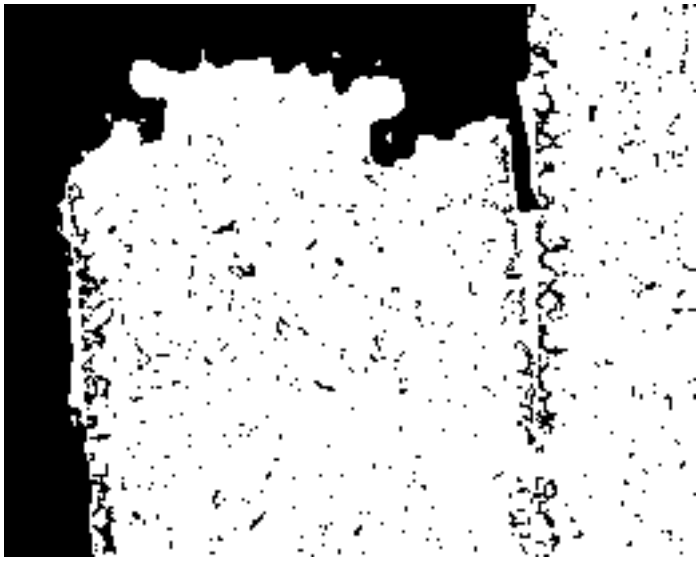


Fig. 33 Liquid-metal-induced embrittlement and cracking evidence that occurred during torch brazing. 2% nital etch. 119×

Alloys that contain a discrete low-melting-point phase rather than an external molten material have evinced LMIE failures. Rapid heating of these materials in the deformation-processed condition may not permit recovery and relaxation of the surrounding matrix prior to liquefaction of the low-melting-point inclusions. The expansion pressure of the liquefying phase forces this liquid metal into the adjacent grain boundaries, resulting in embrittlement or fracture. This has been observed in many alloy systems such as leaded steels or leaded copper alloys, where it is sometimes referred to as fire cracking.

Avoidance/Remediation. The use of susceptible material pairs should be avoided if the constant or transient service temperature is above the melting point of the low-melting-point material. Once it has manifested, LMIE is not reversible and the damage is permanent.

Solid-metal-induced embrittlement (SMIE) occurs in steels as well as many nonferrous materials. This phenomenon appears to be related to LMIE as the materials and couples susceptible to SMIE are those for which LMIE is known to occur. In some instances, contact of a metal with a low-melting-point alloy at a temperature below the melting point of the low-melting-point material has been shown to cause delayed brittle fracture and cracking in addition to decreasing strength and ductility. The relative level of damage increases with temperature up to the melting point of the low-temperature species.

Research suggests SMIE is probably caused by a number of complicated interrelated processes. By whatever mechanisms, empirical evidence shows that the embrittling species can diffuse into the grain boundaries of the base metal, causing embrittlement and cracking. Applied and/or residual stresses above a threshold level are necessary for cracking to occur. In some cases the embrittling species can be an immiscible, discrete solid phase, such as lead in leaded steel grades. Intimate contact between the two materials is necessary, and other factors such as microstructural discontinuities affect SMIE behavior. Solid-metal-induced embrittlement is substantially slower than LMIE, but multiple cracks occur in SMIE.

Avoidance/Remediation. Elevated-temperature service of steels in the presence of solid low-melting-point alloys known to embrittle should be avoided. An intermediate plating with a higher melting point that has been shown to not embrittle can be employed to prevent SMIE.

The Role of Welding in Embrittlement and Overload Behavior. Fusion welding of metallic materials deserves special mention with regard to embrittlement and overload behavior. The combinations of thermal gradients, welding consumables, alloying effects, cooling effects, and residual stresses can result in nearly every manner of embrittlement. Every phase and constituent that can exist in a metal or alloy to its melting point can be present in a weld or surrounding HAZ. Welding-related features contributing to overload failure include: incipient melting/burning, grain coarsening, grain refinement, precipitation effects, hardening, and tempering. In addition, there may be problems associated with lack of fusion and residual stresses created by the thermal environment. Thermal stresses may cause both hot cracking and cold cracking, and cracks may be present in the weld, at the fusion line, in the HAZ, and in the base metal. An example of hot cracking in a tube weldment is shown in Fig. 34. The interdendritic nature of the cracking is evident. The following example shows a variety of welding changes that combined to initiate failure in a heat exchanger.

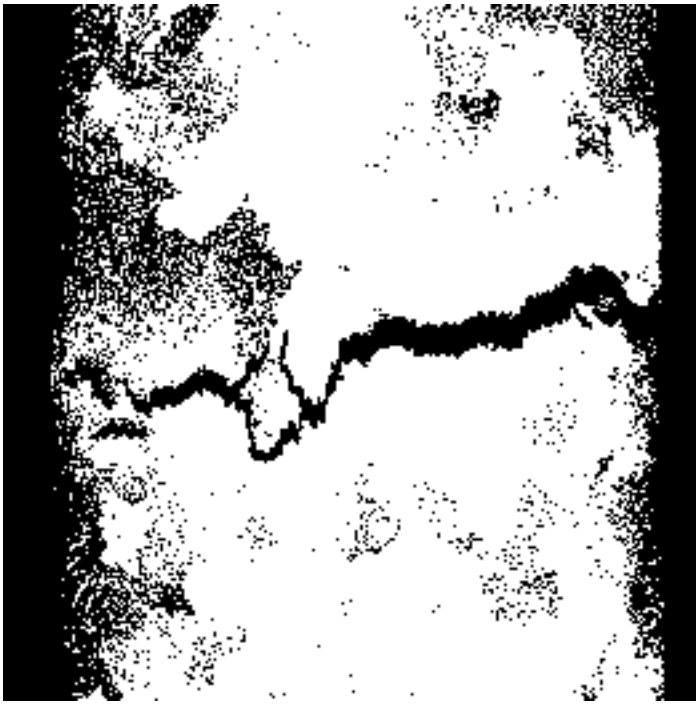


Fig. 34 Hot cracking that developed in a stainless steel tube weld. 10% oxalic acid electrolytic etch. 30×

Example 15: Weld Cracking of a Stainless Steel Heat Exchanger. A welded ferritic stainless steel heat exchanger cracked prior to service. The sheet steel weld joint design is depicted in Fig. 35(a). The welding filler metal was identified as an austenitic stainless steel, and the joining method was gas tungsten arc welding (GTAW).

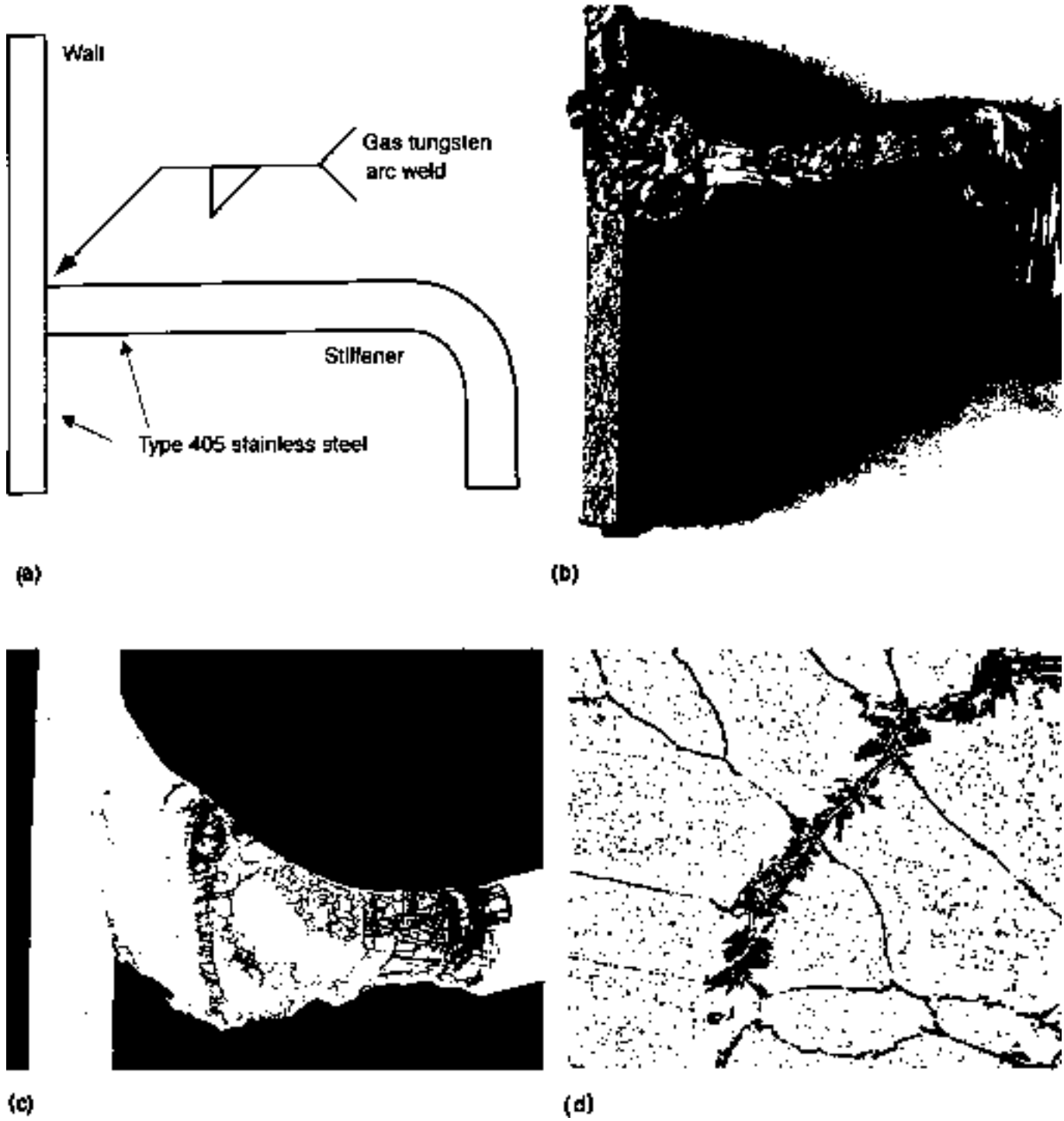


Fig. 35 Cracking of a welded ferritic stainless steel heat exchanger (example 15). (a) Diagram showing the heat-exchanger weld joint design. (b) The transverse crack that occurred through the weld. 5.9×. (c) Metallographic profile of the weld near the cracking, showing melt-through, grain growth, and precipitation. 8.9×. (d) Grain-boundary precipitates in the weld zone. 119×. (c) and (d) were etched with Vilella's reagent followed by electrolytic etching in 10% oxalic acid.

Investigation. Examination of the supplied sample revealed a single crack, transverse to the fillet weld axis. No bulk macroscopic plastic deformation was associated with the cracking. Relatively high heat input during welding was suggested by the extent of the oxide or heat tint discoloration. The underside of the weld joint contained melt through. The crack surface after excision is shown in Fig. 35b. The fracture surface had a faceted appearance with significantly larger facets within the weld and HAZs. Grain growth is not uncommon in arc welding of ferritic stainless steels.

Inspection by SEM confirmed the failure mode as both quasi-cleavage and intergranular fracture. A metallographic cross section through the weld, parallel and adjacent to the crack is shown in Fig. 35c. Dramatic grain coarsening was evident in the weld and HAZ. The microstructure within the resolidified weld was not uniform. Isolated regions of the weld were determined to be austenitic due to the filler metal applied. Grain-boundary phase segregation was apparent, as shown in Fig. 35d.

Conclusions. It was concluded that the heat exchanger cracked due to weld cold cracking or postwelding brittle overload that occurred via flexure during fabrication. The brittle nature of the weld is likely due to a combination of high residual stresses, a mixed microstructure, inclusions, and gross grain coarsening. These synergistic factors resulted from extreme heat input during fillet welding. It was recommended that the welding variables such as voltage, amperage, and travel speed be altered to substantially reduce the heat input.

References cited in this section

5. G.F. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, p 689–736

6. G.F. Vander Voort, Visual Examination and Light Microscopy, *Fractography,* Vol 12, *ASM Handbook,* ASM International, 1987, p 91–165

7. Thermally Induced Embrittlement of Steels, *ASM Specialty Handbook: Carbon and Alloy Steels,* J.R. Davis, Ed., ASM International, 1996, p 309, 312

8. R. Bruscato, Temper Embrittlement and Creep Embrittlement of $2\frac{1}{4}$Cr 1Mo Shielded Metal Arc Weld Deposits, *Weld. J.,* Vol 9 (No. 4), 1970, p 148s–156s

9. Y. Murakami, T. Namura, and J. Watanabe, *Heavy Section $2\frac{1}{4}$Cr-1Mo Steel for Hydrogenation Reactors,* STP 755, ASTM, 1982, p 383–417

10. J.A. Peters et al., On the Mechanisms of Tempered Martensite Embrittlement, *Acta Metall.,* Vol 37, Feb 1989, p 675–686

11. C.L. Briant and S.K. Banerji, Tempered Martensite Embrittlement and Intergranular Fracture in an Ultra-High Strength Sulfur Doped Steel, *Metall. Trans. A,* Vol 12A, Feb 1981, p 310

12. D. Peckner and I.M. Bernstein, *Handbook of Stainless Steels,* McGraw-Hill, 1977, p 44–6, 44–7

13. T. Fushimi, “Life Evaluation of Long Term Used Boiler Tubes,” paper presented at the Conference on Boiler Tube Failures in Fossil Plants (Atlanta, GA), Electrical Power Research Institute, Nov 1987

14. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components,* ASM International, 1989, p 77

15. L.K.L. Tu and B.B. Seth, Effect of Composition, Strength, and Residual Elements on Toughness and Creep Properties of Cr-Mo-V Turbine Rotors, *Met. Technol.,* Vol 5, March 1978, p 79–91

16. G. Sachs and W. Brown, *A Survey of Embrittlement and Notch Sensitivity of Heat Resisting Steels,* ASTM, 1952, p 6–21

17. J.J. Hickey and J.H. Bulloch, The Role of Reverse Temper Embrittlement on Some Low and High Temperature Crack Extension Processes in Low Carbon, Low Alloy Steels: A Review, *Int. J. Pressure Vessel Piping,* Vol 49, 1992, p 339–386

Laboratory Fracture Examination

Among the most important, yet insufficiently used, procedures of the failure analyst is the examination of laboratory fractures at a range of magnifications from visual to more than 1000×. Tensile, impact, or bend testing can provide excellent fractographic comparison samples. This can often be done on specimens used to quantify the mechanical properties. Service conditions, such as elevated or low temperature, can be employed in lab evaluation to help explain the actual failure conditions. In some cases it is possible via laboratory testing to duplicate the fracture appearances to support a proposed set of circumstances causing failure. A laboratory testing program that results in duplication of the appearance of the failed component can be especially helpful when initial evaluation of the failed component does not lead to strong indication of a root cause.

Laboratory fracture study is of particular efficacy in cases of suspected embrittlement. Based on the nature of the embrittlement, it may not always be detectable by electron microscopy, chemical analysis, or metallography. Fractures created in a known manner can also be used to test the null hypothesis, namely that they did not fail in a certain manner. These methods have been used to estimate service temperature after the fact, and other loading conditions may often be deduced. One particular application is in the evaluation of hydrogen embrittlement. Slow-strain-rate tensile fracture features are typically not affected by this static cracking phenomenon.

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

References

1. P.F. Thomason, *Ductile Fracture of Metals,* Pergamon Press, Oxford, England, 1990, p 19–24

2. H. Nelson, “Fractography in the Failure Analysis of Engineering Structures,” Seminar, ASM Annual Meeting, 1997

3. R. Reed and A. Clark, *Materials at Low Temperature,* American Society for Metals, 1983

4. N.J. Petch, The Ductile-Cleavage Transition in Alpha-Iron, *Fracture,* B.L. Averbach, Ed., Technology Press, 1959, p 54–67

5. G.F. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, p 689–736

6. G.F. Vander Voort, Visual Examination and Light Microscopy, *Fractography,* Vol 12, *ASM Handbook,* ASM International, 1987, p 91–165

7. Thermally Induced Embrittlement of Steels, *ASM Specialty Handbook: Carbon and Alloy Steels,* J.R. Davis, Ed., ASM International, 1996, p 309, 312

8. R. Bruscato, Temper Embrittlement and Creep Embrittlement of $2\frac{1}{4}$Cr 1Mo Shielded Metal Arc Weld Deposits, *Weld. J.,* Vol 9 (No. 4), 1970, p 148s–156s

9. Y. Murakami, T. Namura, and J. Watanabe, *Heavy Section* $2\frac{1}{4}$*Cr-1Mo Steel for Hydrogenation Reactors*, STP 755, ASTM, 1982, p 383–417

10. J.A. Peters et al., On the Mechanisms of Tempered Martensite Embrittlement, *Acta Metall.*, Vol 37, Feb 1989, p 675–686

11. C.L. Briant and S.K. Banerji, Tempered Martensite Embrittlement and Intergranular Fracture in an Ultra-High Strength Sulfur Doped Steel, *Metall. Trans. A*, Vol 12A, Feb 1981, p 310

12. D. Peckner and I.M. Bernstein, *Handbook of Stainless Steels*, McGraw-Hill, 1977, p 44–6, 44–7

13. T. Fushimi, "Life Evaluation of Long Term Used Boiler Tubes," paper presented at the Conference on Boiler Tube Failures in Fossil Plants (Atlanta, GA), Electrical Power Research Institute, Nov 1987

14. R. Viswanathan, *Damage Mechanisms and Life Assessment of High-Temperature Components*, ASM International, 1989, p 77

15. L.K.L. Tu and B.B. Seth, Effect of Composition, Strength, and Residual Elements on Toughness and Creep Properties of Cr-Mo-V Turbine Rotors, *Met. Technol.*, Vol 5, March 1978, p 79–91

16. G. Sachs and W. Brown, *A Survey of Embrittlement and Notch Sensitivity of Heat Resisting Steels*, ASTM, 1952, p 6–21

17. J.J. Hickey and J.H. Bulloch, The Role of Reverse Temper Embrittlement on Some Low and High Temperature Crack Extension Processes in Low Carbon, Low Alloy Steels: A Review, *Int. J. Pressure Vessel Piping*, Vol 49, 1992, p 339–386

Overload Failures

Brett A. Miller, Stork Technimet, Inc.

Selected References

- D.G. Altenpohl, *Aluminum: Technology, Applications, and Environment*, 6th ed., The Aluminum Association and The Minerals, Metals & Materials Society, 1998
- G.M. Boyd, *Brittle Fracture in Steel Structures*, Butterworths, London, 1970
- D.N. French, *Metallurgical Failures in Fossil Fired Boilers*, 2nd ed., John Wiley & Sons, 1993
- R. Gibala and R.F. Hehemann, Ed., , *Hydrogen Embrittlement and Stress Corrosion Cracking*, American Society for Metals, 1984
- G.E. Linnert, *Welding Metallurgy*, Vol 1, 4th ed., American Welding Society, 1994
- D.P. Miannay, *Fracture Mechanics*, Springer-Verlag, 1998
- R.D. Port and H.M. Herro, *The Nalco Guide to Boiler Failure Analysis*, McGraw-Hill, 1991
- G.W. Powell and G.F. Vander Voort, Ductile and Brittle Fractures, *Metals Handbook, Desk Edition*, American Society for Metals, 1985, p 32–9 to 32–13
- *Temper Embrittlement of Alloy Steels*, STP 499, American Society for Testing and Materials, 1972
- P.F. Wieser, *The Morphology of Brittle Intergranular Fracture of Steel Castings and the Effects of Processing Variables on its Occurrence*, Steel Founders' Society of America, 1975

Fatigue Failures

Introduction

FATIGUE damage results in progressive localized permanent structural change and occurs in materials subjected to fluctuating stresses and strains. It may result in cracks or fracture after a sufficient number of fluctuations. Fatigue fractures are caused by the simultaneous action of cyclic stress, tensile stress, and plastic strain. If any one of these three is not present, fatigue cracks will not initiate and propagate. The cyclic stress and strain starts the crack; the tensile stress produces crack growth (propagation). Although compressive stress will not cause fatigue cracks to propagate, compression loads may do so.
Many service failures still occur by fatigue, despite the well-defined characterization of fatigue properties in the laboratory. In part, the difficulty is that fatigue behavior is influenced by a great many variables such as the magnitude and frequency of application of the fluctuating stress, the presence of a mean stress, temperature, environment, specimen size and shape, the presence of residual stresses, stress concentration, surface condition, and material imperfections or discontinuities. These factors complicate laboratory simulation of actual application conditions. The problem of fatigue, as a progressive form of damage, also can be sensitive to small imperfections or variations that accumulate into a critical form. Thus, due to its complex and varied conditions, prevention of fatigue failure requires education and awareness in all stages in the life of a part from design and manufacturing to anticipated service, operation inspection, and maintenance.

Fatigue Properties and Design Life

The design of components against fatigue fracture may involve several considerations of irregular loading, variable temperature, and environment. The three design-life methods or philosophies of fatigue are:

Design philosophy	Design methodology	Principal testing data description
Safe-life, infinite-life	Stress-life	*S-N*
Safe-life, finite-life	Strain-life	ε-*N*
Damage tolerant	Fracture mechanics	*da*/*dN*-Δ*K*

These methods have all contributed to a better understanding of both crack nucleation in regions of localized strain and the subsequent crack growth mechanisms outside the plastic zone in the qualification of component designs or to conduct failure analyses.

Fatigue Failures

Infinite-Life Criterion (*S-N* Curves)

The safe-life, infinite-life philosophy is the oldest of the approaches to fatigue. The design method is stress-life, and a general property representation would be *S-N* (stress versus log number of cycles to failure; failure in *S-N* testing is typically defined by total separation of the sample). The applied stresses are described by several parameters used to describe the cyclic-stress conditions (Fig. 1).

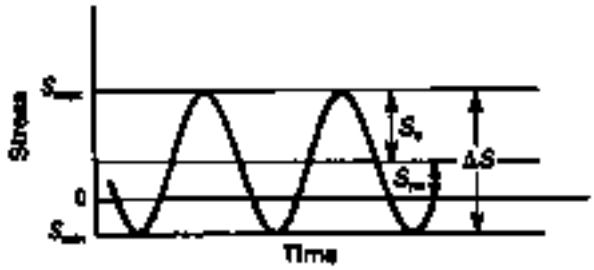


Fig. 1 Schematic showing test parameters for cyclic-stress testing

The stress ratio is the algebraic ratio of two specified stress values in a stress cycle. Two commonly used stress ratios are: the ratio, A, of the alternating stress amplitude to the mean stress ($A = S_a/S_m$) and the ratio, R, of the minimum stress to the maximum stress ($R = S_{min}/S_{max}$). If the stresses are fully reversed, the stress ratio R becomes -1; if the stresses are partially reversed, R becomes a negative number less than 1. If the stress is cycled between a maximum stress and no load, the stress ratio R becomes 0. If the stress is cycled between two tensile stresses, the stress ratio R becomes a positive number less than 1. A stress ratio R of 1 indicates no variation in stress, making the test a sustained-load creep test rather than a fatigue test.

The mean stress, S_m, is the algebraic average of the maximum and minimum stresses in one cycle, $S_m = (S_{max} + S_{min})/2$. In the completely reversed test, the mean stress is 0. The range of stress, S_r, is the algebraic difference between the maximum and minimum stresses in one cycle, $S_r = S_{max} - S_{min}$. The stress amplitude, S_a, is one-half the range of stress, $S_a = S_r/2 = (S_{max} - S_{min})/2$. During a fatigue test, the stress cycle is usually maintained constant, so that the applied stress conditions can be written $S_m \pm S_a$, where S_m is the static or mean stress, and S_a is the alternating stress, which is equal to half the stress range.

Mean Stress Effects. A series of fatigue tests is conducted at various mean stresses, and the results are plotted as a series of *S-N* curves. For design purposes, it is more useful to know how the mean stress affects the permissible alternating stress amplitude for a given life (number of cycles). This is usually accomplished by plotting the allowable stress amplitude for a specific number of cycles as a function of the associated mean stress. At zero mean stress, the allowable stress amplitude is the effective fatigue limit for a specified number of cycles. As the mean stress increases, the permissible amplitudes steadily decrease until, at a mean stress equal to the ultimate tensile strength of the material, the permissible amplitude is zero.

For general consideration of mean stress effects, various models of the mean-amplitude response have been proposed. One type of plot is the Haigh diagram, which plots mean stress (S_m) versus stress amplitude (S_a). The Haigh diagram can be a plot of real data, but it requires an enormous amount of information for substantiation. A slightly more involved, but also more useful, means of showing the same information incorporates the Haigh diagram with S_{max} and S_{min} axes to produce a constant-life diagram.

The conventional plot associated with this problem is produced using the Haigh diagram, with the Goodman line connecting the ultimate strength on S_m, and the fatigue limit, corrected fatigue limit, or fatigue strength on the S_a axis (Fig. 2). This line then defines the boundary of combined mean-amplitude pairs for anticipated safe-life response. The Goodman relation is linear and can be readily adapted to a variety of manipulations.

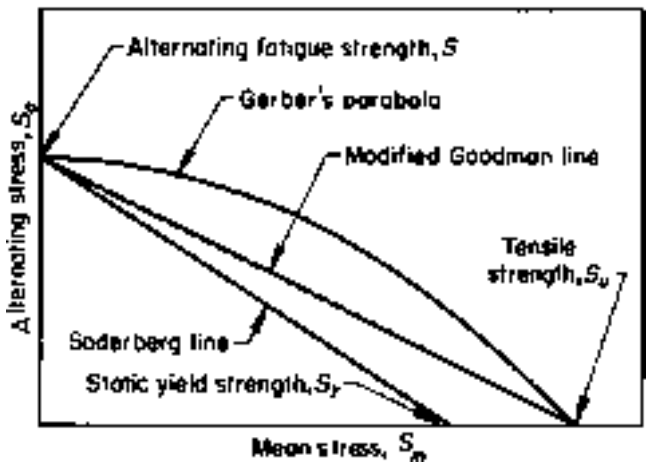


Fig. 2 Effect of mean stress on the alternating stress amplitude, as shown by the modified Goodman line, Gerber's parabola, and Soderberg line. See text for discussion.

The Goodman line is a commonly encountered representation of mean-stress effects, but other models (e.g., Gerber and Soderberg) are also used (Fig. 2). The two straight lines and the curve shown in Fig. 2 represent the three most widely used empirical relations. The straight line joining the alternating fatigue strength to the tensile strength is the modified Goodman law (Eq 1):

$$S_a = S\left[1-\left(\frac{S_m}{S_u}\right)\right] \qquad \text{(Eq 1)}$$

where S_u is the tensile strength. Goodman's original law, which is no longer used, included the assumption that the alternating fatigue limit was equal to one-third of the tensile strength; this has since been modified to the relation shown in Fig. 2, using the alternating fatigue strength determined experimentally.

Gerber found that the early experiments of Wöhler fitted closely to a parabolic relation, and this is known as Gerber's parabola (curve, Fig. 2). Gerber's law is given in Eq 2:

$$S_a = S\left[1-\left(\frac{S_m}{S_u}\right)^2\right] \qquad \text{(Eq 2)}$$

The third relation, known as Soderberg's law (Eq 3), is given in Fig. 2 by the straight line from the alternating fatigue strength to the static yield strength (S_y):

$$S_a = S\left[1-\left(\frac{S_m}{S_y}\right)\right] \qquad \text{(Eq 3)}$$

For many purposes it is essential that the static yield strength not be exceeded, and this relation is intended to fulfill the conditions that neither fatigue failure nor yielding occurs.

An understanding of the Goodman, or constant-life, diagram has resulted in many varied and useful treatments for improving fatigue life. According to a constant-life diagram, increased tension decreases the fatigue life, and increased compression increases it. Because most cracks originate at the surface of the part, placing the surface under compressive stress should be beneficial. Recognition of this has resulted in development of such surface treatments as nitriding, carburizing, shot peening, surface rolling, and overstressing. When these treatments are properly applied, the surface is in a state of residual compression. However, if not properly applied, they can have a detrimental effect on fatigue life.

In many cases Haigh or constant-life diagrams are simply constructs, using the Goodman representation as a means of approximating actual response through the model of the behavior. For materials that do not have a fatigue limit, or for finite-life estimates of materials that do, the fatigue strength at a given number of cycles can be substituted for the intercept on the stress-amplitude axis. Examples of the Haigh and constant-life diagrams are provided in Fig. 3 and 4 (Ref 1). Figure 4 is of interest also because of its construction in terms of a percentage of ultimate tensile strength for the strength ranges included.

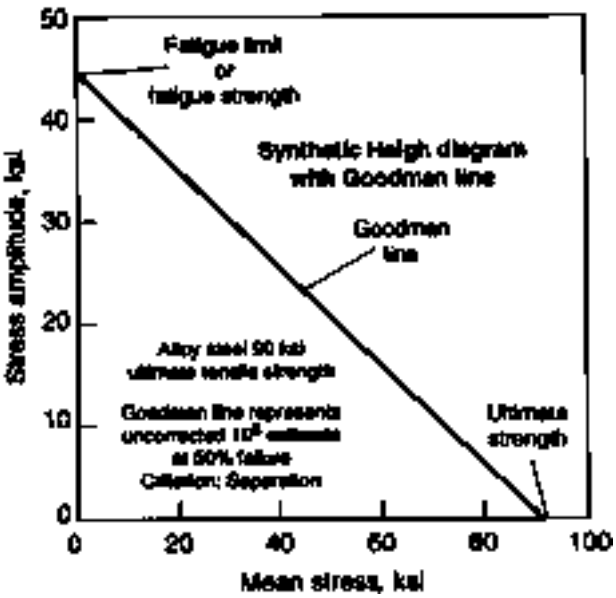


Fig. 3 A synthetically generated Haigh diagram based on typically employed approximations for the axes intercepts and using the Goodman line to establish the acceptable envelope for safe-life, infinite-life combinations

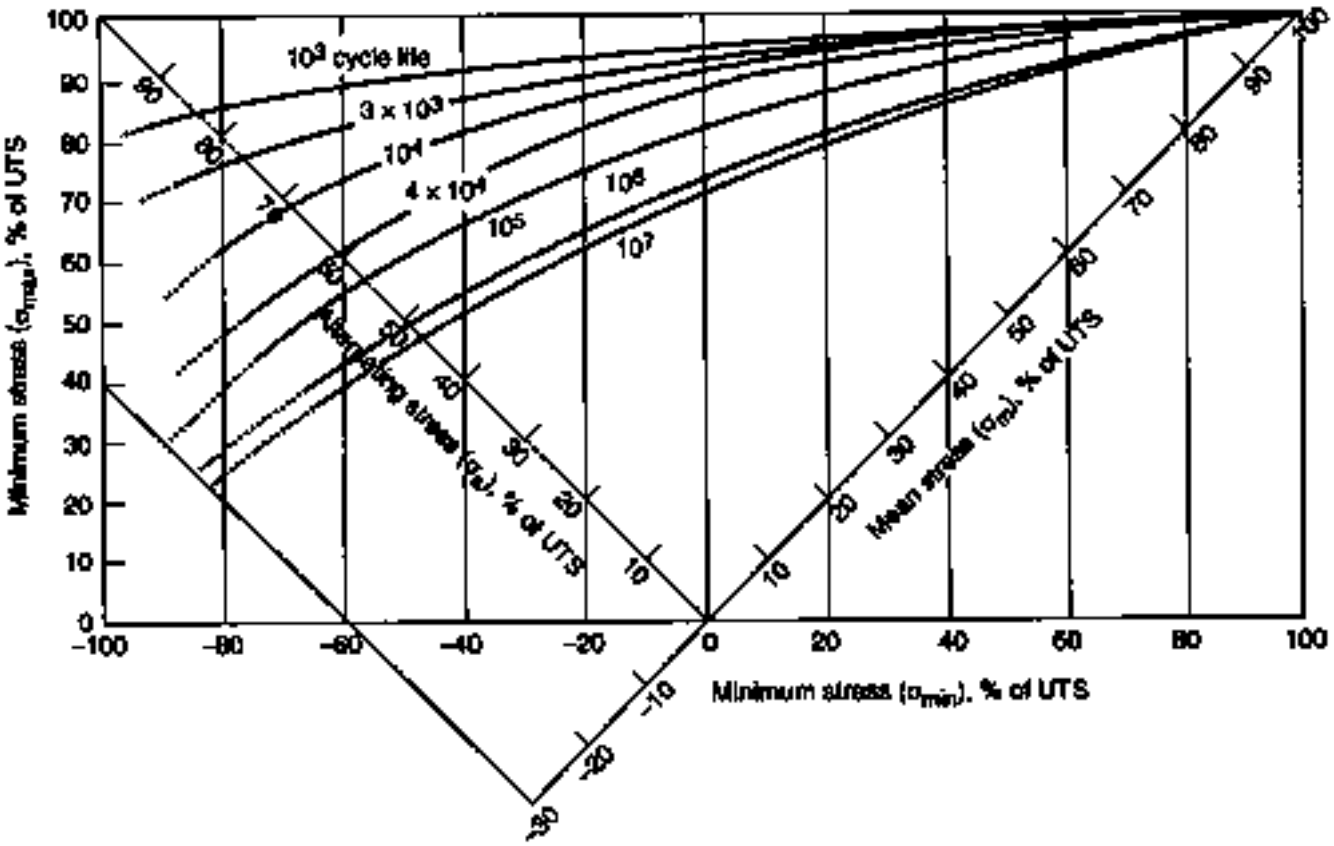


Fig. 4 A constant-life diagram for alloy steels that provides combined axes for more ready interpretation. Note the presence of safe-life, finite-life lines on this plot. This diagram is for average test data for axial loading of polished specimens of AISI 4340 steel (ultimate tensile strength, UTS, 125 to 180 ksi) and is applicable to other steels (e.g., AISI 2330, 4130, 8630). Source: Ref 1

Stress Amplitude. Because stress amplitude will vary widely under actual loading conditions, it is necessary to predict fatigue life under various stress amplitudes. The most widely used method of estimating fatigue under complex loading is provided by the linear-damage law. This is a hypothesis first suggested by Palmgren and restated by Miner, and is sometimes known as Miner's law (Ref 2). Application of n_i cycles at a stress amplitude S_i, for which the average number of cycles to failure is N_i, is assumed to cause an amount of fatigue damage that is measured by the cumulative cycle-ratio n_i/N_i, and failure is assumed to occur when $\sum(n_i/N_i) = 1$. However, this method is not applicable in all cases, and numerous alternative methods of cumulative linear damage have been used, as discussed in more detail in Ref 3.

The effect of varying the stress amplitude (linear damage) may be evaluated experimentally by means of a test in which a given number of stress cycles are applied to a test-piece at a one stress amplitude, and the test is continued to fracture at a different amplitude. Alternatively, the stress may be changed from one stress amplitude to another at regular intervals; such tests are known as block, or interval, tests. These tests do not simulate service conditions, but are useful in assessing the linear-damage law and indicating its limitations. Results of tests by several investigators have shown that initial overstressing reduces both the fatigue limit and the subsequent fatigue life at stresses above the fatigue limit. The results also show that a slight overstress does not markedly reduce the fatigue limit, even if continued for a large proportion of the normal life of a material. However, this is not true for a high overstress.

Advantages and Limitations. The stress-life approach seems best applied to components that look like the test samples and are approximately the same size (this satisfies the similitude associated with the use of total separation as a failure criterion). In addition, the fatigue limit or endurance limit at a high number of cycles (typically > 10^6) of steels under benign environmental conditions provides a useful and beneficial result of *S-N* testing. However, most other materials do not exhibit this infinite life response. Instead, many materials display a continuously decreasing stress-life response, even at a great number of cycles (10^6 to 10^9), which is more correctly described by a fatigue strength at a given number of cycles.

The general limitation of the stress-life method is restricted to circumstances where a no-crack assumption can be applied. Nonetheless, some design guidelines for weldments (which inherently contain discontinuities) offer what amount to residual life and runout determinations for a variety of process and joint types that generally follow the safe-life, infinite-life approach. The advantages of this method are simplicity and ease of application, and it can offer some initial

perspective on a given situation. It is best applied in or near the elastic range, addressing constant-amplitude loading situations in what has been called the *long-life* (hence, *infinite-life*) *regime.*

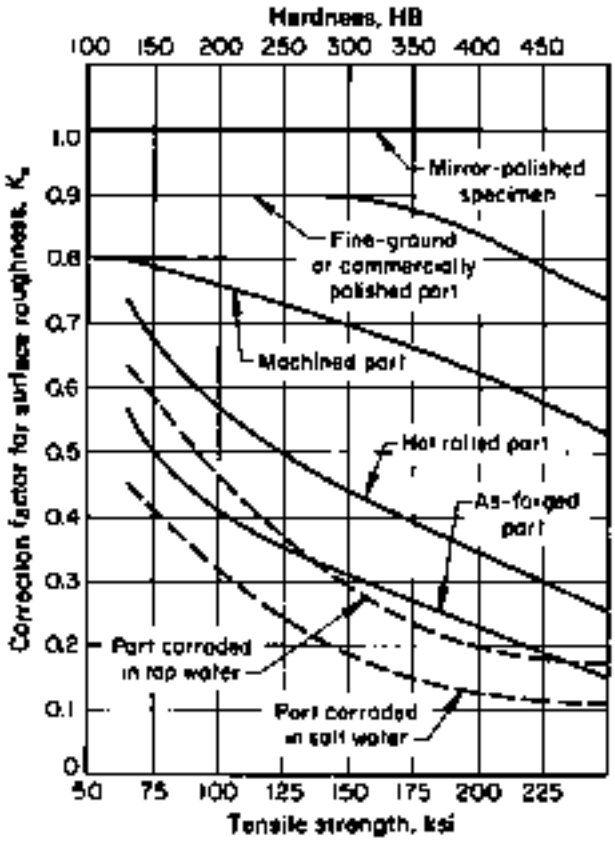
Assessing Fatigue *S-N* Properties. Given the extensive history of the stress-life method, substantial property data are available, but one should beware of the testing conditions employed in producing older data. The usefulness of property data is a critical point due to the numerous variables that influence fatigue results. For example, if a series of tests is conducted at a constant stress ratio ($R = S_{min}/S_{max}$), and the alternating stress amplitude (S_a) is used as the other independent dynamic variable, an *S-N* curve for that situation can be produced, and all dynamic variables can be determined. However, if only one variable is given (e.g., S_a or S_{max}), there is insufficient information to tell what the test conditions were, and the data are virtually useless.

In many cases, insufficient information is available for the effective use of *S-N* data. Many necessary pieces of data are simply missing. A partial list of important questions might be:

- What were the coupon size and geometry?
- Was there a stress concentration?
- What was the temperature?
- Was an environment other than lab air employed?
- What was the specimen orientation in the original material?
- Does the line represent minimum, mean, or median response?
- How many samples were tested?
- What was the scatter?
- If the plot is based on constant-amplitude data, what were the frequency and waveform?
- Was testing performed using variable-amplitude loading? What was the spectrum?
- What was the failure criterion?
- If there were runouts, how were they handled and represented?

An example of what should be considered important as supporting facts can be found in ASTM E 468, “Presentation of Constant Amplitude Fatigue Test Results for Metallic Materials.” It provides guidelines for presenting information other than just final data.

Correction Factors for Test Data. The available fatigue data are normally for a specific type of loading, specimen size, and surface roughness. For instance, the Moore rotating-beam fatigue-test machine uses a 7.6 mm (0.3 in.) diameter (diam) specimen that is free of any stress concentrations because of specimen shape, is polished to a mirror finish, and is subjected to completely reversed bending stresses. It has been suggested that the fatigue-life data used in design calculations be corrected by multiplying the number of cycles (life), N_i' determined in a fatigue test by three factors that account for the variation in the type of loading, part diameter, and surface roughness. Thus, the design life $N_i = K_1 K_d K_s N_i'$ where K_1 is the correction factor for the type of loading, K_d for the part diameter, and K_s for the surface roughness. Values of the three factors for correcting standard Moore fatigue-test data on steel are given in Fig. 5.



Factor	Type of load: Bending	Torsion	Axial
K_1	1.0	0.58	0.9(a)
K_d, where:			
$d \le 0.4$ in.	1.0	1.0	1.0
0.4 in. $< d \le 2$ in.	0.9	0.9	1.0
K_s	From chart above		

(a) A lower value (0.6 to 0.85) may be used to account for known or suspected underestimated bending because of load eccentricity.

Fig. 5 Correction factors for surface roughness (K_s), type of loading (K_1), and part diameter (K_d) for fatigue life of steel parts. See text.

References cited in this section

1. R.C. Juvinall, *Engineering Considerations of Stress, Strain, and Strength,* McGraw-Hill, 1967, p 274
2. M.A. Miner, Cumulative Damage in Fatigue, *Trans. ASME,* Vol 67, 1945, p A159
3. M.R. Mitchell, Fundamentals of Modern Fatigue Analysis for Design, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 227–249

Finite-Life Criterion (ε-*N* Curves)

Strain life is the general approach employed for continuum response in the safe-life, finite-life regime. It is primarily intended to address the low-cycle fatigue area (e.g., from approximately 10^2 to 10^6 cycles). The ε-*N*

method can also be used to characterize the “long-life” fatigue behavior of materials that do not show a fatigue limit.

From a properties standpoint, the representations of strain-life data are similar to those for stress-life data. However, because plastic strain is a required condition for fatigue, strain-controlled testing offers advantages in the characterization of fatigue crack initiation (before subsequent crack growth and final failure). The *S-N* method is based on just one failure criterion: the total separation of the test coupon. In contrast, any of the following may be used as the failure criterion in strain-controlled fatigue testing (per ASTM E 606): separation, modulus ratio, microcracking (initiation), or percentage of maximum load drop. This flexibility can provide better characterization of fatigue behavior.

Testing for strain-life data is not as straightforward as the simple load-controlled (stress-controlled) *S-N* testing. Monitoring and controlling using strain requires continuous extensometer capability. In addition, the developments of the technique may make it necessary to determine certain other characteristics associated with either monotonic or cyclic behavior. Further details on testing are given in the article “Fatigue, Creep Fatigue, and Thermomechanical Fatigue Life Testing” in *Mechanical Testing and Evaluation,* Volume 8 of the *ASM Handbook.*

Damage Tolerance Criterion

In the latter half of the 20th century, the fracture mechanics approach to crack propagation gained acceptance in the prediction of fatigue life. In this approach, crack initiation is neglected and crack propagation to final failure is considered. This method can give a conservative estimate of fatigue life and has had considerable success. However, the treatment of short cracks, the propagation of cracks for negative *R*-ratios and for irregular loading, crack closure effects, and multiaxial loadings provide significant challenges. Reference 4 and *Fatigue and Fracture,* Volume 19 of the *ASM Handbook,* cover these topics in more detail.

Fatigue Crack Growth Rates

The limitation of the *S-N* and ε-*N* methods is the inability of the controlling quantities to make sense of the presence of a crack. In traditional stress analysis of applied loads, basic elasticity calculations show that both stress and strain become astronomical at a discontinuity such as a crack, far exceeding any recognized property levels. Even invoking plasticity still leaves inordinately large numbers or, conversely, extremely low tolerable loads.

The solution to this situation is the characterization and quantification of the stress field at the crack tip in terms of stress intensity (*K*) in linear elastic fracture mechanics. The crack-tip stress-intensity factor, *K*, uniquely characterizes the elastic stress field near a crack and represents the crack driving force that controls the onset of brittle fracture and the rate of subcritical crack growth in structures. Its magnitude depends on the applied nominal stress, geometry, and the crack size. Fracture mechanics methods are thus used to determine the crack-growth rates of pre-existing cracks under cyclic loading, where the subcritical crack growth from fatigue loading, da/dN, is characterized by the stress-intensity range, $\Delta K = K_{max} - K_{min}$.

Variables that influence fatigue-crack-growth behavior are stress ratio, loading frequency (not a factor in benign environments), material chemistry, heat treatment, test temperature, and environment. Fatigue crack propagation behavior of metals generally can be divided into three regions, as shown in Fig. 6 (Ref 5). In region 1, crack growth is negligible for an almost unlimited number of cycles. Region 1 has a fatigue-crack-propagation threshold (ΔK_{th}) where crack-growth rate becomes diminishingly small and a region of low ΔK. As the crack grows, the stress intensity at the crack tip increases, and crack-growth rates are characterized by region 2 with the power-law relationship: $da/dN = C(\Delta K)^n$ where *C* and *n* are the coefficient and exponential terms.

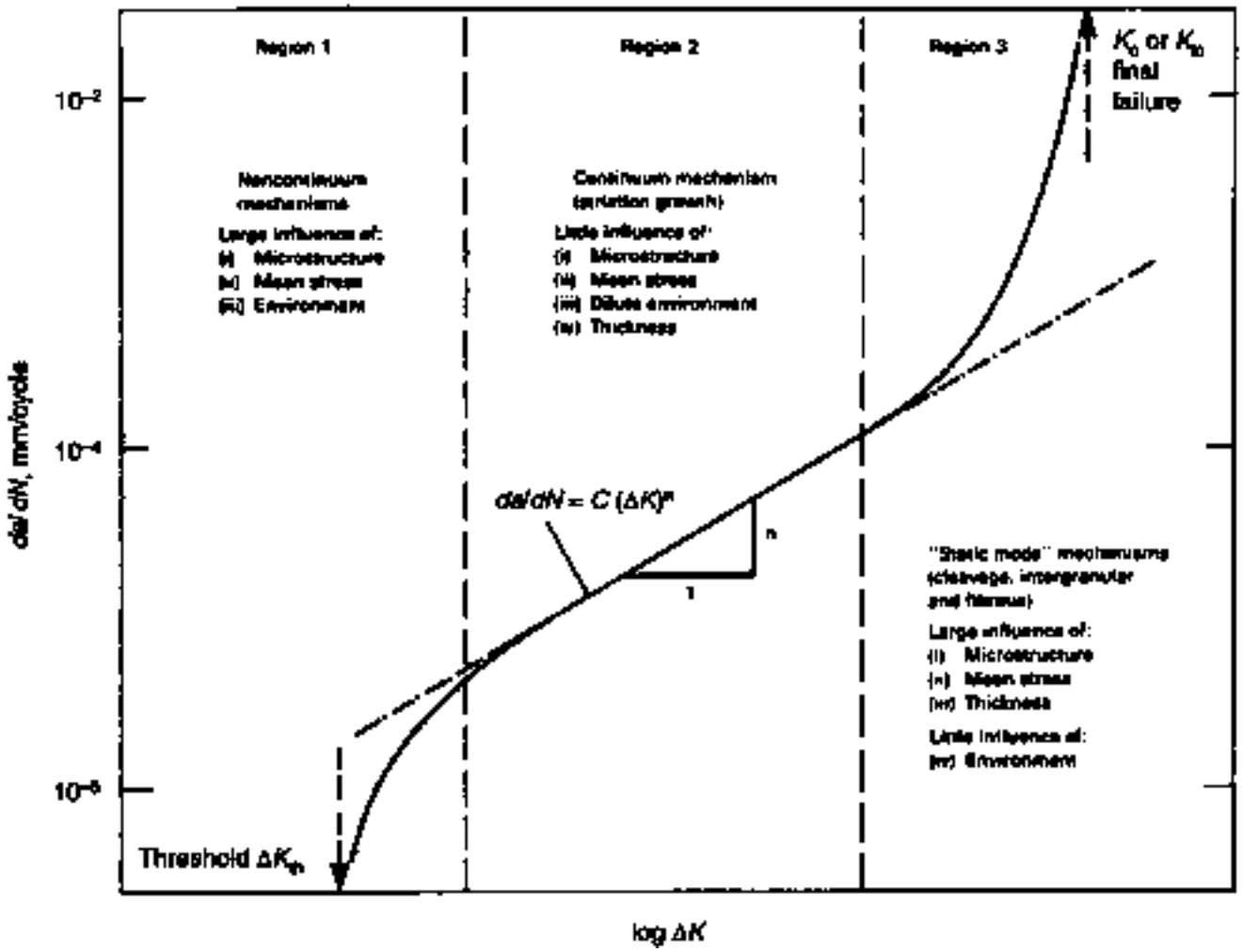


Fig. 6 Schematic illustration of variation of fatigue-crack-growth rate, *da/dN*, with alternating stress intensity, Δ*K*, in steels, showing regions of primary crack-growth mechanisms. Source: Ref 5

The fatigue-crack-growth threshold, ΔK_{th}, is an important design parameter for such applications as rotating shafts involving low-stress, high-frequency fatigue loading where no crack extension during service can be permitted. The fatigue-crack threshold is a function of a number of variables, including the material, the test conditions, the R-ratio, and the environment. ASTM E 647 defines ΔK_{th} as that asymptotic value of ΔK at which da/dN approaches zero. For most materials an operational, although arbitrary, definition of ΔK_{th} is given as that ΔK which corresponds to a fatigue crack growth rate of 10^{-10} m/cycle.

However, short fatigue cracks can grow at stress-intensity levels below the threshold level for macroscopic cracks, as demonstrated by Pearson in 1973 (Fig. 7)(Ref 6). Small cracks have also been observed to grow at non-negligible rates when the nominal applied ΔK is less than the threshold value determined from traditional (large-crack) test methods. Therefore, a structural life assessment based on large-crack analysis methods can be nonconservative if the life is dominated by small-crack growth.

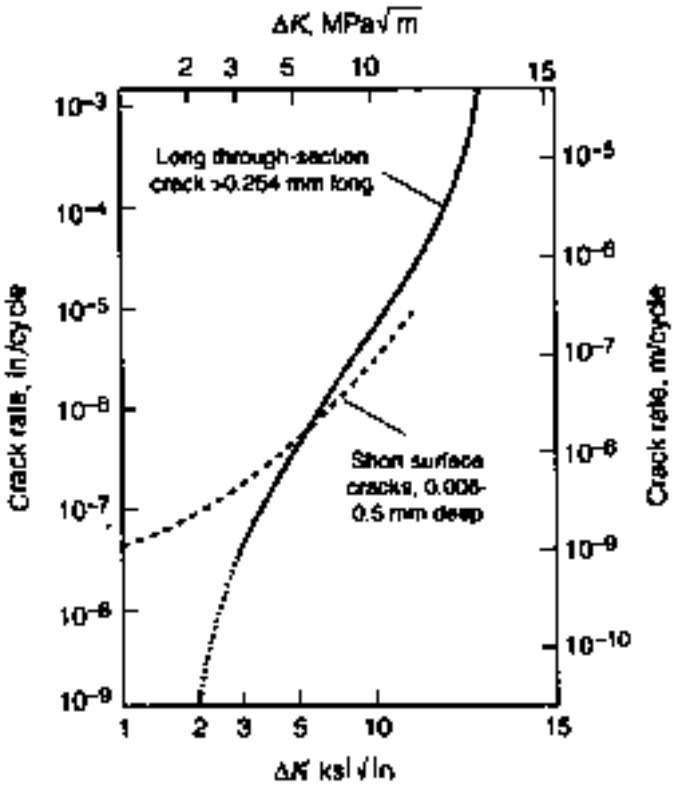


Fig. 7 Pearson plot of crack-growth rate as a function of *K* for short surface cracks and through-cracks. Source: Ref 6

Small-crack behavior is a complex subject, due to the variety of factors that may affect small cracks and the variety of microstructures used in engineering structures. In contrast to large-crack growth rates, which generally increase with increasing ΔK, small-crack growth rates are sometimes observed to increase, decrease, or remain constant with increasing ΔK. From these complex variations, apparent disagreements in data and explanations are not uncommon in the literature.

When small cracks are being evaluated, it is important to recognize that all small cracks are not the same (Ref 7). Different mechanisms are responsible for different types of small-crack effects in different settings. Criteria that properly characterize small-crack behavior in one situation may be entirely inappropriate in another situation. Therefore, before selecting suitable analytical treatments, it is critical to understand the different types of small cracks, such as microstructurally small, mechanically small, and chemically small cracks (Ref 7).

When the surface crack length is larger than 0.5 mm (0.02 in.), then a simple conventional fracture mechanics law can be applied to calculate the threshold condition (Ref 8). However, as a crack under consideration is made smaller and smaller, there is a transition from linear elastic fracture mechanics treatment of long cracks at threshold to an endurance-limit-dominated behavior of short cracks. Kitagawa diagrams are used to depict this type of situation. Below a surface crack length of 0.5 mm (0.02 in.), the threshold stress range departs gradually from its macroscopic value and approaches a constant stress range level that is approximately equal to the fatigue limit of unnotched smooth specimens (Ref 8).

Large Cracks. In traditional (large-crack) methods, conservative estimates of ΔK_{th} for various steels subjected to different stress ratios, R, can be predicted (Ref 9) from:

$$\Delta K_{th} = 6.4(1 - 0.85R) \text{ for } R \geq +0.1 \qquad \text{(Eq 4)}$$

and

$$\Delta K_{th} = 5.5 \text{ for } R < +0.1 \qquad \text{(Eq 5)}$$

where ΔK_{th} is in units of ksi $\sqrt{in}$ for both equations. These equations indicate that the fatigue-crack-propagation threshold for steels is primarily a function of the stress ratio and is essentially independent of chemical or mechanical properties.

Region 2. Behavior in region 2 is relevant to design situations involving a finite number of fatigue cycles. Extensive fatigue-crack-growth-rate data for various steels show that the primary parameter affecting growth rate in region 2 is the stress-intensity factor range, and that the mechanical and metallurgical properties of these steels have negligible effects on the fatigue-crack-growth rate in a room-temperature air environment. The data for martensitic steels fall within a single band (Fig. 8), and the upper bound of scatter can be obtained as follows (Ref 9):

$$da/dN = 0.66 \times 10^{-8}(\Delta K_I)^{2.25} \qquad \text{(Eq 6)}$$

where a is in inches and ΔK_I is in units of ksi $\sqrt{in}$ Similarly, as shown in Fig. 9 (Ref 9), data for ferrite-pearlite steels fall within a single band (different from the band for martensitic steels), and the upper bound of scatter can be calculated from:

$$da/dN = 3.6 \times 10^{-10}(\Delta K_I)^{3.0} \qquad \text{(Eq 7)}$$

where again a is in inches and ΔK_I is in units of ksi $\sqrt{in}$

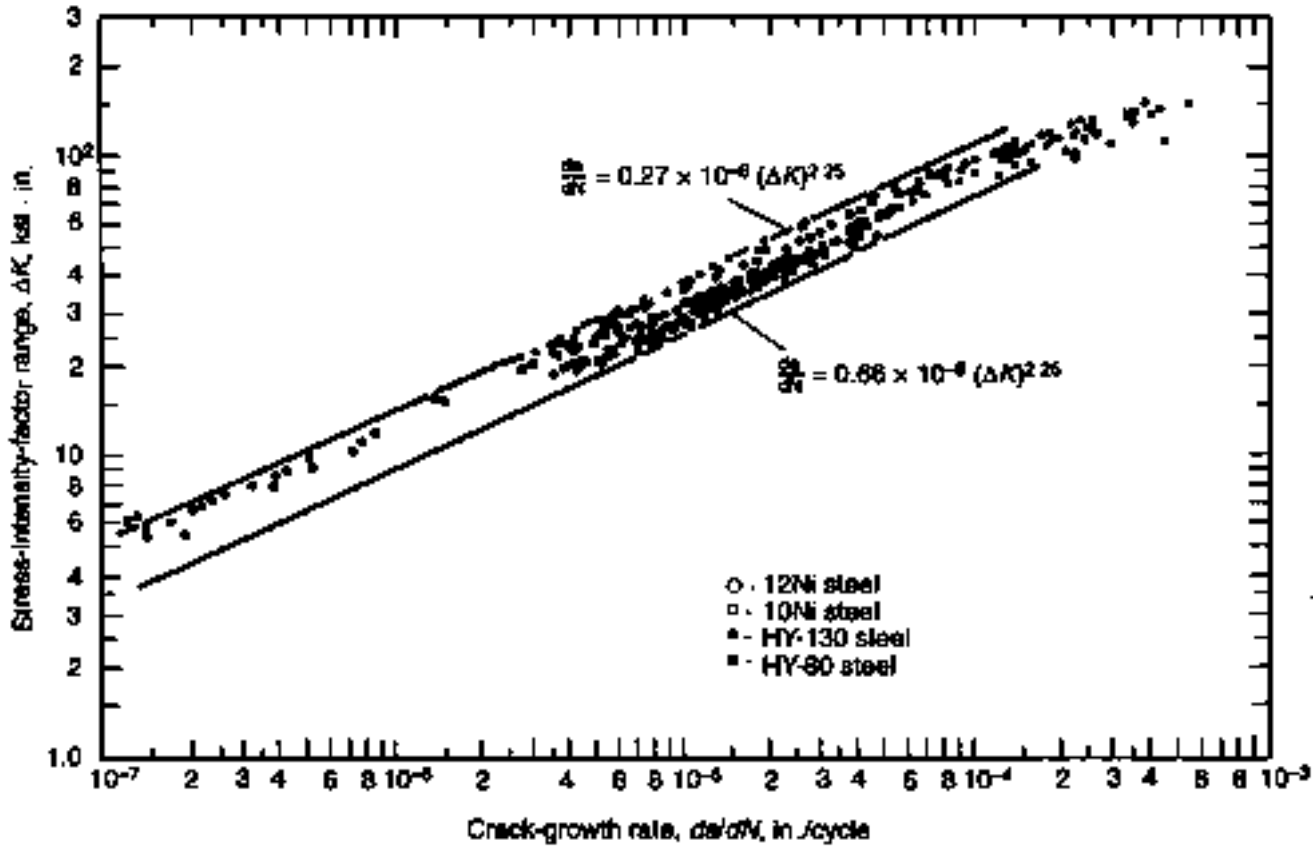


Fig. 8 Summary of fatigue-crack-growth data for martensitic steels. Source: Ref 9

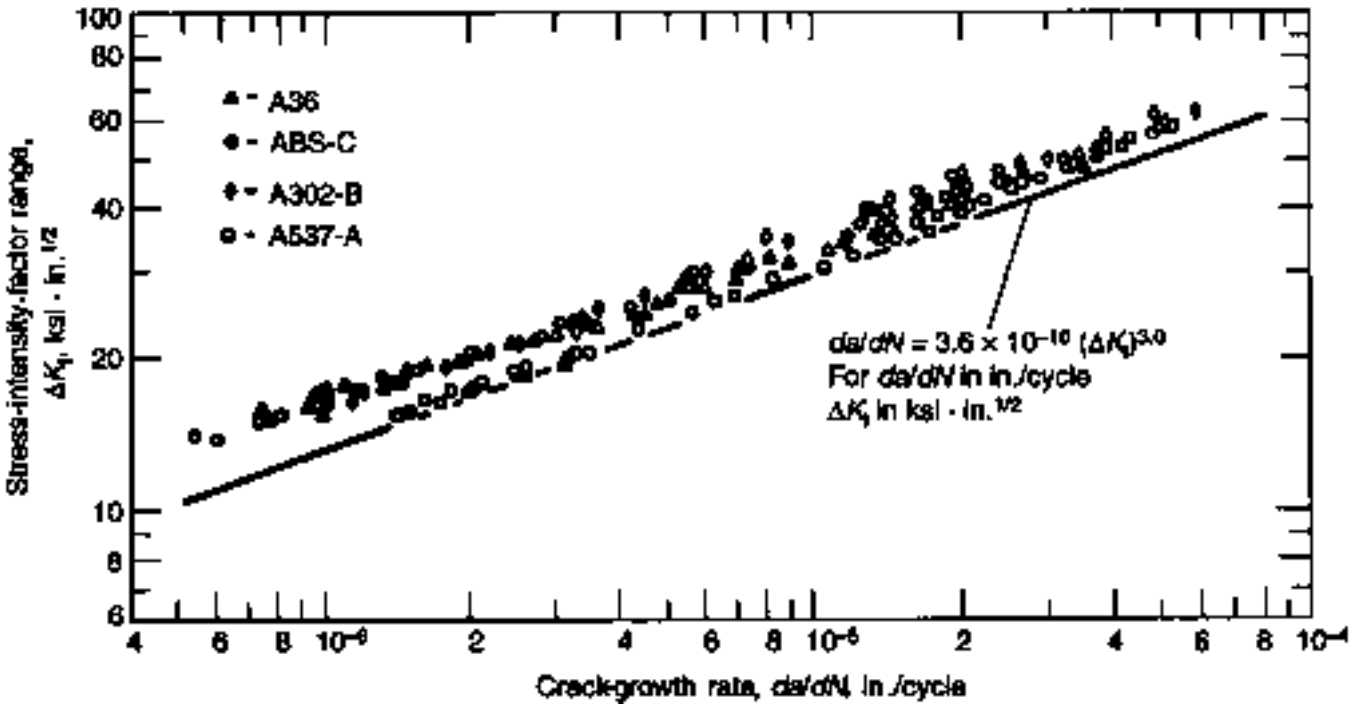


Fig. 9 Summary of fatigue-crack-growth data for ferrite-pearlite steels. Source: Ref 9

The stress ratio and mean stress have negligible effects on the rate of crack growth in region 2. Also, the frequency of cyclic loading and the waveform (sinusoidal, triangular, square, or trapezoidal) do not affect the rate of crack propagation for steels in benign environments (Ref 9).

Region 3. The acceleration of fatigue-crack-growth rates that determines the transition from region 2 to region 3 appears to be caused by the superposition of a brittle or a ductile-tearing mechanism onto the mechanism of cyclic subcritical crack extension, which leaves fatigue striations on the fracture surface. These mechanisms occur when the strain at the tip of the crack reaches a critical value. Thus, the fatigue-rate transition from region 2 to region 3 depends on the maximum stress-intensity factor, on the stress ratio, and on the fracture properties of the material. Region 3 behavior for rapid crack growth is significant only for applications that may experience very few (of the order of ten or fewer) load and unload cycles—for example, pressure vessels in which pressure may be discharged only a few times during the service life.

Environmentally Assisted Fatigue Crack Propagation. The environmentally assisted fatigue-crack growth in metals is a complex phenomenon and is influenced by loading, metallurgical, and environmental variables. In a given alloy, in addition to the environmental effects, crack-growth rates can vary widely with variations in loading parameters, such as stress state, stress-intensity range, and stress-wave shape and frequency. For example, crack size and geometry may alter the electrochemical environment near the crack tip and consequently change the crack-growth rate and may, in extreme cases, change the controlling process and growth mechanism. In one case, the fatigue-crack-growth rate for a small crack in a high-strength (4130) steel in a 3% sodium chloride solution was about ten times faster than that of long cracks at similar stress-intensity levels in the same bulk environment, although the short-crack fatigue data obtained in air do coincide with those for long cracks. The difference, therefore, is purely chemical in nature. The magnitude of the environmental enhancement of crack growth varies, depending on crack size and on applied stress. At a low stress range, crack-growth rates exceed those observed in air by two orders of magnitude and are faster than the values reported for long cracks, while at a high stress range, the rate of cracking is nearly equivalent to that reported for long cracks. For low-strength steels, the crack-size effect does not appear to be as large. These factors and others are discussed in more detail in the article “Mechanisms of Corrosion Fatigue” in *Fatigue and Fracture,* Volume 19 of the *ASM Handbook.*

Life Prediction

The fracture mechanics approach for characterizing fatigue-crack growth can be used in design applications to estimate maximum flaw sizes that allow a part to reach its design life. This approach is also very useful for conducting failure analyses.

Because predicting service fatigue life often involves integrating crack-growth rates over a range of ΔK values, the da/dN versus ΔK relationship can be represented by:

$$\frac{da}{dN} = \frac{A_1}{(\Delta K)^{n_1}} + A_2\left[\frac{1}{(\Delta K)^{n_2}} - \frac{1}{\left[(1-R)K_c\right]^{n_2}}\right] \qquad \text{(Eq 8)}$$

where n_1, n_2, A_1, A_2, and K_c are best fit material constants determined by regression analysis. This equation is general and applies to all three regions of crack growth.

Example 1: Developing Inspection Criteria for Fan Shafts. Fan shafts are typically forged from medium-carbon steels. Primary fatigue stresses are generated from gravity bending and are fully reversed during each revolution at rates of 900 to 1800 revolutions per minute (rpm) or 15 to 30 Hz. Maximum allowable alternating stress on the fan shaft is typically ±52 MPa (±7.5 ksi). Before shipping, shaft forgings are subjected to various types of nondestructive inspections to reveal inclusions and forging bursts. Fatigue-crack propagation analysis can be used to evaluate such defects.

Investigation. Specimens for mechanical testing were obtained by slicing a 580 mm (23 in.) diam shaft forging into 75 mm (3 in.) thick disks. Tensile specimens and Charpy impact specimens were located at various radial distances from the center of the shaft cross section. Compact-type specimens, 25.4 and 6.25 mm (1 and 0.25 in.) thick, were used for fracture-toughness and fatigue-crack-growth rate testing. In addition, 12.5 × 50.8 mm (0.5 × 2 in.) center-crack-tension specimens were also machined from the central region of another disk from the same forging for characterizing fatigue-crack growth at R = -1.0.

Test results at several radial distances from the shaft center indicated that tensile strength increases slightly from the peripheral region of the shaft toward the center. Fatigue-crack-growth testing was conducted on specimens taken from two extreme radial locations of the shaft, but no trend could be attributed to specimen location. Figure 10 shows the wide range of fatigue-crack-growth-rate behavior of the shaft material for load ratios of -1.0, 0.1, and 0.5, which represent a spectrum of typical load histories encountered in service. For a load ratio of -1.0, only the positive half of the load cycle was used to compute ΔK. These data were used to determine the material constants A_1, A_2, n_1, and n_2 using a standard multiple-linear regression computer program.

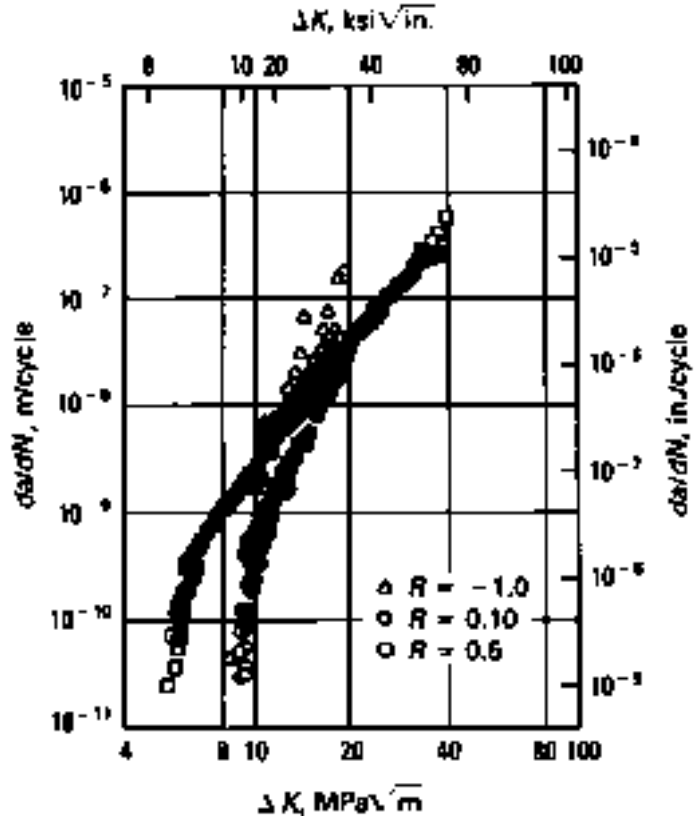


Fig. 10 Room-temperature fatigue-crack-growth behavior of ASTM grade A293 steel. Levels of *R* represent spectrum of typical in-service load history.

Discussion. The number of fatigue cycles that accumulate on the fan shaft during the projected 40-year service life can be about 10^{11} cycles, and little crack extension during service is acceptable. Thus, the ΔK_{th} of the shaft material for completely reversed loading (R = -1.0) is selected as the maximum allowable value of ΔK.

Conclusions. A growth rate of 10^{-15} m (4×10^{-14} in.) per cycle can be considered acceptable for this application because it implies a crack extension of 0.1 mm (0.004 in.) during the 10^{11} fatigue cycles. The ΔK corresponding to this growth rate can be determined from Eq 4 by using the appropriate fitting constants at R = -1.0. This value is 6.4 MPa$\sqrt{m}$ (5.8 ksi$\sqrt{in.}$), and a safety factor of 20% on the ΔK_{th} value reduces the allowable K to 5.1 MPa$\sqrt{m}$ (4.6 ksi$\sqrt{in}$).

Two common defect geometries are shown in Fig. 11 with the flaw shape factor, Q, for various values of the flaw aspect ratio, $a/2c$, and ratios of applied stress to the yield strength of the material. The value of Q influences the magnitude of K according to:

$$K = M_T \sigma \sqrt{\frac{\pi a}{Q}} \qquad \text{(Eq 9)}$$

For fatigue, this relationship can be rewritten as:

$$\Delta K = M_T \Delta\sigma \sqrt{\frac{\pi a}{Q}} \qquad \text{(Eq 10)}$$

where $\Delta\sigma$ is the applied stress range and M_T is the finite width magnification factor. Substituting the allowable value of ΔK yields the critical flaw size, a_{cr}:

$$a_{cr} = \left(\frac{5.1}{M_T \Delta\sigma}\right)^2 \frac{Q}{\pi} \qquad \text{(Eq 11)}$$

For deeply embedded internal flaws, M_T = 1.0, and for small surface defects in large shafts, M_T = 1.1. These values are used to calculate critical flaw sizes of various aspect ratios for several fatigue stress levels (Fig. 12). For surface flaws, critical flaw sizes are a factor of 1.1 lower for equivalent $a/2c$ and stress values.

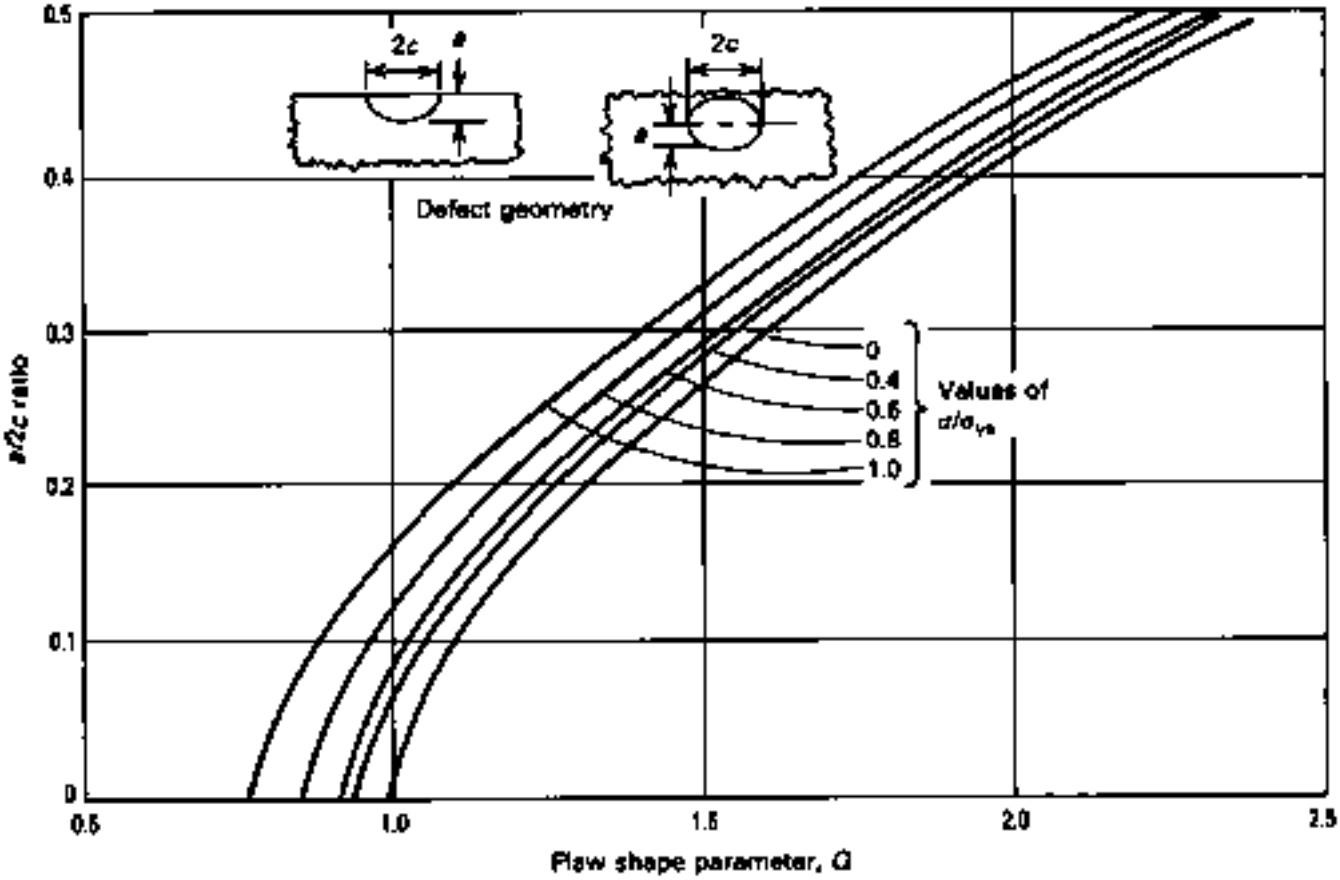


Fig. 11 Flaw shape parameter curves for surface and internal cracks

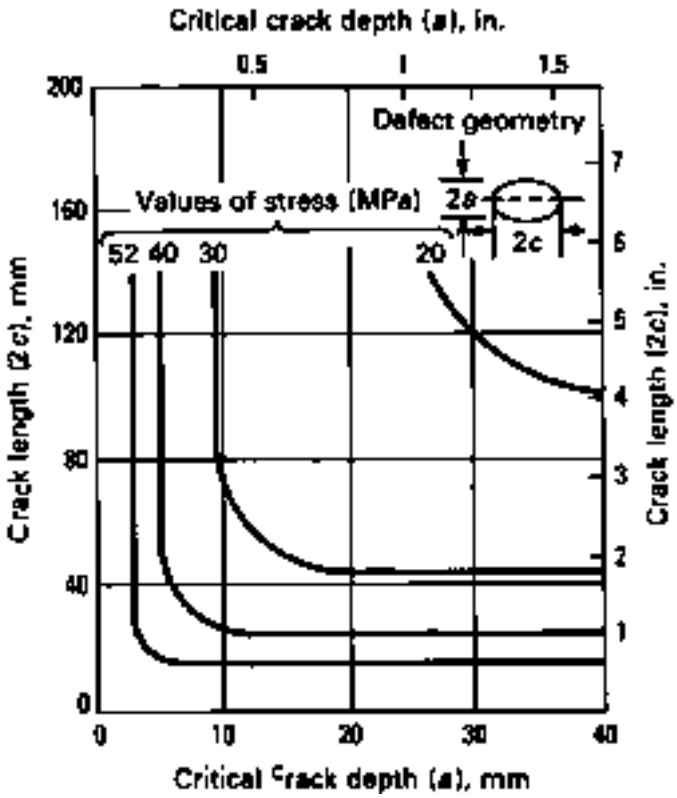


Fig. 12 Critical sizes of internal flaws at various stress levels. Surface flaws are a factor of 1.1 lower.

References cited in this section

4. “Advances in Fatigue Lifetime Predictive Techniques,” STP 1122, M.R. Mitchell and R.W. Landgraf, Ed., ASTM, 1992

5. R.O. Ritchie, Near-Threshold Fatigue-Crack Propagation in Steels, *Int. Met. Rev.,* Vol 24 (No. 5, 6), 1979

6. S. Pearson, Initiation of Fatigue Cracks in Commercial Aluminum Alloys and the Subsequent Propagation of Very Short Cracks, *Eng. Fract. Mech.,* Vol 7 (No. 2), 1975, p 235–247

7. R.C. McClung, K.S. Chan, S.J. Hudak, Jr., and D.L. Davidson, Behavior of Small Fatigue Cracks, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 153–158.

8. A.J. McEvily, Fatigue Crack Thresholds, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 134–152.

9. S.T. Rolfe and J.M. Barsom, *Fracture and Fatigue Control in Structures—Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

Characteristics of Fatigue Fracture

Fatigue cracks generally initiate in a highly stressed region of a component subjected to cyclic stresses of sufficient magnitude. The crack propagates under the applied stress through the material until complete fracture results. On the microscopic scale, the most important feature of the fatigue process is the nucleation of one or more cracks under the influence of reversed stresses that exceed the flow stress, followed by the development of cracks at persistent slip bands or at grain boundaries. Subsequently, fatigue cracks propagate by a series of opening and closing motions at the tip of the crack that produce, within the grains, striations that are parallel to the crack front.
The process of fatigue consists of three stages:

- *Stage I:* Initial fatigue damage leading to crack nucleation and crack initiation
- *Stage II:* Progressive cyclic growth of a crack (crack propagation) until the remaining uncracked cross section of a part becomes too weak to sustain the loads imposed
- *Stage III:* Final, sudden fracture of the remaining cross section

Stage III occurs during the last stress cycle when the cross section cannot sustain the applied load. The final fracture—which is the result of a single overload and which can be brittle, ductile, or a combination of the two—is not a really fatigue process and is discussed in more detail in the article “Overload Failures” in this Volume.
Examination of fatigue-fracture surfaces usually begins visually or by low-magnification optical (light) microscopy. Macroscopic examination of fracture surfaces involves relatively simple techniques; it can often be done at the failure site, requires little or no preparation of the specimen, requires minimal and relatively simple equipment, and does not destroy the specimen or alter the fracture surfaces. Macroscopic examination is particularly useful in correlating fracture-surface characteristics with part size and shape and with loading conditions. The crack origin can best be found by visual examination or by viewing the fracture surface at low magnifications (25 to 100×).
Examination of fatigue fractures by optical microscopy is often difficult because the height of features on the fracture surface may exceed the depth of field of the microscope, especially at high magnifications. Metallographic examinations of cross sections through suspected fatigue failures typically show that the crack path is transgranular, although intergranular fatigue does occur at times.

Crack Initiation

In general, fatigue cracks initiate and propagate in regions where the strain is most severe. Because most engineering materials contain defects, and thus regions of stress concentration that intensify strain, most fatigue cracks initiate and grow from structural defects. Under the action of cyclic loading, a plastic zone—or region of deformation—develops at the defect tip. This zone of high deformation becomes an initiation site for a fatigue crack. The crack propagates under the applied stress through the material until complete fracture results.

Fatigue cracks form at the point or points of maximum local stress and minimum local strength. The local stress pattern is determined by the shape of the part, including such local features as surface and metallurgical imperfections that concentrate macroscopic stress, and by the type and magnitude of the loading. Strength is determined by the material itself, including all discontinuities, anisotropies, and inhomogeneities present. Local surface imperfections, such as scratches, mars, burrs, and other fabrication flaws, are the most obvious flaws at which fatigue cracks start. Surface and subsurface material discontinuities in critical locations also influence crack initiation. Inclusions of foreign material, hard precipitated particles, and crystal discontinuities such as grain boundaries and twin boundaries are examples of microscopic stress concentrators in the material matrix.

Stage I Mechanisms and Appearances. Stage I (the initiation stage) is the submicroscopic alteration of the crystal structure of a metal by repetitive stressing of the specimen or part. In this stage, fatigue cracks nucleate and coalesce by slip-plane fracture, extending inward from the surface at approximately 45° to the stress axis. On the submicroscopic scale, dislocation density, lattice defects, and the orientation of cross-slip planes control the formation of persistent slip bands, intrusions and extrusions, and dislocation cells, which ultimately control nucleation mechanisms in smooth specimens.

Stage I may be difficult to envision or describe because of its submicroscopic nature. Even with electron microscopes, its very difficult to examine and may not be discernible at all, depending on geometry, metal, environment, and stress level. A stage I fracture never extends over more than about two to five grains around the origin. In each grain, the fracture surface is along a well-defined crystallographic plane, which should not be confused with a cleavage plane, although it has the same brittle appearance. There are usually no fatigue striations associated with a stage I fracture surface. In some cases, depending on the material, environment, and stress level, a stage I fracture may not be discernible.

Unnotched Smooth Specimens On a microscopic level, cracks develop at persistent slip bands or at grain boundaries. The local directions of propagation are controlled to some extent by crystallographic planes and may form on or may be parallel to slip bands in grains near the surface. In interior grains, cleavage cracks often form at the intersection of slip bands with grain boundaries. A variety of crystallographic features has been observed to nucleate fatigue cracks. In pure metals, tubular holes that develop in persistent slip bands, extrusion-intrusion pairs at free surfaces, and twin boundaries are common sites for crack initiation.

Cracking may also initiate at grain boundaries in polycrystalline materials, even in the absence of inherent grain-boundary weakness; at high strain rates, the grain boundary seems to be the preferred site for crack nucleation. Nucleation at a grain boundary appears to be purely a geometrical effect related to plastic incompatibility, whereas nucleation at a twin boundary is associated with active slip on crystallographic planes immediately adjacent and parallel to the twin boundary.

The above processes also occur in alloys and heterogeneous materials. However, alloying and commercial production practices introduce segregation, inclusions, second-phase particles, and other features that disturb the structure, and these have a dominant effect on the crack-nucleation process. In general, alloying that enhances cross slip or twinning, or that increases the rate of work hardening will retard crack nucleation. On the other hand, alloying usually raises the flow stress of a metal, which speeds up crack propagation, thus at least partly offsetting the potentially beneficial effect on fatigue-crack nucleation.

In high-strength materials containing spheroidal second-phase particles, secondary crack nucleation ahead of the main crack front (slipless fatigue) occurs around such particles. In slipless fatigue, cracks form along lattice planes that are unfavorably located relative to the maximum tensile stress.

Relation to Environment. In observing locations of crack nucleation, the possibility of environment-related mechanisms—including pitting corrosion, stress-corrosion cracking, and other effects of a hostile environment—must be considered. For example, a great number of fatigue failures in otherwise lightly loaded structures originate in fretted areas. In any structure having joints with some relative motion, fretting provides a possible failure origin. Once a fatigue crack has been nucleated, its rate and direction of growth are controlled by localized stresses and by the structure of the material at the tip of the crack.

Fatigue Crack Propagation

Initial Propagation. The transition from stage I to stage II fatigue fracture is the change of orientation of the main fracture plane. The transition is from one or two shear planes in stage I to many parallel plateaus separated by longitudinal ridges in stage II. The plateaus are usually normal to the direction of maximum tensile stress (Fig. 13a). A transition from stage I to stage II in a coarse-grain specimen of aluminum alloy 2024-T3 is shown in Fig. 13(b). The presence of inclusions rich in iron and silicon did not affect the fracture path markedly. The inclusions, which were fractured, ranged from 5 to 25 μm in diameter. In Fig. 13(b), the stage II area shows a large number of approximately parallel fatigue patches containing very fine fatigue striations that are not resolved at the magnification used. Fine striations are typical in stage II, but are usually seen only under high magnification.

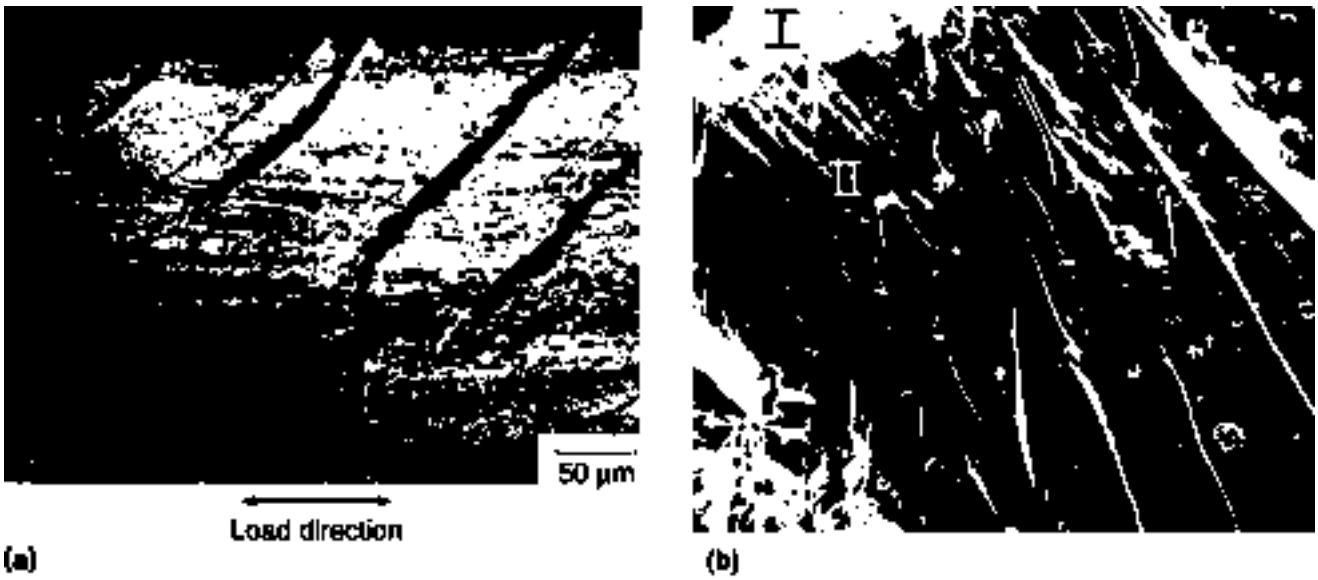


Fig. 13 Transition from stage I to stage II fatigue. (a) Change in fracture path from stage I (top of photo) to stage II (bottom) (b) Transition from stage I to stage II of a fatigue fracture in a coarse-grain specimen of aluminum alloy 2024-T3. Source: Ref 10

In the roots of small fatigue-crack notches, the local stress state is triaxial (plane strain). This reduces the local apparent ductility of the material and helps control the orientation of the crack as long as the crack is small. After a crack has nucleated and propagated to a larger size, it becomes a macroscopic stress raiser and can be more influential than any stress raiser already in the part. At this point, the crack tip will take over control of the fracture direction. Subsequently, the orientation of the crack surface will depend on the stress field at the crack tip and will often follow a series of void coalescences in advance of the crack front. Early crack extension occurs under plane-strain conditions and gives a typical fine-grain, flat-face surface that, when produced under random loading or sequences of high and low stress amplitudes, exhibits characteristic beach marks.

As a general rule, low-stress high-cycle fatigue produces flat-face (plane-strain) fractures. The fracture surface appears fine grained and lightly polished near the crack-nucleation site, where the stress intensification is least. The surface becomes progressively rougher and more fibrous as the crack grows and the intensity of stress increases. On high-stress low-cycle fatigue surfaces, found in certain areas of all complete fatigue fractures, the surface is fibrous, rough, and more typical of plane-stress loading conditions, where the general fracture direction is at 45° to the main tensile load.

Crack-propagation variations in anisotropic or inhomogeneous materials occasionally produce beach marks that are difficult to interpret. In highly anisotropic material such as spring wire, for example, a fatigue-fracture surface may exhibit regions where the crack propagated in typical fatigue and other regions where it propagated preferentially in another mode along planes of weakness. Beach marks or striations, which are the identifying feature of fatigue in many components, are not always present in spring fatigue-fracture surfaces. Beach marks are more likely to be present if the spring material is comparatively soft (40 to 45 HRC), but they are seldom present when higher-strength spring materials fail by fatigue (Ref 11).

In large structural components, the existence of a crack does not necessarily imply imminent failure of the part. Significant structural life may remain in the cyclic growth of the crack to a size at which a critical failure occurs. The

growth or extension of a fatigue crack under cyclic loading is principally controlled by maximum load and stress ratio. However, as in crack initiation, a number of additional factors may exert a strong influence, including environment, frequency, temperature, grain direction, and other microstructural factors.

Beach Marks. The most characteristic feature usually found on fatigue-fracture surfaces is beach marks, which are centered around a common point that corresponds to the fatigue-crack origin. Also called clamshell, conchoidal and arrest marks, beach marks are perhaps the most important characteristic feature in identifying fatigue failures. Beach marks can occur as a result of changes in loading or frequency or by oxidation of the fracture surface during periods of crack arrest from intermittent service of the part or component.

Figure 14 shows an example of beach marks on the fracture surface of an aluminum alloy 7075-T73 forging. The beach marks are produced by changes in crack-growth rates from the constantly changing stress intensity at the crack tip. The initial crack-growth rate is slow, but increases with the increasing stress intensity at the crack tip as the crack extends further. Beach marks are also produced by other factors that cause changes in crack-growth rates during propagation. For example, Fig. 15 shows the surface of a fatigue fracture in a 4130 steel shaft that failed in service; this surface exhibits beach marks produced by oxidation of the fracture when the shaft was idle. Beach marks may also be produced by load variations during service, spectrum-load fatigue testing, or changes in stress intensity from geometric features of the part.

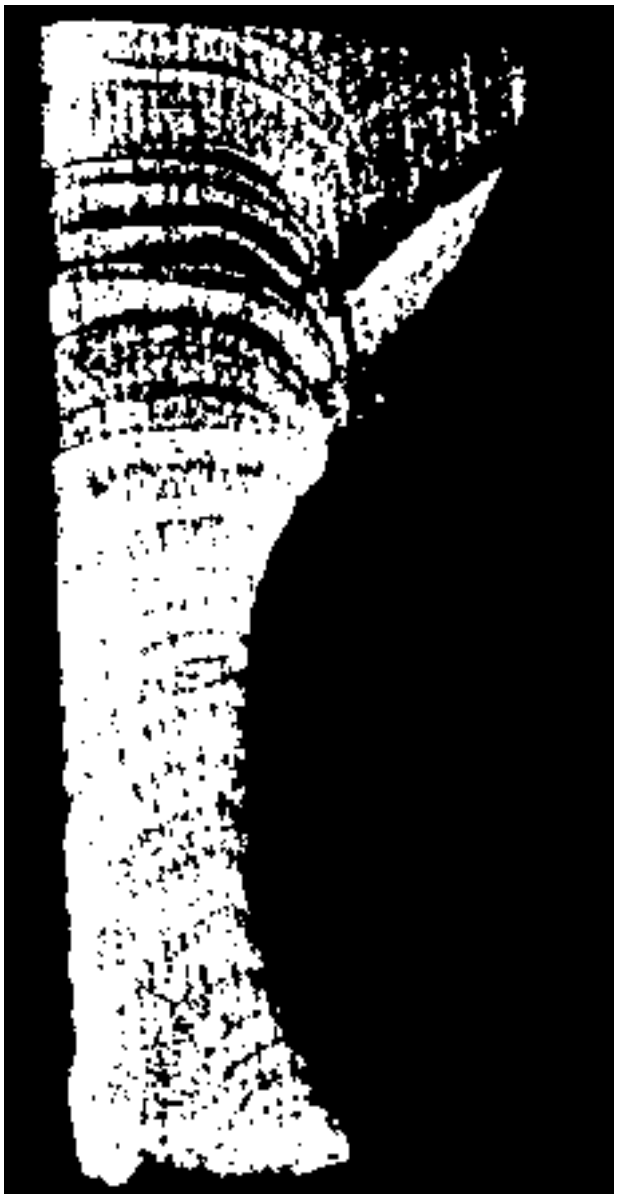


Fig. 14 Beach marks on a fatigue fracture in aluminum alloy 7075-T73 forging. The light-colored reflective bands are zones of fatigue crack propagation. At high magnifications, thousands of fatigue striations can be resolved within each band. The dull, fibrous bands are zones of crack propagation by microvoid coalescence. Light fractograph; actual size

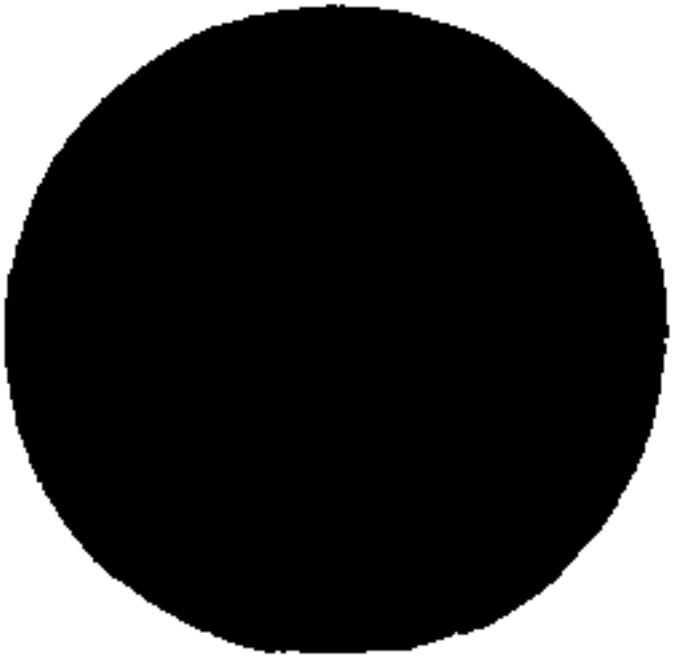


Fig. 15 Fracture surface of steel shaft with beach marks produced by oxidation.

However, many fatigue fractures produced under conditions of uninterrupted crack growth and without load variations do not exhibit beach marks. Moreover, sometimes a fracture surface may have features that look like fatigue beach marks. For example, Fig. 16 is a surface from a stress-corrosion cracking (SCC) fracture that reveals features like fatigue beach marks.

Fig. 16 Beach marks on a 4340 steel part caused by SCC. Tensile strength of the steel was approximately 1780 to 1900 MPa (260 to 280 ksi). The beach marks are a result of differences in the rate of penetration of corrosion on the surface. They are in no way related to fatigue marks. 4×

Ratchet marks are macroscopic features that may be seen on fatigue fractures in shafts and flat-leaf springs; they may also occur in ductile fractures of over-torqued fasteners. In fatigue fractures, ratchet marks are the result of multiple fatigue-crack origins, each producing a separate fatigue crack zone. In shafts, the separately initiated cracks normally are propagated on planes slightly inclined to the plane of the shaft diameter. As two approaching cracks meet, a small step is formed. The small steps around the periphery of a shaft are the ratchet marks. Although ratchet marks are most apparent on the periphery of fractures in shafts, the stepped appearance is characteristic whenever fatigue cracks emanate from several origins and subsequently meet to form one principal crack front (see Fig. 17 in the article “Fracture Appearance and Mechanisms of Deformation and Fraction”).

The occurrence of ratchet marks requires almost simultaneous initiation of several fatigue cracks, a situation that is favored by high stress and to a lesser extent by the presence of high stress-concentration factors, such as corrosion pits or grinding burns.

Striations. In the electron microscope examination of fatigue-fracture surfaces, the most prominent features found are patches of finely spaced parallel marks, called fatigue striations. The fatigue striations are oriented perpendicular to the microscopic direction of crack propagation Fig. 17. In general, two types of striations have been recognized: ductile striations (Fig. 17b) and brittle striations (Fig. 17c). The brittle striations are connected by with what seems to be cleavage fracture along sharply defined facets. Numerous river markings separating these facets run normal to the striations (Fig. 17c). Brittle striations in aluminum alloys usually are indication of a corrosive environment. Ductile striations are more common.

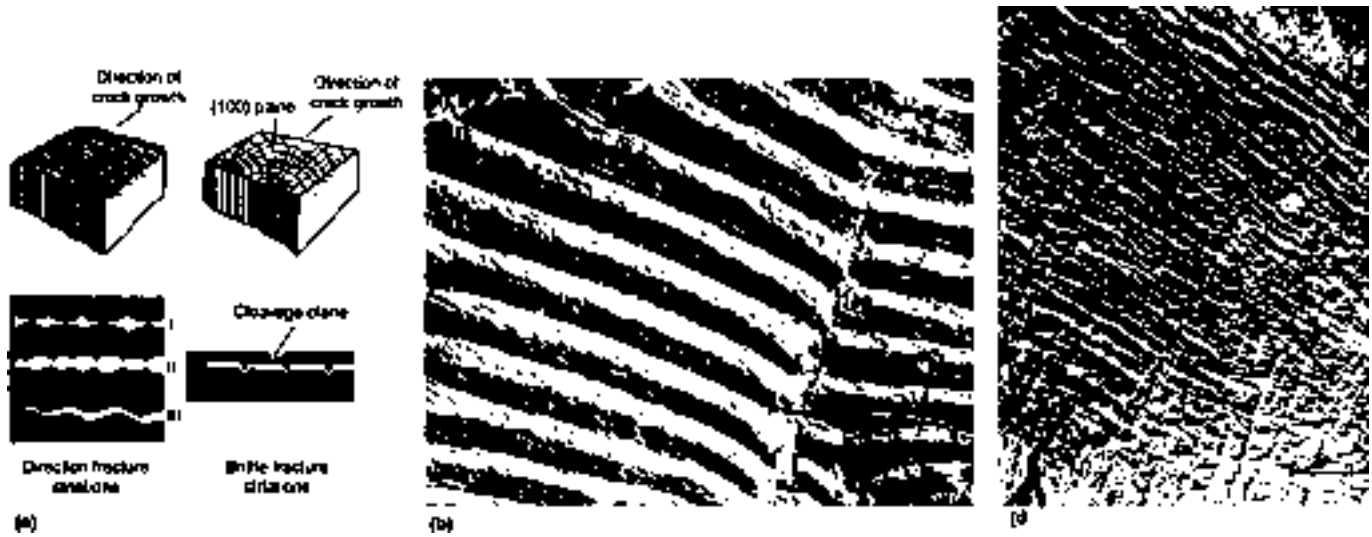


Fig. 17 Ductile and brittle striations. (a) Schematic of different types of ductile and brittle striations. (b) Ductile striations in 718 aluminium alloy. (c) Brittle fatigue striations of 2014 aluminium alloy. Note cleavage facets running parallel to direction of crack propagation and normal to striations. Source: R.M.N Pelloux, *Metals Quarterly*, Nov 1965, p 34

With uniform loading, fatigue striations generally increase in spacing as they progress from the origin of fatigue. Each striation is the result of a single cycle of stress, but every stress cycle does not necessarily produce a striation; striation spacing depends strongly on the level of applied loading. The clarity of the striations depends on the ductility of the material. Striations are more visible at stress levels higher than the fatigue limit, and they are more readily visible in ductile materials. Thus, patches of fatigue striations in high-strength steel are less visible than in an aluminum alloy. Also, the striations“Fracture Appearances and Mechanisms of Deformation and Fracture” are more prominent in stage 2 (Paris-Law region) of fatigue crack growth (see the article).

At high rates of crack growth (0.0025 mm, or 10^{-4} in., per cycle or more), the striations become wavy and develop a rough front. A large plastic zone exists in front of the crack, which may cause extensive secondary cracking. Each secondary crack propagates as a fatigue crack, creating a network of secondary striations. The local crack direction may

differ markedly from the overall direction of crack propagation because of the many changes in direction of the local fracture path.

In steels, fatigue striations that are formed at ordinary crack-growth rates are not always as well defined as they are in aluminum alloys. The striations that are formed in aluminum alloys at very low crack-growth rates (less than 1.3×10^{-4} mm, or 5×10^{-6} in., per cycle) are difficult to resolve and often cannot be distinguished from the network of slip lines and slip bands associated with plastic deformation at and near the crack front as it propagates through the section.

References cited in this section

10. D.A. Ryder, The Elements of Fractography, AGARDograph No. AGARD-AG-155-71, Advisory Group for Aerospace Research and Development of NATO, 1971

11. M. Hayes, Fatigue of Springs, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 363–369

Final Fracture (Stage III)

Final fracture occurs when the crack has grown to the critical size for overload failure. The size of the final-fracture zone depends on the magnitude of the loads, and its shape depends on the shape, size, and direction of loading of the fractured part. The final-fracture zone of a fatigue-fracture surface is often fibrous, resembling the fracture surfaces of impact or fracture-toughness test specimens of the same material.

Two features of the final-fracture zone aid in determining the origin of fracture. First, fatigue usually originates at the surface, and therefore the fatigue origin is not included in the shear-lip zones of the overload region. In tough materials, with thick or round sections, the final-fracture zone will consist of a fracture by two distinct modes: (a) tensile fracture (plane-strain mode) extending from the fatigue zone and in the same plane; and (b) shear fracture (plane-stress mode) at 45° to the surface of the part bordering the tensile fracture.

In thin sheet-metal pieces having sufficient toughness, final fracture occurs somewhat differently. As the crack propagates from the fatigue zone, the fracture plane rotates around an axis in the direction of crack propagation until it forms an angle of about 45° with the loading direction and the surface of the sheet. The fracture plane, inclined 45° to the load direction, can occur on either a single-shear or a double-shear plane (Fig. 18).

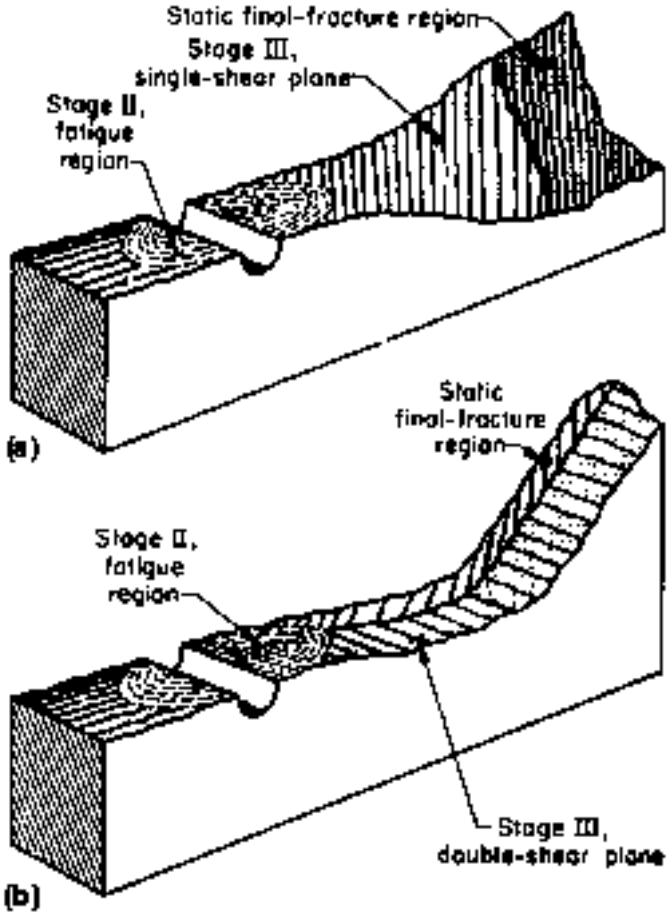


Fig. 18 Fracture planes that are 45° to the direction of loading. (a) Single-shear plane. (b) Double-shear plane

The second characteristic of a fast-fracture zone is chevron marks that point back to the origin of fracture. This is shown in Fig. 19 from a torsional-fatigue fracture of a steel shaft. Beach marks fan out from the origin at a sharp corner in a keyway at the top. Beyond the beach marks, cracking progressed by fast fracture along the hardened perimeter, producing two sets of chevron marks pointing back to the crack origin.

Fig. 19 Surface of a torsional-fatigue fracture in an induction-hardened 1041 (1541) steel shaft. The shaft fractured after 450 hours of endurance testing. $1\frac{1}{4}\times$

Effect of Loading and Stress Distribution

Nucleation and growth of a fatigue crack and the features on the fracture surface are all strongly affected by the shape of the part, the type and magnitude of loading exerted on the part in service, residual stresses, and metallurgical and environmental factors. Fatigue cracks usually initiate at some region of stress concentration resulting from the presence of surface discontinuities (stress raisers), such as a step or shoulder, a screw thread, an oil hole or bolt hole, or a surface flaw. In addition, mechanically fastened joints and welded joints also require special attention.

The magnitude of the nominal stress on a cyclically loaded component is often measured by the amount of overstress, that is, the amount by which the nominal stress exceeds the fatigue limit or the long-life fatigue strength of the material used in the component. The number of load cycles that a component under low overstress can endure is high; thus, the term *high-cycle fatigue* is often applied. Increasing the magnitude of the nominal stress means:

- Initiation of multiple cracks is more likely.
- Striation spacing is increased.
- The region of final fast fracture is increased.

With very high overstress, low-cycle fatigue fractures are produced. The arbitrary but commonly accepted dividing line between high-cycle and low-cycle fatigue is considered to be about 10^5 cycles. A more basic distinction is made by determining whether the dominant component of the strain imposed during cyclic loading is elastic (high cycle) or plastic

(low cycle), which in turn depends on the properties of the metal as well as the magnitude of the nominal stress. In extreme conditions, the dividing line between high-cycle and low-cycle fatigue may be less than 100 cycles.

Design methods should account for the irregular nature of actual load applications, which requires the statistical treatment of the load histories and their translation into load spectra, cycle counting methods, and damage accumulation concepts. For satisfactory correlation with service behavior, full-size or large-scale specimens can be tested under conditions as close as possible to those existing in service. This method is expensive, but it does provide valuable data. A less costly testing procedure is simplified laboratory testing. By using the fatigue information obtained from the testing of standard specimens or by using models and applying the proper correction factors for configuration, surface finish, environment, and various other parameters, an approximation of the lifetime of the component can be estimated.

It should be emphasized, however, that this is just an approximation. For improving production designs, the target load/life test method of evaluating the design of a part has been used by automotive, farm equipment, construction equipment, and aircraft companies. In this method, numerous failures from actual operating machines are observed. From these observations, conclusions can be drawn as to the type of loading that causes the failures. A fatigue test is then set up to apply the appropriate loading. After obtaining failures during the test, certain adjustments can be made in the magnitude and method of applying the loads until the same type of failure is developed in the test as those experienced by the operating machines.

Target lives for the parts in the fatigue test can be established, based on the time-to-failure on the machines in service, the time-to-failure in the fatigue tests, and the desired life in service. For example, if a certain part fails in one-fourth of the desired life of the machine, and if these same parts fail in 25,000 cycles during the fatigue tests, one would not consider any revised parts satisfactory unless they exceeded 100,000 cycles without failing. Once the appropriate loading and number of cycles are established for a part on a production machine, the same test can be used to ensure that similar parts on new machines will be satisfactory.

One application in which this method has been used is the rear axles of automobiles. Maximum tractive torque is applied to the rear axle for 100,000 cycles. If the axle gears withstand this loading, they are considered satisfactory for customer use. The advantages of this procedure are that it makes use of the experience gained from many machines already in use and that it is rather rapid. Its disadvantage is that sufficient machines must be placed in customers' hands to develop the failed parts for study and evaluation of test methods. Moreover, the differences in the operating conditions, environment, fatigue properties of the parts, and processing methods between the production design and new design must be small to produce accurate results. Therefore, this technique is primarily used when there is ample experience, as in improving parts that have a history of failure in service or in evaluating machines that are similar to existing production models.

Fatigue Failures

Load Conditions

Fracture analysis is relatively simple in a beam of simple uniform cross section subjected to axial, bending, or torsional loads. In the case of pure axial loading, the stress is constant across the cross section of the beam, and a fatigue crack may initiate at a discontinuity within the member rather than at the surface. However, pure axial loading is rarely found in service, and the appearance of fatigue cracks caused by cyclic tensile loads is often similar to that described for bending loads. In bending, stresses are greatest at the surface, where fatigue cracks are thus more likely to initiate, as described in more detail for unidirectional, alternating, and rotational bending.

Loading of flat-section components differs from that of long, more cylindrical components. In flat components, biaxial tension is more common, and torsion is less so. Under conditions of biaxial tension, fatigue life depends on the maximum shear stress rather than the principal tensile stresses. Thus, fatigue failure may not occur when the loading conditions result in low shear stresses but high tensile stresses.

In sheet or plate materials, the fatigue-crack front may extend under plane-strain conditions to give a wholly flat-face fracture (Fig. 20a). Fatigue fractures in very thin sheet materials subjected to high-stress intensities may shift from flat face (plane-strain conditions) to shear face (plane-stress conditions), as shown in Fig. 20(b).

Fig. 20 Fatigue-fracture zones in aluminum alloy 7075-T6 plates. (a) Fatigue crack that grew as a flat-face fracture with a shallow convex crack front. (b) Change in orientation of fatigue fracture from plane strain (arrow A) to plane stress (arrow B).

Unidirectional Bending

A beam with a uniform cross section subjected to cyclic unidirectional bending has a bending moment that is uniform along the length of the beam; the tensile fiber stress is also uniform along the length of the beam. Therefore, a fatigue crack may be initiated at any point along the beam; in fact, several fatigue cracks may have been active before one of them became large enough to cause the final fracture.

A related form of loading is cantilever loading, in which the bending moment, and therefore the tensile fiber stress, vary along the length of the beam. Fracture will initiate at the point of highest stress, adjacent to the rigid mounting. Because of the shear-stress component characteristic of cantilever loading, the direction of the maximum tensile stress makes a small angle with the axis of the beam. Consequently, a fatigue crack will usually propagate into that portion of the beam within the fixed mounting.

Figure 21 illustrates that fatigue marks produced from single origins at low and high nominal stresses and from multiple origins at high nominal stresses in a cylindrical bar or shaft subjected to a cyclic unidirectional-bending moment evenly distributed along the length are determined by some minor stress raiser, such as a surface discontinuity. Both the single origin and the smallness of the final-fracture zone in Fig. 21(a) suggest that the nominal stress was low. The crack front, which forms beach marks, is symmetrical relative to the origin and retains a concave form throughout. The larger final-fracture zone in Fig. 21(b) suggests a higher nominal stress.

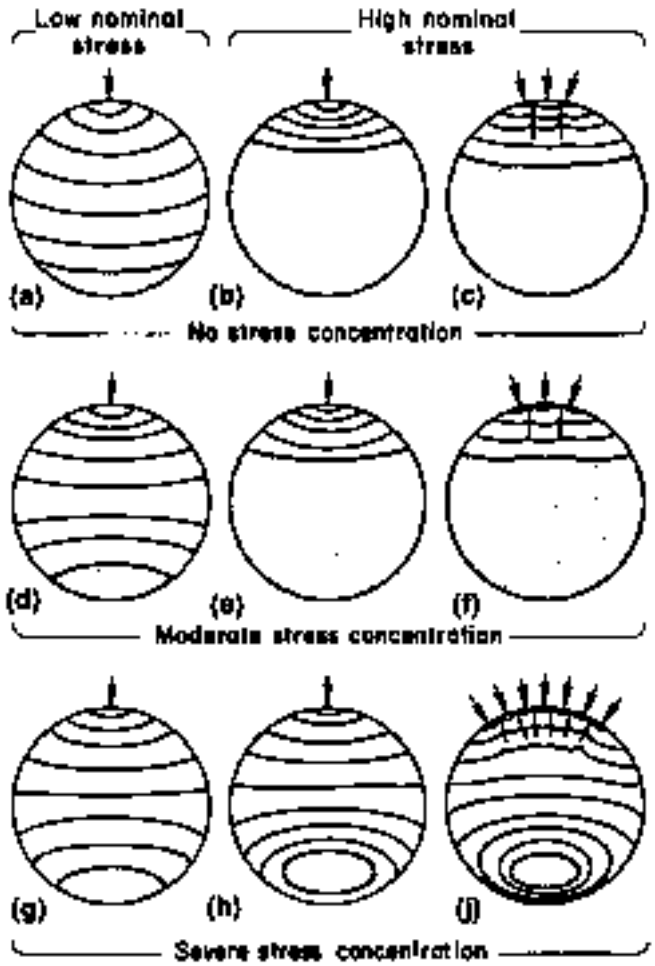


Fig. 21 Fatigue marks produced from single origins at low and high nominal stresses and from multiple origins at high nominal stresses. Fatigue marks are typical for a uniformly loaded shaft subjected to unidirectional bending. Arrows indicate crack origins; final-fracture zones are shaded.

Figure 21(c) shows a typical fatigue crack originating as several individual cracks that ultimately merged to form a single crack front. Such multiple origins are usually indicative of high nominal stress. Radial steps (ratchet marks) are present between crack origins. Figures 21(d), (e), and (f) show typical fatigue beach marks that result when a change in section in a uniformly loaded shaft provides a moderate stress concentration. With a low nominal stress, the crack front changes from concave to convex before rupture (Fig. 21d). At higher nominal stresses, the crack front flattens and may not become convex before final fracture (Fig. 21e and f).

A change in section in a uniformly loaded shaft that produces a severe stress concentration will lead to a pattern of beach marks such as that shown in Fig. 21(g), (h) or (j). An example of a severe stress concentration is a small-radius fillet at the junction of a shoulder and a smaller-diameter portion of a shaft or at the bottom of a keyway. Such a fillet usually results in the contour of the fracture surface being convex with respect to the smaller-section side. The crack-front pattern shown in Fig. 21(g) was produced by a low nominal stress. The crack front in Fig. 21(h) developed more rapidly because of a higher stress in the peripheral zone. Multiple crack origins, high nominal stress, and unidirectional bending usually produce the beach-mark pattern shown in Fig. 21(j).

Figure 22 is an example of a shaft that failed from unidirectional-bending fatigue. The shaft (a free-machining grade of A6 tool steel) was part of a clamping device on a tooling assembly used for tube bending. The shafts were subjected to a tensile stress imposed by the clamping force and a bending stress resulting from the nature of the operation. Unidirectional-bending stresses were imposed on one shaft when a right-hand bend was made in the tubing and on the other shaft when a left-hand bend was made. The tensile stress on the shafts was also cyclic because the clamping force was removed after each bend was made.

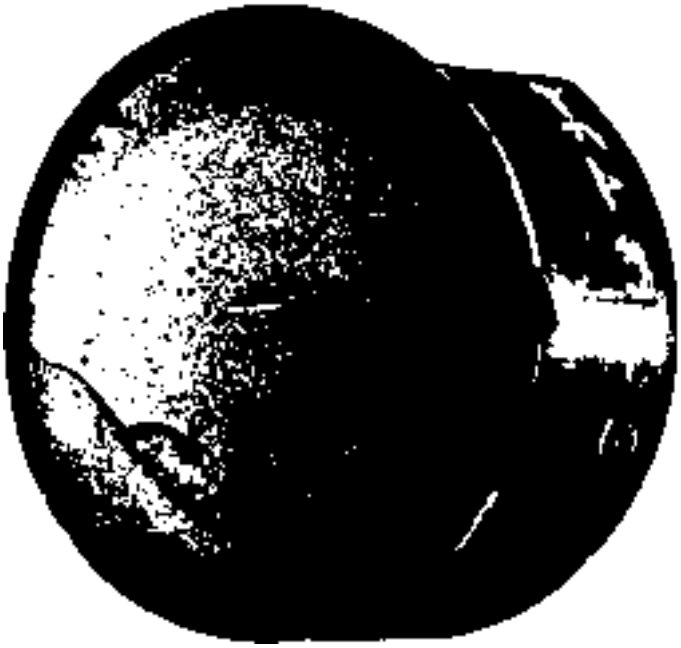


Fig. 22 Steel shaft from a tube-bending machine that failed by fatigue fracture. The fracture surface shows regions of fatigue-crack propagation and final fracture.

Visual examination of the fracture surface revealed both a smooth area and a coarse, granular area (Fig. 22). The dull, smooth area is typical of some fatigue fractures and resulted as the crack was opened and closed by the bending stress. Beach marks on the smooth area of the fracture surface also indicate fatigue fracture. The coarse, bright, crystalline-appearing area is the final-fracture zone. The smooth-textured fatigue zone is relatively large compared with the crystalline-textured final-fracture zone, which indicates that the shaft was subjected to a low overstress. The final-fracture surface at bottom shows that a one-way bending load was involved. The fatigue crack was initiated in a 0.25 mm (0.010 in.) radius fillet at a change in section. Cracking was nucleated by a nonmetallic inclusion that intersected the surface at a critical location in the fillet. The small radius of the fillet at the change in section resulted in a stress concentration that, in conjunction with the oxide-sulfide inclusion that intersected the surface of the fillet, initiated a crack. New shafts were made with a 2.4 mm (0.09 in.) radius fillet at the critical change in section.

Another example is shown in Fig. 23. The loading was mainly by unidirectional fatigue, but secondary origins (C and D in figure) suggest that a small reversed bending load or backlash may have been present.

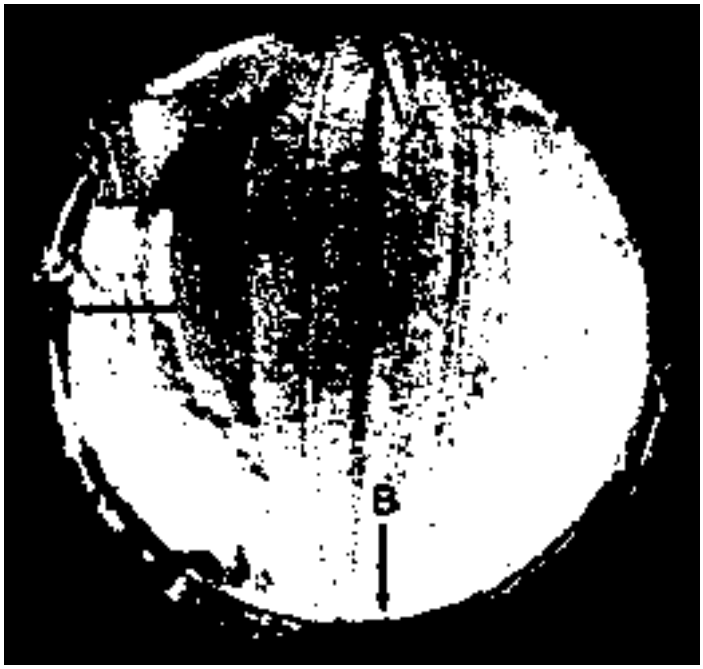


Fig. 23 Fatigue fracture of a steel bolt. Interpretation of the surface indicates that loading was primarily by unidirectional bending. However, secondary origins (C and D) indicate the possibility that a small reversed bending or backlash may have been present. Many closely spaced origins along the thread root (right side, between A and B) occurred from this region of high stress concentration. Nominal stress was low, resulting in small region of fast fracture (at E). $2\frac{1}{2}\times$

Alternating (Reversed) Bending

When the applied bending moment is reversing (alternating), the parties are subjected alternately to tension stress and compression stress; while the points on one side of the plane of bending are in tension, the points on the opposite side are in compression. If the bending moment is of the same magnitude in either direction, two cracks of approximately equal length usually develop from origins diametrically opposite each other and often in the same transverse plane. If the bending moment is greater in one direction than in the other, the two cracks will differ in length. In pure bending, the cracks on opposite sides of the beam or shaft are not necessarily in the same plane; therefore, final fracture may occur at some angle other than 90° to the axis of the beam. The size of the final-fracture zone is indicative of the relative loading of the beam.

There is one significant difference between the appearance of fatigue cracks formed under unidirectional load and those formed under alternating load. With an alternating load, each crack is opened in one half-cycle and closed and pressed together during the other half-cycle, which rubs and polishes the high points on opposite sides of the fatigue crack. Under unidirectional load, the rubbing is less pronounced. Rubbing may sometimes be sufficient to obliterate many of the characteristic marks, and the crack surfaces may become dull or polished. Consequently, beach marks are usually more readily observed after unidirectional loading than after alternating loading.

Figure 24 shows typical fatigue marks on the fracture surface of a stationary (nonrotating) shaft subjected to a reversing bending moment evenly distributed along its length. The crack origins (arrows) are shown diametrically opposite each other, but sometimes they are slightly displaced by minor stress raisers. The pattern shown in Fig. 24(a) is typical of that for a single-diameter shaft with no stress concentration. A large-radius fillet at a change in shaft diameter imposes a moderate stress concentration. Figure 24(b) shows the pattern on the surface of a fracture through such a fillet. A small-radius fillet at a change in diameter results in a severe stress concentration. Figure 24(c) shows the typical pattern on the surface of a fracture through a small-radius fillet. The reason for the fatigue-pattern changes is that the fatigue crack propagates faster in more severe stress concentrations at each end than in the interior.

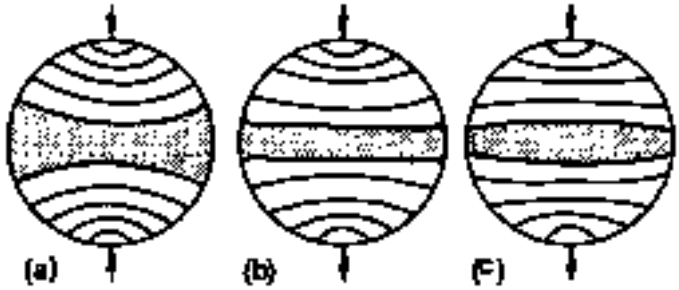


Fig. 24 Typical fatigue marks on the fracture surface of a uniformly loaded nonrotating shaft subjected to reversed-bending stresses. (a) No stress concentration. (b) Moderate stress concentration. (c) Severe stress concentration. Arrows indicate crack origins; shaded areas are final-fracture zones.

A fracture surface from reversed bending is shown in Fig. 25. Stress concentration was mild, indicated by several origins on both sides. Nominal stress was relatively large, indicated by the large region associated with fast fracture.

Fig. 25 Fatigue fracture of steel shaft (50 mm or 2 in., diam) from reversed bending, indicated by fracture initiation sites (arrows) on both sides. The multiple initiation sites indicate a mild amount of stress concentration. Fast fracture region is in area between brackets A.

Rotational Bending

A machine component that is commonly subjected to a bending load is a rotating round shaft. A unique feature of rotational-bending loading is that during one revolution of the shaft, both maximum and minimum loading are exerted around the entire circumference of the shaft in the region of maximum bending moment. Because the loading is axially symmetrical, a fatigue crack can be initiated at any point, or at several points, around the periphery of the shaft. Multiple cracks do not necessarily occur in the same plane and are separated by ridges called *ratchet marks.* The presence of multiple cracks normally indicates a relatively high applied load and rotational bending.

The essential difference between a stationary shaft and a rotating shaft subjected to the same bending moment is that in a stationary shaft the tensile stress is confined to a portion of the periphery only. In a rotating shaft, every point on the periphery sustains a tensile stress, then a compressive stress, once every revolution. Therefore, one difference arising from rotation is the distribution of multiple initiation sites at points on the periphery (Fig. 26b and 26d), depending on the

relative magnitude of the stresses at different locations determined by conditions of balance or imbalance imposed on the shaft. Under low or moderate overloads, a rotating shaft may fail as a result of a single fatigue crack.

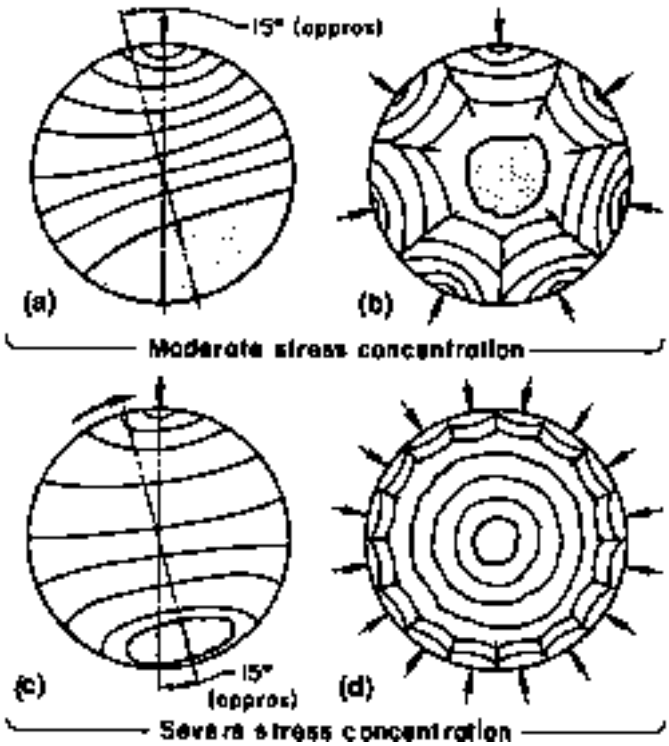


Fig. 26 Typical fatigue marks on the fracture surface of a uniformly loaded rotating shaft. Marks are produced from single and multiple origins (arrows) having moderate and severe stress concentration; shaded areas are final-fracture zones. Shaft rotation is clockwise.

Another important difference introduced by rotation is asymmetrical development of the crack front from a single origin. There is a marked tendency of the crack front to extend preferentially in a direction opposite to that of rotation, typically about 15° or more (Fig. 26a and c). When the shaft is always rotated in the same direction, the crack usually advances asymmetrically, with the beach marks shifted in a direction opposite to that of shaft rotation. Asymmetric fatigue-crack growth indicates rotational bending and the direction of rotation. Under periodically reversing rotating-bending loads, fatigue cracks grow symmetrically. Even without reversing the direction of rotation, a rotating-bending fatigue crack may advance symmetrically.

A common result of final fracture is that slight movement of one side of the crack relative to the other side frequently causes severe damage to the fracture surfaces and tends to obliterate many marks. However, although the high spots on one surface may rub the high spots on the other, the marks in the depressions are retained. Because the depressions are negative images of the damaged high spots on the opposing surface, they provide useful evidence; therefore, it is desirable to examine both parts of a cracked or fractured shaft.

The similarity in macroscopic appearance of fractures in shafts resulting from rotating-bending fatigue and from single overload torsional shear of a relatively ductile metal frequently results in misinterpretation. The fracture surface shown in Fig. 27(a) was the result of fatigue, as evidenced by the ratchet marks around the periphery and the pronounced beach marks. Under the low magnification of the fractograph shown in Fig. 27(b), beach marks are not visible because they were obliterated by rubbing. The presence of ratchet marks around the periphery is also an indication of rotating-bending fatigue. However, the metal smearing apparent on the fracture surface and the twisting deformation of the shaft shown in Fig. 27(c) indicate torsional shear and would preclude mistaking this fracture for a fatigue fracture. The fracture shown in Fig. 27(d) also exhibits a superficial similarity to a fatigue fracture. However, it is evident that this fracture was the result of torsional shear because the entire fracture surface has a smooth texture and no well-defined final-fracture area.

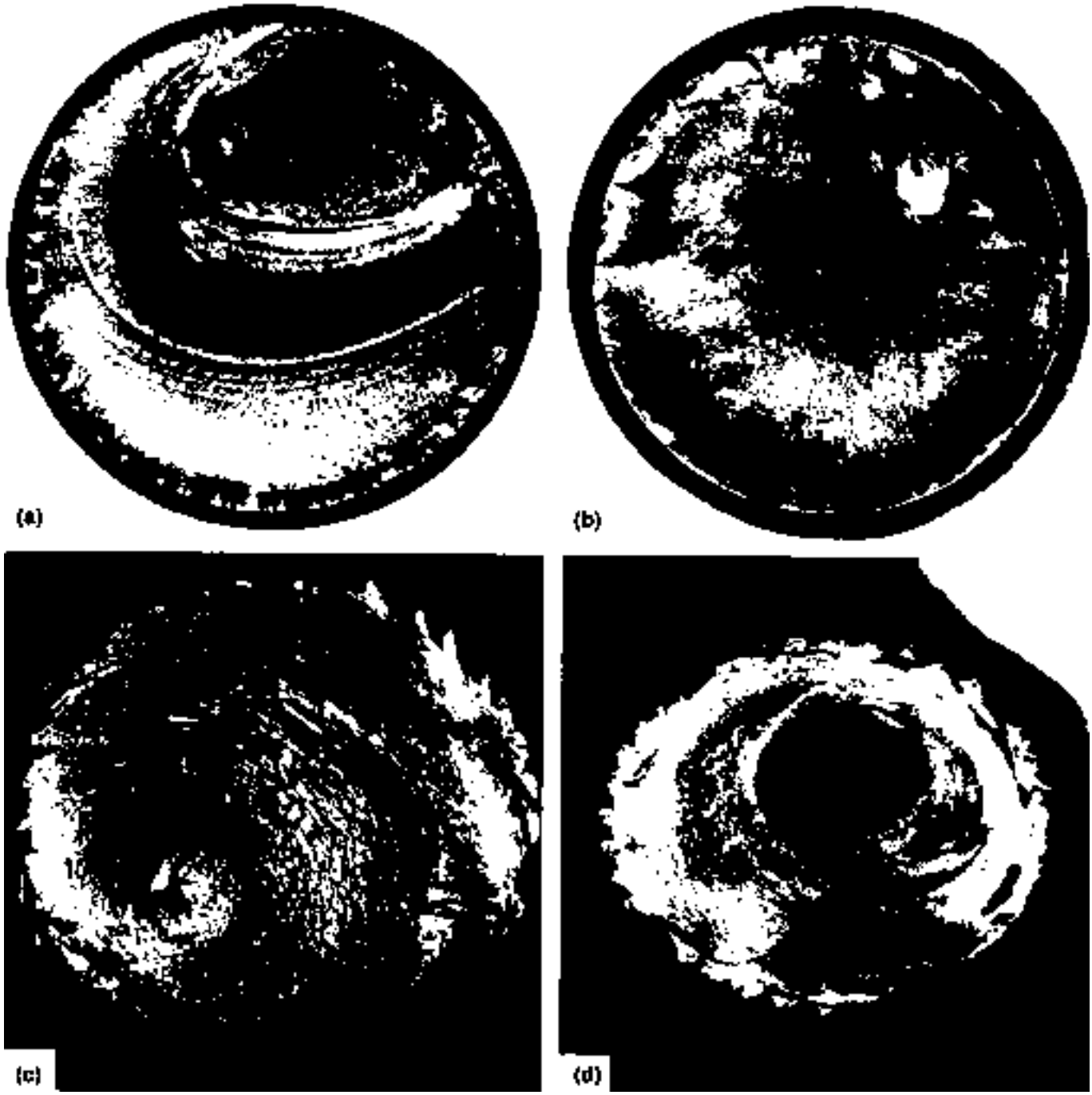


Fig. 27 Fracture surfaces of failed shafts. (a) and (b) Failure by fatigue. (c) and (d) Failure by torsional shear. See text for discussion.

In splined shafts, fracture resulting from torsional shear is frequently accompanied by deformation of the splines not engaged by the mating part. However, the portion of the shaft not engaged by the mating part is sometimes unavailable for examination. When macroscopic examination affords only inconclusive evidence, use of an electron microscope may reveal fatigue striations or elongated shear dimples. Also, metallographic examination of a section through the fracture surface may reveal microdeformation from torsional shear in the rotary direction that would not be present in a fatigue fracture.

Example 2: Failure of an Overhead Crane Drive Shaft Due to Rotating-Bending Fatigue*. A drive shaft used as part of the drive train for an overhead facilities crane fractured due to low-cycle rotating-bending fatigue, initiating and propagating by combined torsional and reverse bending stresses. The part submitted for examination was a principal drive shaft that fractured in the location of a 90° fillet where a wheel hub is machined down to 34.9 mm (1.375 in.) diam from a 60.3 mm (2.375 in.) diam main shaft. A 9.5 mm (0.375 in.) wide by 3.2 mm (0.125 in.) deep keyway was machined into the entire length of the hub, ending approximately 1.6 mm (0.062 in.) away from the 90° fillet (Fig. 28). The hub had fractured, separated from the main shaft, and lodged in the shaft hole of the wheel when the wheel fell away from the shaft. The wheel, together with the fractured hub and the shaft, were submitted for analysis.

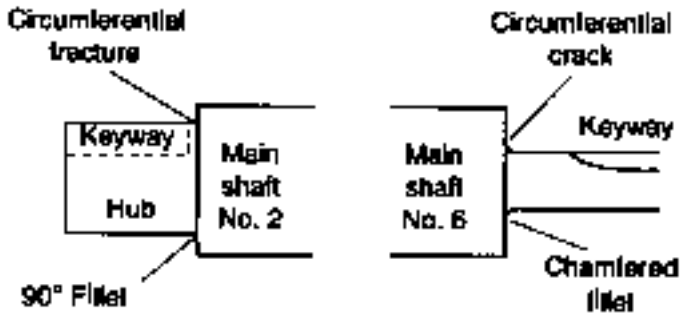


Fig. 28 Locations of observed fractures in a shaft assembly

Investigation. The pre-service and post-failure radiographic and penetrant inspection records were reviewed for the shaft and were found to be satisfactory. No additional cracks or indications were evidenced. The crack that led to the failure was not present when the shaft was installed. It was estimated that installation had been 1 to 2 years prior to the failure. The shaft material was verified as tempered 4150 steel by using wet chemical analysis and metallographic techniques. The hardness and the microstructure were surveyed by preparing metallographic specimens from the fractured hub section. The microstructure was confirmed to be a tempered martensite structure. The surface of the crack was examined at low and high magnifications using optical and scanning electron microscopy.

Two additional shafts had not fractured to failure, but were radiographically and penetrant inspected as part of a verification analysis to determine if the remaining drive shafts were serviceable. One of these shafts was found to have a radial penetrant indication, this time in the chamfered fillet in the 1-inch diam machined motor drive end of the shaft, which is the end opposite to that which fractured in the principal shaft (Fig. 28). This shaft was sectioned diametrically to expose a polished cross section of the indication. One quarter of the radial indication was broken open to expose the crack fracture features for observation in order to determine the mode of fracture. Because of the resultant similarity between this crack and the fracture in the principal shaft, no further forensic work, other than basic fractography, was performed on this second shaft.

Figure 29 shows a magnified view of the fractured hub that caused the wheel to break off the principal drive shaft of the overhead crane. The concentric circles revolving around the final overload, along with the ratchet marks around the outer circumference pointing toward the overload, are typical of a crack growth induced by rotating-bending stresses. The ratchet marks are exposing the fatigue initiation sites around the outer perimeter, as verified by high magnification scanning electron microscopy performed later in the investigation. The individual cracks propagated toward a single crack front and then toward the final overload area. Figure 29 shows the darker ductile dimples of the final overload region slightly off-center of the hub. The final overload was probably off-center due to the rotating-bending stresses that initiated the fracture on one side of the hub (most likely the side opposite the final overload), while at the same time driving the crack around the outer circumference producing multiple inward-growing crack fronts.

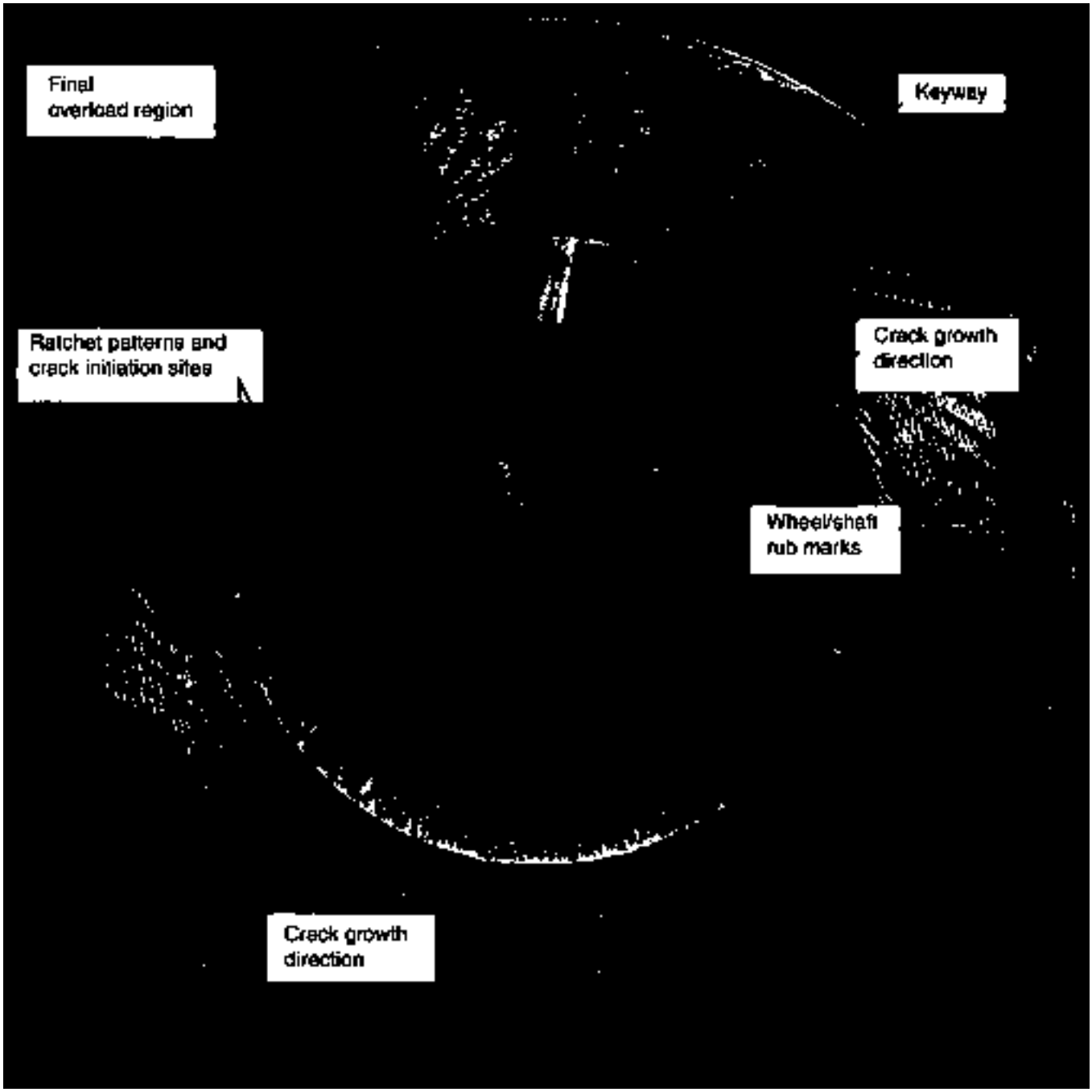


Fig. 29 Fracture section of steel shaft. Courtesy of J.R. Emmons, Rocketdyne Division, Rockwell International

Heavy corrosion was present traversing the fracture surface, extending for an inward distance of 3.2 mm (0.125 in.) from the outer edge, as identified in Fig. 29. The corrosion was confirmed to be rust. This corrosion may signify an older portion of the entire crack, which may have been growing for quite some time. It was originally suspected that the overall fracture had initiated at the keyway, which is usually the case in shafts where keyways are present, but the forensic analysis provided the contributing evidence that the fatigue cracks actually propagated toward the keyway rather than initiating from it.

Conclusions. A drive shaft used to assist in driving an overhead crane for the movement of medium to heavy weight components inside a manufacturing building, fractured at the mating interface with the drive wheel, in the location of a reduced-diameter wheel hub machined at a sharp 90° fillet from the main bulk shaft. The sharp 90° fillet co-existed with a keyway in the vicinity of the fracture, although the keyway did not appear responsible for the failure. A second shaft was cracked in a sharply chamfered region where the main shaft was machined to a reduced diameter to form the motor drive hub opposite the wheel hub end of the shaft. The fracture mode for both shafts was low-cycle rotating-bending fatigue initiating and propagating by combined torsional and reverse bending stresses. All drive shafts were replaced with new designs that eliminated the sharp 90° chamfers in favor of a more liberal chamfer, which reduced the stress concentration in these areas.

Torsional Loading

Under torsional loading of a shaft, the maximum local tensile stresses are at 45° to the axis of the shaft. Under alternating (reversed) torsional loading (S_m = 0, R = -1), two sets of fatigue cracks, perpendicular to each other, may develop. Torsional fatigue cracks can begin in longitudinal shear (Fig. 30a) or transverse shear (Fig. 30b); the relatively equal length of the cracks in each pair in Fig. 30 indicates that the equal and opposite stresses have occurred during loading. At later stages of fatigue-crack growth, one crack of a pair usually grows much faster than the other and eventually causes rupture of the shaft. Under a fluctuating torsional load ($S_m < S_a$, $R > 0$) fatigue cracks may develop normal to the tensile stresses and typically show fatigue cracks in only one direction; the presence of perpendicular fatigue cracks in a component subjected to fluctuating torsion probably indicates the presence of torsional vibration.

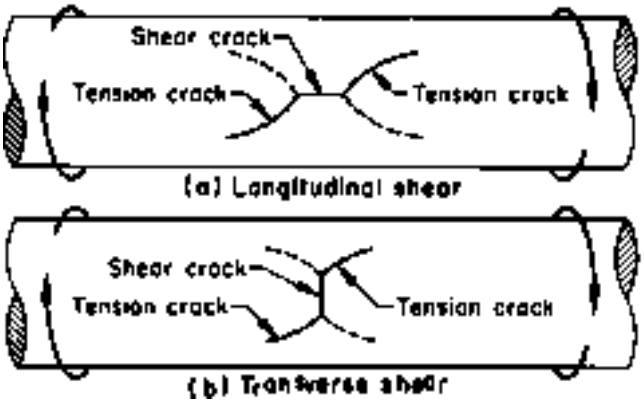


Fig. 30 Schematic of the initiation of torsional-fatigue cracks in shaft subjected to longitudinal shear (a) or transverse shear (b). Dashed lines indicate other cracks that can appear when torsional stresses are reversed.

If the shaft transmits a unidirectional torque but two cracks develop mutually at right angles, it can be presumed that the torque was of a reversing character. The relative extent of development of two torsional-fatigue cracks mutually at right angles can indicate the magnitude of the torque reversals that have been applied. If the cracks are of approximately the same length, the indications are that the torque reversals have been of equal magnitude, but only if the cracks are in a comparatively early stage of development. Beyond this stage, one crack usually takes the lead, and such inferences are no longer justified.

If a bending stress is applied to a shaft that is transmitting torque, the angle at which any fatigue crack develops is modified. In pure torsion loading, the fracture surface is oriented at 45° relative to the shaft axis, as shown in Fig. 31 from a test specimen in a torsional fatigue test. If the angle differs significantly from 45° to the shaft axis, the presence of a bending stress is indicated. Fatigue cracks arising from torsional stresses also show beach marks and ridges.

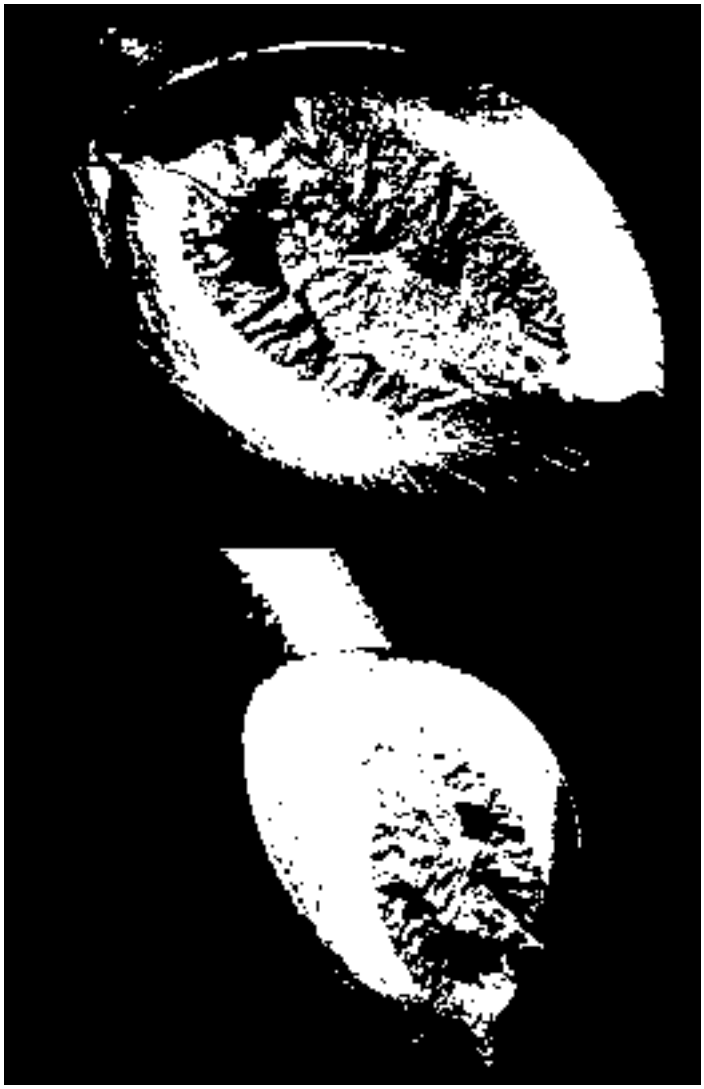


Fig. 31 Fracture surfaces of a torsional fatigue-test specimen. Courtesy of Greg Fett, Dana Corporation

Longitudinal stress raisers are comparatively harmless under bending stresses, but are as important as circumferential stress raisers under torsional loading. This sensitivity of shafts loaded in torsion to longitudinal stress raisers is of considerable practical importance because inclusions in the shaft material are almost always parallel to the axis of rotation. It is not unusual for a torsional-fatigue crack to originate at a longitudinal inclusion, a surface mark, or a spline or keyway corner and then to branch at about 45°.

When a stress raiser such as a circumferential groove is present, different states of stresses exist around the stress raiser, and the tensile stress is increased to as much as four times the shear stress. Therefore, the tensile stress on the 45° plane will exceed the tensile strength of the steel before the shear stress reaches the shear strength of the steel. Fracture occurs normal to the 45° tensile plane, producing a conical or star-shaped fracture surface (Fig. 32b and c)

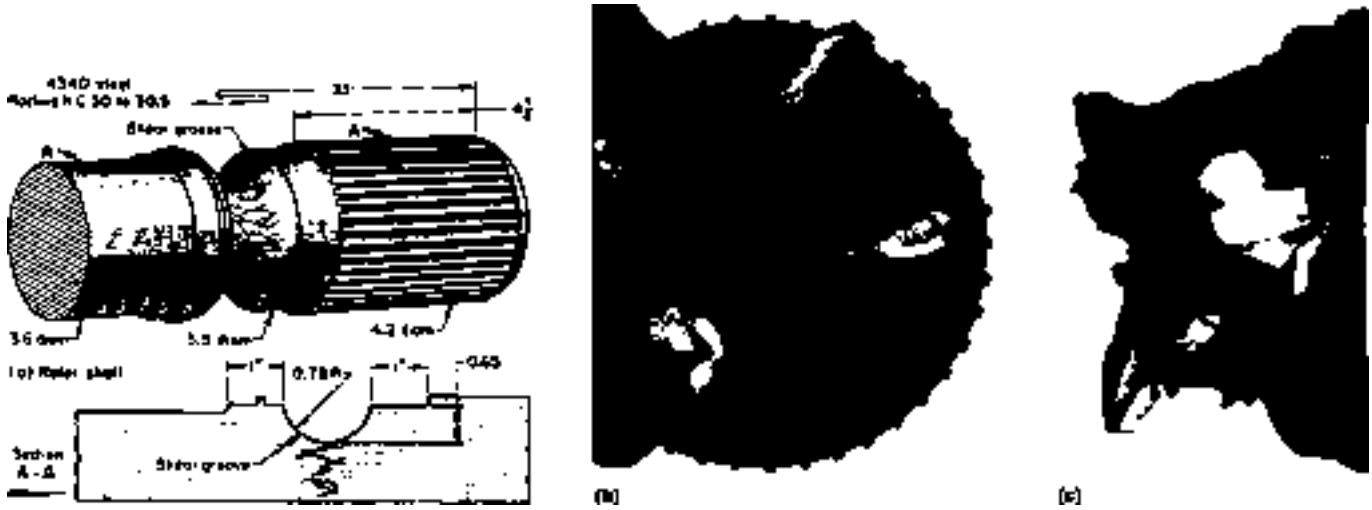


Fig. 32 4340 steel rotor shaft that failed by torsional fatigue. (a) Shear groove designed to protect gear mechanism from sudden overload. Dimensions are in inches. (b) Star-shaped pattern on a fracture surface of the shaft. (c) Longitudinal and transverse shear cracks on the surface of the shear groove, which resulted from high peak loads caused by chatter

Example 3: Torsional-Fatigue Fracture of a Large 4340 Steel Shaft That Was Subject to Cyclic Loading and Frequent Overloads. The 4340 steel shaft shown in Fig. 32(a), which was the driving member of a large rotor subject to cyclic loading and frequent overloads, broke after three weeks of operation. The shaft was also part of a gear train that reduced the rotational speed of the driven member. The driving shaft contained a shear groove at which the shaft should break if a sudden high overload occurred, thus preventing damage to an expensive gear mechanism. The rotor was subjected to severe chatter, which was an abnormal condition resulting from a series of continuous small overloads at a frequency of about three per second.

Investigation. Examination disclosed that the shaft had broken at the shear groove and that the fracture surface contained a star-shaped pattern (Fig. 32b). Figure 32(c) shows the fracture surface with the pieces fitted back in place. The pieces were all nearly the same size and shape, and there were indications of fatigue, cleavage, and shear failure in approximately the same location on each piece. The cracks were oriented at approximately 45° to the axis of the shaft, which indicated that final fracture was caused by a tensile stress normal to the 45° plane and not by the longitudinal or transverse shear stress that had been expected to cause an overload failure.

Examination of the surfaces of one of the pieces of the broken shaft revealed small longitudinal and transverse shear cracks at the smallest diameter of the shear groove. Also, slight plastic flow had occurred in the metal adjacent to these cracks. Cracking occurred at many points in the groove in the shaft before several of the cracks grew to a critical size.

No surface irregularities were present in the shear groove at any of the shear cracks. The structure of the metal was normal, with a uniform hardness of 30 to 30.5 HRC across the section, indicating a strength in the expected range for quenched-and-tempered 4340 steel shafts. A hot-acid etch showed the steel to be free from pipe, segregation, or other irregularities.

Conclusions. The basic failure mechanism was fracture by torsional fatigue, which started at numerous surface shear cracks, both longitudinal and transverse, that developed in the periphery of the root of the shear groove. These shear cracks resulted from high peak loads caused by chatter. Stress concentrations developed in the regions of maximum shear, and fatigue cracks propagated in a direction perpendicular to the maximum tensile stress, thus forming the star pattern at 45° to the longitudinal axis of the shaft. The shear groove in the shaft, designed to prevent damage to the gear train, had performed its function, but at a lower overload level than intended.

Corrective Measures. The fatigue strength of the shaft was increased by shot peening the shear groove, and chatter in the machine was minimized.

Footnote

* *Submitted by J.R. Emmons, Rocketdyne Division, Rockwell International, Canoga Park, CA.

Stress Concentrations

Notches, grooves, holes, fillets, threads, keyways, and splines are common design features. To visualize the distribution of stress at a change in section size or shape, it is helpful to consider the part in terms of electricity flowing through a conductor of similar cross section. In diagrammatic form, stress flow can be represented as a series of parallel lines, where stress concentration is inversely proportional to the distance between the lines, as shown schematically in Fig. 33 for several types stress raisers typically found in parts. A single notch (Fig. 33j) introduces a considerably greater stress-concentration effect than does a continuous thread (Fig. 33k). This is because of the mutual relief afforded by adjacent threads. The stress concentration at the right of the arrow in Fig. 33(k) is very similar to that in the narrow collar in Fig. 33(h). To the left of the arrow, however, the last thread is relieved from one side only; consequently, there is more stress concentration, similar to that of the single notch in Fig. 33(j). This is why bolts so frequently fracture through the last full thread.

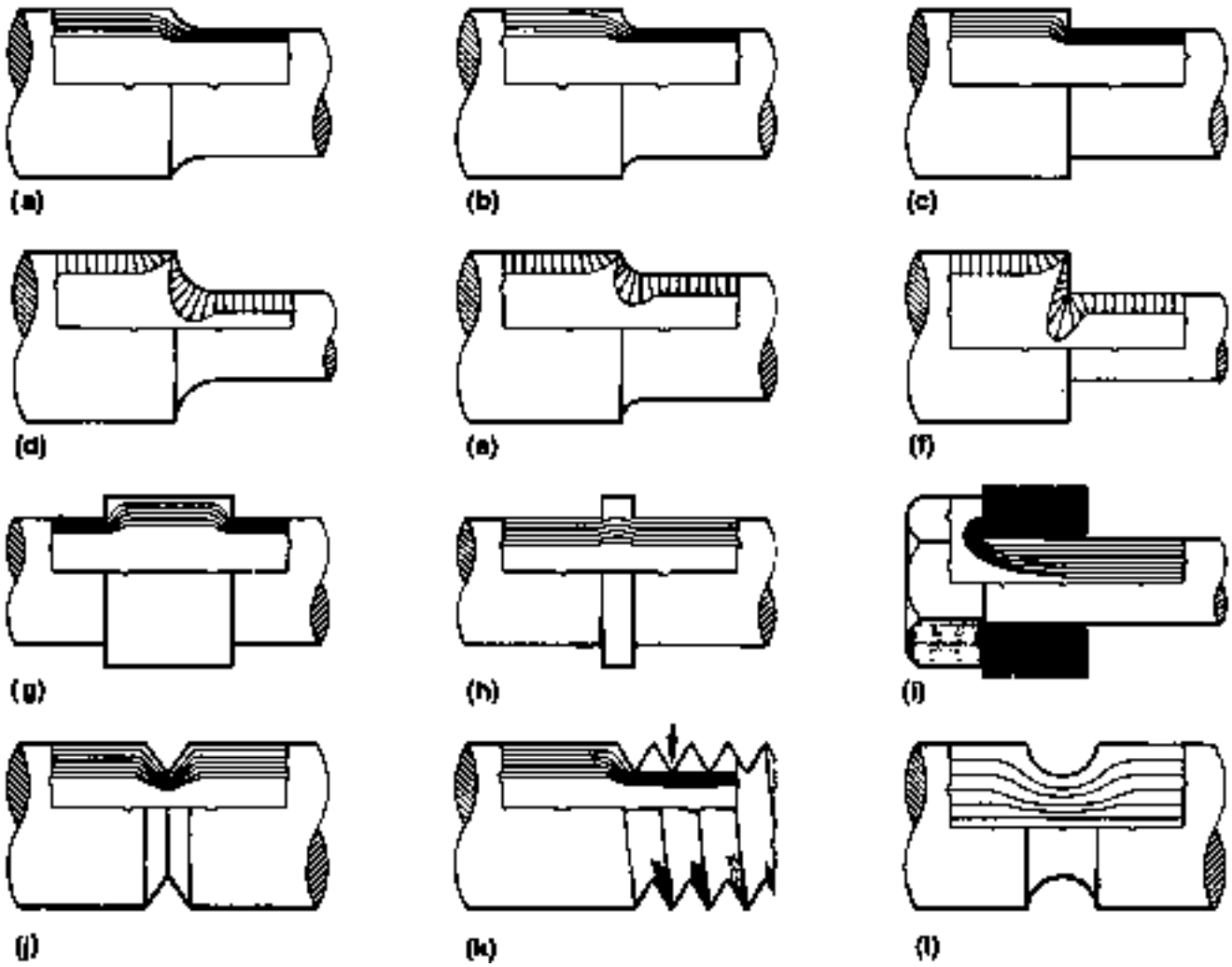


Fig. 33 Effect of stress raisers on stress concentration and distribution of stress at several changes of form in components. (a) to (c) Progressive increases in stress with decreasing fillet radii. (d) to (f) Relative magnitude and distribution of stress resulting from uniform loading. (g) Stress caused by the presence of an integral collar of considerable width. (h) Decrease in stress concentration that accompanies a decrease in collar width. (i) Stress flow at the junction of a bolt head and a shank. (j) Effect of a single sharp notch. (k) Effect of a continuous thread. (l) Effect of a groove or gauge.

All such sectional discontinuities increase the local stress level above that estimated on the basis of minimum cross-sectional area. In addition to the reduction of fatigue strength or fatigue life, increasing the severity of stress concentration has the following effects on fatigue-crack features:

- Initiation of multiple cracks is more likely.
- Beach marks usually become convex toward the point of crack origin.
- Under rotational loading, the beach marks may completely surround the final-fracture zone.
- Combined stress states may be introduced, thereby influencing the direction of crack growth.

In the absence of stress concentrations at the surface, cracks propagate more rapidly near the center of a section than at the surface. This occurs because deformation constraints cause the stresses to be triaxial and more severe away from the surface. However, when there is a stress-concentrating notch at the surface, such as a thread with a sharp root, the stress near this notch may be more severe than it is farther below the surface. Under conditions of severe notching, W-shape crack fronts will sometimes be observed. In regions exposed to high overstress from loading or in the presence of a severe stress concentration, multiple crack origins will be seen. In most cases, the cracks from these origins will eventually unite to form a single crack front. Before the single crack front is formed, the individual microcracks will be separated by small, vertical ledges called ratchet marks. Alternatively, at just above the fatigue limit, or minimum stress for fracture, a single origin will occur, and the entire fracture will emanate from that point.

Stress-Concentration Factor. Stress is concentrated in a metal by structural discontinuities, such as notches, holes, or scratches, which act as stress raisers. The stress-concentration factor, k_t, is the ratio of the area test stress in the region of the notch (or other stress concentrators) to the corresponding nominal stress. For determination of k_t, the greatest stress in the region of the notch is calculated from the theory of elasticity, or equivalent values are derived experimentally.

The stress-concentration factor, k_t, of the discontinuity is a measure of intensity of stress that occurs. In some situations, values of k_t can be calculated using this theory of elasticity or can be measured using photoelastic plastic models. Many of these values are reported in standard reference books (Ref 12, 13, 14, 15, 16). The mathematical theory of elasticity is based on an ideal isotropic material free of any internal discontinuity, a strictly accurate profile, and an increase in stress concentration due solely to the presence of the surface discontinuity. In actual parts, the stress intensification is affected not only by the surface discontinuity but also, to an undetermined extent, by the size of the part, by local readjustments of stress because of plastic yielding, by surface roughness, and by the heterogeneous structure of the material itself, including anisotropy and inherent internal discontinuities. Therefore, the deleterious effect of a stress raiser on a part is usually determined experimentally and expressed in terms of a fatigue notch factor, k_f.

Stress concentrations affect the fatigue behavior of different materials differently. For example, cast irons, containing innumerable internal stress raisers, show little further adverse effects from externally introduced stress raisers. In contrast, relatively brittle materials, such as quenched-and-tempered steels, are more susceptible to the effects of stress raisers than ductile materials, such as normalized or annealed steels. Under static loading conditions in a ductile material, a stress raiser has little or no effect in most situations. If the local stress exceeds the yield strength of the material, plastic deformation occurs and stress is redistributed. Provided the amount of plastic deformation is not excessive, no adverse effect need be anticipated. However, if the part is subjected to fluctuating or alternating stresses and if a fatigue crack nucleates at a stress below the yield strength of the material, stress redistribution by plastic deformation will not occur, and the material will be subject to the full effect of the stress raiser.

The fatigue-notch factor, k_f, is the ratio of the fatigue strength of a smooth (unnotched) specimen to the fatigue strength of a notched specimen at the same number of cycles. In general, experimentally determined values of k_f are somewhat less than the values of k_t calculated for the same specimens.

Fatigue-notch sensitivity, q, for a material is determined by comparing the fatigue-notch factor, K_f, and the stress-concentration factor, K_t, for a specimen of a given size containing a stress concentrator of a given shape and size. A common definition of fatigue-notch sensitivity is $q = (k_f - 1)/(k_t - 1)$, in which q may vary between 0 (where $k_f = 1$) and 1 (where $k_f = k_t$). This value may also be stated as a percentage.

Stress-Intensity Expression of Stress Concentration at a Sharp Notch. It can be shown that, for sharp notches, the maximum-stress range on this element can be related to the stress-intensity factor range, ΔK_I, as follows (Ref 9):

$$\Delta\sigma_{max} = \frac{2}{\sqrt{\pi}}\frac{\Delta K_I}{\sqrt{\rho}} = \Delta\sigma(k_t) \qquad \text{(Eq 12)}$$

where ρ is the notch-tip radius, Δσ is the range of applied nominal stress, and k_t is the stress-concentration factor. Fatigue-crack-initiation behavior of various steels is presented in Fig. 34 (Ref 9) for specimens subjected to zero-to-tension bending stress and containing a smooth notch that resulted in a stress-concentration factor of about 2.5. The data show that $\Delta K_I/\sqrt{r}$, and, therefore, Δσ is the primary parameter that governs fatigue-crack-initiation behavior in regions of stress concentration for a given steel tested in a benign environment. The data also indicate the existence of a fatigue-crack-initiation threshold, $(\Delta K_I/\sqrt{r})_{th}$, below which fatigue cracks would not initiate at the roots of the tested notches. The value of this threshold is characteristic of the steel and increases with increasing yield or tensile strength of the steel. The data show that the fatigue-crack-initiation life of a component subjected to a given nominal-stress range

increases with increasing strength. However, this difference in fatigue-crack-initiation life among various steels decreases with increasing stress-concentration factor (Ref 9).

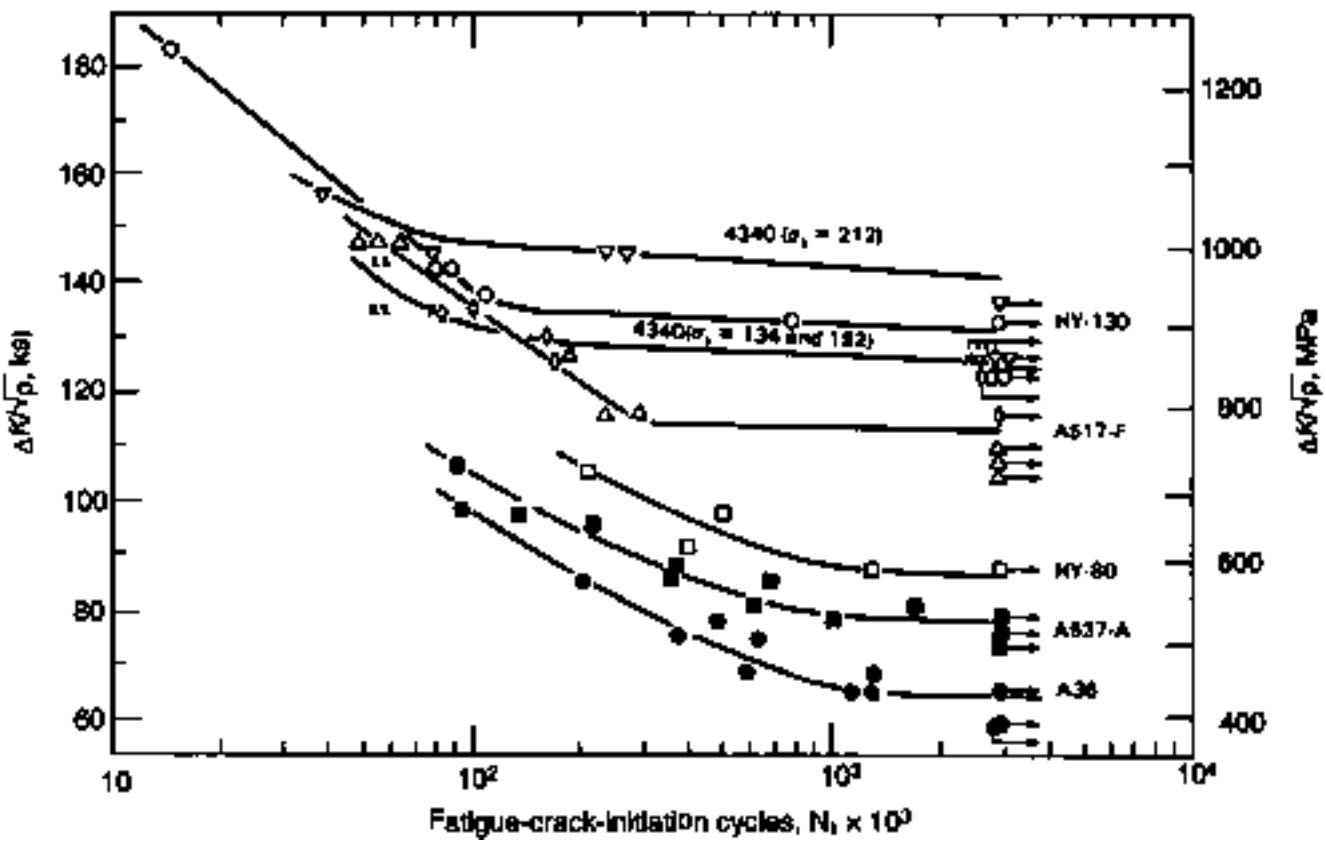


Fig. 34 Fatigue-crack-initiation behavior of various steels at a stress ratio of +0.1. Source: Ref 9

Fatigue-crack-initiation data for various steels subjected to stress ratios (ratio of nominal minimum applied stress to nominal maximum applied stress) ranging from -1.0 to +0.5 indicate that fatigue-crack-initiation life is governed by the total maximum stress (tension plus compression) range at the tip of the notch (Ref 17). The data presented in Fig. 35 (Ref 18) indicate that the fatigue-crack-initiation threshold, $(\Delta K_I/\sqrt{r})_{th}$, for various steels subjected to stress ratios ranging from -1.0 to +0.5, can be estimated from

$$\frac{\Delta K_{total}}{\sqrt{\rho}} = 10\sqrt{\sigma_{ys}} \qquad \text{(Eq 13)}$$

where ΔK_{total} is the stress-intensity-factor range calculated by using the tension-plus-compression stress range, and σ_{ys} is the yield strength of the material.

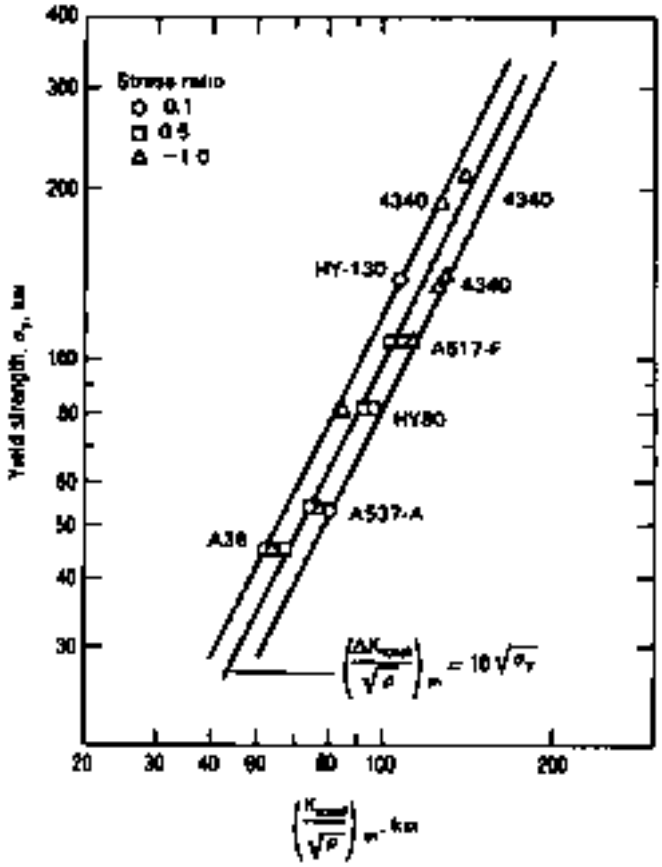


Fig. 35 Dependence of fatigue-crack-initiation threshold on yield strength

Reduction of Stresses. The detrimental effect of stress concentrations can be reduced or eliminated by the use of induced stresses. These stresses may be introduced by plastic deformation of the surface, such as in cold working and shot peening; by phase transformation, such as in case hardening; or by proof loading with a load high enough to cause local plastic flow in the notch but low enough to avoid general yielding.

Extreme or incorrect cold working may have adverse effects under fatigue conditions because it may give rise to minute cracks in the surface of the material, or at least make the surface prone to such cracks. If the correct amount of cold working is applied, however, improvement in the fatigue strength is appreciable. For example, a screw thread produced by rolling is more resistant to fatigue failure than one that has been cut. Surface rolling and shot peening, especially of springs, are other examples of the successful application of cold working. Apart from the increase in fatigue strength of the surface of the material arising from work hardening, the stress distribution is modified considerably.

If the amount of cold working is so adjusted that the residual compressive stress in the part effectively reduces the applied peak tensile stress below the fatigue limit of the material, fatigue failure will not occur. It is difficult in practice, however, to ensure the correct amount of cold working. Cold working intended to improve the endurance of a part should be chosen with appropriate consideration of the affected stress system. Residual compressive surface stresses improve fatigue strength only under conditions involving bending and torsional stresses. Where the stress is purely tensile (although infrequent in real-life engineering applications), the fatigue strength may be reduced because the compressive stress at the surface is balanced by a region of tensile stress just below the surface. In tensile loading, the stress is uniformly distributed across the section and is augmented by the residual tensile stress, and failure is likely to take place by cracking just beneath the surface.

Other methods of reducing the effect of stress concentration are to raise the fatigue strength of the material in the region of high stress and to produce compressive residual stresses by flame hardening, carburizing, or nitriding. If performed correctly, any of the three processes is beneficial in raising the fatigue strength, but the extent of the treated region needs to be carefully chosen; otherwise, failure is likely to occur at one of the junctions between the treated and untreated regions.

In some cases, generous fillet radii can cause interference if the corner radii on the mating parts are too small. Therefore, a large corner radius would result in minimum mismatch at installation.

Residual Stresses. Fatigue fractures generally propagate from the surface. Processing operations, such as grinding, polishing, and machining, that work harden or increase residual stress on the surface can influence the fatigue strength, although there is no generalized formulation that will predict the extent of improved fatigue strength that can be derived from work hardening and residual stress. Compressive residual surface stresses generally increase the fatigue strength, but tensile residual surface stresses do not. There may be a gradual decrease in residual stress if the cyclic stresses

cause some plastic deformation. Compressive residual surface stress provides greater improvements in the fatigue strength of harder materials, such as alloy spring steel; in softer materials, such as low-carbon steel, work hardening effectively improves fatigue strength. This is because the harder material can sustain a high level of residual elastic surface stress, and the tensile strength (and thus the fatigue limit) of the softer material is improved by work hardening.

In a notched high-strength steel, the beneficial effect of prestretching and the detrimental effect of precompression are much greater than in a plain carbon steel because of the type of residual stress present at the notch. A compressive residual stress introduced during quenching from a tempering temperature will increase the fatigue strength, particularly in notched specimens.

In general, residual stresses are introduced by misfit of structural parts, a change in the specific volume of a metal accompanying phase changes, a change in shape following plastic deformation, or thermal stresses resulting from rapid temperature changes such as occur in quenching. The influence of residual stress on fatigue strength is, in principle, similar to that of an externally applied static stress. A static compressive surface stress increases the fatigue strength, and static tensile surface stress reduces it.

Complex Stresses. The criteria of static failure have been applied to fatigue failure. If S_1, S_2, and S_3 are the amplitudes of the principal alternating stresses where $S_1 \geq S_2 \geq S_3$, and if S is the alternating uniaxial fatigue strength, then the following criteria apply:

- Maximum principal-stress criterion $S_1 = S$
- Maximum shear-stress criterion $S_1 - S_3 = S$
- Shear-strain energy $(S_1 - S_2)^2 + (S_2 - S_3)^2 + (S_3 - S_1)^2 = S^2$
- Maximum principal-strain criterion $S_1 = \nu(S_2 + S_3) = S$, where ν is Poisson's ratio

Because fatigue cracks usually propagate from the surface, where one of the principal stresses is zero, only biaxial stresses need to be considered. When the two principal stresses are of the same algebraic sign, the criteria of maximum principal stress and maximum shear stress give the same relationship. If the principal stresses are of opposite algebraic sign, then all criteria give a different relationship. It is often convenient to determine the suitability of the various criteria by comparing the fatigue strength in torsion *t* and bending *b*.

Complex three-dimensional (multiaxial) combinations of stress and strain also occur in the vicinity of surfaces and notches. Basic definitions of multiaxial effective stresses and strains and differences between proportional and nonproportional loading are described in Ref 19, along with some basic correlations for multiaxial fatigue including mean stress effects, sequences of stress/strain amplitude or stress state, nonproportional loading and cycle counting, and HCF fatigue limits. Ref 19 also covers the formation and propagation of small cracks (on the order of several grain sizes in diameter, typically less than 1 mm, or 0.04 in., in length) in initially isotropic, ductile structural alloys.

References cited in this section

9. S.T. Rolfe and J.M. Barsom, *Fracture and Fatigue Control in Structures—Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

12. H. Neuber, “Theory of Notch Stresses: Principles for Exact Calculation of Strength With Reference to Structural Form and Material,” AEC-Tr-4547, Springer Publishing, 1958; available through NTIS, U.S. Department of Commerce.

13. R.E. Peterson, *Stress Concentration Design Factors,* John Wiley & Sons, 1974

14. W. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, 1997

15. W.C. Young, *Roark's Formulas for Stress and Strain,* 6th ed., McGraw-Hill, 1989

16. A. Blake, Practical Stress Analysis in Engineering Design, 2nd ed., Marcel Dekker, 1990.

17. M.E. Taylor and J.M. Barsom, Effect of Cyclic Frequency on the Corrosion-Fatigue Crack-Initiation Behavior of ASTM A 517 Grade F Steel, *Fracture Mechanics: Thirteenth Conf.,* STP 743, ASTM, 1981

18. R. Roberts, J.M. Barsom, J.W. Fisher, and S.T. Rolfe, “Fracture Mechanics for Bridge Design,” FHWA-RD-78-69, Federal Highway Administration, Office of Research and Development, July 1977

19. D.L. McDowell, Multiaxial Fatigue Strength, ASM Handbook, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 263–273

Effect of Load Frequency and Temperature

Exclusive of environmental effects, there are no distinguishing surface features of a fatigue fracture produced at high frequency that differentiate it from other types of fatigue fractures during visual or optical microscope examination. At best, when examined with an optical microscope, a fracture surface created at a high frequency will have a brittle appearance, showing mostly a platelike structure throughout the fatigue zone. Beach marks may or may not be present, depending on whether crack growth was steady or intermittent or if load variations occurred.

The frequency range of 500 to 10,000 cycles/min is generally used in a fatigue test. In this range, the fatigue strength of most materials, based on a given number of cycles to fracture, is little affected by frequency. In general, there is a slight decrease in fatigue strength with a decrease in frequency, because the fatigue limit may be related to the amount of plastic deformation that occurs during the stress cycle. For example, at high frequency there is less relaxation time during each stress cycle for deformation to occur, which results in less damage. For steel, the fatigue limit is not affected between 200 and 5000 cycles/min. However, at frequencies up to 100,000 cycles/min, steels that are 100% ferritic show a marked increase in fatigue strength. In nonferrous metals, the fatigue strength increases continuously with frequency; in plain carbon steel, the fatigue strength reaches a maximum value, then decreases with an increase in frequency.

Elevated Temperatures. Except at very high frequencies, the frequency at which cyclic loads are applied has little effect on the fatigue strength of most metals. The effect, however, becomes much greater as the temperature increases and creep becomes more of a factor. At high temperatures, the fatigue strength often depends on the total time the stress is applied, rather than solely on the number of cycles. This behavior occurs because of continuous deformation under load at high temperatures. Under fluctuating stress, the cyclic frequency affects both the fatigue life and the amount of creep.

Fractures resulting from fluctuating stresses at high temperatures may be similar in appearance to fatigue fractures, stress-rupture fractures, or a mixture of the two, depending on the relative magnitudes of the mean and alternating stresses. An important requirement for fatigue resistance at high temperatures is that the component must have resistance to oxidation and other forms of high-temperature corrosion. Fatigue strength at high temperature can be seriously reduced by surface attack from fuel ash containing vanadium pentoxide, from leaded fuels, and from other contaminants. In general, however, materials are less notch sensitive at high temperatures than at room temperature. Extended operation at high temperatures may result in metallurgical changes in the alloy structure, which may also play a role in reducing fatigue strength. Generally, however, brief exposure to high temperatures, which does not result in such metallurgical changes as recrystallization, tempering, phase changes, coarsening, precipitation, melting, or diffusion, will not have a serious effect on fatigue life upon return to normal operating temperatures.

Effect of Material Condition

Discontinuities within a metal, either at the surface or subsurface, can adversely affect fatigue strength. These discontinuities may arise from melting practices or primary or secondary working of the material, or may be a characteristic of a particular alloy system. Manufacturing practices may also introduce surface irregularities or induce residual stresses that act as crack initiation sites. In wrought metals, the movement of metal during these processes, whether performed at room temperature or at elevated temperatures, makes them common sources of surface discontinuities, such as laps, seams, and cold shuts. Oxides, slivers or chips of the base material, or foreign material can be embedded into the surface by rolling or forging. These surface imperfections produce a notch of unknown severity that acts as a stress raiser under load and adversely affects fatigue strength. Subsurface and core discontinuities can be troublesome, depending on their shape, extent, and location.

Strengthening and Heat Treatments

The microstructure of the material can influence this crack growth by inhibiting or modifying the plastic-deformation process. Under high cyclic strain (low-cycle fatigue), the fatigue life of many metals is independent of grain size. In contrast, under low cyclic strain (high-cycle fatigue), the fatigue life of many metals is increased when grain size is reduced. However, the effect of grain size on high-cycle fatigue properties is difficult to assess because these properties may be altered by the same treatments that alter grain size. In some alloys, an improvement in resistance to high-cycle fatigue brought about by a decrease in grain size may be partly offset by a deleterious effect on another property. For example, a decrease in grain size is thought to raise the smooth-bar fatigue limit in some steels. However, small grain size increases notch sensitivity in these steels; therefore, the net result may constitute no improvement in resistance to notched-bar fatigue. As another example, fine grain size in high-temperature alloys (which may be subjected to both fatigue and creep) may result in high fatigue life but low stress-rupture life at normal service temperatures; thus, an intermediate grain size may afford the longest service life.

The influence of chemical composition on fatigue strength is approximately proportional to its influence on tensile strength. The fatigue limit of plain carbon steel generally increases with carbon content. Molybdenum, chromium, and nickel have a similar effect. The fatigue limit of high-strength steel with tensile strength in the range of 1379 MPa (200 ksi) can be increased by the addition of copper. Although the phosphorus content of steel is generally kept at a minimum to prevent brittleness, steels with high phosphorus contents have greater fatigue strength. A sulfur content of 0.01% has no effect on fatigue limit. Austenitic steels containing nickel and chromium have a high fatigue limit, together with low notch sensitivity and high resistance to corrosion fatigue.

The fatigue strengths of titanium alloys are higher than those of steel. Some titanium alloys maintain considerable strength up to 500 °C (930 °F) and have high resistance to corrosion fatigue. Even though titanium alloys are particularly susceptible to hydrogen embrittlement, the presence of hydrogen does not affect fatigue properties. Aluminum alloys that are solid-solution strengthened show an increase in fatigue strength to about the same degree as the corresponding increase in tensile strength. If the strength of magnesium alloys is increased by solute addition, the fatigue strength is also increased in proportion to the increase in tensile strength.

Second phases that are often present in metallurgical systems affect crack propagation. Second phases have a marked influence on the mechanism and kinetics of crack nucleation and propagation because they can accelerate or inhibit the crack-propagation rate under various circumstances. For example, investigations on aluminum alloys containing a large number of particles of precipitated, intermetallic second phase showed that the particles acted as raisers from which fatigue cracks nucleated. The grain boundaries of age-hardened alloys are generally free of precipitate particles and consequently are relatively soft. Therefore, stress relaxation takes place along the boundaries, resulting in localized strain concentration, and crack nucleation, at grain-boundary triple points.

The instability of coherent precipitate particles is considered to be the most important factor responsible for the low fatigue life of high-strength aluminum alloys. Experimental results indicate that reversion takes place during the early stages of cyclic loading. Reversion under cyclic straining involves the passage of dislocations back and forth through the precipitate particles, which disintegrates the particles to subcritical size. The solute then goes back into solution or is distributed along the dislocation network, thereby softening the slip bands and bringing about crack nucleation.

In alloys that are not strengthened by heat treatment (some copper alloys, aluminum alloys, and stainless steels), fatigue strength can be increased by cold working. Work-hardened copper-aluminum alloys (75 wt% Al) do not soften under cyclic stresses. This results in lower rates of crack propagation in work-hardened alloys, thereby indicating a small amount of deformation at the crack tip during fatigue. Experimental results on annealed and cold-worked α-brass suggest that a work-hardened alloy is harder at the crack tip than an annealed alloy. However, no improvement in fatigue strength is obtained by cold working of age-hardenable alloys. Alloys that are hardened by transformation processes—for example, martensitic steels—exhibit a lesser degree of improvement by cold working than the nonhardenable alloys.

Overheating. In the heat treatment of metals, high temperatures generally cause large grains to develop; depending on the metal and application, large grains may have undesirable attributes. In most metals, large grains will generally reduce fatigue strength. The properties of coarse-grain metals are impaired by the size of the grains and by changes that occur at the grain boundaries, such as the precipitation of solid impurities, which may form weak, continuous grain-boundary films. Damage from overheating is especially evident in high-carbon steels, in which all the harmful effects of coarse grains on properties are combined with an increased probability of cracking during quenching from the hardening temperature.

Eutectic Melting. In aluminum and other age-hardenable alloys, solution heat treatment improves mechanical properties by developing the maximum practical concentration of the hardening constituents in solid solution. The solubilities of these constituents increase markedly with temperature, especially just below the eutectic melting temperature. Consequently, the most favorable temperature for effecting maximum solution is very near that at which melting occurs. Melting in age-hardenable alloys produces an intergranular network of nonductile eutectic products and intragranular circular spots (rosettes), both of which reduce ductility and fatigue strength. Similar effects of melting have been observed in high-temperature alloys, such as the Stellites, and in high-speed tool steels.

Quench cracks are a frequent cause of failure in hardened steel parts. The origin of quench cracks in steel is attributed to sudden volume changes that occur in hardening. The transformation of austenite into martensite is always accompanied by expansion, which under unfavorable conditions, may result in cracking. Conditions that promote the formation of quench cracks are a quenching medium that is too severe, sharp edges and rough finishes, and hardening temperatures that are too high. Sometimes the cracks do not appear immediately but are delayed and take time to become visible. Delayed quench cracks are the result of additional transformation of retained austenite in steel. In highly alloyed steels, delayed cracking can be avoided by tempering immediately after quenching. Quench cracks are generally intergranular. If a quench crack is open to the surface during tempering, the walls of the crack may be covered with scale and may be decarburized.

Decarburization is a loss of carbon from the surface of a ferrous alloy as a result of heating in a medium that reacts with carbon. Unless special precautions are taken, the risk of losing carbon from the surface of steel is always present in any heating to high temperatures in an oxidizing atmosphere. A marked reduction in fatigue strength is noted in steels with decarburized surfaces. The effect of decarburization is much greater on high-tensile-strength steels than on steels with low tensile strength.

Subsurface Discontinuities

Gas Porosity or Shrinkage Porosity. Gas porosity or gas holes are rounded cavities (spherical, flattened, elongated, or partially collapsed) that are caused by the generation or accumulation of gas bubbles in the molten metal as it solidifies. Shrinkage cavities result from varying rates of contraction while the metal is changing from the molten to the solid state. Shrinkage porosity is mainly characterized by jagged holes or spongy areas lined with dendrites. The fatigue strength of cast alloys is only slightly reduced by the presence of shrinkage or gas porosity, but it is much reduced when the porosity extends to the surface, regardless of severity, because of the notch effect.

Inclusions. The presence of nonmetallic inclusions at or close to the surface is detrimental, because the inclusions act as stress raisers and form points for the initiation of fatigue cracks. The larger the number of discontinuities, the greater the possibility of fatigue failure. However, there does seem to be a limit to the number of inclusions that could cause fatigue. If discontinuities such as graphite flakes in cast iron or sulfide inclusions in free-cutting steels are very numerous, mutual stress relief occurs, based on the principle of a single groove versus a continuous thread, and the material tends to become less prone to fatigue.

Inclusions in ferrous alloys are usually oxides, sulfides, and silicates. Many inclusions, however, are of a more complex intermediate composition. Every commercial metal has its own characteristic inclusions. In cast metals, the shape of the inclusions is about the same in all polished sections; in wrought metals, the shape depends on the orientation of the polished surface. Deformation in mechanical working causes inclusions to deform plastically to elongated shapes and to appear in longitudinal sections as stringers or streaks, although in transverse section the shape is globular or flat. Hard, refractory, and very small inclusions, such as alumina inclusions in steel, are not deformed by mechanical work.

Fatigue properties of high-strength alloys are degraded by inclusions, with a more marked effect being apparent on the transverse fatigue properties of wrought alloys than on the longitudinal properties. In general, however, the effect of

inclusions depends on the size and shape of the inclusions, on their resistance to deformation and their orientation relative to the stress, and on the tensile strength of the alloy. Soft steels, for example, are much less affected by inclusions than are hard steels.

The fracture surface of a hardened-steel connecting rod is shown in Fig. 36; large inclusions that intersect the surface are indicated by arrows. Loads on this rod were transverse to the forging grain flow. Inclusions, like most discontinuities, are most damaging when they intersect the surface; however, subsurface inclusions can also be responsible for premature fatigue failure when the stress level at the inclusions is high enough to initiate a crack. Figure 37 shows the surface of a torsional-fatigue fracture that initiated at a large, subsurface nonmetallic inclusion in a hardened steel valve spring.

Fig. 36 Fracture surface of a hardened steel connecting rod. Arrows indicate large inclusions from which fatigue cracking initiated.

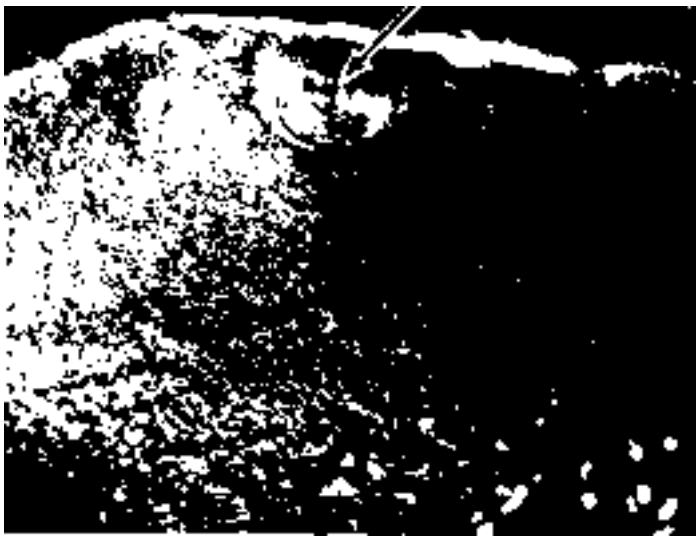


Fig. 37 Fracture surface of a hardened steel valve spring that failed in torsional fatigue. Arrow indicates fracture origin at a subsurface nonmetallic inclusion.

Internal bursts in rolled and forged metals result from the use of equipment that has insufficient capacity to work the metal throughout its cross section. If the working force is not sufficient, the outer layers of the metal will be deformed more than the inside metal, sometimes causing wholly internal, intergranular fissures that can act as stress concentrators, from which fatigue cracks may propagate under tensile, bending, or torsional loading.

Alloy Segregation. Distribution of alloying elements in industrial alloys is not always uniform. Localized deviations from the average composition originate from specific conditions during solidification of the alloys. Hot working and soaking tend to equalize the compositional differences, but these differences sometimes persist into the wrought product. Inhomogeneous distribution of alloying elements in heat-treated alloys is particularly objectionable, because it may lead to the formation of thermal cracks caused by uneven contraction or expansion in heating and cooling.

Banding. When segregation in an alloy occurs in layers, or bands, the alloy is said to have a banded structure. Banding can lead to discontinuities that can in turn cause premature fatigue failure. Figure 38 shows the fatigue fracture of a carburized-and-hardened steel roller. Banded alloy segregation in the metal used for the rollers resulted in heavy, banded retained austenite, particularly in the carburized case, after heat treatment. When the roller was subjected to service loads, the delayed transformation of the retained austenite to martensite caused microcracks near the case/core interface. These internal microcracks nucleated a fatigue fracture that progressed around the circumference of the roller, following the interface between case and core.

Fig. 38 Fracture surface of a carburized-and-hardened steel roller. As a result of banded alloy segregation, circumferential fatigue fracture initiated at a subsurface origin near the case/core interface (arrow).

Flakes are internal fissures in ferrous metals. They are attributed to stresses produced by localized transformation and decreased solubility of hydrogen during cooling after hot working. For alloy steels, flakes are often associated with high hydrogen content in the steel; chemical-element segregation, producing regions of high alloy content; and rapid cooling from the hot-working temperature.

Manufacturing Practices on Fatigue Strength

Machining. Most components subjected to fatigue conditions have been machined. Heavy cuts and residual tool marks from rough machining can promote fatigue failure in a component. Surface irregularities produced by rough machining act as stress raisers. However, a series of parallel grooves, such as results from turning operations, is less severe in its effect than an isolated groove because parallel grooves provide mutual stress relief. Rough machining also damages the metal to an appreciable depth. A machining tool shears off metal, rather than cutting it, with the result that the surface is torn and is work hardened to an extent that depends largely on the depth of cut, the type and shape of the tool, and the characteristics of the metal. Components intended to withstand fatigue conditions should be finished with a fine cut or, preferably, ground, and the direction of the final cut or grind should be parallel to that of the principal tensile load whenever practical.

Most mechanically finished metal parts have a shallow surface layer in residual compression. Apart from the effect it has on surface roughness, the final finishing process will be beneficial to fatigue life when it increases the depth and intensity of the compressively stressed layer and will be detrimental when it decreases or removes the layer. Processes such as electrolytic polishing and chemical and electrochemical machining, which remove metal without plastic deformation, may reduce fatigue properties. Electric discharge machining can be detrimental to fatigue properties without proper control and subsequent processing because of surface and subsurface microstructural changes. Improperly controlled grinding can have similar effects on fatigue properties.

Occasionally, during machining of a component, a tool may scratch or groove the surface. If the part is highly stressed in service, the result can be a premature fatigue fracture nucleating at the tool mark. Stress raisers frequently occur at a change in section, such as the shoulder between two shaft sections of different diameters. Rough-machining marks, as well as steps resulting from improper blending of fillets with shaft surfaces, frequently serve as initiation sites for fatigue cracks.

Drilling. The fatigue strength of components can be reduced merely by the presence of a drilled hole; it is further reduced by failure to remove burrs (incurred during drilling) from the hole edges. Fractures originating at drilled holes are common in complex parts containing internal, intersecting machined passages because of the difficulty and expense of providing adequate break-edge radii at such locations.

Grinding. Proper grinding practice produces a smooth surface that is essentially free of induced residual stresses or sites for the nucleation of fatigue cracks. However, abusive grinding, particularly in steels heat treated to high hardness, is a

common cause of reduced fatigue strength and failure that results from severe induced tensile stresses, intense localized heating, or both. Intense localized heating results in overtempering, formation of untempered martensite, burning, or formation of tight, shallow surface cracks usually referred to as grinding cracks.

Straightening. Components may be unintentionally plastically deformed during manufacture, during shipping, or in service without cracking. These parts can be straightened manually, by heating, in presses, or in roll straighteners. The initial deformation and subsequent working operations can introduce residual stresses or stress raisers. Nicks, scratches, or locally work-hardened surfaces are potential stress raisers. Aircraft propeller blades are frequently damaged by deformation in minor ground accidents and repaired by cold straightening, either manually or in a press. The straightening operations are restricted to bends of less than a specified angle, varying with the location along the blade.

Plating on metal surfaces can be detrimental to fatigue strength of the plated parts. Carbon and low-alloy steels, particularly steels of high hardness, are susceptible to hydrogen damage due to absorption of hydrogen during the plating cycle and the associated acid or alkaline cleaning cycles. A soft plating material, such as cadmium, may inhibit the escape of hydrogen from the steel and lead to hydrogen damage. However, the harmful effects of cadmium plating can be significantly minimized by following recommended procedures, which include plating at high current densities (6 to 8 A/dm^2, or 0.4 to 0.5 $A/in.^2$), followed by baking at about 190 °C (375 °F) for 8 to 24 h. A daily procedure for ensuring that the plating bath will give low hydrogen pickup is highly desirable. Apparatus for such a procedure is available in the form of an electronic-electrochemical instrument that measures the relative amount of hydrogen generated during electroplating. Hard plating materials are usually in a state of tensile stress after they have been deposited on the base metal. Cracks develop in the plating material, act as stress raisers, and produce cracking in the base metal. Several effective procedures for decreasing the residual tensile stress imposed by electroless nickel plating have also been developed, including the addition of saccharin and other organic compounds containing sulfur to the plating bath.

Cleaning is sometimes necessary for the removal of oil, grease, or other contaminants from the surface of a workpiece. For production work, vapor degreasing or a dip into a solvent or a simple chemical cleaning solution may be employed. Many alkaline solutions that will remove grease and oil from aluminum are not satisfactory for most cleaning purposes because they attack the surface. However, alkaline solutions that are suitably inhibited to eliminate the corrosive action can be used successfully. The cleaning treatment should be followed by thorough rinsing in clean water, then drying.

Magnetic-Particle Inspection. Localized regions of untempered martensite can be produced by arc burns that result when the probe used in magnetic-particle inspection touches the surface of the workpiece. For example, fatigue cracking of several knuckle pins occurred within about 1 month in service (Fig. 39). Examination showed that the fractures initiated at circular cracks that apparently existed on the pins before their installation. Micrographs of sections through the cracked areas of both unused and failed knuckle pins revealed a remelt zone and an area of untempered martensite within the region of the cracks. The local areas of overheating were attributed to electric-arc burning that caused circular cracks. The pins subsequently failed by fatigue fracture. The circular cracks were the result of arc burning, attributable to improper technique in magnetic-particle inspection. The conductor should be insulated to prevent arc burning.

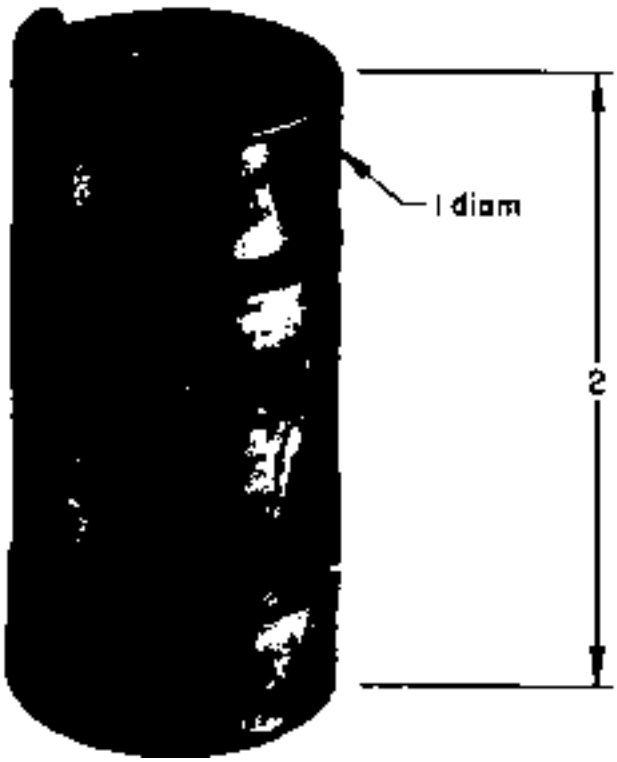


Fig. 39 Steel knuckle pin that failed in service by fatigue cracking that originated at an arc burn at the bottom of a longitudinal oil hole during magnetic-particle inspection

Identification Marking. Excessive stresses may be introduced into components by identification marks (manufacturing date or lot number, steel heat number, size or part number). The location and method of applying the marks can be important. They should not be in areas of high tensile, bending, or torsional stresses.

Raised numerals or letters are preferable to those that are indented, and hot forging methods are recommended over cold forging, stamping, or coining. For marks that are made on machined surfaces, use of a marking ink is generally preferable to use of an electric etching pencil or vibrating mechanical engraver, but either of the latter two is less abusive than using a cold steel stamp. However, using steel stamps for numbers in a nonstressed area may be preferable to some other type of identification mark in a more highly stressed region. Characters with straight-line portions have the greatest tendency to cause cracks, although characters with rounded contours can also cause cracking. If steel stamps must be used, low-stress stamps (stamps with the sharp edges removed from the characters) or dull stamps cause the least trouble, particularly if light impressions are made. The stamping should be located in known low-stress areas.

Corrosion Fatigue

Corrosion fatigue is associated with alternating or fluctuating stresses that occur in a corrosive environment and cause accelerated crack initiation and propagation at a location where neither the environment nor the stress acting alone would be sufficient to produce a crack. The corrosive environment usually introduces stress raisers on the surface. The irregular surface that results is detrimental to the fatigue properties of the part in a mechanical or geometric sense. For parts susceptible to embrittlement by hydrogen or for parts exposed to a fairly continuous corrosive environment with intermittent applications of loading, the cracking mechanism may be somewhat more complex. Although different alloys show differing performance under a given corrosion-fatigue environment, it is customary to protect the surface to achieve adequate performance at low cost.

An important feature of corrosion fatigue is that the stress range required to cause fracture diminishes progressively as the time and number of stress cycles increase. For example, the fatigue limit of mild steel is eliminated in aqueous environments, depending on the electrochemical potential. It is, therefore, impractical and uneconomical to attempt solely to design against corrosion fatigue. The emphasis in corrosion fatigue testing is to measure rates of fatigue-crack propagation in different environments rather than to specify the total life to failure.

Appearances

Fatigue cracking in a corrosive environment is identified by the presence of numerous small cracks adjacent to the fracture and of compacted corrosion product on the fracture surface, which may damage and obscure fine surface detail of the fracture. Corrosion-fatigue cracks generally initiate at the surface unless there are near-surface defects that act as stress-concentration sites and facilitate subsurface-crack initiation. Surface features at origins of corrosion-fatigue cracks vary with the alloy and with specific environmental conditions. In carbon steels, cracks often originate at hemispherical corrosion pits and often contain significant amounts of corrosion products. The cracks are often transgranular and may exhibit a slight amount of branching. Surface pitting is not a prerequisite for corrosion-fatigue cracking of carbon steels, nor is the transgranular fracture path; corrosion-fatigue cracks sometimes occur in the absence of pits and follow grain boundaries or prior-austenite grain boundaries. Crack-propagation paths in steels may be intergranular or transgranular (including quasi-cleavage) depending on the frequency, temperature, and environment. Extensive corrosion often damages fracture surfaces and can make positive identification of the fracture mechanism difficult. At low frequencies and low temperatures, fatigue-crack growth in hydrogen, for example, follows primarily prior-austenite grain boundaries, although some quasi-cleavage fracture is also observed. At high frequencies and temperatures, crack growth tends to follow a transgranular path.

In aluminum alloys exposed to aqueous chloride solutions, corrosion-fatigue cracks frequently originate at sites of pitting or intergranular corrosion. Initial crack propagation is normal to the axis of principal stress. This is contrary to the behavior of fatigue cracks initiated in dry air, where initial growth follows crystallographic planes. Initial corrosion-fatigue cracking normal to the axis of principal stress also occurs in aluminum alloys exposed to humid air, but pitting is not a requisite for crack initiation. Corrosion-fatigue cracking susceptibility in the aluminum alloys also does not appear to be sensitive to grain directionality and grain shape. This is in contrast to the propagation of stress-corrosion cracks in susceptible wrought aluminum alloys, which shows greater susceptibility in the short-transverse direction.

Corrosion-fatigue cracks in copper and various copper alloys initiate and propagate intergranularly. Corrosive environments have little additional effect on the fatigue life of pure copper over that observed in air, although they change the fatigue-crack path from transgranular to intergranular. Alternatively, copper-zinc and copper-aluminum alloys exhibit a marked reduction in fatigue resistance, particularly in aqueous chloride solutions. This type of failure is difficult to distinguish from SCC except that it may occur in environments that normally do not cause failures under static stress, such as sodium chloride or sodium sulfate solutions.

Both corrosion fatigue and mechanical fatigue may exhibit multiple origins that ultimately join to form a single crack front on a single plane. However, it is sometimes possible to distinguish corrosion fatigue from purely mechanical fatigue by the number of cracks propagating through the part. Frequently, several corrosion-fatigue cracks form and propagate simultaneously along parallel paths. On the other hand, mechanical-fatigue cracks may initiate at several points in the same general region on a part, but one crack usually becomes dominant, or several cracks join to form a single front, before cracking has progressed very far into the part. Environmental effects can usually be identified by the presence of corrosion damage or corrosion products on fracture surfaces or within growing cracks. Corrosion products, however, may not always be present. For example, corrosion-fatigue cracking of high-strength steel exposed to a hydrogen-producing gas, such as water vapor, may be difficult to differentiate from some other forms of hydrogen damage, such as hydrogen-induced slow-strain-rate embrittlement and sustained-load cracking. At sufficiently high frequencies, the fracture-surface features produced by corrosion-fatigue crack initiation and propagation do not differ significantly from those produced by fatigue in nonaggressive environments.

Examination

Analysis of fracture-surface deposits can be vital to the identification of corrosion fatigue. The first step in the macroscopic investigation should be to strip the fracture surface with cellulose acetate tape or plastic replicating material to entrap any deposit for later microanalysis. Then, whether complex electron optics or simple photographic documentation is to be used, the piece should be cleaned and degreased in an ultrasonic cleaner containing reagent-grade acetone, or similar organic solvent, as the tank medium.

The main advantage of reagent-grade acetone is that it leaves a minimal amount of organic residue on the fracture surface yet removes almost all traces of the cellulose acetate left from stripping. Also, acetone is an excellent solvent for wet magnetic-particle suspensions and for most liquid penetrants, which are often used in the field for nondestructive detection of cracks. If the part is too large or otherwise not amenable to ultrasonic cleaning, a solvent may be sprayed onto the surface of the part using a squeeze bottle or an aerosol can; the solvent should not be applied with a brush.

Features at the origin of a corrosion-fatigue fracture are often indistinct because the compression portion of each stress cycle has forced mating fracture surfaces together and has formed an extremely rubbed, discolored origin. Also, the area of the origin is exposed to the environment for the longest time and thus may exhibit more extensive residues of a corrosion product than the rest of the fracture surface. A corrosion product at fatigue origins can be misleading when viewed macroscopically. The oxide may be flat and tenacious, globular, or nodular. At magnifications of less than 60×, globular or nodular oxide particles are easily mistaken for intergranular facets; therefore, no conclusions should be drawn until other metallographic or fractographic

techniques confirm the mechanism of crack initiation. Occasionally, oxidation is so severe that no information other than location of origin can be obtained.

Other features that can be observed macroscopically are secondary cracks, pits, and fissures, all of which are often adjacent to the main origin of a particular fracture. In corrosion-fatigue failures, cracks and fissures adjacent to the primary origin are indicative of a uniform state of stress at that location. Therefore, the primary fracture simply propagated from the flaw that either induced the most severe stress concentration or was exposed to the most aggressive local environment. Sometimes the primary crack origin is related more to heterogeneous microstructure than to the stress distribution.

Contact Fatigue

Rolling-contact fatigue is a type of surface damage that results from repeated rolling, or rolling and sliding contact between curved metal surfaces. The significant stress in rolling-contact fatigue is the maximum alternating shear stress that undergoes a reversal in direction during rolling. In pure rolling, this shear stress occurs on a plane slightly below the surface and can lead to initiation of fatigue cracks in the subsurface material (see the article “Rolling Contact Fatigue” in this Volume). As these cracks propagate under the repeated loads, they reach the surface and produce pits.

When sliding is imposed on rolling, the tangential forces and thermal gradient caused by friction alter the magnitude and distribution of stresses in and below the contact area. The alternating shear stress increases in magnitude and is moved nearer to the surface by sliding forces. Thus, initiation of contact-fatigue cracks in gear teeth, which are subjected to significant amounts of sliding adjacent to the pitchline, is found to be in the surface material. These cracks propagate at a shallow angle to the surface, and pits result when the cracks are connected to the surface by secondary cracks. If pitting is severe, the bending strength of the tooth may be decreased to the point at which fracturing can occur.

The existence of several distinctly different modes of contact-fatigue damage has been recognized, but the factors that control the nucleation and propagation of each type are only partially understood. The contact fatigue of hardened steel has been the subject of extensive research using analytical and experimental methods, and publications have revealed a large variation in the contact fatigue strength of similar materials when tested under different conditions of contact geometry, speed, lubrication, temperature, and sliding in combination with rolling. These variations result from a poor understanding of the basic mechanism of contact fatigue. Although the cause of the fatigue and the locations of the origins are somewhat different, these are true fatigue fractures that are caused by shear-stress origins due to contact stresses between two metal surfaces.

In addition to an incomplete understanding of the basic mechanism, the bearing and gear industries have, in the past, used completely different nomenclature to describe identical fatigue damage modes. In 1996, American National Standard ANSI/AGMA 1010-E96 was published after several years' work to standardize nomenclature for failure modes that is compatible with both gear and bearing industries. Reference 20 also uses nomenclature for bearings that is similar to that used in ANSI/AGMA 1010-E96 for gears. In order to improve communication, it is recommended to maintain compatible nomenclature per ANSI/AGMA 1010-E96 such that:

- The term *macropitting* should be used instead of *spalling. Spalling* has been used in the past to describe several different failure modes and therefore is confusing. It should be avoided where possible.
- The term *micropitting* should be used instead of *peeling. Peeling* is not a good name for micropitting because it does not describe either the appearance or the mechanism of the failure mode. *Micropitting* is a much better name because it describes both appearance and mechanism and links the failure mode to macropitting, which is similar except for scale.

Table 1 lists terminology used by the bearing and gearing industries to describe the fracture appearance of contact-fatigue failed components. The following sections describe the types of damage from contact fatigue based on the categories in Table 1 as previously published in Ref 21. However, general prevention and characteristics are forms of damage that represent the limit of load-carrying capacity with the following origins:

- *Surface-origin pitting:* This is the type of pitting characteristic of combined rolling and sliding, such as occurs on gear teeth slightly away from the exact pitch line, where only rolling is present. Also, cam and cam-follower assemblies suffer this type of failure on the cam follower itself. The maximum shear stress is no longer buried within the metal, but is now brought to the surface because of the friction at the interface. Because the character of the surface is intimately involved, the lubricant becomes critical. It is for this reason that transmissions, final drives, and various other types of gear systems require lubricants with special additives. Everything possible must

be done to reduce surface friction, expand the contact area, and reduce the load in order to reduce the contact stress and provide a hard, fatigue-resistant structure in the metal. Unfortunately, if those things are already near optimum, little can be done to improve the fatigue life. Only minimal improvements, if any, can be expected. Improvements can best be obtained, if at all, by changes in the lubricant.

- *Subsurface-origin pitting:* Subsurface-origin pitting is the type of pitting that destroys the integrity of contacting surfaces where one surface rolls under pressure across another. This action is typical of antifriction bearings, such as ball, roller, tapered roller, and needle bearings, as well as other rolling devices such as roller cams. It is prevented by eliminating the hard, brittle, angular inclusions. A principal application for vacuum-treated steels is the manufacture of antifriction, or rolling-element, bearings and other critical parts and is now common practice for such components.
- *Subcase fatigue:* The previous types of contact-stress fatigue result in small, often tiny, pits in the contacting surfaces. Subcase-origin fatigue is different because very large cavities may be formed, apparently in a very short time. The fatigue originates near the case/core interface where the shear stresses. Because the basic cause is inadequate metal strength in the subcase region, the logical solution should be to strengthen the metal there. This can be done by increasing the case depth and/or the core hardness. This must be done carefully, however, because the changes could lead to through-hardening and brittle fracture, particularly on gear teeth.

Table 1 List of contact fatigue terminology used to describe the different fatigue mechanisms

Macroscale	
Macropitting	**Pitting** **Initial pitting** **Destructive pitting** **Flaking** **Spalling** **Scabbing** **Shelling** **Fatigue wear**
Subcase fatigue	**Case crushing**
Microscale	
Micropitting	**Microspalling** **Frosting** **Glazing** **Gray staining** **Surface distress** **Peeling**

References cited in this section

20. C.A. Moyer, “Comparing Surface Failure Modes in Bearings and Gears: Appearances versus Mechanisms,” AGMA Tech. Paper, 91 FTM 6, 1991

21. R.S. Hyde, Contact Fatigue of Hardened Steel, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 691–703

Macropitting

Macropitting is a general term that includes spalling and other forms of macroscale damage (Table 1) caused by Hertzian contact fatigue. Macropitting of bearing raceways or gear teeth is generally due to contact fatigue, which occurs from localized plastic deformation, crack initiation, and finally macropitting from crack propagation in and near the contact surface.

Macropitting results from the subsurface growth of fatigue cracks, which may have a surface or subsurface origin. Figure 40 shows macropit formation by subsurface growth of surface-origin pitting on a gear tooth. The bottom of inclusion-origin macropits form first, and so forth. When circumstances cause surface-origin pitting, crack growth often occurs within and beyond the range of maximum shear stresses in the contact stress field (so-called *hydraulic-pressure propagation crack paths* that may be due to chemical reactions at the crack tip as well as to lubricant viscosity effects).

Fig. 40 Fractograph showing the advanced stages of macropitting fatigue for a helical pinion tooth

When fully developed, the craters exist at a depth comparable to that of the maximum alternating Hertzian shear stress. A macropit does not result from the gradual enlargement of a small cavity, but rather it results from the subsurface growth of a fatigue crack, which eventually separates from the main body of material. In the example of a gear damaged by crater formation due to surface-origin macropitting (Fig. 40), the craters are generally steep walled and essentially flat bottomed, as the walls and bottom are formed by fracture surfaces. The bottoms of macropits usually form first, and then the crack extends through the contact surface, creating the walls of the crater. However, there are circumstances in which macropits are initiated at or near the contact surface and the crack propagates through the Hertzian shear stress field, forming the walls and then the bottom of the crater (Ref 22).

Macropitting fatigue life is inherently statistical because defect severity and location are randomly distributed among macroscopically identical contact components. Two major classes of macropitting damage are distinguished according to the location of the initiating defect: subsurface- and surface-origin macropitting.

Subsurface-origin macropitting results from defects in the bulk material subjected to the Hertzian cyclic stress field. Subsurface-origin macropits are characterized by a smooth area parallel to the contact surface with a steep wall exit (inclined by more than 45° with respect to the contact surface). Sometimes the subsurface defect that caused the damage remains visible after the macropit forms. The defects that cause subsurface-origin macropits are inclusions and/or material microstructure alterations. The defects that most likely produce this type of macropit are located above the depth of maximum alternating Hertzian shear stress. Variables governing the life of a component with respect to this type of macropit are Hertzian shear stress level, material matrix fatigue resistance, defect severity, and defect location.

Inclusion-origin fatigue is normally initiated at nonmetallic inclusions below the contact surface and is the primary mode of contact fatigue damage in antifriction bearings. The initiation and propagation of fatigue cracks are the result of cyclic stresses that are locally intensified by the shape, size, and distribution of nonmetallic inclusions in steels. Oxide inclusions

from deoxidation, reoxidation, or refractory sources are the most frequently observed origins of inclusion-origin fatigue damage. Sulfide inclusions alone are rarely associated with contact fatigue cracks. The ductility of sulfide inclusions at hot working temperatures and a relatively low interfacial energy make them less effective stress concentrators than the harder, incoherent oxide inclusions. In addition, oxide inclusions are often present as stringers, or elongated aggregates of particles, which provide a much greater statistical probability for a point of stress concentration to be in an area of high contact stress.

When a nonmetallic inclusion as described above results in significant stress concentration, two situations are possible: (a) localized plastic flow occurs; and/or (b) a fatigue crack begins to propagate with no plastic flow as observed microscopically. If localized plastic deformation occurs, microstructural alterations begin to develop and the effective stress concentration of the inclusion is reduced. The probability of fatigue crack initiation/propagation is less than before the plastic flow occurred, so a macropit may not form. However, if a fatigue crack forms due to the inclusion, prior to any plastic deformation, the stress concentration is increased and crack propagation results in macropitting fatigue. This hypothesis includes the possibility of a fatigue crack originating from an inclusion that is associated with some sort of microstructural alteration, but also proposes that microstructural alterations are not a necessary step in the initiation/propagation of inclusion origin fatigue.

Early macroscopic propagation of inclusion-origin fatigue takes on a characteristic pattern. Many branching subsurface cracks are formed around the inclusion, with propagation often being most rapid perpendicular to the rolling direction, with a resulting elliptical shape for the incipient macropit. In some instances, the cracked material around the inclusion macropits early and further propagation is relatively slow, leaving a small macropit with very few subsurface cracks to cause further damage. However, when material has macropitted from the contact surface, the result may be debris bruising and denting of other contact areas with subsequent surface-origin macropitting.

Surface-origin macropitting is caused by defects in the immediate subsurface material subject to asperity scale cyclic stress fields and aggravated by surface tractive forces. These defects are either preexisting defects, such as nicks, dents, grinding furrows, surface discontinuities, and so forth, or microscale pits, which are described subsequently. The distinguishing features of a surface-origin macropit are in the entrance zone of the macropit. The entrance zone may exhibit a shallow-angle entry wall (inclined less than 30° to the contact surface), an arrowhead configuration (Fig. 41), the presence of a visible surface defect, or an association with a stress concentration due to design geometry. Surface-origin macropits have been classified by the nature of the defect: geometric stress concentration (GSC) and point surface origin (PSO).

Fig. 41 Gear with arrow-shaped surface-origin pit. Source: Ref 23

Geometric stress concentration macropits are distributed on the contact surface at the ends of line contact. When the contact geometry, deflection under applied loads, and alignment cause the contact stress to be higher at the end of line contact, fatigue occurs within a narrow band in which the contact stresses are more severe than those associated with inclusions. The GSC macropits can propagate more rapidly in either direction—parallel or transverse to the rolling direction. This mode of contact fatigue is more frequently observed in life tests accelerated due to overload conditions because contact geometry is designed to produce uniform loading under expected service loads.

Crowning of bearing rollers and gear teeth in line contact situations have been shown to be an effective method of prolonging life or preventing the GSC mode of macropitting. The fact that GSC macropits can occur simultaneously at both ends of a line contact indicates that this mode of contact fatigue is truly the result of end stress concentrations and not simply misalignment. However, misalignment can aggravate the situation by increasing end contact stresses. Early stages of GSC macropitting exhibit cracks extending to the surface, but it is not clear whether the first cracks always originate at the surface or slightly subsurface. However, the origins are rarely associated with a nonmetallic inclusion, but occasionally an inclusion at the end of line contact will serve as the nucleus for a GSC fatigue macropit.

Point surface origin macropits occur randomly distributed on the contact surface, similar to inclusion-type macropits. However, PSO macropits are different in two important aspects: first, there is no consistent association with nonmetallic inclusions; second, the origin is located at or near the contact surface, whereas with inclusion-origin macropits, the initiation site is located subsurface. These characteristics result in a distinct arrowhead appearance, in which the tip of the arrowhead is the origin and the advanced stages of the macropit develop in a fanned-out appearance with respect to the tip in the rolling direction. Figure 42 shows an example of a tapered roller-bearing cone that was damaged due to PSO macropitting.

Fig. 42 Fractograph showing the advanced stages of point-surface-origin macropitting fatigue of a bearing raceway in which the origin is still visible

References cited in this section

22. T.E. Tallian, *Failure Atlas for Hertz Contact Machine Elements,* American Society of Mechanical Engineers, 1992

23. D.G. McPherson and S. Rao, Mechanical Testing of Gears, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000, p 861

Micropitting

Micropitting is the preferred term for peeling fatigue, which is defined as microscale spalling fatigue. It is damage of rolling/sliding contact metal surfaces by the formation of a glazed (burnished) surface, asperity scale microcracks, and asperity scale micropit craters. Micropitting damage is a gradual type of surface fatigue damage that is a complex function of surface topography and its interaction with the lubricant. (Even under high contact loads, a thin film of lubricant between the contacting surfaces from elastohydrodynamic, or EHD, lubrication is present.) Cumulative plastic deformation first results in a surface with a mirrorlike reflection known as a glazed surface. The surface and immediate subsurface material of a glazed surface are heavily deformed and with continued cyclic stressing, the ductility of the material decreases and microcracks form. The cracks tend to propagate parallel to the contact surface at depths comparable to that of asperity scale shear stresses. A microcracked surface is not usually distinguishable from a glazed surface by the unaided eye, but with microscopy, microcracks opening to the surface may be visible in glazed areas. As the microcracks multiply and propagate, the surface becomes undermined (in asperity dimensions), and multiple microscopic pits form. A micropitted surface appears frosted to the unaided eye, with black spots representing the micropits. An example is shown in Fig. 43. If extensive, the micropitting is considered a failure.

Fig. 43 Gear with micropitting (frosting) failure. Source: Ref 23

The occurrence of micropitting is controlled by the EHD film-thickness-to-roughness ratio and surface microgeometry. In the presence of a sufficiently high EHD film-thickness-to-surface-roughness height ratio, micropitting does not occur because the film prevents high-contact microstress in asperity interactions. However, film thickness may be sufficient to prevent generalized micropitting, but not local micropitting in the vicinity of surface defects such as nicks and dents that locally depressurize and thereby thin the EHD film. Tribology features that affect micropitting are both the height and sharpness (RMS height and slope) of the asperities. There is also some evidence that the normal wear in boundary-lubricated gear contacts removes incipient surface-origin pitting, which is one reason gears may be less prone to macropitting from surface-origin macropitting under some circumstances.

Micropitting (or peeling) is characterized by the limited depth of cracking and propagation over areas rather than propagation in depth. Micropitting fatigue cracks usually propagate parallel to the surface at depths from 2.5 to 13 mm (0.1 to 0.5 mil) and rarely as deep as 25 mm (1 mil). The high-stress gradient that exists in the vicinity of asperities explains why such cracks do not propagate to greater surface depths of the macro-Hertzian contact stress field. In essence, micropitting removes the asperity ridges and relieves the high-stress gradient. An additional visual characteristic is that the micropitting occurs more or less in the rolling direction, following finishing lines.

Macropitting fatigue is the most hazardous consequence of micropitting (peeling). The principal operating factor influencing the rate of macropitting from micropits is tractive interfacial stress. When friction is high, macropits form rapidly from the micropits on the surface. If early macropitting does not terminate the operation of a component that has suffered micropitting, then surface material delamination may occur. The original surface is thereby lost over wide areas, and the component becomes inoperative through loss of dimensional accuracy, noise, or secondary damage.

In many cases, micropitting (or peeling) is not destructive to the bearing raceway or gear-tooth surface. It sometimes occurs only in patches and may stop after the tribological conditions have been improved by run-in. Run-in is defined as an initial transition process occurring in newly established wearing contacts, often accompanied by transients in coefficient of friction, wear rate, or both, that are uncharacteristic of the long-term behavior of the given tribological system. Gears appear less sensitive to the consequences of surface distress than rolling-element bearings. This may be due to the greater ductility of lower-hardness materials.

Reference cited in this section

23. D.G. McPherson and S. Rao, Mechanical Testing of Gears, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000, p 861

Subcase Fatigue

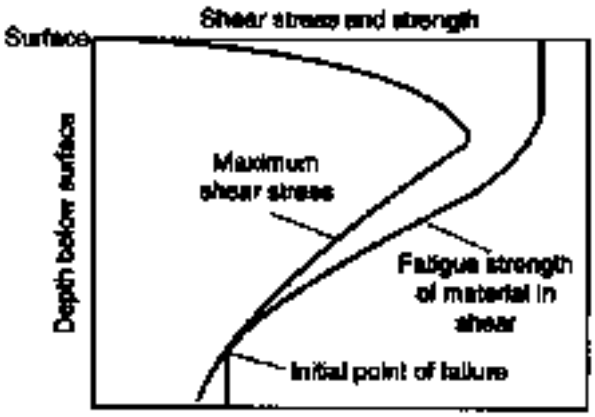
Subcase fatigue is fracture of case-hardened components by the formation of cracks below the contact surface within the Hertzian stress field. However, the depth at which the cracks form is much greater than those that cause macropitting fatigue, and it is a function of material strength in conjunction with the alternating Hertzian shear stresses. Subcase fatigue is also sometimes referred to as *case crushing,* but because it results from a fatigue crack that initiates below the effective case depth or in the lower-carbon portion of the case, the former nomenclature is used.

In case-hardened components—produced by carburizing, nitriding, or induction hardening—a gradient of decreasing hardness (shear strength) exists from the case to the core. If the material shear strength gradient is steeper than the gradient in Hertzian shear stress, then the applied-stress-to-strength ratio is least favorable at some depth below the normal maximum Hertzian shear-stress region. This depth is usually located at (or near) the case/core interface due to shear strength and stress gradients, in addition to unfavorable tensile residual stresses. The reversal in the sign of the residual stress from compression in the case to tension in the core occurs at or very near the case/core interface, and therefore the Hertzian shear stresses are increased due to the tensile residual stresses.

The fatigue originates near the case/core interface where the shear stresses, which are still fairly high, exceed the strength of the relatively low-strength core metal in that region. Fatigue cracks propagate parallel to the surface but underneath the case, then they join and form cracks that come up through the case to the surface, resulting in macropit formation. Figure 44 shows micrographs of a gear tooth and a roller specimen in which subcase fatigue cracks had formed and propagated during testing. Typical example of this type of failure in gear teeth is shown in Figure 45.

(a)

(b)


Shear stress and strength
Surface
Depth below surface
Maximum shear stress
Fatigue strength of material in shear
Initial point of failure

(c)

Fig. 44 Subcase fatigue cracking. The micrographs show etched cross sections of (a) a carburized cylindrical test specimen (1.9×) and (b) a carburized gear tooth in which subcase fatigue cracks initiated and propagated during testing. Subcase fatigue, also known as case crushing, shows the early stages of fatigue cracking in the lower carbon portion of the carburized case. (c) The diagram indicates how the stress and strength gradients combine to cause the weakest condition at the subcase location where fatigue cracking begins.

Fig. 45 Pinion with several very large cavities where metal from the surface down to the depth of the case has fallen out due to subcase fatigue. Source: Ref 24

Subcase fatigue is usually caused by too thin a case, such as on a nitrided gear tooth, by insufficient core hardness on carburized or induction-hardened parts, or in heavily loaded case-hardened rolling-element bearings. Factors that tend to promote the various types of contact fatigue are listed in Table 2. It should be noted that three of the five modes of fatigue damage are influenced by lubrication, including all surface-origin damage modes. Figures 46(a) to (d) show Hertzian shear stresses and material shear strengths as a function of depth below the contact surface. These are an adaptation of the subcase fatigue analysis by Pederson and Rice (Ref 25). Because fatigue occurs when the applied shear stress is greater than the shear strength of the material, the different modes of contact fatigue can be described by determining where these stress profiles cross each other, shown schematically in Fig. 46.

Table 2 List of contact fatigue modes and their controlling factors

Mode of failure	Factors that control occurrence
Subsurface	
Inclusion origin	**Size and density of oxide or other hard inclusions** **Absence of other modes of failure**
Subcase fatigue	**Low core hardness** **Thin case depth relative to radius of curvature and load**
Surface	
Point-surface origin	**Low lubricant viscosity** **Thin EHD film compared with asperities in contact** **Tangential forces and/or gross sliding**
Geometric stress concentration	**End contact geometry** **Misalignment and deflections**

	Possible lubricant film thickness effects
Micropitting	**Low lubricant viscosity** **Thin EHD film compared with asperities in contact** **Loss of EHD pressure** **Slow operating speeds**

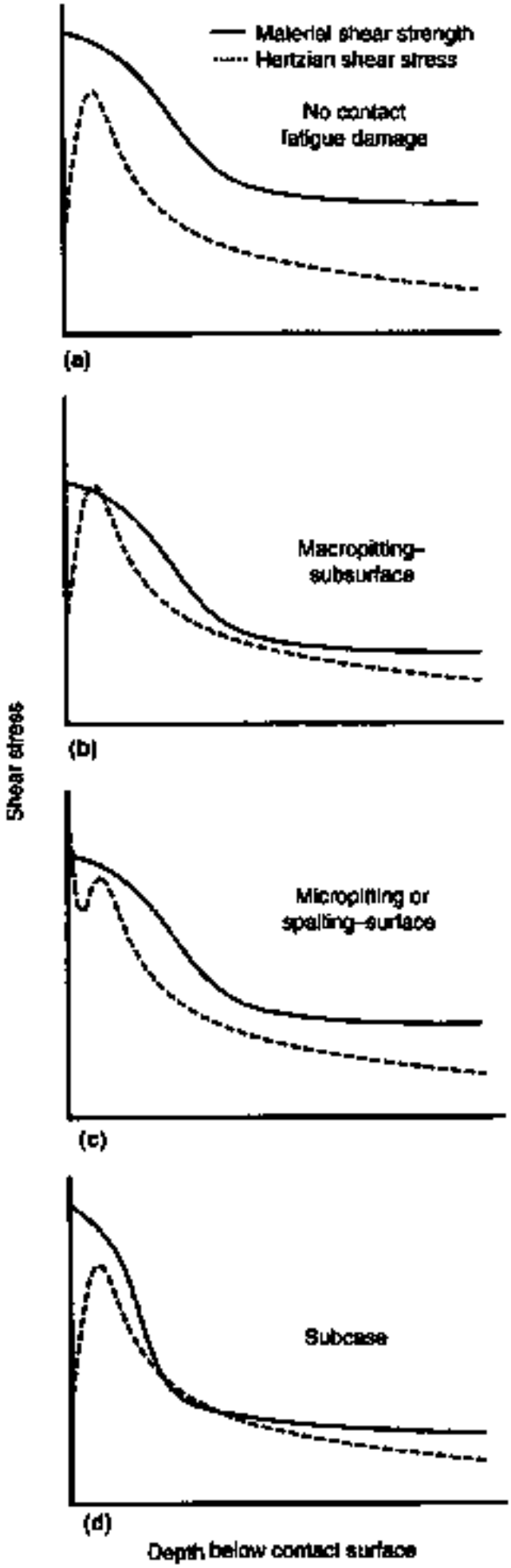

Material shear strength
Hertzian shear stress
No contact
fatigue damage
(a)
Macropitting–
subsurface
(b)
Shear stress
Micropitting or
spalling–surface
(c)
Subcase
(d)
Depth below contact surface

Fig. 46 Applied shear stress and material shear strength as a function of depth representing types of fatigue damage. (a) No damage. (b) Subsurface-origin, macropitting fatigue. (c) Micropitting or surface-origin macropitting fatigue. (d) Subcase fatigue.

References cited in this section

24. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999, p 197

25. R. Pederson and S.L. Rice, Case Crushing of Carburized and Hardened Gears, *SAE Trans.,* Vol 68, 1960, p 187

Thermal Fatigue

Thermal-fatigue failure is the result of temperature cycling, as opposed to fatigue at high temperatures caused by strain cycling. Two conditions necessary for thermal fatigue are some form of mechanical constraint and a temperature change. Thermal expansion or contraction caused by a temperature change acting against a constraint causes thermal stress. The constraint may be external—for example, imposed by rigid mountings for pipes—or it may be internal, in which case it is set up by a temperature gradient within the part. In thick sections, temperature gradients are likely to occur both along and through the material, causing highly triaxial stresses and reducing material ductility, even though the uniaxial ductility often increases with increasing temperature. Reduction in the ductility of the material gives rise to fractures that have a brittle appearance, often with many cleavagelike facets in evidence.

Identification

Identifying features of low-cycle thermal-fatigue failures are multiple initiation sites that join randomly by edge sliding to form the main crack, transverse fractures, an oxide wedge filling the crack, and transgranular fracture. Cracks having similar characteristics but distinguished by intergranular fracture are caused by stress-rupture phenomena (long periods at elevated temperature under high static tensile load). The primary failure mechanism involved in stress rupture is grain-boundary sliding. In thermal-fatigue cracks, slip processes and cleavage operate much as they do in failure at normal temperature, but the evidence is often destroyed by oxide formation, flame polishing, and melting processes.

True thermal fatigue occurs in such components as internal-combustion engines, in which thick-section cast materials are used, and in applications such as heat exchangers, in which thin wrought material is used. Another prominent example occurs in jet engine turbine blades. Uniform sections, mild strain gradients, and short-flake graphite are desirable design features in cast materials. On oil-fired and gas-fired furnace heat exchangers, the thermal cycle is important because it controls temperature gradients, and in thin sections, external constraints are of minor importance.

Under certain circumstances, thermal-fatigue and stress-rupture failures blend into each other. Thermal fatigue is the basic mechanism in failures that occur because of numerous short heating and cooling cycles. Stress rupture becomes an important consideration as the cycle times increase and therefore is primarily a long-term rate process. Most thermal-fatigue fractures are of the low-cycle, high-strain type; the fracture surfaces are rough and faceted at or near the initiation sites and are more fibrous and with shear lips at 45° angles in the final-fracture area.

Acknowledgments

This article was adapted from the article “Fatigue Failures” in the 8th and 9th Edition *Metals Handbook* Volumes on Failure Analysis.

Fatigue Failures

References

1. R.C. Juvinall, *Engineering Considerations of Stress, Strain, and Strength,* McGraw-Hill, 1967, p 274

2. M.A. Miner, Cumulative Damage in Fatigue, *Trans. ASME,* Vol 67, 1945, p A159

3. M.R. Mitchell, Fundamentals of Modern Fatigue Analysis for Design, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 227–249

4. “Advances in Fatigue Lifetime Predictive Techniques,” STP 1122, M.R. Mitchell and R.W. Landgraf, Ed., ASTM, 1992

5. R.O. Ritchie, Near-Threshold Fatigue-Crack Propagation in Steels, *Int. Met. Rev.,* Vol 24 (No. 5, 6), 1979

6. S. Pearson, Initiation of Fatigue Cracks in Commercial Aluminum Alloys and the Subsequent Propagation of Very Short Cracks, *Eng. Fract. Mech.,* Vol 7 (No. 2), 1975, p 235–247

7. R.C. McClung, K.S. Chan, S.J. Hudak, Jr., and D.L. Davidson, Behavior of Small Fatigue Cracks, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 153–158.

8. A.J. McEvily, Fatigue Crack Thresholds, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 134–152.

9. S.T. Rolfe and J.M. Barsom, *Fracture and Fatigue Control in Structures—Applications of Fracture Mechanics,* 3rd ed., ASTM, 1999

10. D.A. Ryder, The Elements of Fractography, AGARDograph No. AGARD-AG-155-71, Advisory Group for Aerospace Research and Development of NATO, 1971

11. M. Hayes, Fatigue of Springs, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 363–369

12. H. Neuber, “Theory of Notch Stresses: Principles for Exact Calculation of Strength With Reference to Structural Form and Material,” AEC-Tr-4547, Springer Publishing, 1958; available through NTIS, U.S. Department of Commerce.

13. R.E. Peterson, *Stress Concentration Design Factors,* John Wiley & Sons, 1974

14. W. Pilkey, *Peterson's Stress Concentration Factors,* 2nd ed., John Wiley & Sons, 1997

15. W.C. Young, *Roark's Formulas for Stress and Strain,* 6th ed., McGraw-Hill, 1989

16. A. Blake, Practical Stress Analysis in Engineering Design, 2nd ed., Marcel Dekker, 1990.

17. M.E. Taylor and J.M. Barsom, Effect of Cyclic Frequency on the Corrosion-Fatigue Crack-Initiation Behavior of ASTM A 517 Grade F Steel, *Fracture Mechanics: Thirteenth Conf.,* STP 743, ASTM, 1981

18. R. Roberts, J.M. Barsom, J.W. Fisher, and S.T. Rolfe, "Fracture Mechanics for Bridge Design," FHWA-RD-78-69, Federal Highway Administration, Office of Research and Development, July 1977

19. D.L. McDowell, Multiaxial Fatigue Strength, ASM Handbook, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 263–273

20. C.A. Moyer, "Comparing Surface Failure Modes in Bearings and Gears: Appearances versus Mechanisms," AGMA Tech. Paper, 91 FTM 6, 1991

21. R.S. Hyde, Contact Fatigue of Hardened Steel, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 691–703

22. T.E. Tallian, *Failure Atlas for Hertz Contact Machine Elements,* American Society of Mechanical Engineers, 1992

23. D.G. McPherson and S. Rao, Mechanical Testing of Gears, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000, p 861

24. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999, p 197

25. R. Pederson and S.L. Rice, Case Crushing of Carburized and Hardened Gears, *SAE Trans.,* Vol 68, 1960, p 187

Creep and Stress Rupture Failures

Introduction

HIGH-TEMPERATURE and stress are common operating conditions for various parts and equipment in a number of industries. The principal types of elevated-temperature mechanical failure are creep and stress rupture, stress relaxation, low-cycle or high-cycle fatigue, thermal fatigue, tension overload, and combinations of these, as modified by environment. Generally, the type of failure is established by examination of fracture surfaces and comparison of component-operating conditions with available data on creep, stress-rupture, tension, elevated-temperature fatigue, and thermal fatigue properties. Such an analysis is usually sufficient for most failure investigations, but a more thorough analysis may be required when stress, time, temperature, and environment have acted to change the metallurgical structure of the component (see, for example, the section "Metallurgical Instabilities" in this article).

This article briefly reviews the applied aspects of creep-related failures, where mechanical strength of a metal becomes limited by creep rather than by the elastic limit for a given alloy. Creep deformation may produce sufficiently large changes in the dimensions of a component to either render it useless for further service or cause fracture. When excessive creep deformation causes the material to reach or exceed some design limit on strain, the term *creep failure* is used. For example, a creep failure of a cobalt-base alloy turbine vane is shown in Fig. 1. The bowing is the result of a reduction in creep strength at the higher temperatures from overheating.

Fig. 1 Creep damage (bowing) of a cobalt-base alloy turbine vane from overheating

Creep deformation may also result in the complete separation or fracture of a material and the breach of a boundary or structural support. Fracture may occur from either localized creep damage or more widespread bulk damage caused by the accumulation of creep strains over time. Structural components that are vulnerable to bulk creep damage typically are subjected to uniform loading and uniform temperature distribution during service. The life of such a component is related to the creep-rupture properties, and the type of failure is referred to as stress rupture or creep rupture. Stress or creep rupture is apt to occur when damage is widespread with uniform stress and temperature exposure, as in the situation of thin-section components (such as steam pipes or boilers tubes).

On the other hand, creep damage can also be localized, particularly for thick-section components that are subjected to gradients in stress (strain) and temperature. An example of a creep-related crack is shown in Fig. 2. Cracks may develop at a critical location and propagate to failure before the end of the predicted creep-rupture life. Creep cracking may also originate at a stress concentration or at preexisting defects in the component. In these cases, most of the life of the component is spent in crack growth. This involves assessment of fracture resistance rather than a strength assessment based on bulk creep rates and time to stress rupture. Therefore, creep life assessment may involve evaluation of both creep strength (i.e., creep rate, stress rupture) and resistance to fracture under creep conditions.

Fig. 2 Creep crack in a turbine vane. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

Failures from creep deformation depend on the alloy, the time-temperature exposure, loading conditions, component geometry, and also environmental and metallurgical factors. It is also not unusual to find other contributing factors, such as corrosion, fatigue, or material defects, involved in creep and stress-rupture failures. This article briefly reviews these factors as they relate to creep behavior and associated failures of alloys used in high-temperature applications. The complex effects of creep-fatigue interaction are also discussed, although more detail on this is described in the article "Thermomechanical Fatigue: Mechanisms and Practical Life Analysis" in this Volume. Life assessment is also covered in the article "Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing" in this Volume.

Bulk Creep Behavior

Some of the key material properties at high temperature are thermal expansion coefficient, stress rupture, elastic modulus, fatigue life, and oxidation resistance. Total strain at temperature is given by the sum of elastic stress-strain modulus, thermal expansion strain, and creep strain. Change in elastic modulus with temperature cannot be neglected at higher homologous temperatures. Parametric modeling of creep or stress-rupture behavior often normalizes the stress with the modulus.

One way to avoid creep failure is to select the proper alloy based on parametric extrapolation of creep properties. In general, creep occurs in any metal or alloy at a temperature where atoms become sufficiently mobile to allow time-dependent rearrangement of structure. Creep behavior of a polycrystalline metal or alloy often is considered to begin at approximately one-third to one-half of its melting point (~0.3 to 0.5 T_M) measured on an absolute temperature scale (degrees Kelvin or Rankine). However, this rule of thumb is not necessarily the criteria for engineering design. Actually, creep deformation becomes important when mechanical strength of a metal becomes limited by creep rather than by yield strength. This transition in engineering design is not directly related to melting temperature; consequently, the temperature at which the mechanical strength of a metal becomes limited by creep, rather than by elastic limit, must be determined individually for each metal or alloy.

Approximate temperatures at which creep behavior begins for several metals and alloys are listed in Table 1. Low-melting-point metals (such as lead, tin, and high-purity aluminum) may deform by creep at or a little above room temperature. In contrast, refractory body-centered cubic metals (such as tungsten and molybdenum) and nickel-base superalloys require temperatures near 1000 °C (1830 °F) to activate the onset of creep-deformation engineering significance. Typical materials and application temperatures of some creep-resistant alloys are listed in Table 2.

Table 1 Approximate temperatures for onset of creep behavior for selected alloy groups

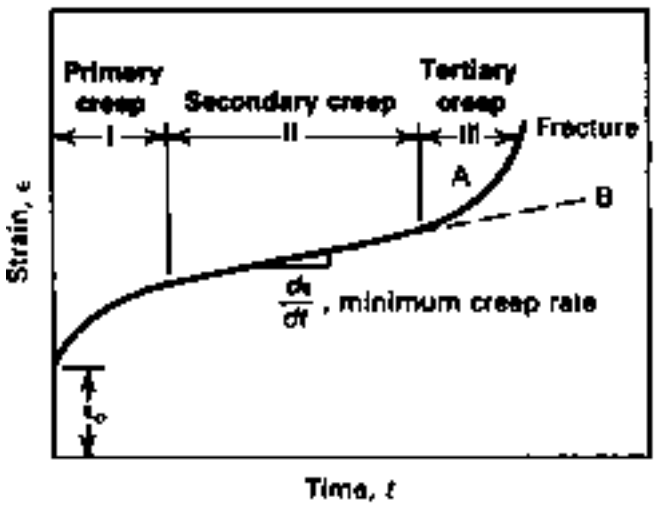
Metal or alloy	Temperature (T) for onset of creep		
	°C	**°F**	T as ratio of T_M, K
Aluminum alloys	150–200	300–400	**0.48–0.54 T_M**

Titanium alloys	315	600	**0.3 T_M**
Low-alloy steels	370	700	**0.36 T_M**
Austenitic, iron-base, heat-resisting alloys	540	1000	**0.49 T_M**
Nickel- and cobalt-base heat-resisting alloys	650	1200	**0.56 T_M**
Refractory metals and alloys	**980–1540**	**1800–2800**	**0.4–0.45 T_M**

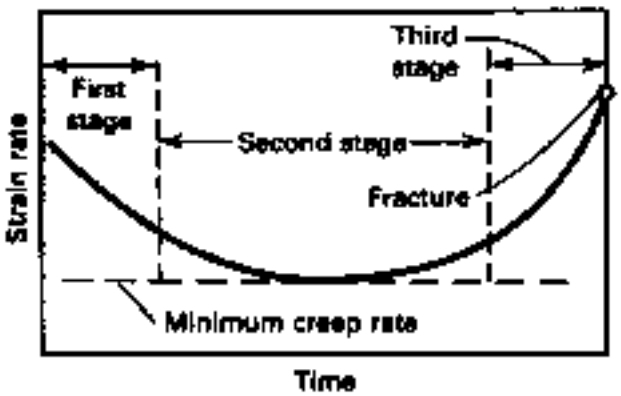
Table 2 Typical elevated temperatures in engineering applications

Application	**Typical materials**	**Typical temperatures, K**	Homologous temperatures, T/T_M
Rotors and piping for steam turbines	Cr-Mo-V steels	825–975	**0.45–0.50**
Pressure vessels and piping in nuclear reactors	316 stainless steel	650–750	**0.35–0.40**
Reactor skirts in nuclear reactors	316 stainless steel	850–950	**0.45–0.55**
Gas turbine blades	Nickel-base superalloys	775–925	**0.45–0.60**
Burner cans for gas turbine engines	**Oxide dispersion-strengthened nickel-base alloys**	**1350–1400**	**0.55–0.65**

Constant-load bulk deformation creep curves typically (but not always) consist of three distinct stages (Fig. 3a). The first stage (called primary creep) is the region of the initial instantaneous elastic strain from the applied load, followed by a region of increasing plastic strain at a decreasing strain rate (Fig. 3b). Following the first stage of primary creep is the region of secondary creep, where the creep rate is nominally constant at a minimum rate, generally known as the minimum creep rate. The stage of secondary creep is sometimes referred to as the region of steady-state creep, because the creep rate is often nominally constant for some period of time. Finally, there is a region of drastically increased strain rate with rapid extension to fracture in the third stage of tertiary creep. Constant-stress tests (as opposed to constant-load tests) often do not show tertiary behavior.



(a)



(b)

Fig. 3 Stages of creep deformation. (a) Strain curve for the three stages of creep under constant-load testing (curve A) and constant-stress testing (curve B). (b) Relationship of strain rate, or creep rate, and time during a constant-load creep test. The minimum creep rate is attained during second-stage creep

Creep and stress-rupture data usually are obtained under constant-load test condition, and therefore, the stress in the gage section varies with time. However, it is sometimes desirable or necessary to obtain test data under constant-stress conditions, where the applied load is adjusted as the length of the specimen changes to maintain constant stress on the specimen. Constant-stress testing is used to accurately determine differences between the temperature dependence and the stress dependence of creep behavior. In general, the form of creep strain (in the form of primary, secondary, and tertiary stages) under constant-stress conditions is the same as those for constant load. However, the period of time spent in primary and secondary creep under constant stress can be much longer than under an identical engineering stress (constant load). Hence, rupture life is longer under constant-stress conditions.

It is also important to note that the region of initial elastic strain under load (ε_0 in Fig. 3a) is not identified as a stage of creep deformation, because it is common practice to ignore this contribution to total strain when plotting creep curves. Consequently, creep curves generally show only the time-dependent plastic strain that follows the initial elastic (and possibly plastic) strain. Although this procedure is acceptable for research or investigation, the initial strain, which may amount to a substantial fraction of the total strain, should not be omitted from design studies and failure analyses. Therefore, as stated previously, modulus at temperature is of importance. Thermal expansion may impose additional strain and must also be considered.

During creep, significant microstructural changes occur on all levels. On the atomic scale, dislocations are created and forced to move through the material. This leads to work hardening as the dislocation density increases and the dislocations encounter barriers to their motion. At low temperatures, an ever-diminishing creep rate results; however, if the temperature is sufficiently high, dislocations rearrange and annihilate through recovery events. During creep deformation, the material also is progressively degraded or damaged as the amount of creep strain increases over time.

The classical pattern of creep deformation fits many materials under various test conditions. However, even though it is relatively easy to quantify damage in laboratory creep tests conducted at constant temperature and load (or stress), components in service hardly ever operate under constant conditions. Start-stop cycles, reduced power operation, thermal gradients, and other factors result in variations in stresses and temperatures. Therefore, damage rule procedures are used to estimate cumulative damage under changing exposure conditions. In addition, creep behavior can depart from classical creep curves even under controlled laboratory testing. The following sections briefly review the microstructural changes and bulk mechanical behavior of classical creep behavior and some instances of nonclassical behavior. This is followed by a brief description of microstructural changes and damage from creep deformation.

Classical Creep Behavior

Primary creep, also known as transient creep, represents a stage of adjustment within the metal during which rapid thermally activated plastic strain occurs. The combined action of hardening and recovery processes during primary creep can lead to the formation of a stable distribution of subgrains or loose, three-dimensional dislocation networks in some materials, or an approximately uniform dislocation distribution without subgrains in other materials. These stable dislocation configurations are maintained and are characteristic of second-stage creep. Primary creep occurs in the first few moments after initial strain and decreases in rate as crystallographic imperfections within the metal undergo realignment. This realignment leads to secondary creep.

Because a variety of metallurgical processes can be involved, and because the rates and direction of these processes can vary with time and temperature, departures from classical creep curves may be overlooked, unless accurate creep readings are taken at sufficiently close intervals, particularly during early stages of creep.

Secondary creep is an equilibrium condition between the mechanisms of work hardening and recovery; it is also known as steady-state creep. As previously noted, secondary creep is essentially a transition between primary and tertiary creep, where the creep rate reaches a minimum value (Fig. 3b). This often occupies the major portion of the duration of the creep test, and the strain rate in this region for many creep-resistant materials is sufficiently constant to be considered as a steady-state creep rate. For these materials, the minimum creep rate is a steady-state value that can be empirically related to rupture life and is widely used in research and engineering studies.

However, the concept of a steady creep rate during secondary creep is more of an empirical concept than a mechanistic description of material behavior. Nonuniform deformation and changes in stress and metallurgical conditions at high temperature are factors that may influence the creep rate. The process of secondary creep thus may be more complicated than a constant, steady-state mechanistic condition. Nonetheless, the observation of steady creep rates over a large portion of creep-test curves is an important empirical result with practical value. The concept of a steady-state creep rate should only be viewed as an experimental result, not as an underlying material condition. Variations from metallurgical instabilities and changing stress-temperature conditions are additional factors as well.

Tertiary Creep. Primary creep has no distinct endpoint, and tertiary creep has no distinct beginning. Tertiary creep refers to the region of increasing rate of extension that is followed by fracture. Principally, it may result from metallurgical changes, such as recrystallization under load, that promote rapid increases in deformation, accompanied by work hardening that is insufficient to retard the increased flow of metal. It also may correspond to the onset of necking in some alloys.

In service or in creep testing, tertiary creep may be accelerated by a reduction in cross-sectional area resulting from cracking or necking. Environmental effects, such as oxidation, that reduce cross section may also initiate tertiary creep or increase the tertiary-creep rate. In many commercial creep-resistant alloys, tertiary creep is apparently caused by inherent deformation processes and occurs at creep strains of 0.5% or less. In designing components for service at elevated temperatures, data pertaining to the elapsed time and extension that precede tertiary creep are of the utmost importance; design for creep resistance is based on such data. However, the duration of tertiary creep is also important, because it constitutes a safety factor that may allow detection of a failing component before catastrophic fracture.

Nonclassical Creep Behavior

Although the classical pattern of creep deformation can be made to fit many materials and test conditions, the relative duration of the three periods differs widely with materials and conditions. For example, in many superalloys and other materials in which a strengthening precipitate continues to age at creep temperatures, brief primary creep often shows transition to a long, upward sweep of creep rate, with only a point of inflection for the secondary period.

Aging of normalized and tempered 0.5Cr-0.5Mo-0.25V steel during creep under 80 MPa (11.6 ksi) stress at 565 °C (1050 °F) has been reported to cause the creep curve to effectively exhibit only a continuously increasing creep rate to fracture (Ref 1). For twice the amount of stress, the creep curve in this case followed the classical trends. In other alloys, such as titanium alloys, with limited elongation before fracture, the tertiary stage may be brief and may show little increase in creep rate before rupture occurs. A more obvious departure from classical behavior develops during the early portion of many tests, when precise creep measurements are taken. When 34 ferritic steels were studied for as long as 100,000 h at temperatures ranging from 450 to 600 °C (842 to 1112 °F), step-form irregularities were observed, with an extended period of secondary creep preceded by a lower creep rate of shorter duration during primary creep (Ref 2).

Under certain conditions, some metals also may not exhibit all three stages of plastic extension. For example, at high stresses or temperatures, the absence of primary creep is not uncommon, with secondary creep or, in extreme cases, tertiary creep following immediately upon loading. At the other extreme, notably in cast alloys, no tertiary creep can be observed, and fracture may occur with only minimum extension.

Oxide and Nitride Strengthening. An entirely different source of variation from classical creep may occur from environmental reactions at high temperature. For example, tests longer than 50 h with 80Ni-20Cr alloys at 816 and 982 °C (1500 and 1800 °F) showed a deceleration of creep after the normal tertiary stage was reached, resulting in a second period of steady-state creep and later, another period of last-stage creep (Ref 3). This behavior is due to oxide strengthening, which prolonged rupture life and caused a slope decrease in curves of log stress versus log rupture life. Oxides and nitride formed on the surfaces of the intercrystalline cracks that occur extensively during tertiary creep. Observed interconnection of the bulk of these cracks added substantially to strengthening against creep deformation in the late stages of the tests. This effect also has been observed in 99.8% Ni tested at 816 °C (1500 °F) under 20.7 MPa (3000 psi) stress (Ref 4).

Microstructural Changes during Creep

During creep, significant microstructural changes occur on all levels. On the atomic scale, dislocations are created and forced to move through the material. This leads to work hardening as the dislocation density increases and the dislocations encounter barriers to their motion. At low temperatures, an ever-diminishing creep rate results; however, if the temperature is sufficiently high, dislocations rearrange and annihilate through recovery events. Creep deformation also produces microstructural changes, such as slip bands, grain-boundary sliding, cavity formation and growth, and cracking (grain boundary, interphase boundary, and transgranular). There are others, some of which are very important. One is in situ graphitization that occurs in iron-base alloys without adequate carbide stabilizing elements present. Iron carbide (cementite) decomposes to graphite and iron. Normally, chromium is added to prevent this. In situ graphitization is also affected strongly by molybdenum content and deoxidation practice. The extent of these microstructural changes is generally increased near fracture sites, compared to other regions.

The microstructure of a creep or stress-rupture specimen after elevated-temperature exposure rarely resembles the initial microstructure prior to exposure. Of course, most materials are not thermodynamically stable; hence, prolonged exposure under creep conditions can result in microstructural changes, such as the precipitation of new phases, dissolution or growth of desired phases, grain growth, and so on. Many of the structural changes can be duplicated through simple heat treatment. However, some microstructural changes only occur under the combined influence of stress and temperature. For example, the cube-shaped γ-(Ni_3Al) strengthening phase in nickel-base superalloy NASAIR 100, under normal isothermal annealing, undergoes Ostwald ripening, characterized by an increase in particle size without any shape change. However, during creep testing, the individual precipitates grow together rapidly and form thin γ plates, where

the long dimensions of each plate are perpendicular to the stress in tensile creep and parallel to the applied stress in compressive creep (Ref 5).

The changes in microstructure that occur during creep testing affect properties. Although such changes may be unavoidable, in many cases, thermomechanical processing schedules can be established to influence the changes so that they tend to strengthen the material or minimize the overall effect. For example, if a heterogeneous precipitate is formed during creep, a simple heat treatment or cold work followed by annealing prior to testing should give a homogeneous distribution of precipitates.

Microstructural changes due to the combined influence of temperature and stress can be difficult to evaluate. Therefore, simulation of creep exposure prior to actual use is sometimes necessary, where tested and untested materials are examined and compared. Complete microstructural examination should be an essential part of any creep experiment and failure investigation. At a minimum, microstructure and phase identification should be compared for specimens near and further away from the fracture site. Comparison of these may help identify relevant deformation mechanism, environmental effects, or microstructural changes. Such information is vital for interpreting and understanding creep behavior.

Generally, creep (distortion) failures are recognized by local ductility and multiplicity of intergranular cracks (Fig. 4). However, creep deformation of engineering significance can also occur before intergranular fracture initiates. Stress-rupture data (log stress versus log time to failure) typically show an inflection when a change in fracture mechanism occurs. This includes significant formation of sigma and a change in fracture from transgranular to intergranular, plus many others. Measurements of fracture strain also show a minimum when the mechanism changes.

Fig. 4 Typical creep deformation and intergranular cracking in a jet-engine turbine blade. Courtesy of J. Schijve

Cavitation Damage. The most common form of microstructural change is the accumulation of nucleation and growth of voids. Void growth is well understood, because voids grow by the same mechanisms that cause creep deformation. In

contrast, void nucleation is not fully understood, although it appears to be the result of strain “misaccommodation” similar to crack nucleation in some low-temperature fractures (Ref 6). For the high-temperature case, two adjacent grains may move at different creep rates. If the rates cannot be made to match, a gap forms between the grains. In general, higher creep rates cause voids to form earlier during the creep process. Thus, any material alteration that leads to a lower creep rate also improves creep strength.

Creep damage begins as small holes or cavities that typically form at grain boundaries or second phases. With time and stress, these holes or cavities can link up and form cracks, by both void growth (diffusion controlled) and by shear strain on the grain boundary, that eventually lead to failure of the component. The damage is progressive and may occur in bulk or, more predominately, in localized regions. The formation of creep cracks is usually very localized in thick-section components; they form in welds, bends, or other highly stressed regions. Determining the extent of this progressive damage is thus an important imperative for regular inspections of components exposed to elevated temperatures. Various damage rules are also used to approximate remaining time to rupture for components that have undergone prolonged high-temperature exposure. Damage rules and postservice creep testing are discussed further in the article “Assessment and Use of Creep-Rupture Properties” in *Mechanical Testing and Evaluation,* Volume 8 of *ASM Handbook* (Ref 7).

Replication Methods for Assessment of Creep Damage. Surface replication is a well-known sample preparation technique that can be used to assess the condition of high-temperature power plant and petrochemical components from creep damage. The usual method of metallographic investigation involves cutting large pieces from components, which thus renders the component unfit for service. In contrast, surface replication allows examination of microstructural damage without cutting sections from the component. Plastic replicas lend themselves to in-plant nondestructive examination because of their relative simplicity and short preparation time. Plastic replicas can be examined with the light optical microscope, the scanning electron microscope, and the transmission electron microscope, depending on the resolution required.

Replication techniques are sufficiently sophisticated to allow classifications of microstructural damage (Fig. 5) (Ref 8) that can be directly correlated to life fractions. Bulk damage from creep deformation typically occurs by the nucleation and growth of voids either within grains or, as is more common, along grain boundaries. Internal voids first nucleate during creep deformation and then grow. A distinct correlation exists for these data, such that a minimum and maximum remaining life fraction can be specified (Table 3) (Ref 9). The qualitative-quantitative relation is advantageous, because data from surface replication can be predictive in terms of generating a conservative minimum- and maximum-life estimate. The maximum life is useful in a predictive maintenance environment, because it would dictate the planning of future repairs or replacement.

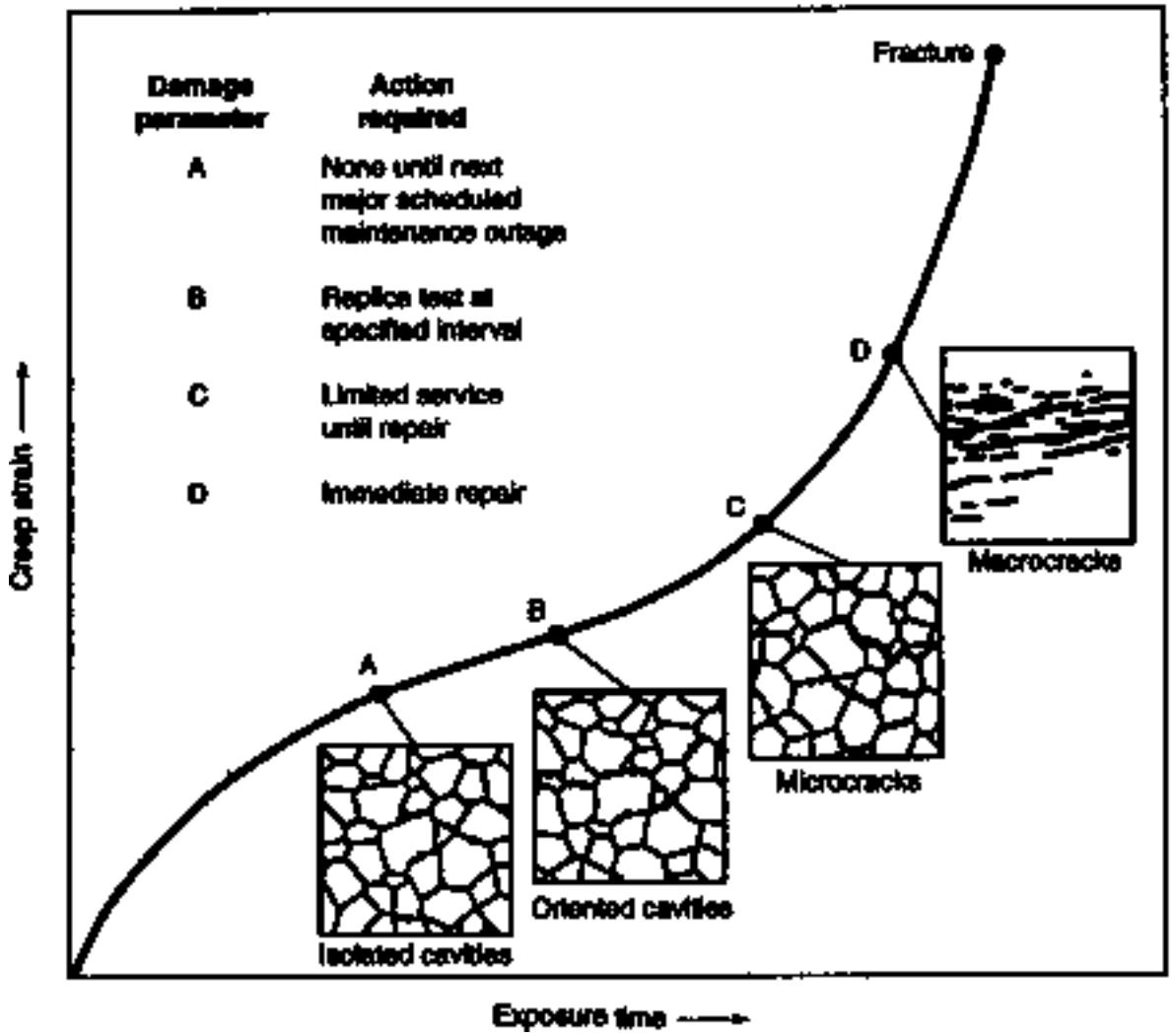


Fig. 5 Creep life assessment based on cavity classification in boiler steels

Table 3 Correlation of damage level and life fraction consumed

Damage level	**Consumed life fraction Range *X***	Remaining life factor (1/*X* - 1)	
		Minimum	Maximum
1	0.00–0.12	7.33	**Unknown**
2	0.04–0.46	1.17	**24.00**
3	0.3–0.5	1.0	**2.33**
4	0.3–0.84	0.19	**2.33**
5	**0.72–1.00**	**0 = failed**	**0.39**

Source: Ref 9

References cited in this section

1. K.R. Williams and B. Wilshire, Effects of Microstructural Instability on the Creep and Fracture Behavior of Ferritic Steels, *Mater. Sci. Eng.,* Vol 28, 1977, p 289–296

2. M. Wild, Analyse des Zeitstandverhaltens warmfester ferritischer Stahle bei Temperaturen von 450 bis 600 C, *Arch. Eisenhüttenwes.,* Vol 34, Dec 1963, p 935–950

3. R. Widmer and N.J. Grant, The Role of Atmosphere in the Creep-Rupture Behavior of 80Ni-20Cr Alloys, *J. Basic Eng.* (*Trans. ASME D,*) Vol 82 (No. 4), Dec 1960, p 882–886

4. P. Shahinian and M.R. Achter, Creep-Rupture of Nickel of Two Purities in Controlled Environments, *Joint International Conference on Creep* (London), Institute of Mechanical Engineers, 1963, p 7-49 to 7-57

5. M.V. Nathal and L.J. Ebert, Gamma Prime Shape Changes during Creep of a Nickel-Base Superalloy, *Scr. Metall.,* Vol 17, 1983, p 1151–1154

6. T.H. Courtney, Fundamental Structure-Property Relationships in Engineering Materials, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 354

7. H. Voorhees and M. Prager, Assessment and Use of Creep-Rupture Properties, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000, p 383–397

8. B. Neubauer and U. Wedel, Rest-of-Life Estimation of Creeping Components by Means of Replicas, *Advances in Life Prediction Methods,* American Society of Mechanical Engineers, 1983, p 307–314

9. J.M. Brear et al., *Possibilitic and Probabilistic Assessment of Creep Damage,* ICM 6, Pergamon, 1991

Stress Rupture

In the design of components for service at elevated temperature, both strength and stress-rupture ductility must be considered. Normally, these properties are determined quantitatively by smooth-bar stress-rupture tests. In most service applications, components are subjected to complex stresses, including stress concentrations arising from inherent flaws in the metal as well as those stemming from design. Results of smooth-bar tests do not account for such stress conditions. Fortunately, these conditions can be partly reproduced in the laboratory by use of the notched-bar test. Notched-bar stress-rupture tests are of principal interest because of the relationship between stress-rupture ductility and the ratio of notched-bar to smooth-bar stress-rupture strength.

Prevention of creep-induced fracture thus involves the consideration of both material strength and damage tolerance, much as in static design. The chromium-molybdenum steels, stainless steels, and Cr-Mo-V steels, which are commonly used in high-temperature service, are considered to be creep-ductile materials with the ability to sustain significant amounts of crack growth before failure. However, most engineering alloys can lose ductility during high-temperature service, because the diffusion of impurities to the grain boundaries becomes more pronounced. Formation of precipitates can also influence creep behavior, as discussed in the section “Metallurgical Instabilities” in this article.

Stress-rupture tests evaluate rupture strength by determining the time to fracture as a function of temperature and applied stress. Logarithmic (log-log) plots of normal stress (σ) versus rupture life (RL) provide an indication of metallurgical instabilities. These instabilities are delineated by the occurrence of new straight-line segments of changed slope on the log-log plot, as shown by inflection points A, B, N, O, and Y in Fig. 6 (Ref 10) for a well-known nickel-base heat-resisting alloy. Stress rupture is a variable-rate process; straight lines on log σ-log RL plots are simply approximations of a continuously varying curve.

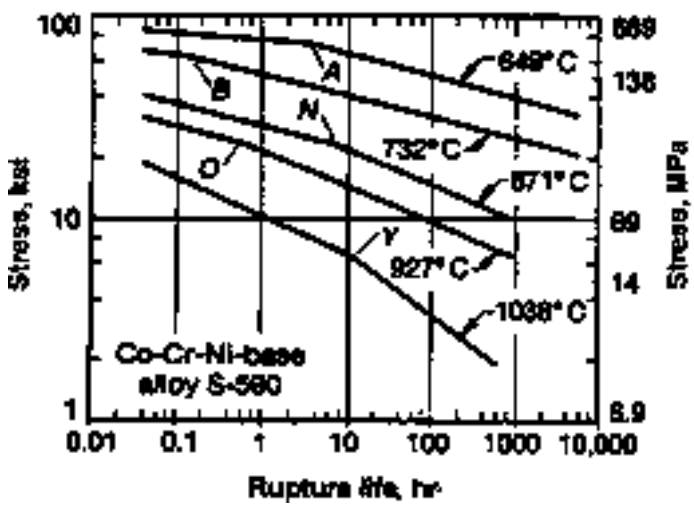


Fig. 6 Logarithmic plot of stress-rupture stress versus rupture life for Co-Cr-Ni-base alloy S-590. The significance of inflection points A, B, N, O, and Y is explained in the text. Source: Ref 10

Several parameters have been used for comparison and interpolation of stress-rupture data. One used very widely is the Larson-Miller parameter (LMP), defined by F.R. Larson and J. Miller. The LMP is useful to determine the remaining rupture life at a given applied stress, and it is defined by the following equation:

$$\text{LMP} = T(\text{R}°)[C + \log t\ (\text{hours})] \times 10^{-3}$$

where T is the test temperature in Rankine (°F + 460), t is the operating time in hours, and C is a material constant, approximately 20 for a low-alloy steel. Other parametric methods are discussed in the article "Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing" in this Volume and in the article "Assessment and Use of Creep-Rupture Properties" in *Mechanical Testing and Evaluation,* Volume 8 of *ASM Handbook.*

Stress-Rupture Ductility. While creep strength and rupture strength are given considerable attention as design and failure parameters, rupture ductility is an important mechanical property when stress concentrations and localized defects, such as notches, are a factor in design. Rupture ductility, which varies inversely with creep and rupture strength, influences the growth of cracks or defects and can be compared in terms of notch sensitivity.

In many service conditions, the amount of deformation is not critical, and relatively high rupture ductility can be used in design. Under such conditions, with the combined uncertainties of actual stress, temperature, and strength, it may be important that failure not occur without warning and that the metal retain high elongation and reduction in area throughout its service life. In the oil and chemical industries, for instance, many applications of tubing under high pressure require high long-time ductility, so that impending rupture is evident from the bulging of the tubes.

The two common measures of ductility are elongation and reduction of area. However, in tension-creep curves, there are two meaningful measures of elongation. First, there is true elongation, which is defined as the elongation at the end of the second stage of creep. Second, there is total elongation, which is the elongation at fracture. In some instances, elongation at fracture is mainly extension caused by crack separation. In others, it is mainly necking extension or extension resulting from other tertiary-creep processes. Uniform elongation, on the other hand, nominally consists of extension resulting from nonlocalized creep processes only, although it frequently includes some elongation due to intergranular void formation.

Ductility data from stress-rupture tests are generally erratic, even in replicate tests with data scatter at the same temperatures and stress, and variation in the measured strain at fracture varies with both stress and temperature. Ductility data are more erratic for castings than for wrought products. In some metals and alloys, the values of total elongation follow a smooth curve that either increases or decreases with increasing rupture time and temperature. Ductility data typically show a minimum, initially increasing above room-temperature ductility, then going through a minimum, then increasing again, and finally dropping very fast near the melting point. The increase above the minimum in the curve is often assumed to be due to dynamic recrystallization.

Data on total elongation and reduction of area at various stress levels for a low-alloy steel at 540 °C (1000 °F) and a stainless steel at 705 °C (1300 °F) are given in Table 4. Despite some difficulties in the interpretation of ductility data, common practice is to plot elongation versus rupture life, as shown in Fig. 7. Here are shown values for both total and true elongation at two different testing temperatures for the nickel-base alloy S-590. At both temperatures, an appreciable scatter of data points exists for total elongation (open circles in Fig. 7).

Table 4 Typical elevated-temperature ductility of a 1.25Cr-0.5Mo low-alloy steel and type 316 stainless steel

Time to rupture, h	**Total elongation in 38 mm (1.5 in.), %**	Reduction in area, %
1.25Cr-0.5Mo low-alloy steel at 540 °C (1000°F)		
8.7	19.6	**41**
47.0	12.1	**29**
259.4	14.0	**20**
660.6	10.5	**13**
2162.0	17.6	**32**
Type 316 stainless steel at 705 °C (1300 °F)		
3.1	26.6	**26**
3.7	24.4	**27**
37.5	17.8	**28**
522.6	56.2	**41**
881.6	**39.4**	**33**

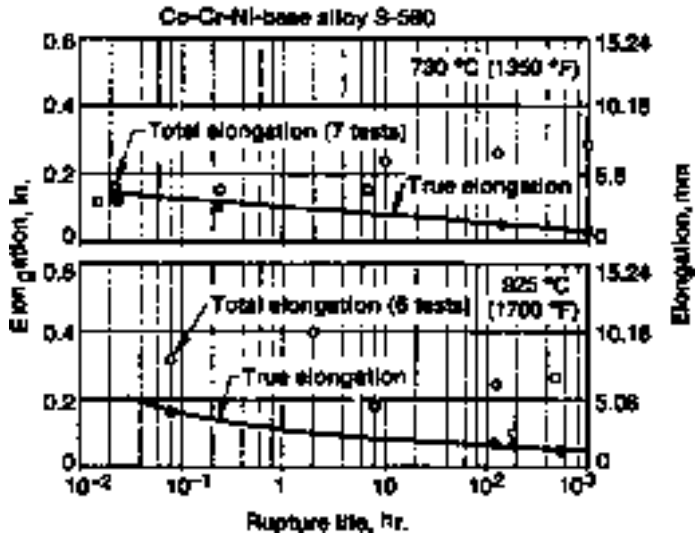


Fig. 7 Relation of elongation and rupture life for Co-Cr-Ni-base alloy S-590 tested at two temperatures at different stresses. Source: Ref 10

Generally, true elongation provides the more accurate representation of ductility in metals at elevated temperatures. The large differences between total and true elongation, such as those shown in Fig. 7, are a function of crack volume and distribution in tertiary creep. The differences in crack initiation and growth can make only relatively small differences in total elongation. Replicate tests may thus exhibit large differences in total elongation but very small differences in true elongation.

Stress-rupture ductility is an important factor in alloy selection. In conventionally cast nickel-base heat-resisting alloys, for example, a stress-rupture ductility of approximately 1% is common at 760 °C (1400 °F), compared to values above 5% for the strongest wrought heat-resisting alloys. Because component designs are frequently based on 1% creep, a low stress-rupture ductility may preclude use of an alloy to its full strength potential. As shown by the schematic creep curves in Fig. 8, a higher rupture ductility for the same load and temperature conditions means a higher safety margin. Premature failures have resulted from lack of ductility during tertiary creep.

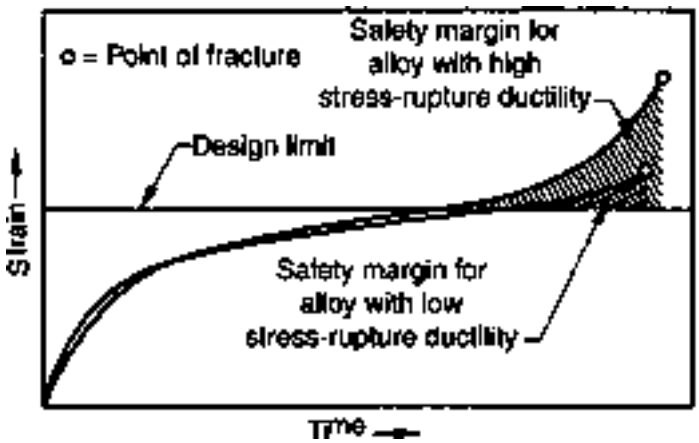


Fig. 8 Schematic creep curves for alloys having low and high stress-rupture ductility, showing the increased safety margin provided by the alloy with high stress-rupture ductility. Source: Ref 10

Creep Crack Growth. Components with localized damage, which is a result of nonuniform stress and temperature distribution found most commonly in large structures, are more prone to fail as a result of creep crack propagation rather than stress rupture. Despite the sophisticated methods of flaw detection that are available, defects and impurities are commonly present in all large components and can potentially escape detection. In the high-temperature regime, components fail by the accumulation of time-dependent creep strains at these defects, which, with time, evolve into cracks, causing eventual failure.

Most engineering alloys are highly plastic at elevated temperatures, and any preexisting crack quickly becomes blunted. Under such conditions, creep is generally considered a net-section effect. In this case, the process of subcritical creep crack growth would not be characterized by the fracture-mechanic stress-intensity (*K*) parameter. Other methods are

used to characterize creep crack growth in ductile materials used at high temperature, as described in more detail in the article “Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing” in this Volume.
However, some high-temperature materials (such as intermetallics and some precipitation-hardened superalloys) are less ductile and may have creep crack growth that depends on the crack tip stress field. In these cases, the creep crack growth rate (*da*/*dt*) can be correlated with *K,* with a power-law dependence as follows:

$$\frac{da}{dt} = \alpha K^n$$

where α and *n* are the curve-fitting parameters. The value of *n* varies between 1 and 8, depending on the stress dependence of the specific creep mechanism (Ref 11). A distinct threshold stress intensity is usually not observed, as shown in the general example in Fig. 9.

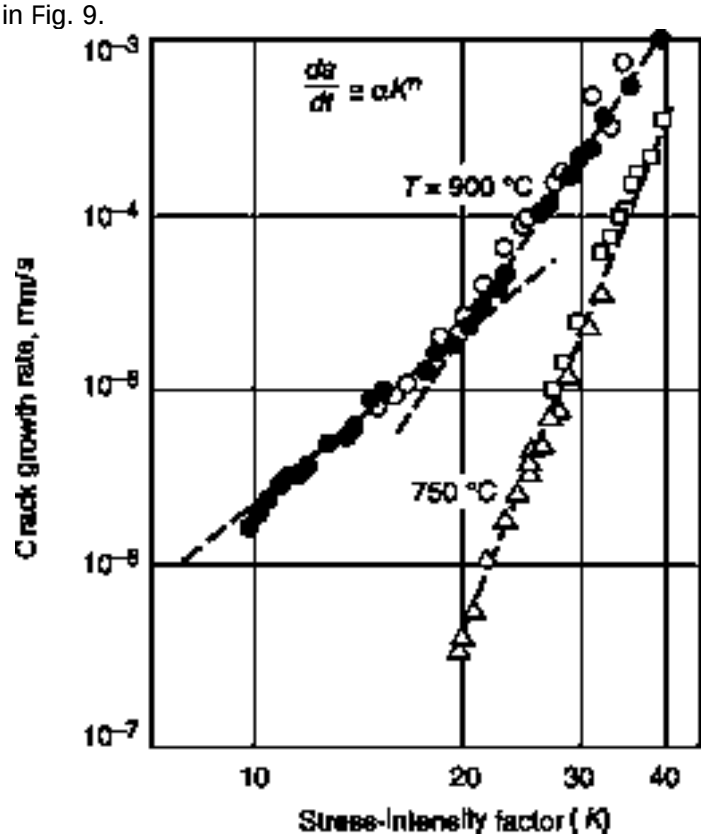


Fig. 9 Power-law dependence of creep crack growth with *K* in less ductile materials. Source: Ref 11

References cited in this section

10. D.J. Benac, Elevated-Temperature Failures, Lesson 7, *Principles of Failure Analysis,* Materials Engineering Institute, ASM International

11. H. Nelson, “Fractography in the Failure Analysis of Engineering Structures,” ASM annual meeting, 1997

Stress-Rupture Fractures

Stress-rupture fractures are typically characterized by a multiplicity of creep voids adjacent to the main fracture. The voids are generally easy to identify by optical examination of microsections. Figure 10 shows an example of a heater tube that failed by stress rupture and microvoids that formed near the fracture surface. Another example is shown in Fig. 11 for a failure from short-term creep.

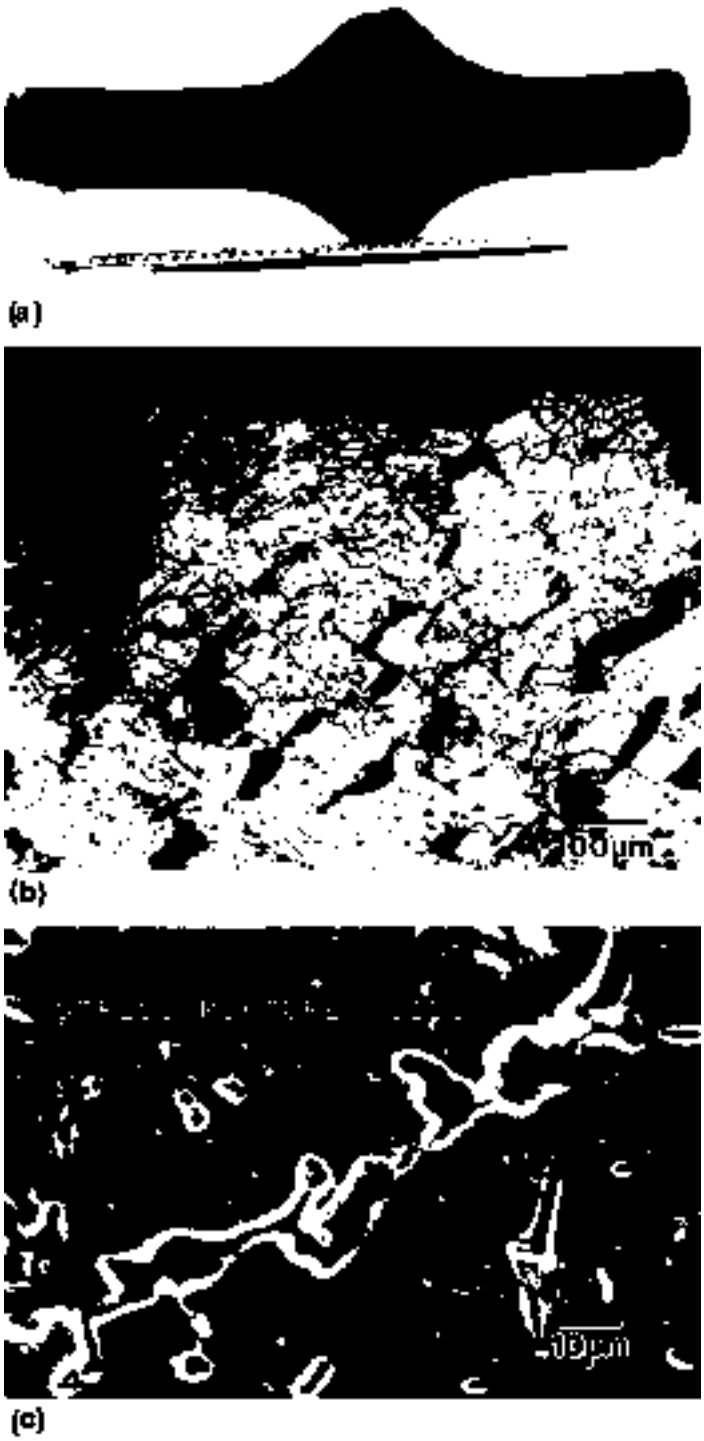


Fig. 10 Stress rupture of heater tube. (a) Heater tube that failed due to stress rupture. (b) and (c) Stress-rupture voids near the fracture. Source Ref 10

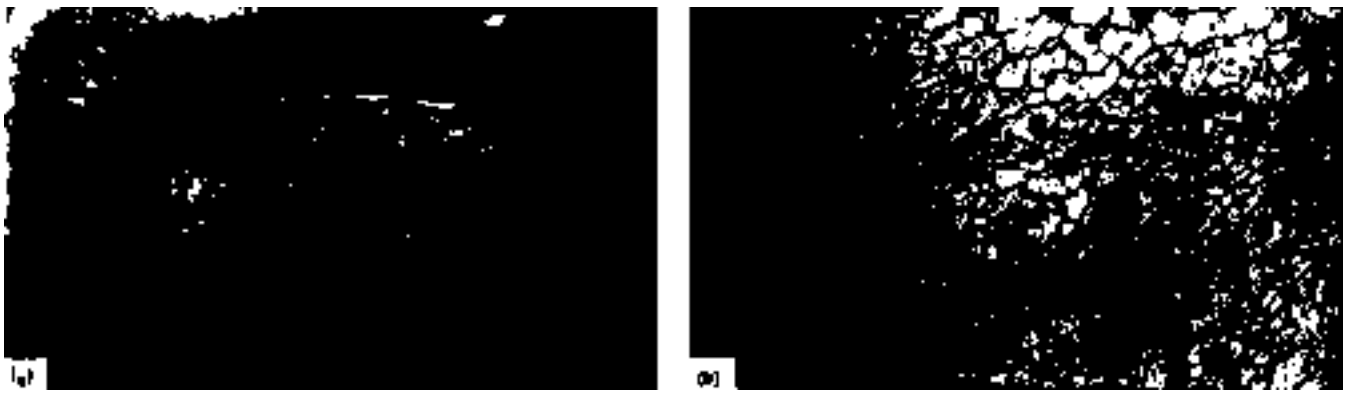


Fig. 11 Creep-induced failure of a boiler plate. (a) A polished cross section of the plate that shows necking, a feature of short-term creep. 2×. (b) Intergranular voids (dark areas) in an area near the fracture surface. Courtesy of B. Gabriel, Packer Engineering Associates, Inc.

Depending on the alloy, temperature, and strain rate, stress-rupture fracture may be macroscopically either ductile or brittle. A macroscale brittle fracture usually is intergranular and occurs with little elongation or necking. In general, lower creep rates, longer rupture times, or higher temperatures promote intergranular fractures. Ductile fractures are transgranular and are typically accompanied by more pronounced elongation and necking. Transgranular creep ruptures, which generally result from high applied stresses (high strain rates), fail by a void-forming process similar to that of microvoid coalescence in dimple rupture. Fractographs of some stress-rupture fractures exhibit both transgranular and intergranular fracture paths. In such instances, it is usually found that the transgranular fractures were initiated by prior intergranular fissures that decreased the cross-sectional area and raised the stress.

Intergranular (IG) creep fracture is the more common fracture path for most stress-rupture failures, depending on stress and temperature. Low stress and/or high temperature increases the likelihood of IG fracture. Sometimes, however, IG cracking may not be readily discernible on the surface of a brittle stress-rupture fracture, due to buildup of surface oxides. If oxide buildup is removed, such cracking usually is visible. It can readily be seen in a polished longitudinal section and may extend beyond the fracture zone. Intergranular creep fracture depends on the nucleation, growth, and subsequent linking of voids on grain boundaries to form two types of cavities: wedge-type (w-type) cavities or isolated, rounded-type (r-type) cavities. Under low-stress conditions, IG fractures occur by void formation at the grain boundaries. These cavities form along grain edges rather than at grain corners. Because they appear to be round or spherical on metallographic cross sections, these voids are sometimes referred to as r-type cavities.

Wedge or w-type are usually associated with cracking at grain-boundary triple points and are also referred to as triple-point or grain-corner cracks in the literature. Wedge-type cracks form at triple points due to grain-boundary sliding and may be promoted by decohesion at interfaces between grain-boundary precipitates and the matrix (e.g., see Fig. 690 from the "Atlas of Fractographs" in *Fractography,* Volume 12 of *ASM Handbook.*) High stresses and lower temperatures promote wedge-crack formation, and wedge cracking at triple points produces a rough fracture surface with identifiable grain-boundary precipitates.

Other significant microstructural features also are associated with stress-rupture fracture. When IG cracking occurs, not only do the crack paths follow grain boundaries at and beneath the fracture surfaces, but also the grains appear equiaxed even after considerable plastic deformation and total elongation; this implies dynamic recrystallization. In contrast, transgranular fracture frequently exhibits severely elongated grains in the vicinity of the fracture—that is, no recrystallization. One can also see severely distorted annealing twins and deformation bands in the face-centered cubic materials. The degree of elongation varies with certain metallurgical factors: notably, prior condition of the metal and its susceptibility to recrystallization or grain-boundary migration under given service or test conditions.

Reference cited in this section

10. D.J. Benac, Elevated-Temperature Failures, Lesson 7, *Principles of Failure Analysis,* Materials Engineering Institute, ASM International

---

Metallurgical Instabilities

Stress, time, temperature, and environment may act to change the metallurgical structure during service and hence to change the strength. The microstructure of a creep or stress-rupture specimen after elevated-temperature exposure rarely resembles the initial microstructure, and prolonged exposure under creep conditions can result in microstructural changes, such as the precipitation of new phases, dissolution or growth of desired phases, grain growth, and so on.

Sharp change of slope in a rupture-life curve can occur from metallurgical instabilities, such as transgranular-intergranular fracture transition, recrystallization, aging or overaging (phase precipitation or decomposition of carbides, borides, or nitrides), intermetallic-phase precipitation, delayed transformation to equilibrium phases, order-disorder transition, general oxidation, IG corrosion, stress-corrosion cracking, slag-enhanced corrosion, and contamination by trace elements. Other metallurgical changes (such as spheroidization and graphitization) and corrosion effects may also

occur during long-term exposure at elevated temperature. Therefore, tests after long-term exposure may be useful in determining the effect of these metallurgical changes on short-term or long-term properties.

Transgranular-Intergranular Fracture Transition. The principal metallurgical factor in stress-rupture behavior is the transition from transgranular fracture to IG fracture. The temperature at which the transition occurs is called the equicohesive temperature (ECT). The transition occurs because the properties of grain-boundary regions differ from those of grains. At low temperatures, grain-boundary regions are stronger than grains, and thus, deformation and fracture are transgranular. At high temperatures, grain boundaries are weaker than grains, and deformation and fracture are largely IG.

The ECT varies with exposure time and stress. For each combination of stress and rupture life, there is a temperature above which all stress-rupture fractures are IG. For shorter rupture lives and lower stresses, this transgranular-to-intergranular transition occurs at higher temperatures. Longer rupture lives and higher stresses occur at lower temperatures. This effect is illustrated by points A and B in Fig. 6; points N, O, and Y represent other types of metallurgical instability. Under certain conditions, both transgranular fracture and IG fracture are found. Consequently, an analysis of rupture-life data or component failure is not complete without a thorough metallographic examination to establish the initial failure mechanism.

Aging. Age-hardening alloys, which include certain stainless steels, heat-resisting alloys, and aluminum alloys (also many others), are characteristically unstable when they are in a state of transition to the stable (equilibrium) condition. Overaging in alloys can occur when they are exposed to elevated operating temperatures and without being loaded. Under creep conditions, temperature-induced and stress-induced atomic migration usually cause aging to continue, resulting in reduced strength. The extent and nature of this change depends on several factors, including the condition of the alloy prior to creep and the temperature, stress, and time of exposure.

Some of the more common high-temperature structural alloys that harden as a result of decomposition of highly supersaturated solid solutions include the Nimonic alloys (Ni-Cr-Al-Ti), austenitic steels that do not contain strong carbide-formers, and secondary-hardening ferritic steels. These alloys are widely used for their creep resistance, but they are not immune to reduced rupture life due to overaging.

Intermetallic-Phase Precipitation. Topologically close-packed phases (commonly known as tcp), such as sigma, mu, and Laves phases, form at elevated temperatures in austenitic, high-temperature alloys. The morphology (shape) of the precipitates determine the effect they have on creep strength. For example, needlelike precipates reduce the toughness and creep strength. Not all of the effects of such phases on rupture life are well known, but some are well understood, such as the behavior of certain nickel-base alloys. The effect of sigma phase (a hard, brittle intermediate phase) on alloy U-700 creep strength at 815 °C (1500 °F) is shown in Fig. 12. Here, a pronounced “break” is shown in the rupture curve starting at approximately 1000 h. Sigma phase was identified in this alloy system and was clearly associated with the failure. However, it was found that sigma did not have a similar effect on certain other nickel-base alloys. It must therefore be concluded that sigma phase does not seem to have a universally deleterious effect on stress-rupture behavior. The amount, location, and shape of sigma-phase precipitation determine whether sigma strengthens or weakens an alloy, or whether it has no effect. Sigma and other intermetallics can severely reduce ductility and toughness on subsequent cooling to room temperature.

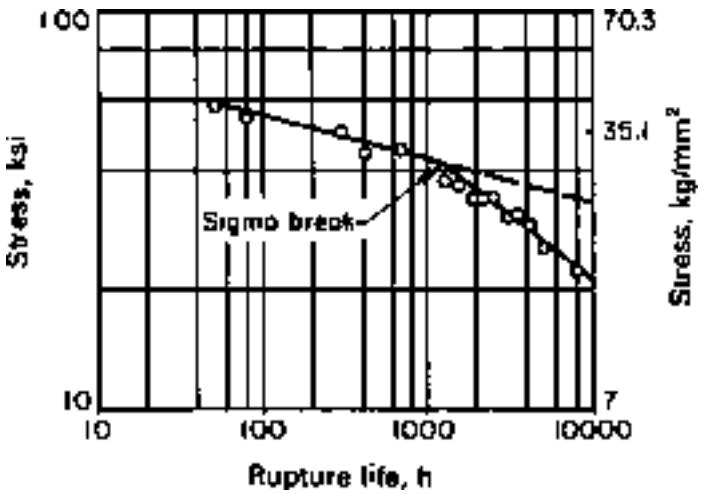


Fig. 12 Logarithmic plot of stress-rupture stress versus rupture life for nickel-base alloy U-700 at 815 °C (1500 °F). The increasing slope of the curve to the right of the sigma break is caused by sigma-phase formation.

The inconsistency of the effect of sigma-phase formation on creep and stress-rupture properties may arise from the simultaneous presence of other phases, such as carbides. The shape, distribution, and chemistry of carbide particles can influence crack initiation and propagation (and hence the resultant stress-rupture ductility and rupture life) in a

pronounced manner. It is improbable that sigma phase affects ductility to a significant degree at low strain rates. Consequently, sigma need not always result in deterioration of creep and stress-rupture properties, unless it is present in relatively large amounts. Because the presence of sigma does not automatically result in decreased rupture life, careful metallographic work must be performed on failures to ensure discrimination between sigma-promoted failures and other types of failures in which sigma or other tcp phases are merely present.

Carbide Reactions. Several types of carbides are found in steels as well as in heat-resisting alloys. Although temperatures and stress affect both carbides within grains and in the grain boundaries, the effects of grain-boundary carbides usually are a more significant factor in altering creep behavior. The presence of carbides is considered necessary for optimal creep and stress-rupture resistance in polycrystalline metals, although subsequent alteration in their shape or breakdown and transition to other carbide forms may be a source of property degradation and resulting failure.

In acicular form, grain-boundary carbides do not appear to act as brittle notch formers that might directly affect rupture life at elevated temperatures, but they may reduce impact strength. Indirectly, compositional changes in the vicinity of carbides can alter rupture strength. In general, acicular M_6C carbides are not believed to affect the properties of nickel-base high-temperature alloys greatly, unless the alloying elements involved in the carbide reaction alter the matrix composition noticeably. Carbide films formed at grain boundaries can decrease rupture life. Figure 13 shows an electron micrograph of grain-boundary carbide films in a Waspaloy forging; these films substantially reduced the stress-rupture life of the alloy.

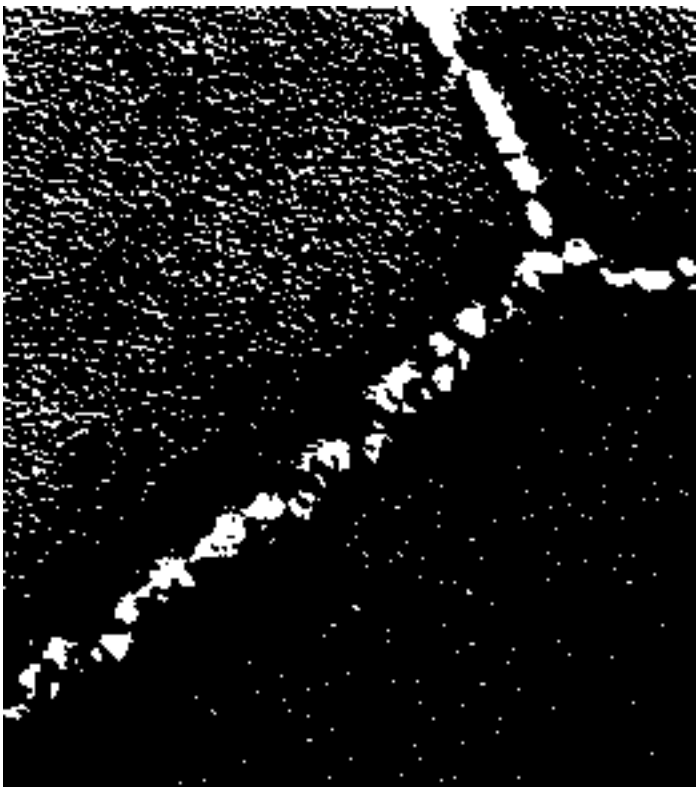


Fig. 13 Grain-boundary carbide films in a Waspaloy forging. The films substantially reduced stress-rupture life. The specimen was electropolished before replication in a solution containing (by volume) 100 parts hydrochloric acid, 50 parts sulfuric acid, and 600 parts methanol. Transmission electron micrograph. 4000×

Carbides, for example, MC, form in steel when carbon combines with reactive elements, such as titanium, tantalum, hafnium, and niobium. During heat treatment and service, the MC carbide can decompose and generate other carbides, such as $M_{23}C_6$, which tends to form along the grain boundary. Carbide formation and microstructural catalogues are discussed in more detail in the article “Elevated-Temperature Life Assessment for Turbine Components, Piping, and Tubing” in this Volume. Specific transitions are known for common Fe-Cr-Mo alloys, such as the sequence of carbide formation in Fig. 14 (Ref 12) for $2\frac{1}{4}$Cr-1Mo steel.

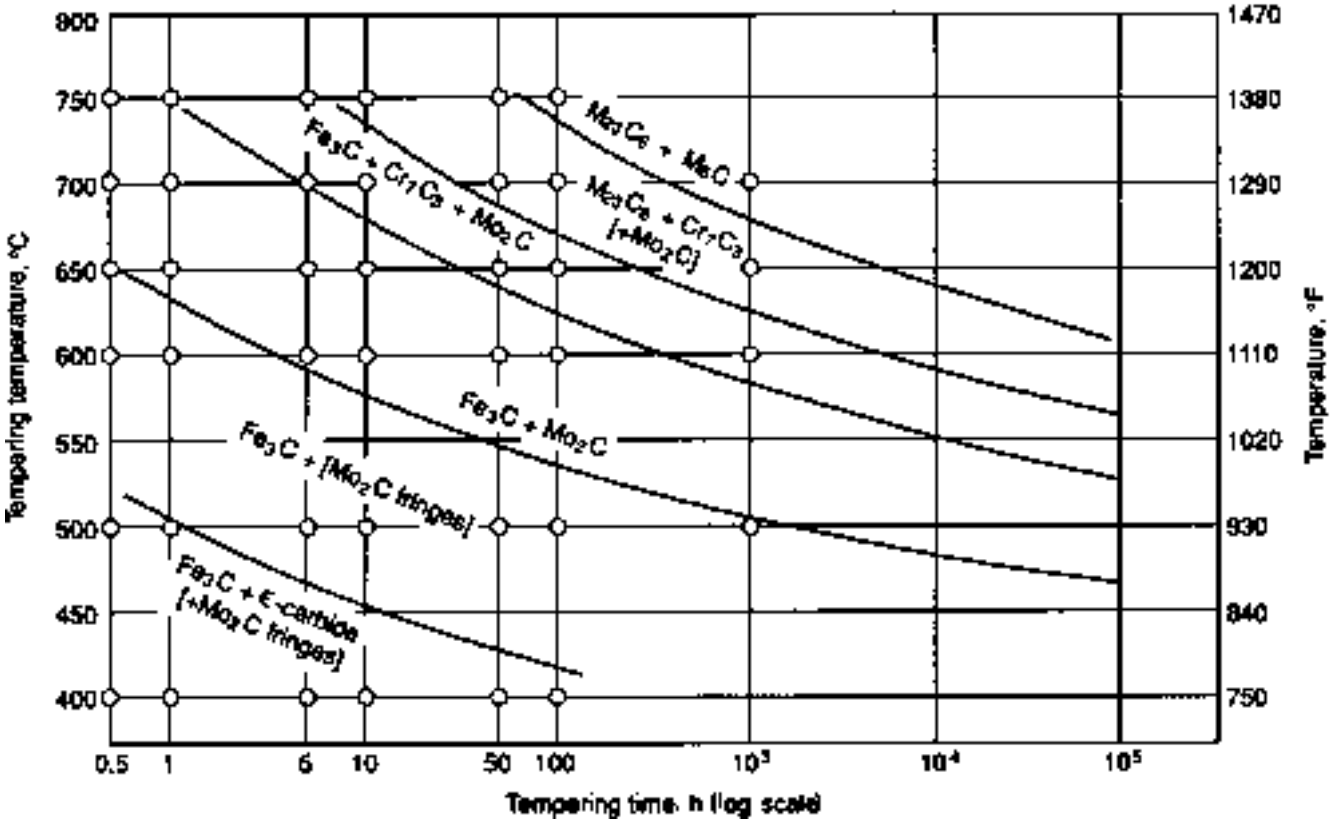


Fig. 14 Isothermal diagram showing the sequence of carbide formation on tempering of normalized 2$\frac{1}{4}$Cr-1Mo steel. Source: Ref 12

Long-term exposure to elevated temperature may affect either short-term or long-term properties. For example, the initial microstructure of creep-resistant chromium-molybdenum steels consists of bainite and ferrite containing Fe_3C carbides, ε-carbides, and fine M_2C carbides. Although a number of different carbides may be present, the principal carbide phase responsible for strengthening is a fine dispersion of M_2C carbides, where M is essentially molybdenum. With increasing aging in service, or tempering in the laboratory, a series of transformations of the carbide phases takes place that eventually transform M_2C into M_6C and $M_{23}C$ (where the M in the latter two metal carbides is mostly chromium). Such an evolution of the carbide structure results in coarsening of the carbides, changes in the matrix composition, and an overall decrease in creep strength. The effect of exposure on the stress-rupture strength of two chromium-molybdenum steels is shown in Fig. 15.

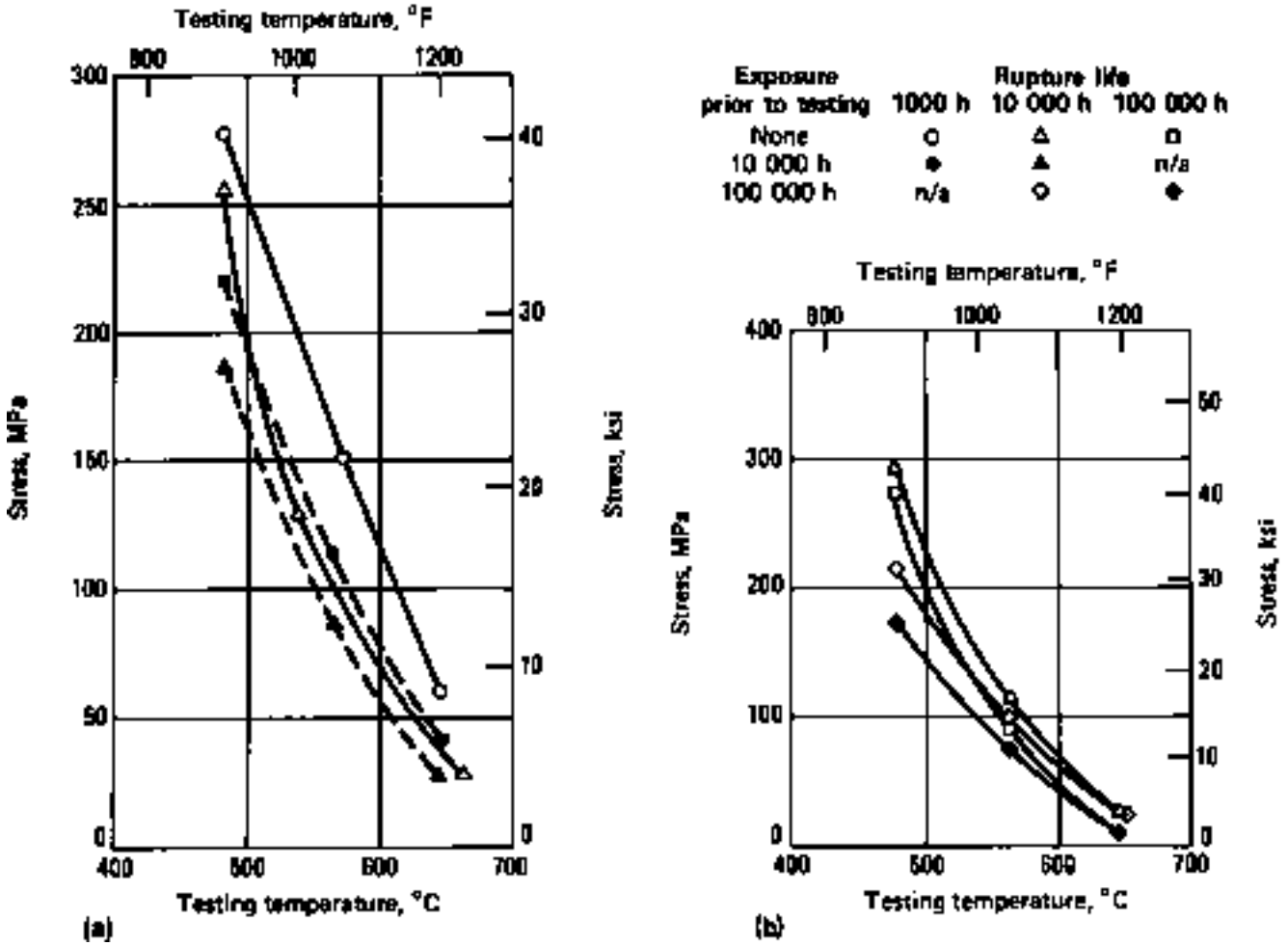


Fig. 15 Effect of elevated-temperature exposure on stress-rupture behavior of (a) normalized and tempered 2Cr-1Mo steel and (b) annealed 9Cr-1Mo steel. Exposure prior to stress-rupture testing was at the indicated test temperatures (without stress) and was 10,000 h long for the 2Cr-1Mo steel and 100,000 h long for the 9Cr-1Mo steel. n/a, data not available at indicated exposure and rupture life

Interaction of Precipitation Processes. The inconsistent behavior of phase precipitation in alloys at elevated temperature is very often due to the fact that carbide and tcp phases do not always react independently but may interact with each other. Because of these reactions, which can become extremely complex, a thorough microstructural examination by the failure analyst must be made (usually with the aid of an electron microscope) before accurate conclusions can be drawn.

Creep-Rupture Embrittlement. The accumulation of creep damage, in conjunction with metallurgical changes, can cause embrittlement and a reduction in rupture ductility. In low-alloy steels, it has similarities with temper embrittlement but is irreversible. In low-alloy steels, the temperature range of creep embrittlement is 425 to 590 °C (800 to 1100 °F) and appears closely related to formation of fine IG precipitates during creep (Ref 13). Impurities such as phosphorus, sulfur, copper, arsenic, antimony, and tin have been shown to reduce rupture ductility.

References cited in this section

12. R.G. Baker and J. Nutting, *J. Iron Steel Inst.,* Vol 192, 1959, p 257–268

13. J.J. Hickey and J.H. Bulloch, *Int. J. Pressure Vessels Piping,* Vol 49, 1992, p 339–386

Thermal Fatigue and Creep Fatigue

In many high-temperature applications, mechanical conditions and creep are not the only sources of strain. Transient thermal gradients within a component can induce plastic strains; if these thermal gradients are applied repeatedly, the resulting cyclic strain can induce failure. This process is known as thermal fatigue, where gradual deterioration and eventual cracking of a material occurs from alternating heating and cooling, during which free thermal expansion is partially or fully constrained.

In the past, thermal fatigue traditionally has been treated as synonymous with isothermal low-cycle fatigue (LCF) at the maximum temperature of the thermal cycle. Life-prediction techniques also have evolved from the LCF literature. More recently, advances in finite-element analysis and testing have made it possible to analyze complex thermal cycles and to conduct thermomechanical fatigue (TMF) tests under controlled conditions. However, isothermal LCF tests and TMF tests should not be considered equivalent, because it has been shown that for the same total strain range, the TMF test can be more damaging under certain conditions than the pure LCF test.

Thermomechanical and combined creep-fatigue loads can result in substantially decreased life at elevated temperatures as compared with that anticipated in simple creep loading or isothermal LCF tests. This effect must be considered in failure analysis. When both creep strains and cyclic (i.e., fatigue) strains are present, damage can accumulate from both and can seriously reduce fatigue life and/or creep life. This complex effect is discussed in more detail in the next article "Thermomechanical Fatigue: Mechanisms and Practical Life Analysis." The following is a brief description of thermal-fatigue cracks, as distinguished from creep-rupture cracks. Fatigue case studies, in which practical systems are analyzed from a fracture mechanics perspective, are also presented in Ref 14.

Thermal-Fatigue Fractures. Thermal-fatigue cracks usually initiate at the surface and then propagate inward, perpendicular to the surface and to the plane of maximum normal stress. They may occur singly or in multiples but often occur in multiples. Because the crack initiates externally, the amount of corrosion or oxidation along the surfaces of a thermal-fatigue crack is inversely proportional to the depth of the crack. In stress-rupture cracks, on the other hand, oxidation is more uniform. Thermal-fatigue cracks usually progress transgranularly.

Excessive creep produces considerable distortion and numerous subsurface cracks, whereas thermal fatigue produces little or no distortion and surface-initiated cracks that usually are multitudes in number. Typically, fracture surfaces of components cracked in stress rupture are irregular and discontinuous, in contrast to the planar, continuous surfaces of thermal-fatigue fractures.

Reference cited in this section

14. M.H. Aliabadi, Ed., *Thermomechanical Fatigue and Fracture,* WIT Press, 2002

Acknowledgments

This article is a revision of the article "Elevated Temperatures" in *Failure Analysis and Prevention,* Volume 11 of *ASM Handbook,* with additional content from *Fatigue and Fracture,* Volume 19, *ASM Handbook* and Materials Engineering Institute Lesson 7, "Elevated Temperature Failure" (Ref 10).

Reference cited in this section

10. D.J. Benac, Elevated-Temperature Failures, Lesson 7, *Principles of Failure Analysis,* Materials Engineering Institute, ASM International

References

1. K.R. Williams and B. Wilshire, Effects of Microstructural Instability on the Creep and Fracture Behavior of Ferritic Steels, *Mater. Sci. Eng.,* Vol 28, 1977, p 289–296

2. M. Wild, Analyse des Zeitstandverhaltens warmfester ferritischer Stahle bei Temperaturen von 450 bis 600 C, *Arch. Eisenhüttenwes.,* Vol 34, Dec 1963, p 935–950

3. R. Widmer and N.J. Grant, The Role of Atmosphere in the Creep-Rupture Behavior of 80Ni-20Cr Alloys, *J. Basic Eng.* (*Trans. ASME D,*) Vol 82 (No. 4), Dec 1960, p 882–886

4. P. Shahinian and M.R. Achter, Creep-Rupture of Nickel of Two Purities in Controlled Environments, *Joint International Conference on Creep* (London), Institute of Mechanical Engineers, 1963, p 7-49 to 7-57

5. M.V. Nathal and L.J. Ebert, Gamma Prime Shape Changes during Creep of a Nickel-Base Superalloy, *Scr. Metall.,* Vol 17, 1983, p 1151–1154

6. T.H. Courtney, Fundamental Structure-Property Relationships in Engineering Materials, *Materials Selection and Design,* Vol 20, *ASM Handbook,* ASM International, 1997, p 354

7. H. Voorhees and M. Prager, Assessment and Use of Creep-Rupture Properties, *Mechanical Testing and Evaluation,* Vol 8, *ASM Handbook,* ASM International, 2000, p 383–397

8. B. Neubauer and U. Wedel, Rest-of-Life Estimation of Creeping Components by Means of Replicas, *Advances in Life Prediction Methods,* American Society of Mechanical Engineers, 1983, p 307–314

9. J.M. Brear et al., *Possibilitic and Probabilistic Assessment of Creep Damage,* ICM 6, Pergamon, 1991

10. D.J. Benac, Elevated-Temperature Failures, Lesson 7, *Principles of Failure Analysis,* Materials Engineering Institute, ASM International

11. H. Nelson, “Fractography in the Failure Analysis of Engineering Structures,” ASM annual meeting, 1997

12. R.G. Baker and J. Nutting, *J. Iron Steel Inst.,* Vol 192, 1959, p 257–268

13. J.J. Hickey and J.H. Bulloch, *Int. J. Pressure Vessels Piping,* Vol 49, 1992, p 339–386

14. M.H. Aliabadi, Ed., *Thermomechanical Fatigue and Fracture,* WIT Press, 2002

Thermomechanical Fatigue: Mechanisms and Practical Life Analysis

Stephen D. Antolovich, Washington State University; Ashok Saxena, Georgia Institute of Technology

Introduction

THERMOMECHANICAL FATIGUE (TMF) refers to the process of fatigue damage under simultaneous changes in temperature and mechanical strain. Fatigue damage at high temperatures develops as a result of inelastic deformation where the strains are nonrecoverable. Therefore, TMF damage is complex, as it may accumulate over a range of temperatures and strains under both steady-state and/or transient conditions. Thermomechanical fatigue is also often a low-cycle fatigue issue, because relatively infrequent transients (compared to steady-state exposures) can contribute to the accumulation of TMF damage in various kinds of equipment, such as jet engines, land-based turbines for power generation, and pressure vessels.

For example, during straight and level flight, aircraft jet engines have essentially constant temperatures and imposed loads, where steady-state creep (and the environment) are the primary damage mechanisms. During takeoff and landing, however, the transient demand for more power output induces load and temperature changes, which thus impose fatigue damage. In this case, it is customary to consider each flight to be one fatigue cycle with an imposed hold time, so in addition to creep and fatigue acting independently, there are creep-fatigue interactions (which are also influenced to a very significant degree by the environment). Obvious transients are also introduced during start-up and shut down. Here it is clear that there are strong thermal gradients and independently applied mechanical loads, which may also occur during flight during times at which operating conditions change.

Furthermore, the damage modes may interact in a TMF cycle in ways that are never encountered in conventional fatigue testing, which makes life prediction under TMF conditions particularly challenging. Nevertheless, techniques have been proposed, validated, and introduced into practice that provide reasonable assessment of remaining life. Life prediction in the TMF regime is particularly important in high-temperature turbines (jet engines and land-based turbines), the power-generation industry (both conventional and nuclear), and in the automotive industry, to name a few. Significant advances can be made by combining mechanistic studies with increasingly sophisticated finite-element thermal stress analysis. Development of more refined engineering-based models should result in less reliance on component simulation, which is subject to some uncertainty.

It has been customary to characterize materials in terms of their fatigue and creep properties (including creep/fatigue/environment properties) and to compute the life of components such as turbine blades using experimental data and models that implicitly or explicitly describe damage evolution as a function of the number of cycles (Ref 1, 2, 3). While such experimental and analytical information is essential for life prediction, it is not always sufficient. It has been recognized that some of the most damaging events occur during transient operation when temperature as well as loads and/or strains are changing independently. This is the process of TMF, which is briefly reviewed in this article with a practical example of life assessment.

References cited in this section

1. H.D. Solomon, G.R. Halford, L.R. Kaisand, and B.N. Leis, Ed., *Low Cycle Fatigue*, STP 942, American Society for Testing and Materials, l988

2. S.D. Antolovich, S. Liu, and R. Baur, Low Cycle Fatigue Behavior of René 80 at Elevated Temperatures, *Metall. Trans.*, Vol 12A, 1981, p 473–481

3. S.D. Antolovich and B. Lerch, Cyclic Deformation and Fatigue in Ni-Base Alloys, *Superalloys, Supercomposites, and Superceramics*, J.K. Tien and T. Caufield, Ed., Academic Press, 1989, p 363–411

Stephen D. Antolovich, Washington State University; Ashok Saxena, Georgia Institute of Technology

Basic Descriptions of TMF

In thermomechanical fatigue, the path is defined by plotting the mechanical strain (i.e., the total strain minus the thermal strain) as the independent variable and the load (stress) as the dependent variable. There are two possible types of loading: in-phase and out-of-phase cycling. For in-phase cycling, the direction of imposed mechanical strain and temperature change in the same direction (e.g., the temperature increases with increasing mechanical strain). In out-of-phase cycling, one variable decreases with an increase in the other (e.g., the temperature decreases with increasing mechanical strain). These two cases are illustrated schematically in Fig. 1. More complex paths are used to simulate practical, real-world situations. The next level of complexity may be represented by a so-called "diamond" cycle, which has both in-phase and out-of-phase cycle character, (Fig. 2).

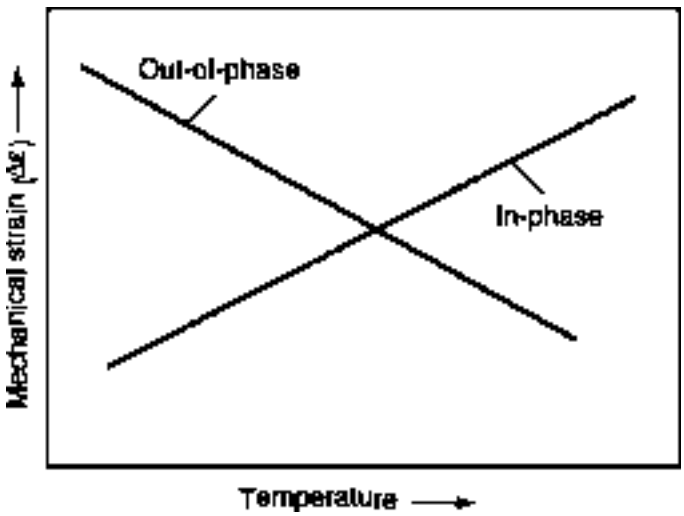


Fig. 1 In-phase and out-of-phase thermomechanical fatigue cycles. The term "phase" refers to the nature of the relationship between the mechanical strain and the temperature.

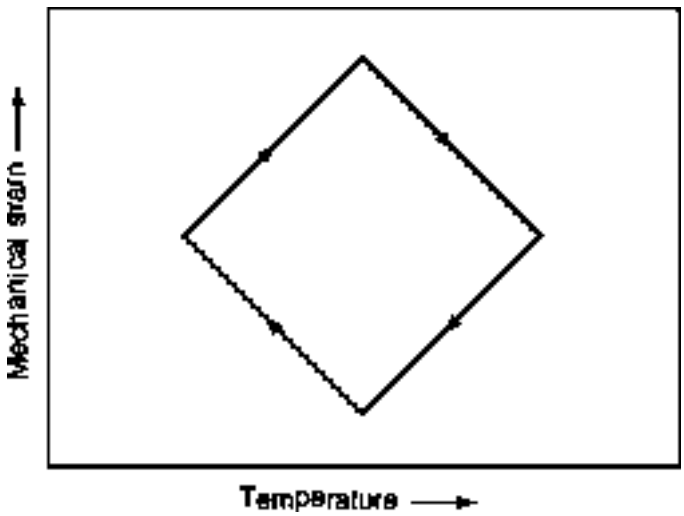


Fig. 2 "Diamond" cycle, showing both in-phase and out-of-phase character

In general, experiments are carried out to simulate the strain-temperature path of critical components in critical locations. The fundamental problem is that different thermomechanical cycles will produce different damage interactions, making it difficult to establish empirical representations that have a wide regime of applicability. In addition, susceptibility to TMF

damage can be influenced by the complex nature of manufacturing forms, such as the interface between a high-temperature coating and the substrate. For example, Fig. 3 is a micrograph of TMF cracks that initiate from the coating-substrate interface. This is a particular problem of TMF damage. It is thus important to understand what kinds of damage can occur in a given material and what the damage interactions will be for a given TMF cycle.

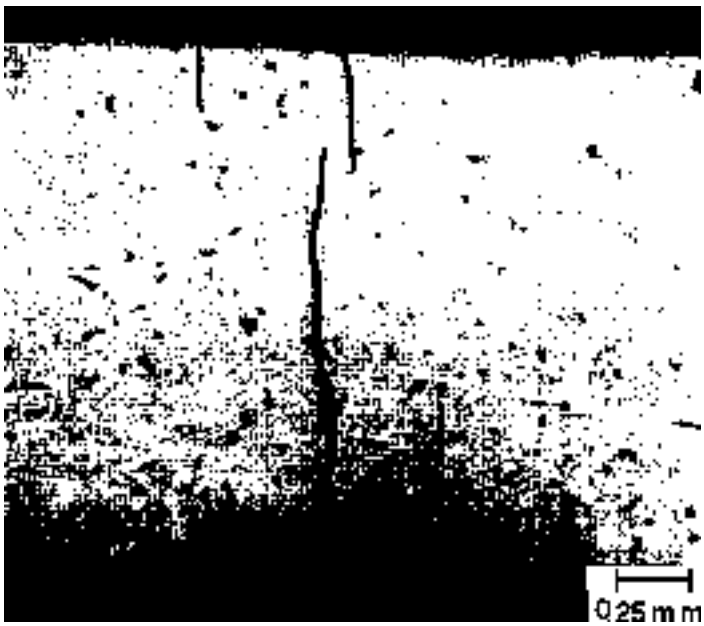


Fig. 3 Coating-initiated TMF cracks propagating into the substrate material of a thin-walled blade. Courtesy of Ashok Koul, Life Prediction Technologies, Inc.

With this knowledge it should be possible to:

- Define the conditions for which empirical, experimentally based, life-prediction equations may be used
- Develop an appropriate form for life-prediction equations based on fundamental mechanisms

---

Thermomechanical Fatigue: Mechanisms and Practical Life Analysis

Stephen D. Antolovich, Washington State University; Ashok Saxena, Georgia Institute of Technology

---

TMF Mechanisms

In many cases, the out-of-phase cycle is most damaging. This may be understood in terms of a very strong environmental effect. At high temperatures, significant environmental damage occurs even though the strain is low. For example, oxygen may penetrate along boundaries (grains, twins, subgrains) and cause a general weakening. As the temperature is lowered, the strain (and stress) increases and the weakened boundary cracks some distance into the metal. This process is repeated until final failure. The effect of the nature of the TMF cycle was illustrated in a simple, yet revealing experiment (Ref 4). In this experiment, René 80, a polycrystalline nickel-base turbine blade was used. In one experiment, meant to simulate out-of-phase cycling in a very simple way, a specimen was fatigued to half of its life (i.e., 272 cycles) at 1033 K, a plastic strain range of 0.5%, and an *R*-ratio ($\varepsilon_{min}/\varepsilon_{max}$) of 0. Cycling was continued at 298 K under the same conditions: it failed on the first loading cycle, even though the life of such a specimen is predicted to be thousands of cycles without the high-temperature precycling. The reason for the reduced life is that oxidized boundaries were created that cracked upon loading at room temperature.

In other materials, that is, those that are more sensitive to creep damage than to environmental degradation, the in-phase cycle may be more damaging. Here the high imposed strain at the high temperature can cause more creep-type damage such as voids on the grain boundary. As the temperature is lowered, the strain (and stress) is also reduced, and the damage that was induced at high temperature does not propagate or induce further microcracking. This process is essentially uniformly distributed, although the number and size of voids may increase with time and cycle (i.e.,

continuous nucleation and growth) or the number may be constant (instantaneous nucleation) with void size increasing with time and cycle. The process continues until the boundaries have lost their structural integrity.

In this section, ways in which damage can interact at high and low temperatures are pointed out, and microstructurally based models are developed and put into parametric form. A more detailed discussion is provided in Ref 5 and 6. These examples are meant to be illustrative and not exhaustive; they provide a way in which to view the TMF process and a methodology for developing engineering models.

Stable Precipitate Structure and Plasticity/ Oxidation. In certain applications, such as disks and turbine blades in jet engines, the temperatures and stresses are such that the microstructure remains stable. The essential features of the microstructure of blade alloys are:

- Large cuboidal γ′ precipitates for creep resistance
- Small spherical γ′ precipitates to provide strength
- Large grains to reduce the intrinsic creep rate by reducing the grain-boundary area and therefore the grain-boundary sliding
- Grain-boundary carbides (in polycrystalline materials) to reduce the creep rate by inhibiting grain-boundary sliding

The way in which cracking occurs could be envisioned as oxidation of a grain boundary at high temperatures and cracking by slip-band impingement at low temperatures (Ref 7). A crack initiates when the extent of penetration is about the same order of magnitude as the slip-band spacing. For Waspaloy, it has been shown (Ref 8) that the slip-band density is given by:

$$i = A\Delta\varepsilon_p^{-\delta} \qquad \text{(Eq 1)}$$

where i is slip-band spacing, $\Delta\varepsilon_p$ is plastic strain range at the lower temperature, and A and δ are experimentally determined parameters.

Oxidation proceeds according to the approximate formula:

$$x = \sqrt{Dt} \qquad \text{(Eq 2)}$$

where x is penetration distance, D is diffusion coefficient, and t is time to form a microcrack.

These equations (along with some simple substitutions) may be combined to yield the following result:

$$N_i = \frac{A^2 \nu}{D_0} \cdot \exp\frac{Q}{RT_{eff}} \cdot \Delta\varepsilon_p^{-2\delta} \qquad \text{(Eq 3)}$$

In Eq 3, D_o is the pre-exponential part of the diffusion coefficient Q is, the activation energy for diffusion of oxygen, and v is frequency. It is worth noting that this equation will have the same form regardless of whether the first crack to initiate can grow in a stable manner or if the initiation process must be repeated N_i times to form a stable microcrack since the same basic kinetic equation (i.e., Eq. 1) applies.

From an engineering point of view, Eq 3 can be represented by a set of parameters that can be experimentally derived from isothermal experiments:

$$N_i = C_1(m, \nu, T) \cdot \exp\frac{B}{T_{eff}} \cdot \Delta\varepsilon_p^{-\beta} \qquad \text{(Eq 4)}$$

where C_1 is a material-frequency-, and temperature-dependent parameter; B is a material-dependent parameter (no temperature dependence); and T_{eff} is the effective temperature over TMF cycle.

The effective temperature is the temperature at which the diffusional rate is the same as the average rate over the temperature range from the minimum temperature in the TMF cycle, T_L, to the maximum temperature, T_H. The result, after some straightforward integration, is:

$$T_{eff} = \frac{-B}{-\ln(\Delta T) + \ln\left(\int_{T_L}^{T_H} \exp\left(\frac{-B}{T}\right) dT\right)} \qquad \text{(Eq 5)}$$

By the nature of diffusional laws T_{eff} would be expected to be close to the maximum temperature in the TMF cycle. The variation of T_{eff} with the maximum temperature is shown in Fig. 4 for three cases. In two cases (curves 1 and 2), everything is the same except that the minimum temperature is different. In curve 3, everything is the same as for curve 2 except that the activation energy is changed from 100 to 20 kcal/mol. It can be seen that, regardless of the assumptions that are invoked, the effective temperature is close to the maximum. Furthermore, the variation in T_{eff} versus T_H is approximately linear. These trends are most closely followed for high activation energies but hold fairly well even for activation energies as low as 20 kcal/mol. However, it is well known that the activation energies of most of the damaging diffusional processes of interest lie in the range of 50 to 100 kcal/mol. Consequently, the curves shown in Fig. 4 provide an envelope for virtually all cases of interest.

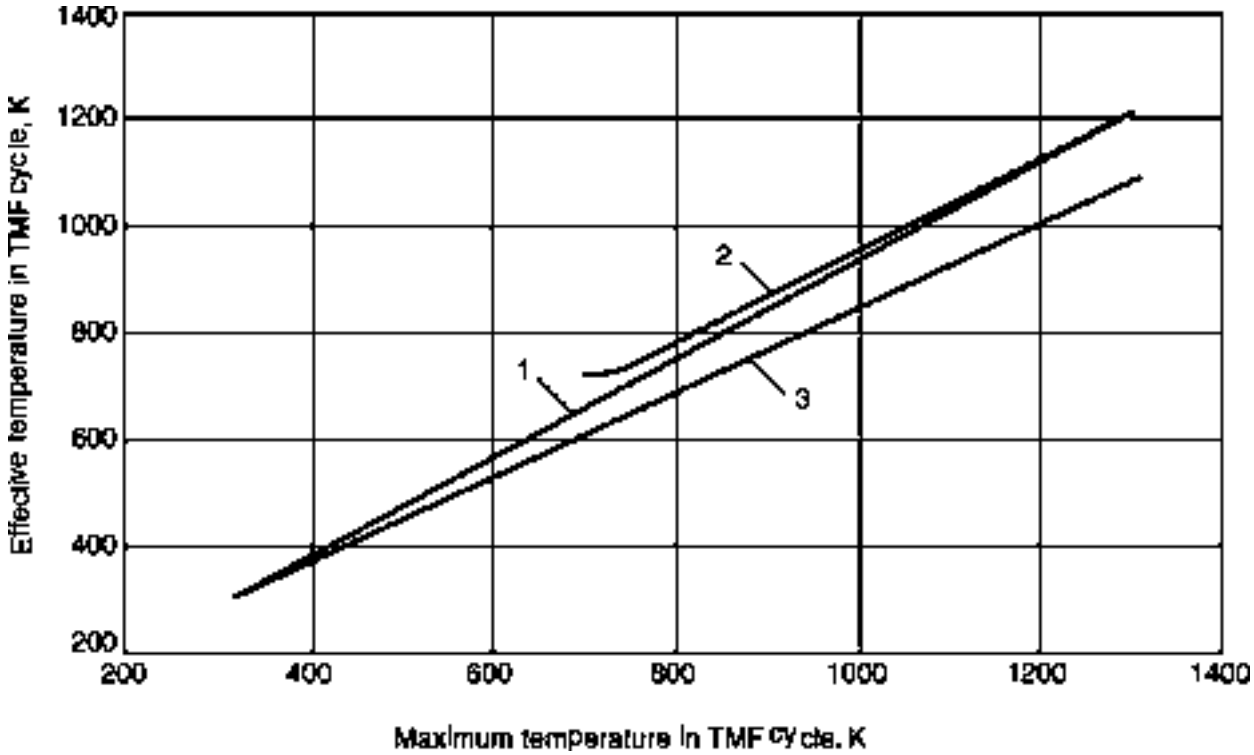


Fig. 4 Effective temperature in a TMF cycle as a function of the maximum temperature. For curves 1 and 2, the activation energy is 100 kcal/mol, but the minimum temperature is different. Curve 3, everything is the same as for curve 2 except that the activation energy is changed from 100 to 20 kcal/mol. These curves form an envelope for virtually all conditions encountered in practice.

For Eq 4 to apply, the main condition that must be satisfied is that the stress at the tip of the slip band must exceed the cohesive strength of the oxygen-degraded grain boundary. That condition is simply stated as:

$$\sigma_0 = \left(\frac{\sigma_c a_0 E}{2\pi L}\right)^{1/2} \cdot \frac{1}{1-\nu} \qquad \text{(Eq 6)}$$

where σ_0 is the stress at the low temperature in the cycle, E is Young's modulus, σ_c is the oxide cohesive strength, a_0 is the atomic spacing, and L is the grain size of the alloy.

Equation 6 represents a plasticity-based Griffith equation. It is interesting to note that the model described previously is more likely to apply to coarse-grained materials than to fine-grained ones since the spacing of slip bands will be greater and the stresses at the tip of the slip band will be higher.

Thermomechanical Fatigue Life Model with Stable Precipitate Structure and Plasticity/Oxidation. Conditions for this situation are assumed to be:

- In-phase cycle
- Stresses at low temperature below oxide cohesive strength
- Damage occurring at the high temperature
- No interaction between the high- and low-temperature modes

In this case, there are very few slip bands formed at the lower temperatures because of the in-phase cycle, and damage in this case may be viewed as that which would occur at the highest temperature in the absence of any low-temperature damage. The life for this situation has been computed (Ref 2, 5), and the result is expressed as:

$$N_i = C_2\left(\frac{\nu}{1+\nu t_h}\right)\exp\frac{Q}{RT}\cdot\Delta\varepsilon_p^{-8n'} \qquad \text{(Eq 7)}$$

where all terms have been previously defined and where t_h is the hold period and n' is the cyclic strain hardening exponent. Note that the expression is given in terms of plastic strain range.

Systems to which this model would be likely to apply would be those that have a stable precipitate structure (i.e., low mismatch between matrix and precipitate) and that are prone to damage by oxidation of grain boundaries. Examples

would be Waspaloy, a disk alloy, and Udimet 500, a blade alloy. It would also be expected to apply to alloys with high mismatch values such as René 80 or Mar-M 246 as long as the maximum temperature was lower than the temperature at which coarsening begins (i.e., ~850 °C, or 1560 °F).

Thermomechanical Fatigue Life Model with Metastable Precipitates and Plasticity/ Oxidation. Conditions for this situation are again assumed to be:

- In-phase cycle
- Stresses at low temperature below oxide cohesive strength
- Damage occurring at the high temperature
- No interaction between the high- and low-temperature modes

To develop Eq 7, it was necessary to assume that the structure was stable and the stress in a cycle was also stable. At high temperatures, however, especially when there is a hold at the maximum strain, there is a continual coarsening that occurs, Fig. 5, and the stress decreases continually.

Fig. 5 Precipitate structure of René 80 cycled in strain-controlled low-cycle fatigue at 1254 K. Note the "globular" appearance and coarsening of the precipitates and the dislocation networks around the precipitates. The small precipitates have been consumed by growth of the large ones. Source: Ref 2

A fruitful topic for research would be to compute from first principles the decay in the maximum stress (σ_{max}) as a function of cycle number. In a very approximate way, this can be attempted by assuming that the coarsening kinetics follow the classical ripening equation and that dislocation motion is controlled by Orowan looping. The fundamental damage equation is assumed to be of the form first proposed for high-temperature low-cycle fatigue (LCF) (Ref 2):

$$\sigma_{max} \cdot \ell_i^{\alpha} = C_3 \qquad \text{(Eq 8)}$$

where ℓ_i is the diffusion length at initiation and α is an empirical exponent. For Orowan looping:

$$\sigma_{max} = \frac{2Gb}{\ell} \qquad \text{(Eq 9)}$$

where ℓ is the distance between barriers, *G* is the shear modulus, and b is the Burgers vector. If it is assumed that the precipitates form a square array of cubes of size *S* and if it is further assumed that during coarsening the precipitates retain a cubic array, then the precipitate size is related to the interprecipitate spacing by:

$$\ell = SV_f^{-1/3} \quad \text{(Eq 10)}$$

where V_f is the volume fraction of precipitates.
Combining Eq 8, 9, 10 gives the following result for the life under TMF conditions:

$$N_s = \nu C_3^{1/\beta} \cdot \frac{D_A^{K/\beta}}{D^{\alpha/2\beta}} \cdot \left(\frac{\alpha_\square}{2GbV_f^{1/3}} \right)^{1/\beta} \quad \text{(Eq 11)}$$

Again, this equation can be put in parametric form for use in engineering situations with the constants determined by regression analysis and parameterized in terms of the important variables. The engineering equation would be:

$$N = \nu C(m) \exp \frac{A(m)}{T_{eff}} \quad \text{(Eq 12)}$$

As for the preceding models, the parameters can be determined by experimentation.
TMF Life Model with Carbide Precipitation and Cracking. Conditions for this situation are assumed to be:

- Carbides precipitate as $M_{23}C_6$ on grain boundaries at high temperature.
- Carbide coarsening obeys Oswald ripening.
- Carbides are cracked at low temperatures by impingement of slip bands where the stress is concentrated.
- Loading type is out-of-phase cycling.

A situation such as that described previously could arise for a coated component in which the toughness of the coating exceeds that of the carbides on the boundaries. For this situation equations very similar to those for damage by oxidation of the boundaries can be deduced. The results are given by:

$$N_s = \frac{AV^3}{\alpha_\tau D_\square} \cdot \exp \frac{Q}{RT_{eff}} \cdot \Delta\varepsilon_p^{-\lambda\delta} \quad \text{(Eq 13)}$$

where all quantities have been defined previously. An equation similar to Eq 7 would define the limits of applicability for Eq 13 except that the carbide cohesive strength would replace the cohesive strength of the oxide. This model would apply to coated alloys that form grain-boundary films of $M_{23}C_6$ at elevated temperatures and that are protected by a tough coating. Examples would be René 80 and René 77, both of which have been used for blades in the past.

References cited in this section

2. S.D. Antolovich, S. Liu, and R. Baur, Low Cycle Fatigue Behavior of René 80 at Elevated Temperatures, *Metall. Trans.*, Vol 12A, 1981, p 473–481

4. B. Boursier, M.S. Thesis, University of Cincinnati, 1982

5. D.L. McDowell, S.D. Antolovich, and R.L.T. Oehmke, Mechanistic Considerations for TMF Life Prediction of Nickel-Base Superalloys, *J. Nucl. Energy Des.*, Vol 133, 1992, p 383–389

6. H. Sehitoglu, Thermal and Thermomechanical Fatigue of Structural Alloys, *Fatigue and Fracture*, Vol 19, *ASM Handbook*, ASM International, 1996, p 527–556

7. B.A. Lerch, S.D. Antolovich, and N. Jayaraman, A Study of Fatigue Damage Mechanisms in Waspaloy from 25°C–800°C, *Mater. Sci. Eng.*, 1984, p 151–165

8. D. Diederich, B. Lerch, and S.D. Antolovich, Observations of Low Cycle Fatigue and Fatigue Crack Propagation Substructures of Waspaloy, *Proc. VIII Inter-American Conference on Materials*, 25–29 June 1984, p 7.1–7.6

Stephen D. Antolovich, Washington State University; Ashok Saxena, Georgia Institute of Technology

Experimental Techniques

Equipment. The basic equipment required is a low thermal mass programmable furnace, a closed-loop servohydraulic mechanical test machine, and suitable extensometry. A radio frequency (RF) heat source has essentially no thermal inertia and is thus ideal, assuming that the material being tested couples to RF energy. High-quality closed-loop test equipment is available from several manufacturers, and high-temperature extensometers can readily be constructed in the lab or purchased commercially.

Control Logic. The temperature signal is processed and used as the command signal for the strain-control channel to produce a desired strain-temperature profile. However, it is customary to have a cycle in which temperature and mechanical strain are correlated. To accomplish this, the gage displacement corresponding to the thermal expansion of the specimen at the current temperature is computed through knowledge of the coefficient of thermal expansion, the temperature, and the gage length. It is then transformed into an equivalent electrical signal to which a signal corresponding to the total mechanical strain is added to get the total signal corresponding to the desired mechanical strain. In essence an adaptive (control "on the fly"), combinatorial control (mixing of two signals) system is used. The output may be read in terms of the mechanical strain by suitable signal conditioning.

Component-Simulation Testing. Life calculation for high-temperature fatigue is challenging in itself. When the complication of simultaneously varying temperature and strain is imposed, the difficulties are increased significantly. For that reason, component simulation testing is often carried out. In this, the TMF cycle is established either by sophisticated calculation or, in some cases, by experimental observation. It is apparent, however, that it is not always possible to carry out an experiment with a high degree of confidence. The high-temperature rotating components in a jet engine come readily to mind.

Provided the loads/strains and temperatures can be accurately ascertained, a fairly reliable component test should be possible. However, note that in the definition of the TMF cycle time is not an explicit variable. Tests are usually conducted so as to maintain constant temperature in the sample at any instant, and cooling is done so that gradients are avoided. However, in practice, gradients are usually present. It is well known (Ref 9) that such gradients can be the source of very large stresses. Thus if these gradients are not duplicated, then the simulation will produce more benign results than are warranted based on the real service conditions. In addition, high-temperature damage accumulation mechanisms such as oxidation and creep depend not only on temperature and stress, but also on time. Thus it is important to precisely duplicate cycle times when carrying out a simulation. For these reasons, in many cases it appears as if the life computed from a sophisticated model can yield results that are at least as reliable as component or simulation testing. In the following sections, sophisticated models are used to calculate the life under TMF conditions.

Reference cited in this section

9. B.A. Boley and J.H. Weiner, *Theory of Thermal Stresses*, John Wiley and Sons, 1960

Thermomechanical Fatigue: Mechanisms and Practical Life Analysis

Stephen D. Antolovich, Washington State University; Ashok Saxena, Georgia Institute of Technology

Case Study: Prediction of Residual Life in a Turbine Casing

The following is a practical application of TMF life prediction. Additional information on TMF and TMF testing is contained in Ref 6 and 10, respectively. Fatigue case studies, in which practical systems are analyzed from a fracture mechanics perspective are also presented in Ref 11.

Casings for a ship service turbine generator (SSTG) are cast from 2.25Cr-1Mo steel and are designed to provide service for 30 years. The regions of the casings in the vicinity of the steam chest and the steam passageways are subjected to high transient thermal stresses during start-up and moderate steady-state mechanical stresses during sustained operation. The metal temperature during steady-state operation is approximately 538 °C (1000 °F), which is in the creep regime; therefore, creep-fatigue crack initiation and crack growth are the dominant damage mechanisms. Cracks have been found to occur in casings that have been removed from service after being in operation for 25 years or longer (Ref 12). Thus, to ensure the integrity of the casings, it is necessary to inspect them periodically during service and estimate their remaining life or the next inspection interval. A methodology for estimating the remaining-life/inspection interval of such turbine casings was developed by Saxena et al. (Ref 13) and is briefly described in this example.

Figure 6 shows the geometry and approximate dimensions of the SSTG casing including a schematic of the steam passageways where the majority of the service cracks were found during inspection. The SSTG start-up loading diagram is shown in Fig. 7(a) (Ref 14). The transient temperature differentials computed for the initial loading portion (0 to 20 min) of this diagram are most severe. A typical SSTG unit experiences 120 cold starts per year and once started, runs on the average of 72 h before being shut down. This loading cycle is used in the estimation of remaining life and inspection intervals. A schematic of the stress history that results from this turbine-loading procedure is shown in Fig. 7(b). Note that the turbine casings experience significant compressive stresses during start-up. The main focus of this analysis was on predicting the crack growth behavior because it was concluded that the majority of the casings that were of concern were in service long enough to develop cracks. The casing that is used as an example here was retired from service after 27 years of operation due to cracks that were found during inspection.

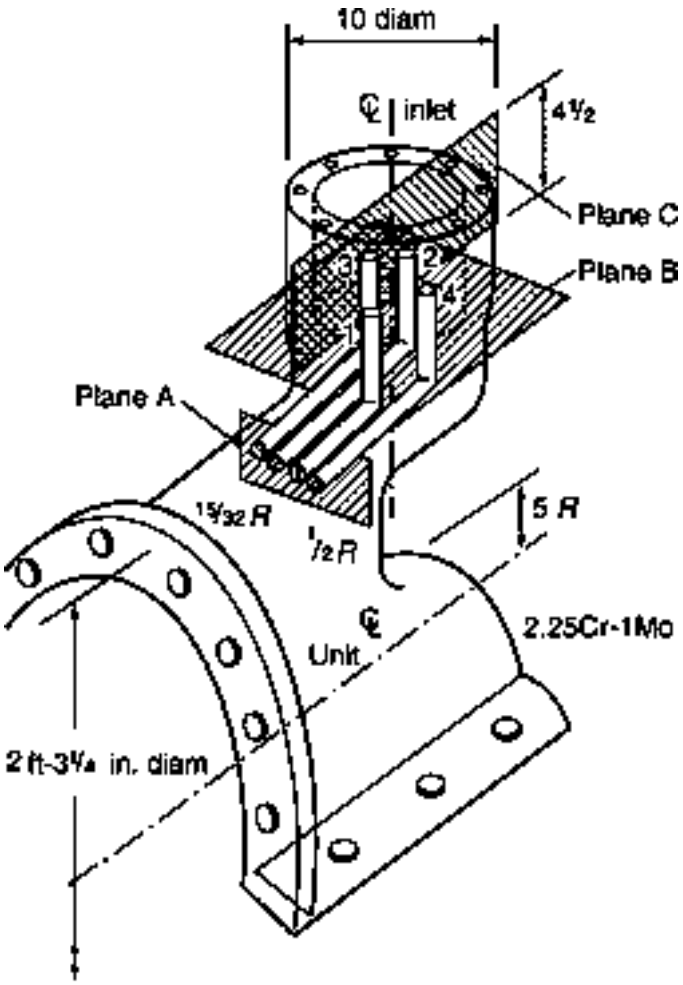


Fig. 6 Ships's service turbine generator casing showing critical planes A, B, and C on which stresses were determined by finite-element analysis during transient and steady-state conditions. Dimensions given in inches. Source: Ref 13

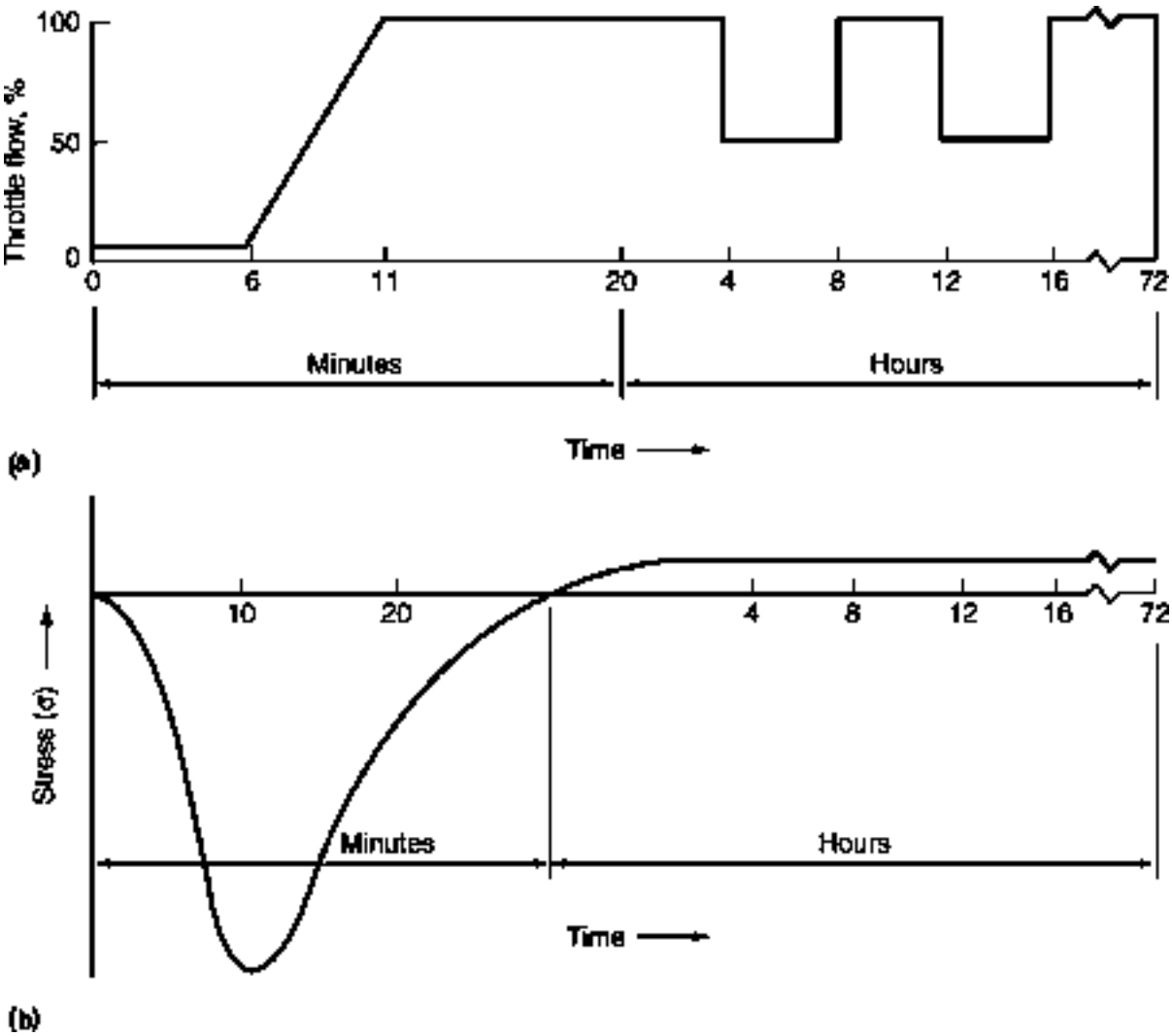


Fig. 7 Ships's service turbine generator. (a) Load diagram showing the steam throttle flow in percent. (b) A schematic of the resultant transient and steady-state stresses as a function of time. The periodic reduction in throttle flow is due to reduced power requirements during some portions of service duty cycle. However, these fluctuations are shown to have negligible effect on the magnitude of the stresses. Source: Ref 14

Stress Analysis of the Casing. The stresses during the transient and steady-state portions of the SSTG operation were estimated using finite-element analysis on the three critical planes labeled A, B, and C in Fig. 6. The details of the analysis are described elsewhere (Ref 15), but the key results are summarized and used in the analysis presented here. In planes A and B, the principal stresses along several directions (or planes) were obtained as a function of distance from the passageway bores. A typical elastic stress distribution is shown in one direction within plane A. In the analysis described here, the most important is the maximum principal stress that occurs in approximately the hoop direction. Therefore, only the hoop stress is considered in the analysis. Since most of the passageway cracks found in the casing were indeed along the radial-axial plane, the choice of the hoop stress seems appropriate for the crack growth analysis. The stress distribution along plane 3L (Fig. 8) is used to analyze the behavior of cracks coming out of the two outermost holes. Stress distributions were available for various times during the transient period. However, for a fatigue analysis, the most relevant stress distribution is the one corresponding to the instance when the bore surface stress is at its peak value. Both the instance at which the stress is maximum and the instance at which it is minimum are of interest because they define the range of the maximum fatigue stresses that initiates and drives the cracks. The peak elastic stress in Fig. 8 is consistently above the yield strength of the 2.25Cr-1Mo steel at the operating temperatures and the elastic stress range exceeds twice the cyclic yield strength. Therefore, the material in the vicinity of the steam passageway bores undergoes plastic deformation during each cold start. An elastic-plastic analysis was conducted that led to the stress range, $\Delta\sigma$, distributions shown in Fig. 9 for several other directions within the plane A.

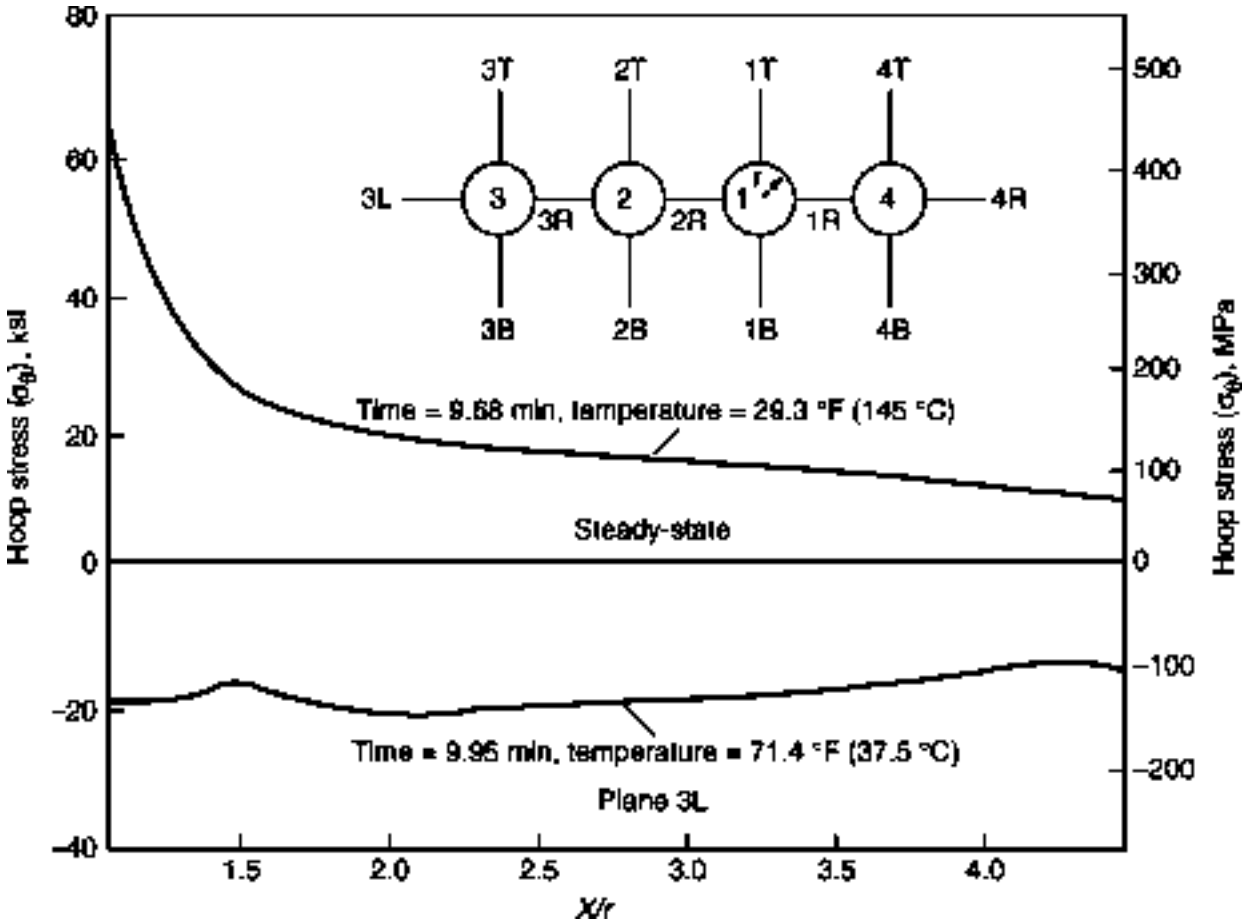


Fig. 8 Results of the elastic finite-element analysis for direction 3L along plane A. *X* is the distance from the bore hole, and *r* is the radius of the bore hole used to normalize the distance. Source: Ref 13

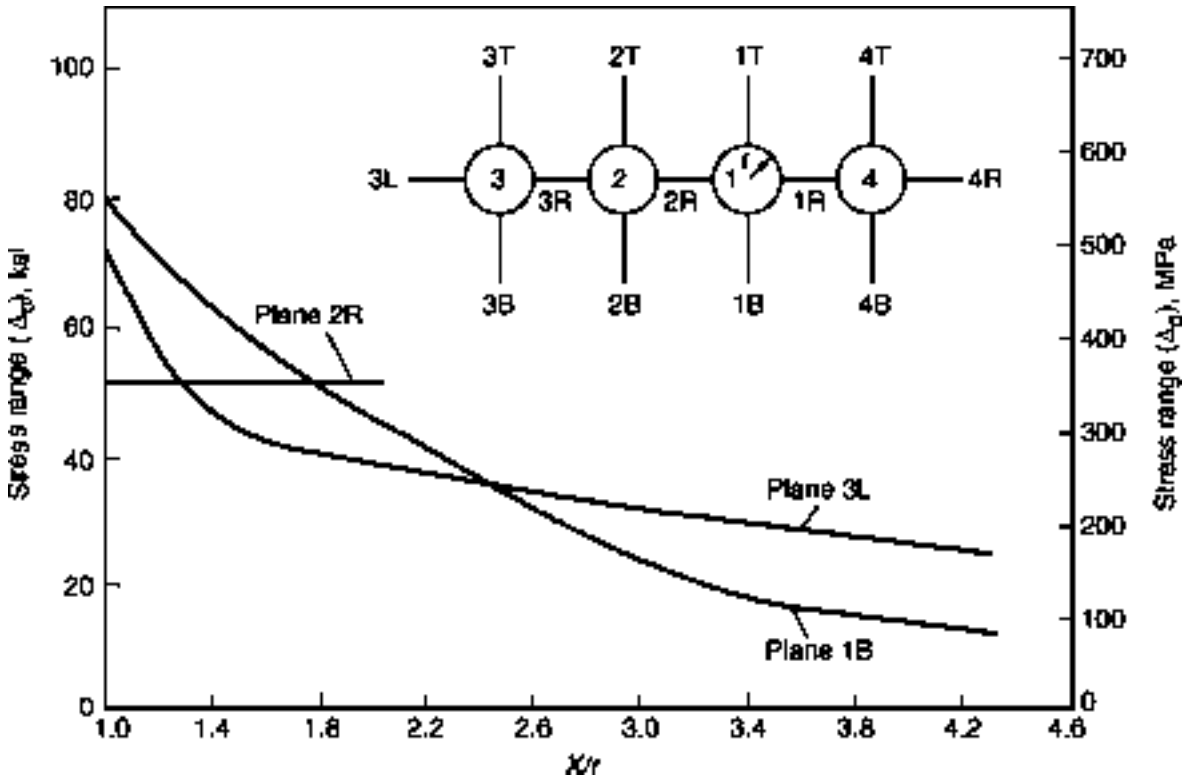


Fig. 9 Results of elastic-plastic analysis showing the hoop stress range distribution along several directions on plane A. Source: Ref 13

Fracture Mechanics Analysis. Figure 10 shows the cracked configurations used in the fracture mechanics analysis. These configurations have been selected on the basis of the observed cracks in a casing that was examined destructively (Ref 12) and also on the basis of cracks that are likely to be most detrimental to the operation of the casing, including posing safety risks to the personnel working in the area. These cracks are designated A, B, C, D, E, and F for analysis purposes. Crack type A can grow into the floor of the steam chest and cause control problems. Crack types B, C, and D can cause steam leakage to the outside of the casing and therefore present a safety hazard. Crack types E and F can cause leakage between passageways and reduce efficiency of the turbine. Cracks A, B, and C will propagate under stresses determined along 1B and D under stresses along 3L and E and F under stresses along 2R that were shown in Fig. 9.

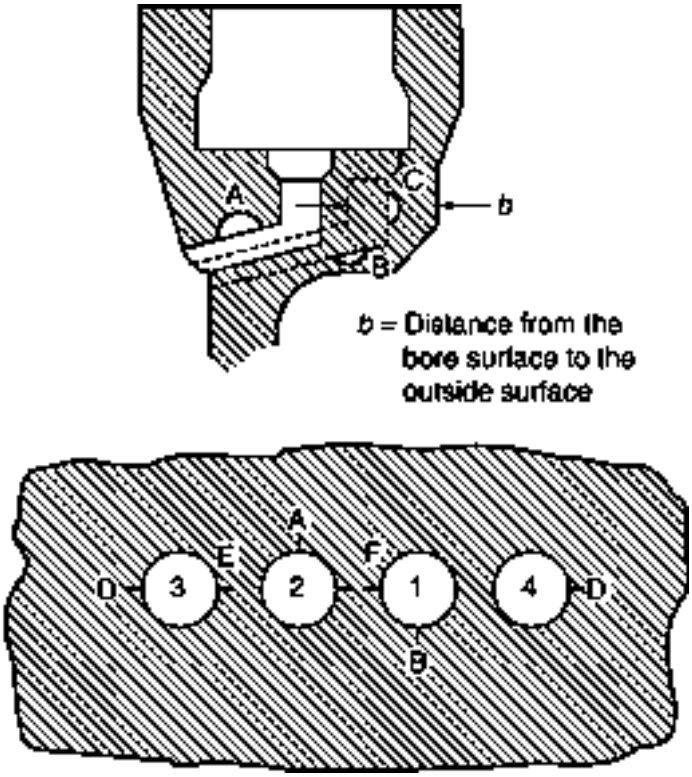


Fig. 10 Crack configurations considered in the fracture mechanics analysis of the turbine casing. Source: Ref 13

In order to estimate the remaining life or appropriate inspection intervals, in the presence of cracks, the fatigue crack growth behavior for the above configurations of cracks must be predicted, which requires estimating the magnitude of the relevant crack-tip parameters. Under linear elastic conditions, this parameter is ΔK, the cyclic stress-intensity parameter. However, one must be aware of the limitations of ΔK. These limitations for high-temperature conditions are: the region of crack-tip plasticity should be small in comparison to the ligament size, b, and the crack depth, *a*. The latter condition is definitely violated in SSTGs because of the small crack sizes being considered. Thus, nonlinear analysis using the *J*-integral concepts (Ref 16, 17) is needed.

The cyclic version of the *J*-integral, ΔJ, was first introduced by Dowling and Begley (Ref 16, 17), who used it to correlate the fatigue crack growth behavior of specimens of different geometries under elastic-plastic and fully plastic conditions. They proposed a relationship between the fatigue crack growth rate per cycle, da/dN, and ΔJ:

$$da/dN = C(\Delta J)^q \qquad \text{(Eq 14)}$$

where *C* and *q* are material parameters that are obtained from regression of the fatigue crack growth data. Also, under elastic conditions, the following relationship holds (Ref 16, 17, 18):

$$\Delta J = \Delta K^2/E \qquad \text{(Eq 15)}$$

It is also shown that if a plastically adjusted value of ΔK is obtained using Eq 15, under elastic-plastic and fully plastic conditions, the fatigue crack growth behavior under conditions ranging from linear elastic to fully plastic can all be

consolidated into a single correlation. Therefore, laboratory data under linear-elastic conditions can be used to predict the behavior of the component under elastic-plastic and fully plastic conditions and vice versa (Ref 16, 17, 18).

Estimation of ΔJ for Turbine Casing Cracks. The general expression for estimating ΔJ is given by:

$$\Delta J = \Delta K_e^2 / E + \Delta J_p \qquad \text{(Eq 16)}$$

where E is elastic modulus and ΔK_e is the elastic stress-intensity parameter given by:

$$\Delta K_e = \Delta\sigma\sqrt{\pi a}\,F(a/b) \qquad \text{(Eq 17)}$$

where $F(a/b)$ is a dimensionless function dependent on crack depth geometry and the loading configuration and is listed in handbooks for several crack geometry/load combinations (Ref 19). ΔJ_p is the plastic contribution to ΔJ that can be obtained from the following equation for single-edge notch configuration under pure tension (Ref 20):

$$\Delta J_p = \frac{kah_1}{(1-a/b)^n(\Delta\sigma/1.455\eta)^{n+1}} \qquad \text{(Eq 18)}$$

where h_1 is a dimensionless function given in Ref 20 and k and n are defined by:

$$\Delta\varepsilon_p/2 = k(\Delta\sigma/2)^n \qquad \text{(Eq 19)}$$

where $\Delta\varepsilon_p$ is the plastic strain range, and k and n are the constants related to the cyclic stress-strain properties:

$$\eta = \left[1 + \left(a/b(1-a/b)\right)^2\right]^{1/2} - a/b(1-a/b) \qquad \text{(Eq 20)}$$

Equations 16, 17, 18, 19, 20 are for uniform tension. Therefore, some assumptions are necessary to convert a nonuniform stress distribution into an equivalent uniform stress distribution along the direction 1B. For plane 2R, the stress distribution is very nearly uniform, 359 MPa (52 ksi), so Eq 16, 17, 18, 19, 20 are directly applicable to the cases E and F. Along 1B a uniform stress value of 445 MPa (64.5 ksi) was selected because it was the average value of stress (see Fig. 9) over a radial distance of 12.5 mm (0.5 in.) from the surface of the bore.

The stresses presented in Fig. 9 were computed assuming no cracks were present. These stresses are appropriate when analyzing for crack initiation or when cracks are very small, for example less than 0.5 mm (0.02 in.). As the cracks grow, the thermal stresses are expected to relax due to loss of stiffness of the system. Therefore, crack growth will result in decreasing stress magnitude of the relaxed stress due to crack growth. Thermal stresses closely approximate constant deflection conditions. Therefore, it was assumed that the crack mouth displacement range, $\Delta\delta$, shown in Fig. 11, was held constant during crack growth. The initial value of $\Delta\delta$ corresponded to the uncracked stress level acting on a crack depth of 0.64 mm (0.025 in.) that was defined as the approximate crack size at initiation. The generalized expression for $\Delta\delta$ for the configuration shown in Fig. 11 as a function of stress range is given by:

$$\Delta\delta = 4\Delta a(1-\nu^2)V_1/E + \left(kah_2/(1-a/b)^n\right) \times (\Delta\sigma/1.455\eta)^n \qquad \text{(Eq 21)}$$

where

$$V_1 = 1.46 + 3.42(1-\cos\pi a/2b)(\cos\pi a/2b)^{-2} \qquad \text{(Eq 22)}$$

h_2 is obtained from the elastic-plastic fracture handbook (Ref 20), v is Poisson's ratio, and the other parameters are as defined before. Substituting $\Delta\sigma$ from the uncracked stress analysis and using a = 0.64 mm (0.025 in.) and b = 41.3 mm (1.625 in.) for crack configurations A, B, and C, $\Delta\delta$ can be determined. This value remains constant during the subsequent crack growth. A new value of crack length is then selected, and all quantities except $\Delta\sigma$ are determined in Eq 21 for the new crack size. A nonlinear equation in $\Delta\sigma$ is thus generated that can be solved iteratively to find the $\Delta\sigma$ value. This $\Delta\sigma$ value is used in Eq 18 to determine ΔJ_p, which in turn is used in Eq 16 to obtain ΔJ. Thus, a plasticity-modified value of ΔK can be obtained as a function of crack depth for all configurations. This procedure led to the results shown in Fig. 12 for flaw types A, B, and C. Similar plots can be made for other crack configurations. The life-limiting cases were found to be A, B, and C, and therefore the following discussion focuses on those cases.

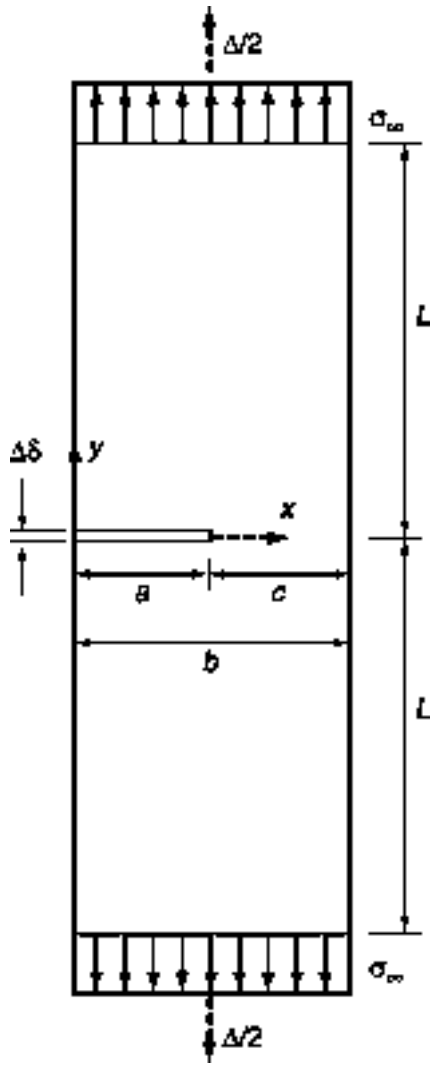


Fig. 11 Edge cracked plate under uniform tension showing the crack mouth displacement, Δδ, that is held constant during crack growth

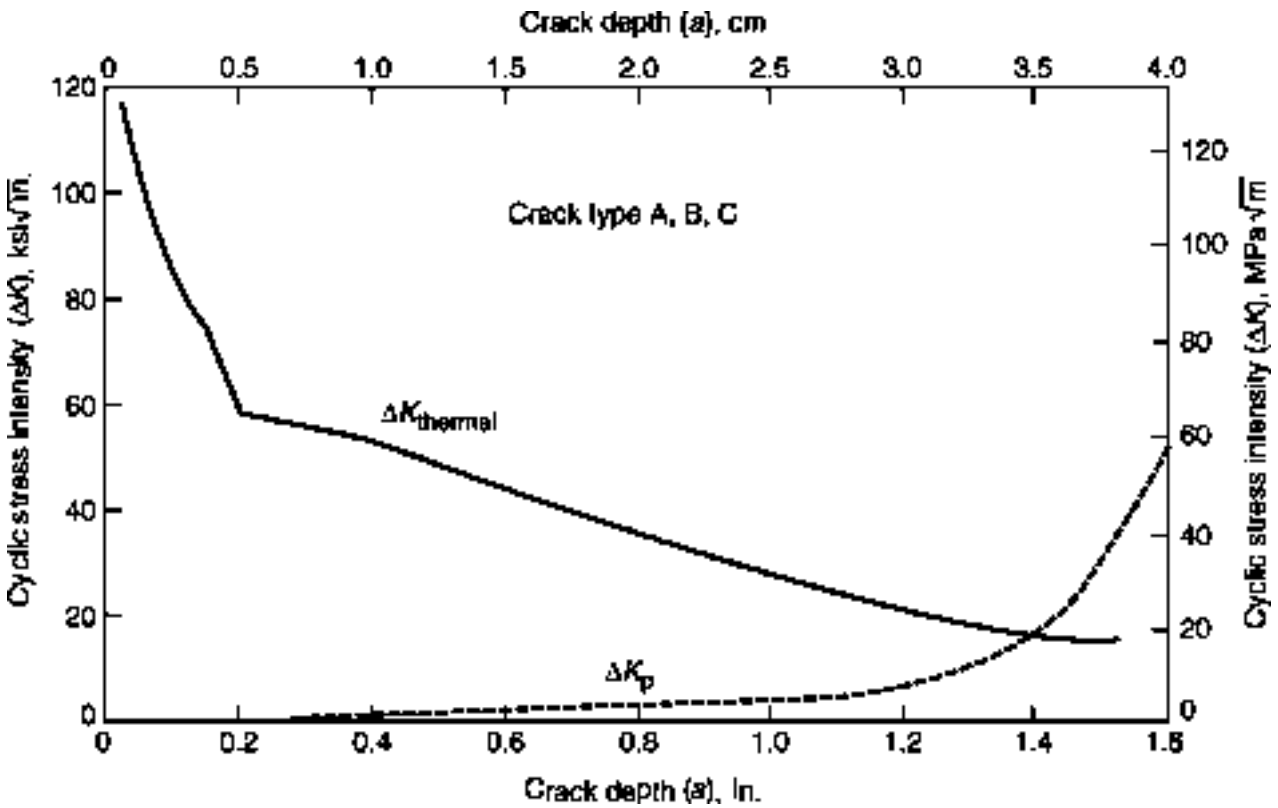


Fig. 12 Cyclic stress intensity (ΔK) as a function of crack size for crack types A, B, and C

When crack depths become large, a contribution to the crack driving force will also arise from the pressure forces acting on the crack surfaces. The ΔK value from the resulting pressure forces can also be estimated and is shown to be insignificant until the crack depth becomes almost 25 mm (1 in.). Figure 12 shows the ΔK versus crack depth for crack sizes A, B, and C. The ΔK due to thermal stresses is labeled $\Delta K_{thermal}$, and due to pressure stresses, ΔK_p; these values are plotted separately in Fig. 12.

Selecting the Final Crack Size. In fracture mechanics analyses, the final crack size depends on the fracture toughness of the material. However, in the case of the turbine casings, the ΔK value decreases as the crack grows and the potential for a burst-type condition does not exist. Thus, an alternate criterion is needed to determine the final crack size in this case. One criterion for final crack size could be when the passageway cracks grow large enough to penetrate through to the other free surface. For the case of cracks of the type A, B, or C, this value corresponds to b = 41.3 mm (1.625 in.). It is, however, noted that at crack size of 33 mm (1.3 in.), the contribution of pressure forces to the value of ΔK becomes significant and then cracks can grow by a creep mechanism. Since in the analysis creep is not explicitly accounted for, 33 mm (1.3 in.) is assumed to be the terminal crack size for calculating the remaining life. This assumption leads to conservative results, desirable in such applications. For other types of cracks, the ligament size is assumed to be the final crack size.

Remaining Life Calculation. The equation governing the fatigue crack growth behavior as a function of hold time for 2.25Cr-1Mo steel at 538 °C (1000 °F) was derived from tests conducted at elevated temperature described in Ref 14. The data are presented in Fig. 13, and the equation relating fatigue crack growth rate to the applied magnitude of ΔK and the hold time, t_h, expressed in hours is given by:

$$da/dN = 8.6 \times 10^{-10} \Delta K^{3.8} + 8.85 \times 10^{-5} \Delta K^{0.94} t_h^{-0.53} = f(\Delta K) \quad \text{(Eq 23)}$$

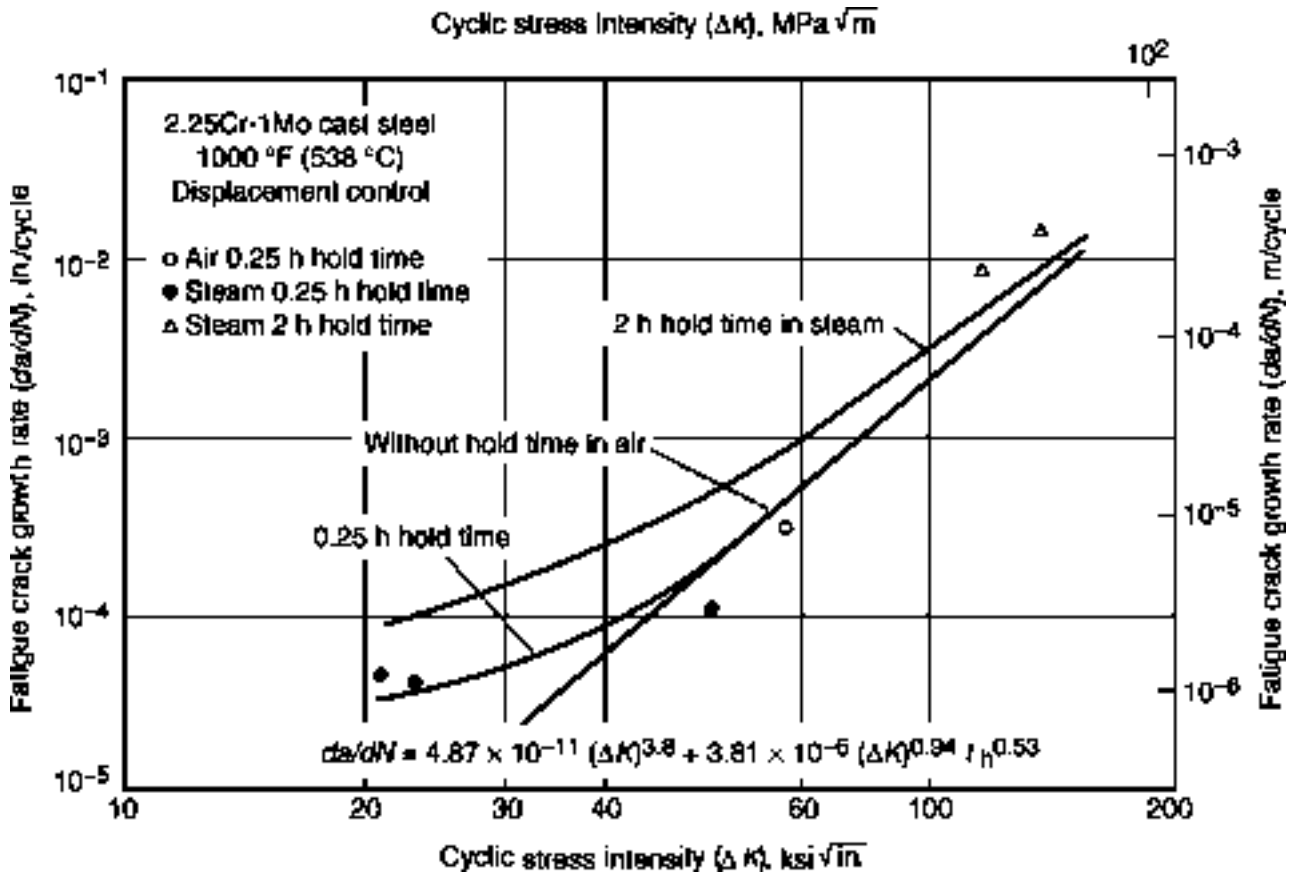


Fig. 13 Effect of hold time on the fatigue crack growth rate properties of 2.25Cr-1Mo cast steel. The 2 h hold time tests were performed in steam at 538 °C (100 °F). Source: Ref 13

In Eq 23, t_h = 72 h, the value of ΔK is taken in MPa $\sqrt{m}$, and da/dN in millimeters/cycle. The remaining life, N, from the inspection crack size, a_i, to final crack size, a_f, is given by:

$$N = \int_{a_i}^{a_f} da / f(\Delta K) \quad \text{(Eq 24)}$$

where ΔK can be expressed in terms of crack size as in Fig. 12. Therefore, $f(\Delta K)$ can be obtained at different crack depths, and by selecting different inspection crack depth values, a plot between a_i and N can be generated as seen in Fig. 14 for crack types A, B, and C. Since several regions of the turbine casing experience temperatures less than 538 °C (1000 °F), the remaining life was also calculated using data at 427 °C (800 °F) and plotted in Fig. 14. There are approximately two orders of magnitude difference in predicted remaining lives between regions of the casing exposed to 538 °C (1000 °F) and those exposed to 427 °C (800 °F). Even for regions exposed to 538 °C, the predicted remaining life for an inspection crack depth of 6.25 mm (0.25 in.) is approximately 800 cycles (16 years), which is quite significant and must be considered in making run/retire decisions.

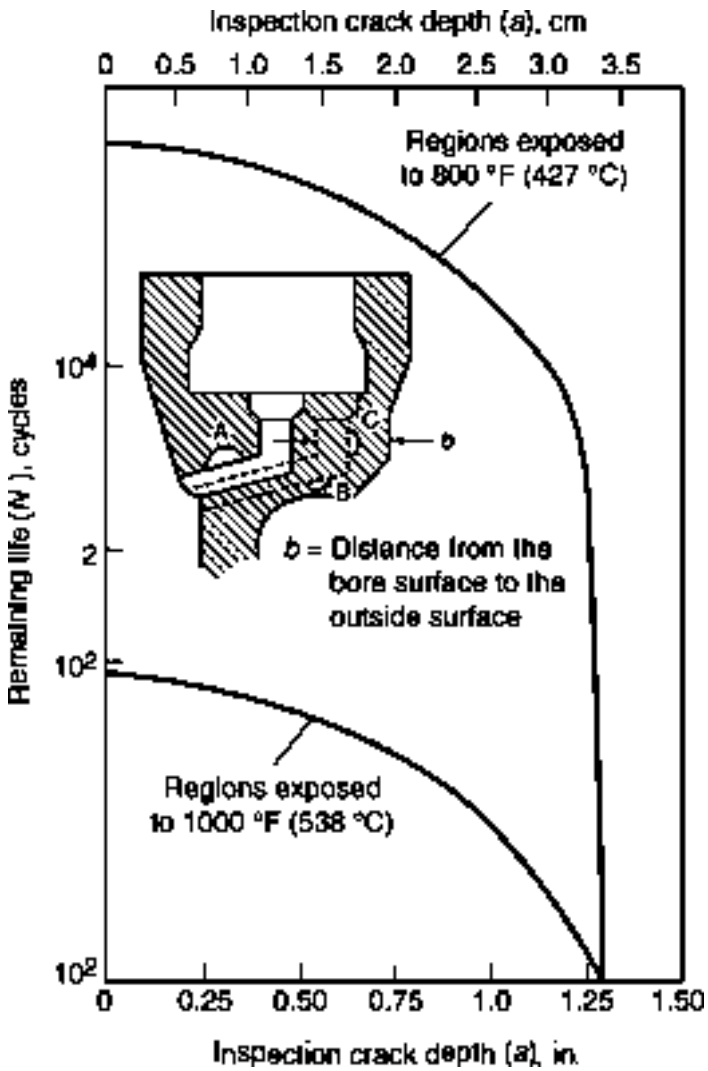


Fig. 14 Remaining life as a function of crack depth for cracks of types A, B, and C

Figure 15 shows the crack initiation life as a function of strain range at 427 °C and at 538 °C for 2.25 Cr-1 Mo wrought material (Ref 21). The estimated strain range at the bore surface of the passageways is approximately 0.0125. This corresponds to a cyclic life of 1200 cycles at 538 °C. Assuming an average of 50 cycles/year, this is approximately 24 years of life. This implies that the cracks grew only for three out of the 27 years of operation experienced by the turbine casing, in other words for 150 cycles of start-up and shutdown. From Fig. 14, cracks could conceivably grow to 10 mm (0.4 in.) depth in 150 cycles. In the casing in which crack sizes were characterized destructively, cracks as deep as 10 mm were in fact found (Ref 12). Thus, predictions from Fig. 14 are in good agreement with the limited service data.

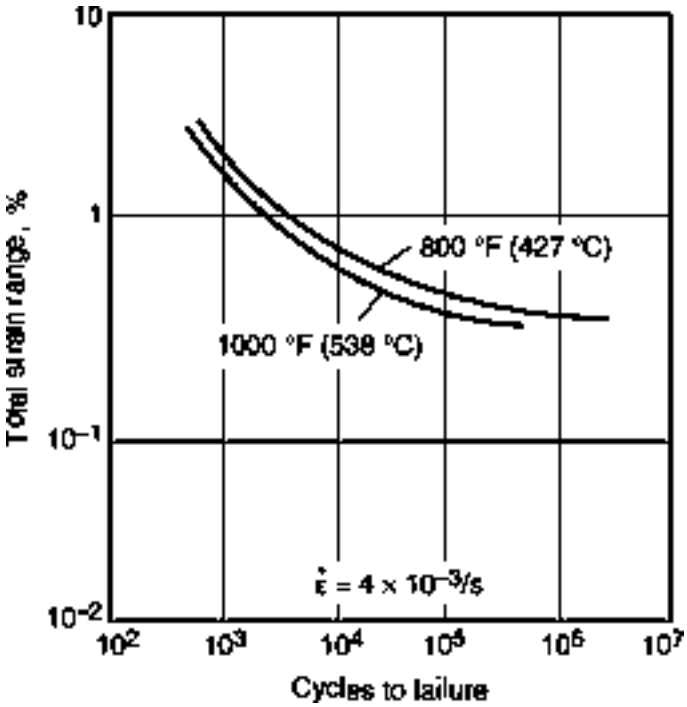


Fig. 15 Total strain range versus cycles to failure for isothermally annealed 2.25Cr-1Mo steel at 427 and 538 °C (800 and 1000 °F). Source: Ref 21

References cited in this section

6. H. Sehitoglu, Thermal and Thermomechanical Fatigue of Structural Alloys, *Fatigue and Fracture*, Vol 19, *ASM Handbook*, ASM International, 1996, p 527–556

10. G.R. Halford, B.A. Lerch, and M.A. McGaw, Fatigue, Creep, and Thermomechanical Fatigue Life Testing, *Mechanical Testing and Evaluation*, Vol 8, *ASM Handbook*, 2000, p 687–716

11. M.H. Aliabadi, Ed., *Thermomechanical Fatigue and Fracture*, WIT Press, 2002

12. V.E. Hulina, "Preliminary Results on the Mode of Failure Analysis and Methods to Combat Crack Propagation in SSTG Upper Casing," unpublished results, Westinghouse Electric Corp., 1984

13. A. Saxena, P.K. Liaw, W.A. Logsdon, and V.E. Hulina, Residual Life Prediction and Retirement for Cause Criteria for SSTG Casings—II. Fracture Mechanics Analysis, *Eng. Fract. Mech.*, Vol 25, 1986, p 289–303

14. W.A. Logsdon, P.K. Liaw, A. Saxena, and V.E. Hulina, Residual Life Prediction and Retirement for Cause Criteria for SSTG Upper Casings—I. Mechanical and Fracture Mechanics Properties Development, *Eng. Fract. Mech.*, Vol 25, 1986, p 259–288

15. P.J. Cronin, V.E. Hulina, and S.N. Pagay, "Residual Life Prediction and Retirement for Cause Criteria for SSTG Upper Casings," unpublished research, Westinghouse Electric Co., 1984

16. N.E. Dowling and J.A. Begley, Fatigue Crack Growth Under Gross Plasticity and J-Integral, *Mechanics of Crack Growth*, STP 590, American Society for Testing and Materials, 1976, p 82–103

17. N.E. Dowling, Geometry Effects and J-Integral Approach to Elastic-Plastic Fatigue Crack Growth, *Cracks and Fracture*, STP 601, American Society for Testing and Materials, 1976, p 19–32

18. A. Saxena, *Nonlinear Fracture Mechanics for Engineers*, CRC Press, 1998

19. H. Tada, P.C. Paris, and G.R. Irwin, *The Stress Analysis of Cracks Handbook*, 3rd ed., American Society for Mechanical Engineers, 2000

20. V. Kumar, M.D. German, and C.F. Shih, "Elastic-Plastic Fracture Analysis," EPRI-NP-1931, Electric Power Research Institute, Palo Alto, CA, 1981

21. C.R. Brinkman et al., "Interim Report on the Continuous Cycling Elevated Temperature Fatigue and Subcritical Crack Growth Behavior of 2.25 Cr-1 Mo Steel," ORNL-TM-4993, Oak Ridge National Laboratory, Oak Ridge, TN, 1975

Thermomechanical Fatigue: Mechanisms and Practical Life Analysis

Stephen D. Antolovich, Washington State University; Ashok Saxena, Georgia Institute of Technology

---

References

1. H.D. Solomon, G.R. Halford, L.R. Kaisand, and B.N. Leis, Ed., *Low Cycle Fatigue*, STP 942, American Society for Testing and Materials, l988

2. S.D. Antolovich, S. Liu, and R. Baur, Low Cycle Fatigue Behavior of René 80 at Elevated Temperatures, *Metall. Trans.*, Vol 12A, 1981, p 473–481

3. S.D. Antolovich and B. Lerch, Cyclic Deformation and Fatigue in Ni-Base Alloys, *Superalloys, Supercomposites, and Superceramics*, J.K. Tien and T. Caufield, Ed., Academic Press, 1989, p 363–411

4. B. Boursier, M.S. Thesis, University of Cincinnati, 1982

5. D.L. McDowell, S.D. Antolovich, and R.L.T. Oehmke, Mechanistic Considerations for TMF Life Prediction of Nickel-Base Superalloys, *J. Nucl. Energy Des.*, Vol 133, 1992, p 383–389

6. H. Sehitoglu, Thermal and Thermomechanical Fatigue of Structural Alloys, *Fatigue and Fracture*, Vol 19, *ASM Handbook*, ASM International, 1996, p 527–556

7. B.A. Lerch, S.D. Antolovich, and N. Jayaraman, A Study of Fatigue Damage Mechanisms in Waspaloy from 25°C–800°C, *Mater. Sci. Eng.*, 1984, p 151–165

8. D. Diederich, B. Lerch, and S.D. Antolovich, Observations of Low Cycle Fatigue and Fatigue Crack Propagation Substructures of Waspaloy, *Proc. VIII Inter-American Conference on Materials*, 25–29 June 1984, p 7.1–7.6

9. B.A. Boley and J.H. Weiner, *Theory of Thermal Stresses*, John Wiley and Sons, 1960

10. G.R. Halford, B.A. Lerch, and M.A. McGaw, Fatigue, Creep, and Thermomechanical Fatigue Life Testing, *Mechanical Testing and Evaluation*, Vol 8, *ASM Handbook*, 2000, p 687–716

11. M.H. Aliabadi, Ed., *Thermomechanical Fatigue and Fracture*, WIT Press, 2002

12. V.E. Hulina, "Preliminary Results on the Mode of Failure Analysis and Methods to Combat Crack Propagation in SSTG Upper Casing," unpublished results, Westinghouse Electric Corp., 1984

13. A. Saxena, P.K. Liaw, W.A. Logsdon, and V.E. Hulina, Residual Life Prediction and Retirement for Cause Criteria for SSTG Casings—II. Fracture Mechanics Analysis, *Eng. Fract. Mech.*, Vol 25, 1986, p 289–303

14. W.A. Logsdon, P.K. Liaw, A. Saxena, and V.E. Hulina, Residual Life Prediction and Retirement for Cause Criteria for SSTG Upper Casings—I. Mechanical and Fracture Mechanics Properties Development, *Eng. Fract. Mech.*, Vol 25, 1986, p 259–288

15. P.J. Cronin, V.E. Hulina, and S.N. Pagay, "Residual Life Prediction and Retirement for Cause Criteria for SSTG Upper Casings," unpublished research, Westinghouse Electric Co., 1984

16. N.E. Dowling and J.A. Begley, Fatigue Crack Growth Under Gross Plasticity and J-Integral, *Mechanics of Crack Growth*, STP 590, American Society for Testing and Materials, 1976, p 82–103

17. N.E. Dowling, Geometry Effects and J-Integral Approach to Elastic-Plastic Fatigue Crack Growth, *Cracks and Fracture*, STP 601, American Society for Testing and Materials, 1976, p 19–32

18. A. Saxena, *Nonlinear Fracture Mechanics for Engineers*, CRC Press, 1998

19. H. Tada, P.C. Paris, and G.R. Irwin, *The Stress Analysis of Cracks Handbook*, 3rd ed., American Society for Mechanical Engineers, 2000

20. V. Kumar, M.D. German, and C.F. Shih, "Elastic-Plastic Fracture Analysis," EPRI-NP-1931, Electric Power Research Institute, Palo Alto, CA, 1981

21. C.R. Brinkman et al., "Interim Report on the Continuous Cycling Elevated Temperature Fatigue and Subcritical Crack Growth Behavior of 2.25 Cr-1 Mo Steel," ORNL-TM-4993, Oak Ridge National Laboratory, Oak Ridge, TN, 1975

Analysis and Prevention of Corrosion-Related Failures

S.R. Freeman, Millennium Metallurgy, Ltd.

Introduction

CORROSION is the deterioration of a material by a reaction with its environment. In a study by the U.S. Federal Highway Administration in cooperation with the National Association of Corrosion Engineers (NACE), it was estimated that the annual cost of corrosion was between $121 and $138 billion in 1998 in the United States. These costs included cost of corrosion-control methods, equipment, and services; cost of labor attributed to corrosion management; cost of use of more expensive materials to lessen corrosion damage; and cost of lost revenue, loss of reliability, and loss of capital due to corrosion deterioration. Only selected industrial sectors were analyzed in the study. When extrapolated to all U.S. industries, the total cost estimate is $276 billion, or more than 3% of the U.S. gross domestic product (Ref 1). This great cost is a measure of the importance of corrosion management and an indication of the significance of potential cost saving that corrosion abatement can yield.

Reference cited in this section

1. "Corrosion Costs and Preventive Strategies in the United States," FHWA-RD-01-156, Federal Highway Administration, 2002

Analysis and Prevention of Corrosion-Related Failures

S.R. Freeman, Millennium Metallurgy, Ltd.

Electrochemical Nature of Corrosion

The articles in this Section are devoted to the identification and analysis of corrosion-related failures, the categorization of corrosion failures by form and mechanism, and the application of preventive measures. The mechanisms of corrosion are described in more detail in *Corrosion,* Volume 13 of the *ASM Handbook.* However, as a brief introduction, the electrochemical nature of corrosion can be illustrated by the attack on zinc by hydrochloric acid. When zinc is placed in dilute hydrochloric acid, a vigorous reaction occurs; hydrogen gas is evolved and the zinc dissolves, forming an acidic aqueous solution of zinc chloride. The reaction is:

$$Zn + 2HCl \rightarrow ZnCl_2 + H_2 \qquad \text{(Eq 1)}$$

Since the chloride ion is not involved in the reaction, this equation can be written in the simplified form:

$$Zn + 2H^+ \rightarrow Zn^{2+} + H_2 \qquad \text{(Eq 2)}$$

Zinc reacts with the hydrogen ions of the acid solution to form zinc ions and hydrogen gas. Equation 2 shows that during the reaction, zinc is oxidized to zinc ions and hydrogen ions are reduced to hydrogen. Thus, Eq 2 can be conveniently divided into two reactions: the oxidation of zinc and the reduction of hydrogen ions:

$$\text{Oxidation (anodic reaction) } Zn \rightarrow Zn^{2+} + 2e \qquad \text{(Eq 3)}$$

$$\text{Reduction (cathodic reaction) } 2H^+ + 2e \rightarrow H_2 \qquad \text{(Eq 4)}$$

An *oxidation* or *anodic* reaction is indicated by an increase in valence or a release of electrons. A decrease in valence charge or the consumption of electrons signifies a *reduction* or *cathodic* reaction. Equations 3 and 4 are *partial* reactions; both must occur simultaneously and at the same rate on the metal surface. If this were not true, the metal would spontaneously become electrically charged, which is clearly impossible.

The corrosion of zinc in hydrochloric acid is an electrochemical process. That is, any reaction that can be divided into two or more partial reactions of oxidation and reduction due to the transfer of electrical charge is termed *electrochemical.* Dividing corrosion or other electrochemical reactions into partial reactions makes them simpler to understand. Iron and aluminum, like zinc, are also rapidly corroded by hydrochloric acid.

Thus, the problem of hydrochloric acid corrosion is simplified because in every case the cathodic reaction is the evolution of hydrogen gas according to Eq 4. This also applies to corrosion in other acids such as sulfuric, phosphoric, hydrofluoric, and water-soluble organic acids such as formic and acetic. In each case, only the hydrogen ion is active, the other ions such as sulfate, phosphate, and acetate do not participate in the electrochemical reaction.

When viewed from the standpoint of partial processes of oxidation and reduction, all corrosion can be classified into a few generalized reactions. The anodic reaction in every corrosion reaction is the oxidation of a metal to its ion. This can be written in the general form:

$$M \rightarrow M^{n+} + ne \qquad \text{(Eq 5)}$$

where n is the number of electrons released.

A few examples are:

$$Ag \rightarrow Ag^{+} + e \qquad \text{(Eq 6)}$$

$$Zn \rightarrow Zn^{2+} + 2e \qquad \text{(Eq 7)}$$

$$Al \rightarrow Al^{3+} + 3e \qquad \text{(Eq 8)}$$

In each case the number of electrons produced equals the valence of the ion.

There are several different cathodic reactions that are frequently encountered in metallic corrosion. The most common cathodic reactions are: Hydrogen evolution:

$$2H^{+} + 2e \rightarrow H_2 \qquad \text{(Eq 9)}$$

Oxygen reduction (acid solutions):

$$O_2 + 4H^{+} + 4e \rightarrow 2H_2O \qquad \text{(Eq 10)}$$

Oxygen reduction (neutral or basic solutions):

$$O_2 + 2H_2O + 4e \rightarrow 4OH^{-} \qquad \text{(Eq 11)}$$

Metal-ion reduction:

$$M^{3+} + e \rightarrow M^{2+} \qquad \text{(Eq 12)}$$

Metal deposition:

$$M^{+} + e \rightarrow M \qquad \text{(Eq 13)}$$

Hydrogen evolution is a common cathodic reaction because acid or acidic media are frequently encountered. Oxygen reduction is very common, because any aqueous solution in contact with air is capable of producing this reaction. Metal-ion reduction and metal deposition are less-common reactions and are most frequently found in chemical process streams. All of the above reactions are quite similar; they consume electrons.

The above partial reactions can be used to interpret virtually all electrochemical corrosion problems. Consider what happens when iron is immersed in water or seawater that is exposed to the atmosphere (an automobile fender or a steel pier piling are examples). Corrosion occurs. The anodic reaction is:

$$Fe \rightarrow Fe^{2+} + 2e \qquad \text{(Eq 14)}$$

Since the medium is exposed to the atmosphere, it contains dissolved oxygen. Water and seawater are nearly neutral, and thus the cathodic reaction is:

$$O_2 + 2H_2O + 4e \rightarrow 4OH^{-} \qquad \text{(Eq 15)}$$

Remembering that sodium and chloride ions do not participate in the reaction, the overall reaction can be obtained by adding Eq 9 and 12:

$$2Fe + 2H_2O + O_2 \rightarrow 2Fe^{2+} + 4OH^- \rightarrow 2Fe(OH)_2 \qquad \text{(Eq 16)}$$

Ferrous (Fe^{2+}) hydroxide precipitates from solution. However, this compound is unstable in oxygenated solutions and is oxidized to the ferric (Fe^{3+}) salt:

$$2Fe(OH)_2 + H_2O + \frac{1}{2}O_2 \rightarrow 2Fe(OH)_3 \qquad \text{(Eq 17)}$$

The final product is the familiar *rust.*

The classic example of a replacement reaction, the interaction of zinc with copper sulfate solution, illustrates metal deposition:

$$Zn + Cu^{2+} \rightarrow Zn^{2+} + Cu \qquad \text{(Eq 18)}$$

or, viewed as partial reactions:

$$Zn \rightarrow Zn^{2+} + 2e \qquad \text{(Eq 19)}$$

$$Cu^{2+} + 2e \rightarrow Cu \qquad \text{(Eq 20)}$$

The zinc initially becomes plated with copper and eventually, given enough time and reactants, the products are copper sponge and zinc sulfate solution.

During corrosion, more than one oxidation and one reduction reaction may occur. When an alloy is corroded, its component metals go into solution as their respective ions. More importantly, more than one reduction reaction can occur during corrosion. Consider the corrosion of zinc in aerated hydrochloric acid. Two cathodic reactions are possible: the evolution of hydrogen and the reduction of oxygen. Since the rates of oxidation and reduction must be equal, increasing the total reduction rate increases the rate of zinc solution. Therefore, acid solutions containing dissolved oxygen normally will be more corrosive than air-free acids. Oxygen reduction simply provides another means of electron disposal. The same effect is observed if any oxidizer is present in acid solutions. A frequent impurity in commercial hydrochloric acid is the ferric ion (Fe^{3+}), present as ferric chloride. Metals corrode much more rapidly in such impure acid because there are two cathodic reactions, hydrogen evolution and ferric ion reduction:

$$Fe^{3+} + e \rightarrow Fe^{2+} \qquad \text{(Eq 21)}$$

The anodic and cathodic reactions occurring during corrosion are mutually dependent, and it is possible to reduce corrosion by reducing the rates of either reaction. In the above case of impure hydrochloric acid, it can be made less corrosive by removing the ferric ions and consequently reducing the total rate of cathodic reduction. Oxygen reduction is eliminated by preventing air from contacting the aqueous solution or by removing air that has been dissolved. Iron is nearly inert in air-free water or seawater because there is limited cathodic reaction possible.

If the surface of the metal is coated with paint or other nonconducting film, the rates of both anodic and cathodic reactions will be greatly reduced and corrosion will be retarded. A *corrosion inhibitor* is a substance that, when added in small amounts to a corrosive, reduces its corrosivity. Corrosion inhibitors function by interfering with either the anodic or cathodic reactions, or both. Many of these inhibitors are organic compounds; they function by forming an impervious film on the metal surface or by interfering with either the anodic or cathodic reactions. High-molecular-weight amines retard the hydrogen-evolution reaction and subsequently reduce corrosion rate. Adequate conductivity in both the metal and the electrolyte is required for continuation of the corrosion reaction. Of course, it is not practical to increase the electrical resistance of the metal because the sites of the anodic and cathodic reactions are not known, nor are they predictable. However, it is possible to increase the electrical resistance of the electrolyte and thereby reduce corrosion. Very pure water is much less corrosive than impure or natural waters. The low corrosivity of high-purity water is due to its high electrical resistance and few reducible cations.

Passivity. Essentially, *passivity* refers to the loss of chemical reactivity experienced by certain metals and alloys under particular environmental conditions. That is, certain metals and alloys become essentially inert and act as if they were noble metals such as platinum and gold. Fortunately, from an engineering standpoint, the metals most susceptible to this kind of behavior are the common engineering and structural materials, including iron, nickel, silicon, chromium, titanium, and alloys containing these metals. Also, under limited conditions, other metals such as zinc, cadmium, tin, uranium, and thorium have also been observed to exhibit passivity effects.

Passivity, although difficult to define, can be quantitatively described by characterizing the behavior of metals that show this unusual effect. First, consider the behavior of what can be called an *active* metal, that is, a metal that does not show passivity effects. The lower part of the curve in Fig. 1 illustrates the behavior of such a metal. Assume that a metal is immersed in an air-free acid solution with an oxidizing power corresponding to point A and a corrosion rate corresponding to this point. If the oxidizing power of this solution is increased, say, by adding oxygen or ferric ions, the corrosion rate of the metal will increase rapidly. Note that for such a metal, the corrosion rate increases as the oxidizing power of the solution increases. This increase in rate is exponential and yields a straight line when plotted on a semilogarithmic scale as in Fig. 1. The oxidizing power of the solution is controlled by both the specific oxidizing power of the reagents and the concentration of these reagents. Oxidizing power can be precisely defined by electrode potential, but is beyond the scope of this discussion.

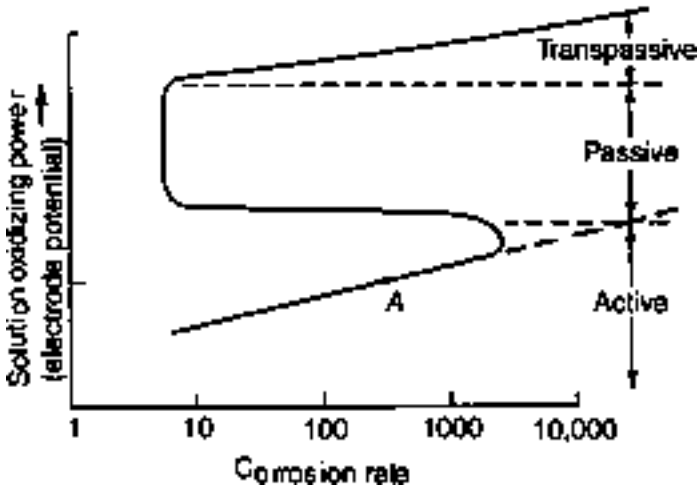


Fig. 1 Corrosion characteristics of an active-passive metal as a function of solution oxidizing power (electrode potential)

The behavior of this metal or alloy can be conveniently divided into three regions: active, passive, and transpassive. In the *active* region, slight increases in the oxidizing power of the solution cause a corresponding rapid increase in the corrosion rate. However, at some point, if more oxidizing agent is added, the corrosion rate shows a sudden decrease. This corresponds to the beginning of the *passive* region. Further increases in oxidizing agents produce little if any change in the corrosion rate of the material in the passive region. Finally, at very high concentrations of oxidizers or in the presence of very powerful oxidizers, the corrosion rate again increases with increasing oxidizing power. This region is termed the *transpassive* region.

It is important to note that during the transition from the active to the passive region, a 10^3 to 10^6 reduction in corrosion rate is usually observed. Passivity is due to the formation of a surface film or protective barrier that is stable over a considerable range of oxidizing power and is eventually destroyed in strong oxidizing solutions. Under conditions in which the surface film is stable, the anodic reaction is stifled and the metal surface is protected from corrosion. For example, stainless steel owes its corrosion-resistant properties to a passive surface film. The naturally occurring passive film is usually enhanced with immersion in a hot nitric acid solution or steam. For example, stainless steel surgical implants develop this passive layer when the implants are sterilized in steam. The exact nature of this barrier that forms on the metal surface is not well understood. It may be a very thin, transparent oxide film or a layer of adsorbed oxygen atoms. However, for the purposes of engineering application, it is not necessary to understand completely the mechanism. To summarize, metals that possess an active-passive transition become passive (very corrosion-resistant) in moderately to strongly oxidizing environments. Under extremely strong oxidizing conditions, these materials lose their corrosion-resistant properties.

S.R. Freeman, Millennium Metallurgy, Ltd.

Analysis of Corrosion-Related Failures

Corrosion can be categorized as being uniform or localized. Uniform corrosion is the most common form of corrosion. Its mechanism is electrochemical and is identified as a consistent loss of material over the entire exposed surface of the material in question. This is the type of corrosion that affects the greatest number of structures, such as buildings, bridges, pipelines, and outdoor equipment. It attacks the most surface area and is the most costly. Uniform corrosion is also the easiest to analyze, to predict corrosion rates for, and to establish preventive measures and maintenance schedules for. Preventive techniques include painting (coatings), inhibition, and cathodic or anodic protection. The service life of the affected component can normally be estimated with a reasonable degree of accuracy, and catastrophic failures can be avoided.

Localized corrosion, such as crevice and pitting corrosion, intergranular corrosion, selective leaching, erosion-corrosion, and stress-corrosion cracking (SCC), act on a small portion of a component. The rate of localized corrosion is often orders of magnitude greater than that of uniform corrosion. Evidence of localized corrosion, by virtue of its physical size and location, is often difficult to detect. These factors combine to make localized corrosion mechanisms insidious. If undetected they can lead to catastrophic system failures.

Certain environments induce specific corrosion failure modes. Hydrogen damage, liquid- and solid-metal-induced embrittlement, biologically and microbiologically influenced corrosion (MIC), and high-temperature corrosion are also addressed in articles in this Section.

While categorizing local corrosion by form, mechanism, or environment aids understanding, failures are often attributed to a combination of modes. The combination of failure modes makes prediction of localized corrosion failures more difficult.

Corrosion failure involves the same general steps as other failures, as described in the article “Practices in Failure Analysis” in this Volume. However, one potential difference between a general failure and a failure related to corrosion is the need with the latter for immediate preservation and protection of all evidence. Corrosion failure also may require sampling and testing of corrosion products immediately, such as in the case of MIC, where viable cultures can provide the most meaningful result. When possible and cost effective, a site visit is most desirable. A site visit may provide the investigator with information that otherwise may have been omitted or overlooked.

Corrosion failures often relate to the material selection and the environment. Details of the material specifications, quality-assurance records, installation and maintenance records, and a history of the environment are all useful resources in resolving corrosion failures. Information regarding system upsets or diversions from the normal operating environment should be provided. A comparison of the actual material utilized with the material design specifications should also be performed.

History of Failed Parts. Obtaining a history of the operation of the failed component is crucial to solving the cause of failure. The operating environment, any changes in the environment, and temperature excursions should all be obtained. Any notation of previous failures or operating anomalies is useful. When possible, engineering drawings and sketches should be reviewed.

Information should be provided regarding any testing that was performed by the plant personnel. For instance, liquid-penetrant examination can result in chemical contamination of a surface. The use of paints or dyes to mark components may also alter the corrosion resistance of a part and the chemical composition of a corrosion product. It is also important that the investigator retains pertinent corrosion evidence before movement or disassembly of a component. The chain of events leading to, and occurring after, the failure should be documented.

On-Site Examination and Sampling. On-site examination should include a walking tour of the failure area. Photographic documentation should be made to depict the conditions after the failure. If applicable, viewing a similar undamaged assembly or operation on the site where the failure occurred may be instructive.

Photographic documentation must be performed with special attention given to capturing the true colors of the corrosion products. Macrographs of the corrosion deposit should include any layering effects in the deposits.

Laboratory photography in a controlled setting should also be performed to guarantee accurate reproduction of color tones and surface textures.

The type and extent of on-site sampling depends on the environment and its availability. Piping corrosion in a domestic water service would require water samples from the source (incoming water supply) and from the end use (for example, a faucet). Microbiologically influenced corrosion may apply in certain cases. On-site testing or sample removal is necessary to retain the most accurate information regarding the type and number of bacteria involved. In most cases, removal of corrosion deposits from the surface may be performed in the field. If the sample is undergoing extensive laboratory examination, it may be prudent to carefully retain the corrosion in place for laboratory documentation and removal.

If samples are removed on site, care must be taken to avoid any contamination. Sealed bags, latex gloves, tools for sample removal, buffered containers for water samples, MIC kits, and adhesive tape are useful for sample removal. (See the article “Biological Corrosion Failures” in this Volume for listings of commercially available kits.) The analyst should avoid touching any corrosion product directly with bare hands to prevent contamination.

Physical removal of samples must be done in a manner to avoid further damage to the failed component and to avoid disturbance of corrosion products. In general, cutting must be done with care to avoid alterations of the metallurgical condition of the material and corrosion deposits. Saw cutting is generally preferred over torch cutting, since heating of the sample can affect the material and the corrosion product. If torch cutting is performed, a distance of 75 to 150 mm (3 to 6 in.) should be maintained from the area of interest. Saw cutting should be performed at a slow rate to avoid overheating. Use of lubricants and coolants should be avoided, if possible, to minimize contamination.

The proper shipment and storage of samples lessens the possibility of destruction of pertinent evidence. Wrapping and sealing of the failed component will generally retain evidence and prevent further oxidation or contamination. A neutral environment may be required to reduce activity. Any fracture surfaces should be protected from the potential of rubbing and contact.

Laboratory Examination. While each type of failure may have unique tests, some general steps should be taken in investigating all corrosion failures:

- All samples must be properly identified and their origin, handling, and processing within the laboratory documented.
- Photographs should document samples in the “as-received” condition.
- Stereomicroscopic examination of the involved areas should follow. Photographic documentation during stereomicroscopy should also be performed.
- Nondestructive test methods should now be considered. The key to performing nondestructive testing is to avoid any disturbance of scale product until corrosion samples have been removed. Radiography to document casting quality or to evaluate cracking is usually acceptable. However, the use of liquids or dyes is not acceptable until the corrosion samples have been removed.
- The removal of corrosion deposits for evaluation is the next step. The samples should be removed with a noncontaminating tool such as a stainless steel pick. Corrodent samples should be placed in clean, clearly marked containers.
- The corrosion deposits should be analyzed. One of the most common analysis techniques is energy-dispersive spectroscopy (EDS). This method, used in conjunction with scanning electron microscopy (SEM), provides information regarding the elemental composition of the corrosion deposit.
- Based on the visual examination, a corrosion sample may need to be subjected to microbiological analysis.
- These steps may be followed by cleaning and/or other tests.

After the corrosion deposits selected for examination have been secured, the failure sample may be cleaned. However, in some cases it may be desirable to retain the corrosion product intact for metallographic examination. For instance, a component subject to hydrogen damage or caustic corrosion may benefit from an analysis of the corrosion product layering effect. Cross sections should be taken prior to removal of corrosion products.

Precautions must be taken during the cleaning process to avoid any destruction to the base metal. Generally, the least aggressive cleaning method is initiated first, such as brushing with a soft brush or light air pressure.

Ultrasonic cleaning in acetone is considered a nonaggressive approach. This cleaning method will remove some light surface deposits. Deposits that have been exposed to elevated temperatures will generally require a more aggressive approach. Inhibited dilute acid solutions and citric acid cleaners can be used to clean adherent corrosion deposits. In cases where the protection of the fracture features is critical, softened acetate tape can be used to remove adherent deposits. This method also retains the deposits removed for further examination if necessary.

If the failure analysis does not involve preservation of a fracture surface, fine sandblasting of the base metal may be useful to remove scale deposits as in the evaluation of the pitting.

After the corrosion deposits have been removed, additional nondestructive testing may prove useful. Magnetic particle examination, penetrant examination, eddy current testing, and ultrasonic testing are a few techniques that can be employed to explore the quality and condition of the failure and material.

Microscopic Examination. As noted previously, stereomicroscopy is performed to document the corrosion product, and it also is an appropriate tool for analysis of the fracture surface (if one exists) or other surface after cleaning. This analysis will provide information regarding the failure initiation, cracking, and corrosion-surface patterns such as pitting and wear/erosion. Photographic documentation of the surface should be performed. Based on the results of the stereomicroscopy, areas can be selected for SEM to characterize the fracture features. A determination of the fracture features, such as intergranular cracking, cleavage, and ductile dimples, will provide valuable information pointing to the likely cause of failure and the corrosion mechanism.

Metallography is an essential tool for the examination of the failure specimens. Selection of the most informative cross-section locations is important. Metallurgical examination of a cross section requires mounting, surface grinding, polishing, and examining the sample in the unetched and etched condition under a microscope. Microstructural features and conditions such as cracking and crack progression, pit morphology, selective leaching, surface features, and other characteristics can all be examined. These features provide key evidence regarding the cause of failure and extent of damage.

As discussed earlier, in some cases it is useful to examine a cross section of the corrosion product with the base material. There are various techniques to retain the scale without pullout during the polishing process. In some cases, plating over the scale prior to mounting the sample or impregnating the mount with resin after preparation can retain the scale for examination. Additional information is provided in the articles “Practices in Failure Analysis” and “Metallographic Techniques in Failure Analysis” in this Volume.

Corrosion testing includes several categories of tests. Normally, corrosion testing is considered a long-term approach to investigations regarding material selection. Simulated testing or in situ testing is often performed when a given environment may be unique or where materials may experience unique flow conditions. For example, piping containing fluid may experience unique flow conditions that may result in erosion-corrosion failures. In that case, sections of pipe of different alloys may be placed on line and monitored. Other common methods of testing include accelerated tests, simulated or pilot testing, and electrochemical tests.

Typically, a variety of materials are selected for evaluation. A standard, such as carbon steel, may be used to verify the corrosiveness of the environment. Multiple samples of the same material should be tested to verify reproducibility. Sources for corrosion test methods are ASTM International, NACE, and internationally, ISO, EN, and JIS standards. Other testing methods and inspection procedures have been developed on an industry-specific basis by industrial and government organizations, such as in the pulp and paper and the electric power industries, to specifically address corrosion issues relating to their specific operating conditions and environments.

Accelerated tests are performed when an expedited answer is needed to solve a serious corrosion problem and to demonstrate an acceptable corrosion-resisting lifetime, provided that correlations can be established. Accelerated testing usually involves the exposure to extreme conditions, relative to the actual service environment, to accelerate the corrosion process. The process may be accelerated with high-stress conditions, elevated temperatures, or highly aggressive solutions.

While accelerated testing does provide test data sooner, extrapolation of results may be misleading. The acceleration of certain factors during the testing may cause different corrosion mechanisms to become active, and thus the results may not reflect those that would be obtained under actual service conditions. Care must be taken when evaluating results so that economical and acceptable choices are not eliminated due to their performance in unrealistic environments.

Simulated-use tests are primarily set up in the laboratory with an environment as close as possible to the environment of interest. In these cases, sufficient time must be allowed for an appropriate response; a short time

exposure likely will not produce enough corrosion activity to evaluate the corrosion behavior of the material. Nevertheless, by simulating the exact conditions of operation, an accurate assessment of the material can generally be produced with an adequate exposure time.

ASTM publishes standard test methods and analytical procedures for corrosion and wear testing (Ref 2).

Electrochemical testing is performed for general information regarding the passivity or anodic protection of a material against corrosion and to determine the critical breakdown or pitting potential. This type of testing is performed by two methods: controlling the current or controlling the potential. Standard methods for electrochemical testing are published in Ref 2.

As the name suggests, in the controlled-current test method, the current is controlled and the resulting corrosion potential is measured. Polarization curves are generated. Galvanostatic and galvanodynamic polarization measurements are used to plot anodic and cathodic polarization curves. The assumption that corrosion rates remain constant with time can produce inaccurate results with this test method.

In the controlled-potential method, instrumentation is available for both constant-potential (potentiostatic) and variable-potential (potentiodynamic) testing to determine overall corrosion-rate profiles for metal-electrolyte systems over a range of potentials.

Corrosion Rates and Types. Knowing and understanding corrosion rates and the types of corrosion is essential to the evaluation of corrosion failures and to the communication of the results to others. The articles that follow in this Section discuss the forms, mechanisms, and relative rates of corrosion. *Corrosion,* Volume 13 of the *ASM Handbook,* provides detailed information regarding types of corrosion, corrosion testing, corrosion failures, and industry- and alloy-specific corrosion considerations. Volume 13 also provides information regarding the use of specific alloys in given environments, corrosion prevention, and the use of nonmetallic materials. The *Handbook of Corrosion Data,* 2nd edition (ASM International, 1995) is a compilation of corrosion data from published sources. Corrosion rates of various alloys are provided with a general discussion of the corrosion resistance of alloy groups in particular environments.

Internet web sites published by ASM International, ASTM, NACE, the Nickel Development Institute, and the Copper Development Association provide the ability to search libraries of data for a given request.

Analysis of Incomplete Data. Incomplete or inconsistent data may occur in certain instances when the failure piece has been contaminated by an unknown source. Improper handling of the failure sample can introduce contamination on the sample. Testing of contaminated samples may produce misleading data and erroneous results. For example, the sampling of a deposit removed from a fracture surface that experienced stress corrosion may not reveal the corrodent that caused the SCC. Often the fracture surface is flushed with water or cleaned prior to testing. Liquid penetrants, cleaning fluids, cutting fluids, and solvents may alter the chemical composition of the surface deposits.

Results from laboratory testing may provide extraneous results. The tests may not model the service conditions. Laboratory testing cannot easily model flow conditions such as turbulence, erosion, and localized attack. Care must be taken when evaluating the laboratory data to be certain that the conclusions drawn are an accurate assessment of the operating environment.

ASTM G 16, “Applying Statistics to Analysis of Corrosion Data” (Ref 3), provides a guide for handling data; it refers to ASTM E 178, “Practice for Dealing with Outlying Observations” (Ref 4) for the treatment of data that appear inconsistent with the bulk of the findings.

References cited in this section

2. *Wear and Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* ASTM

3. “Applying Statistics to Analysis of Corrosion Data,” G 16, *Annual Book of ASTM Standards,* ASTM

4. “Practice for Dealing with Outlying Observations,” E 178, *Annual Book of ASTM Standards,* ASTM

Examples of Corrosion Failure Analysis

Example 1: Analysis of Pitting and MIC of Stainless Steel Piping. Type 316L (UNS S31603) austenitic stainless steel piping was installed as part of a collection system for a storm sewer treatment system used in a manufacturing facility. Within six months of start-up, leaks were discovered.

On-Site Examination. A series of tests were performed on site to eliminate stray currents as a possible cause of failure in the piping system. The piping system was inspected for possible sources of ac or dc current flow. An external battery was used to impress a voltage. Voltage and current were recorded with the external power supply applied and removed. There was no evidence of stray currents or electrical discharge.

It was noted that the ambient temperature and slow or stagnant flow conditions present in the piping were ideal for bacteria growth.

On-Site Sampling. Water samples were obtained for corrosivity and MIC testing. Commercially available field MIC kits were used. Samples of the damaged pipe were removed.

Laboratory Testing. High levels of chlorides, as high as 20,000 ppm, were reported in the water sample. The total dissolved solids level of 10,000 ppm was also high. The MIC water samples were shipped to the laboratory within 24 h of removal for viable culture testing. The testing showed high levels of aerobic, acid-producing, and low-nutrient bacteria.

Laboratory Examination. Perforated pipe samples were provided for metallurgical evaluation. Figure 2(a) shows the leak area as viewed from the outside-diameter surface. The sample was cut dry to avoid contamination. The inside-diameter surface is shown in Fig. 2(b). The pit appeared larger on the inside-diameter surface, indicating pit initiation occurred at the inside surface and at the bottom of the pipe. A rusty discoloration was apparent along the bottom length of the pipe. This discoloration corresponded to the area of the pitting. There was no corrosion deposit associated with the pitting. The discolored areas and other areas were evaluated using EDS. The EDS revealed contaminants consisting of chlorine, sulfur, sodium, silicon, and potassium. The area of discoloration revealed iron and oxygen only.

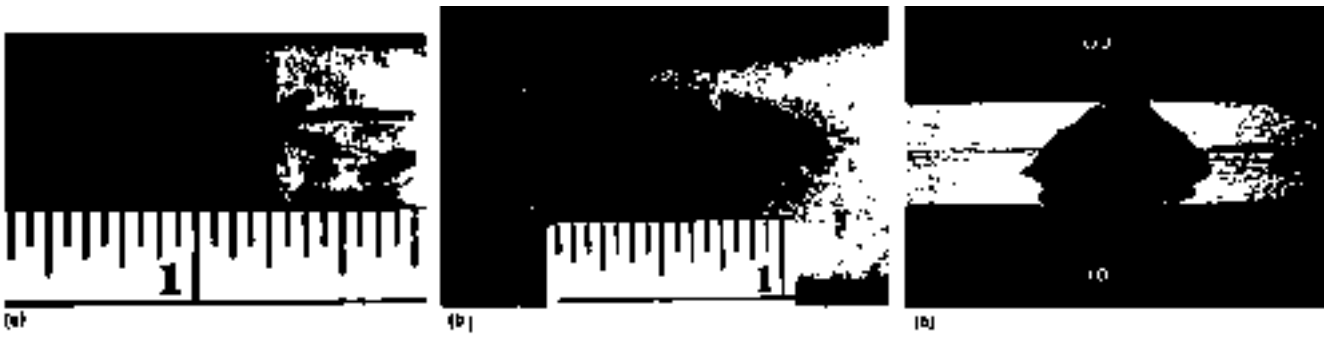


Fig. 2 Pitting corrosion of 316L stainless steel pipe. (a) View of pitting on the outside-diameter surface at the leak location. (b) View of the inside-diameter surface, where the pit size was larger at the leak location. There was a rusty discoloration along the bottom of the pipe. (c) Cross section of pipe wall through the perforation. The sample was etched in ASTM 89 reagent to delineate the microstructure. Uniform wall thickness is approximately 2.9 mm (0.11 in.). 5×. Courtesy of S.R. Freeman, Millennium Metallurgy, Ltd.

A metallurgical cross section was prepared through the pitted region. Figure 2(c) shows a 10× magnification of the cross section through the pitted region after etching with ASTM 89 reagent. The pit was not associated with a welded region, and the microstructure appeared normal. There was no evidence of general wall loss (uniform corrosion).

Conclusion. The pitting in the austenitic stainless steel pipe is believed to be caused by damage to the passive layer brought about by a combination of MIC, high chloride levels, and high total dissolved solids. The low-flow and stagnant conditions present in the piping are primary contributors to the pit progression. This type of material will perform significantly better when there is more flow of water.

Retesting of the water indicated similar high levels of aerobic bacteria, so these must be considered a part of the design basis.

Recommendations. Due to the extensive amount of piping, the pinhole size of the leaks, and environmental consequences of leaks, repair of the existing pitted pipe was impractical. Replacement of the pipe was recommended. Several alloys, nonmetallic materials, and lined materials were proposed for coupon testing to determine which is the best in this particular environment.

Example 2: Analysis of a Corrosion Failure of an Aboveground Storage Tank. A failure of an aboveground storage tank occurred due to external corrosion of the tank floor. The liquid asphalt tank operated at elevated temperatures (approximately 177 °C, or 350 °F) and had been in service for six years. Cathodic protection (rectifiers) had been installed since start-up of the tank operation. It was noted, however, that some operational problems with the rectifier may have interrupted its protection.

On-Site Examination. The underside of the floor plates showed extensive localized wall thinning. Figure 3 shows a pit in one of the tank floor plates. While it was expected that the ground under the tank was near operational temperatures, it was determined that the temperatures were below 104 °C (220 °F) at most areas evaluated in the ground. Thus, it was possible for moisture to accumulate at the tank floor, producing intermittent wetting and drying conditions. This provided an explanation of the circumferential corrosion pattern observed several feet in from the periphery of the tank. Heating was provided near the center of the tank and water in the backfill below the tank floor should be minimal. Perforations of the tank were large, and thinning was apparent in the adjacent areas.

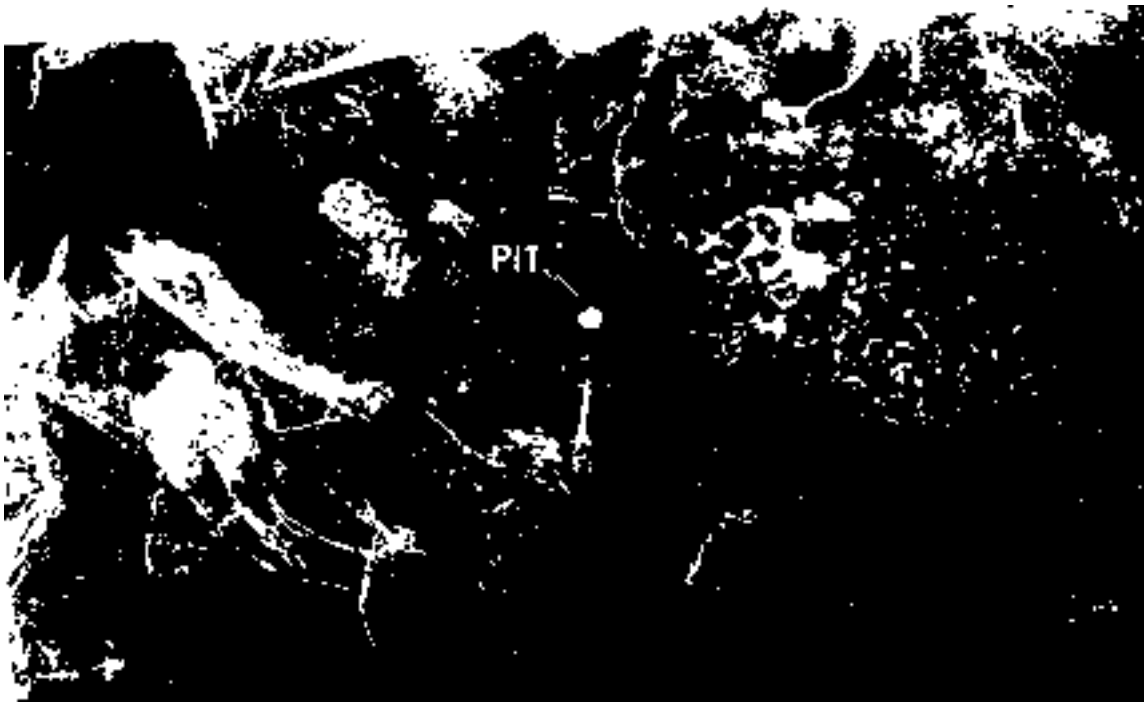


Fig. 3 View of an isolated pit on the outside of the floor tank of a liquid asphalt tank. The asphalt had been leaking for some time as the surrounding area was covered with deposits and the wall thinning was significant. Courtesy of S.R. Freeman, Millennium Metallurgy, Ltd.

On-Site Sampling. Soil samples were taken and tested, but did not indicate high corrosivity. Sample of the scale deposits from the outer surface of the tank floor were taken.

Laboratory Examination. Scale samples analyzed by EDS showed the deposits consisted of primarily Fe_2O_3 (iron oxide). The rapid exfoliation, scale buildup, and pitting of the tank floor from the external (underside) surface also prompted MIC testing of the soil and of the deposits in the area of wall thinning. Although the temperatures were somewhat high for MIC to thrive, the results of the samples confirmed high levels of various

bacteria including aerobic and low nutrient. Certain types of bacteria (thermophiles) have been found to survive significantly high temperatures.

Conclusion. An evaluation of the cathodic protection system showed the design to be adequate provided it was reliably powered. Cathodic protection is one means of controlling MIC, although interruption in the cathodic protection service will adversely affect its performance. The effectiveness of the cathodic protection on established microbial deposits is questionable.

Recommendations. The tank floor was ultrasonically tested, and portions of the floor replaced based on the remaining wall thickness. Doubling the wall thickness of the replacement floor plates was recommended.

Example 3: Preventive Maintenance for Buried Pipelines. Corrosion is a major contributor to the deterioration of buried infrastructure such as pipelines in the United States. Figure 4 shows the result of an unexpected corrosion failure of a 100-year-old riveted steel water transmission main. The snowplow and icing of the power lines in the photograph indicate that repairs had to be made under less than ideal weather conditions. The costs associated with the evaluation, modifications, and cathodic protection of a water system typically represent less than 10% of the total cost to replace the buried infrastructure. These savings can be obtained through the implementation of a comprehensive corrosion-control program.

Fig. 4 Corrosion failure of 100-year-old riveted steel water transmission main. Courtesy of S. Paul, CorrTech, Inc.

To ensure that the proper corrosion-control method(s) are employed and the full infrastructure service life is realized, local factors contributing to corrosion must be understood. Where history or analyses indicate that the environment is aggressive, corrosion-control methods need to be developed and initiated. To evaluate corrosion-exposure conditions for metallic piping networks, field conditions such as soil type, soil pH, stray current effects, and electrical continuity require quantification by field investigation.

Based on preliminary survey results, excavations of the pipe system are made to inspect the pipe for corrosion. At each location, the pipe is cleaned and evaluated for uniform and pitting corrosion. Of concern for pipelines are the deepest pits and penetrations. Statistical analysis of the system is performed. Typically, greater than 95% of the pipe is in good condition with sufficient original wall thickness. By defining the corrosion exposure and understanding the exposure, personnel can then make informed decisions about extending the life of the buried infrastructure or developing replacement options. By applying proven technologies such as cathodic protection, corrosion can be arrested and infrastructure can remain in service for 50-plus years without failures and downtime.

Example 4: Structural Epoxy Rehabilitation of Steel and Reinforced Concrete Structures. Hydrogen sulfide gas is very aggressive to reinforced concrete structures in wastewater treatment plants. Figure 5 provides views of severe corrosion of a wastewater tunnel structure. Note the failed lining repairs. Structures of concern are sludge storage tanks, chambers, and head works structures. Typically, up to 15 cm (6 in.) of concrete corrosion loss has been found on many structures, affecting both structural integrity and function.

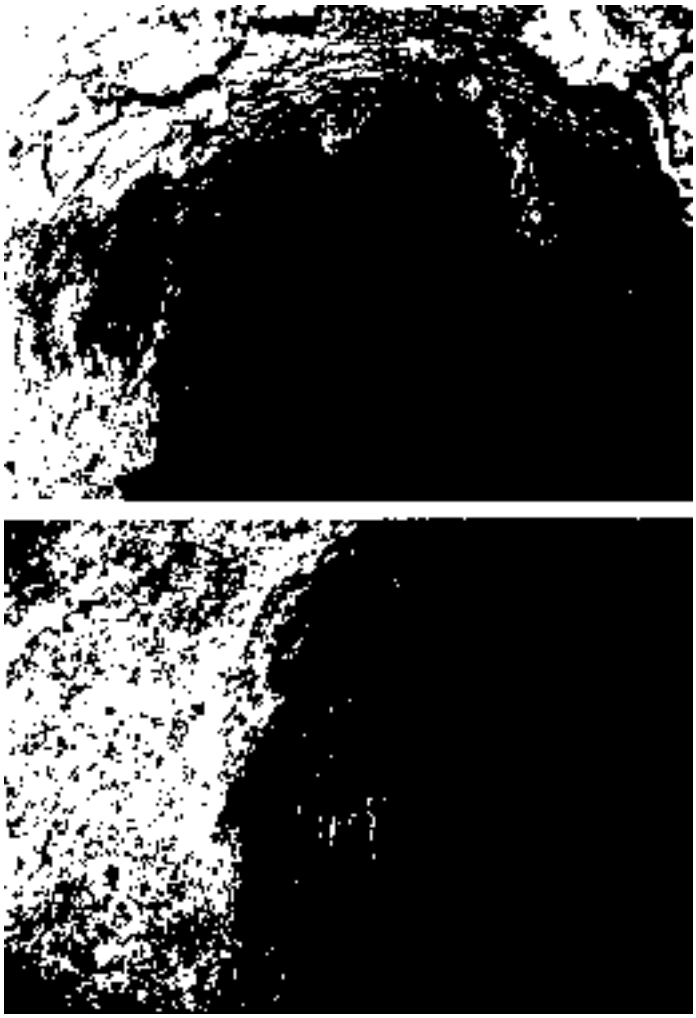


Fig. 5 Two views of severe corrosion of wastewater tunnel structure. Note the failed lining repairs. Courtesy of S. Paul, CorrTech, Inc.

The extent of deterioration and factors contributing to it need to be defined by inspection of pipe barrel interiors, manholes and risers, flow splitting chambers, sludge tanks, clarifiers, and tunnels associated with wastewater treatment systems. The purpose of this task is to document the location and extent of coating and concrete deterioration and corrosion of reinforcing steel.

Much of the existing infrastructure, both of steel and reinforced concrete construction, can be internally lined or externally coated to prevent deterioration. Linings can be used to control corrosion of the structure as well as provide structural integrity where leaks exist. Properly selected and applied high-performance epoxy can protect H_2S headspace corrosion on wastewater treatment facilities and related structures. The spray-on or trowel-on high-performance structural epoxy is a proven system for these applications in the water/wastewater,

power generation, and other industries. Figure 6 provides a view of a severely deteriorated sludge storage tank during repairs.

Fig. 6 Severely deteriorated sludge storage tank during repairs. A structural epoxy will be applied to the reinforced concrete. Courtesy of S. Paul, CorrTech, Inc.

Prevention of Corrosion-Related Failures of Metals

Following analysis of corrosion failures, corrective actions are recommended to prevent or delay recurrence of failures. The approaches to corrective and preventive actions include:

- Change in the environment
- Change in the alloy or heat treatment
- Design change
- Use of galvanic protection (cathodic protection and anodic protection)
- Use of inhibitors
- Use of nonmetallic coatings and liners
- Application of metallic coatings
- Use of surface treatments, thermal spray, or other surface modifications
- Corrosion monitoring
- Preventive maintenance

Change in the Environment. Environmental control can be a complicated task. A complete understanding of the environment is key to changing or treating it. A detailed analysis of the environment should be performed prior to the implementation of any modifications or treatments.

Safeguards must be put in place to avoid unintended consequences of altering the environment, such as the introduction of a contaminant. In some cases, an environment is subject to seasonal changes such as changes in water quality and temperature. Controls must then be seasonally adjusted as well.

Controlling the pH to lower corrosion rates is a common practice in municipal and industrial water supplies. The chemicals that may be introduced are limited for public health reasons.

Pourbaix diagrams are used to predict material behavior with the adjustment of the pH. Pourbaix diagrams show the conditions of potential and pH under which the metal either is nonreactive (immune) or can form oxides. The diagrams do not provide any information regarding corrosion rates, however. Pourbaix diagrams also provide information regarding the formation of barrier films; however, conclusions regarding their effectiveness in the presence of specific ions cannot be drawn. General information can be retrieved from the diagrams indicating the potential for a corrosive situation in a given environment at a specific pH and how to adjust the pH to avoid corrosion.

Changing the environment can be performed in various ways without the use of additives. Ozone treatment is one method used in municipal water systems to treat for MIC. Stirring a tank of fluid may be all that is required to avoid pitting corrosion.

Change of Alloy and Heat Treatment. Alloy changes are often employed to combat corrosion problems, particularly when the replacement material is cost effective. For example, a copper alloy prone to selective leaching can easily be replaced by an inhibited copper alloy that is not susceptible to this kind of corrosion failure.

While proper selection of the material prior to use is the ideal, a complete understanding of the operating electrochemical environment, stresses induced by the operation, inherent stresses in the material from fabrication, and cyclic operating temperatures may not be possible. A failure and subsequent analysis or research of case histories of failures in similar situations may be needed to provide the necessary material selection information. Reviewing corrosion data for a specific environment or performing corrosion testing is helpful in determining the best material for the application.

Design changes involving components, processes, or systems may be necessary to alleviate corrosion. Poor designs can cause materials highly resistant to uniform corrosion attack to fail catastrophically to local attack. Conditions that may cause an acceleration of corrosion include stagnation and rapid fluid flow. Designs that minimize the possibility of crevice corrosion, erosion-corrosion, cavitation, and impingement are desired. Figure 7 shows good and poor examples of how design affects local corrosion (Ref 5).

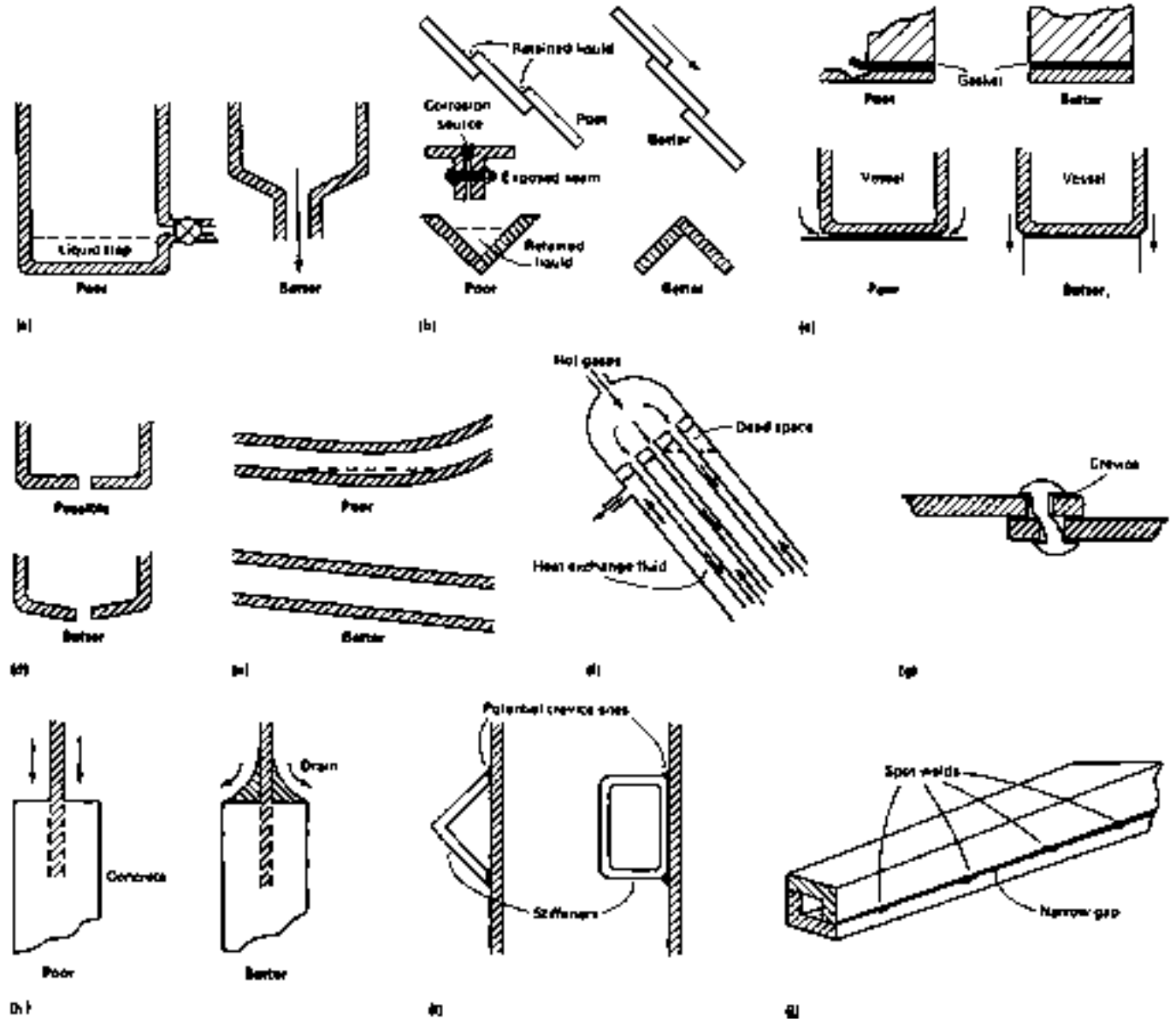


Fig. 7 Examples of how design and assembly can affect localized corrosion by creating crevices and traps where corrosive liquids can accumulate. (a) Storage containers or vessels should allow complete drainage; otherwise, corrosives can concentrate in the bottom of vessel, and debris may accumulate if the vessel is open to the atmosphere. (b) Structural members should be designed to avoid retention of liquids; L-shaped sections should be used with the open side down, and exposed seams should be avoided. (c) Incorrect trimming or poor design of seals and gaskets can create crevice sites. (d) Drain valves should be designed with sloping bottoms to avoid pitting of the base of the valve. (e) Nonsloping tubing can leave pools of liquid at shutdown. (f) Nonvertical assembly of the heat exchanger permits a dead space that may result in overheating if very hot gases are involved. (g) Nonaligned assembly distorts the fastener, which creates a crevice and may result in a loose fitting that can contribute to vibration, fretting and wear. (h) Structural supports should allow good drainage; use of a slope at the bottom of the member allows liquid to run off, rather than impinging directly on the concrete support. (i) Continuous welding is necessary for horizontal stiffeners to prevent the formation of traps and crevices. (j) Square sections formed from two L-shaped members require continuous welding to seal out the external environment. Source: Ref 5

Conditions that cause serious corrosion problems include airborne contamination, dust, pollution, acid rain, and proximity to a seacoast. Designing to avoid retention of such contaminants is critical. Thus, design specifications for construction projects must include normal and extremes of environment conditions.

Design change may also include the isolation of dissimilar metals. The article “Forms of Corrosion” in this Volume includes a discussion of galvanic corrosion and protection. Isolation may include the use of coatings and nonmetallic insulators.

Corrosion is sometimes caused or aggravated by applied stress. For instance, SCC of stainless steel will occur when the steel is exposed to chlorides in the presence of tensile stresses. The magnitude of stresses obviously is affected by the fabrication techniques and the component design. Changes in the design of the components in areas exposed to chlorides that serve to reduce the stress concentrations may be enough to alleviate the possibility of SCC. Areas of stress concentration such as sharp radii or joints are vulnerable. Surface conditions, such as rough surfaces, dents, or scratches may also provide preferential sites for corrosion to occur. Designing to minimize these conditions will reduce corrosion.

Galvanic protection can be either cathodic protection or anodic protection. Cathodic protection reduces the corrosion rate by shifting the corrosion potential of the electrode toward a less oxidizing potential through use of an external electromotive force. Anodic protection is defined as a technique to reduce the corrosion rate of a metal by polarizing it into its passive region, where dissolution rates are low.

Cathodic protection systems require an anode, a cathode, a continuous electric circuit between the cathode and the anode, and the presence of an electrolyte. The two types of cathodic protection are sacrificial-anode protection and impressed-current protection.

Sacrificial-anode protection is the simplest method. A material that is more anodic than the metal to be protected becomes sacrificial. A typical cathodic protection design is shown in Fig. 8 (Ref 6).

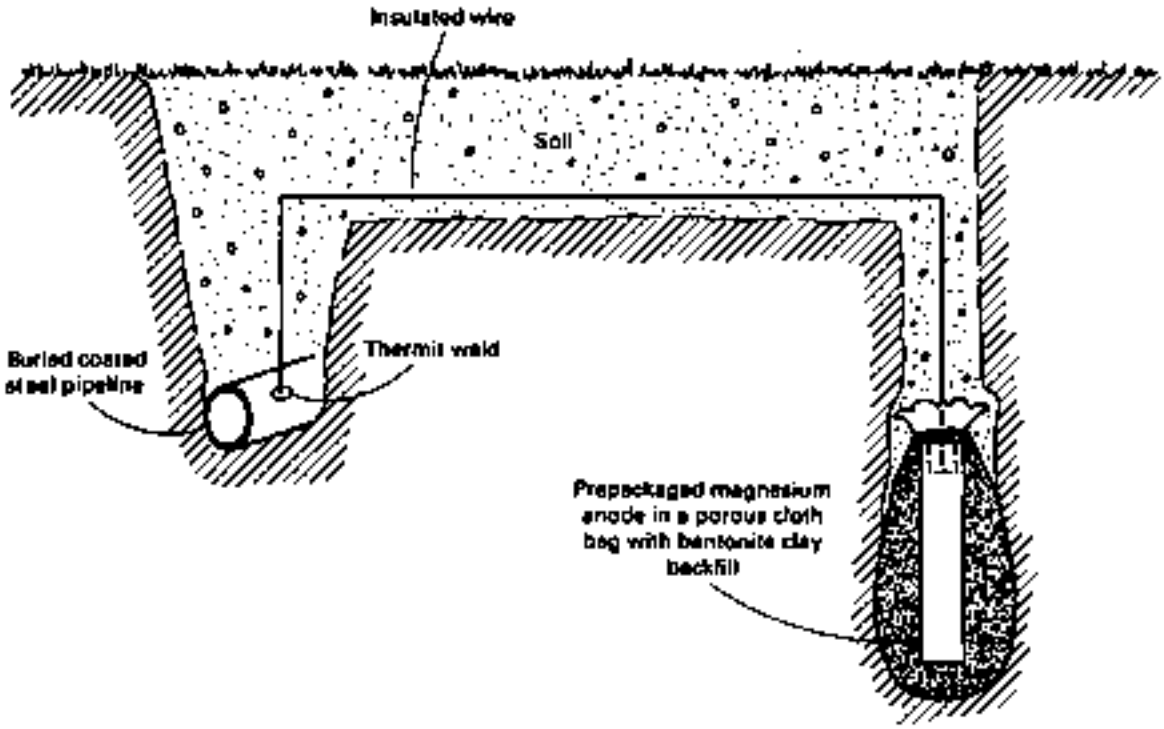


Fig. 8 Cathodic protection of a buried pipeline using a buried magnesium anode. Source: Ref 6

Magnesium, zinc, and aluminum alloys are common sacrificial anodes. Magnesium anodes are most commonly used for buried soil applications. Zinc is most often used for freshwater and saltwater marine applications. Aluminum alloys are most often used for offshore structures. Prevention of passivation of aluminum is key for effective protection. Alloying elements of tin, antimony, and mercury are used for this purpose.

Impressed-current protection requires buried anodes and an electrical connection between the protected structure and the current source. A power rectifier supplies direct current to the buried electrodes (the structure and the anode). The direct-current source reverses the natural polarity and allows the materials to act as anodes.

A typical impressed-current system of a buried pipeline is shown in Fig. 9 (Ref 6). Impressed-current anodes must be corrosion resistant and durable in the environment of use. There are several materials that are used for impressed-current anodes: high-silicon cast irons, graphite, polymeric-coated wires, precious metals, lead alloys, and ceramics. Each material has specific applications. Reference 6 provides detailed information regarding the applications of the impressed-current anodes. Impressed-current cathodic protection must have an external power source. This power source is most often rectified alternating-current power, but batteries, solar cells, and other alternative power supplies can be used in remote locations.

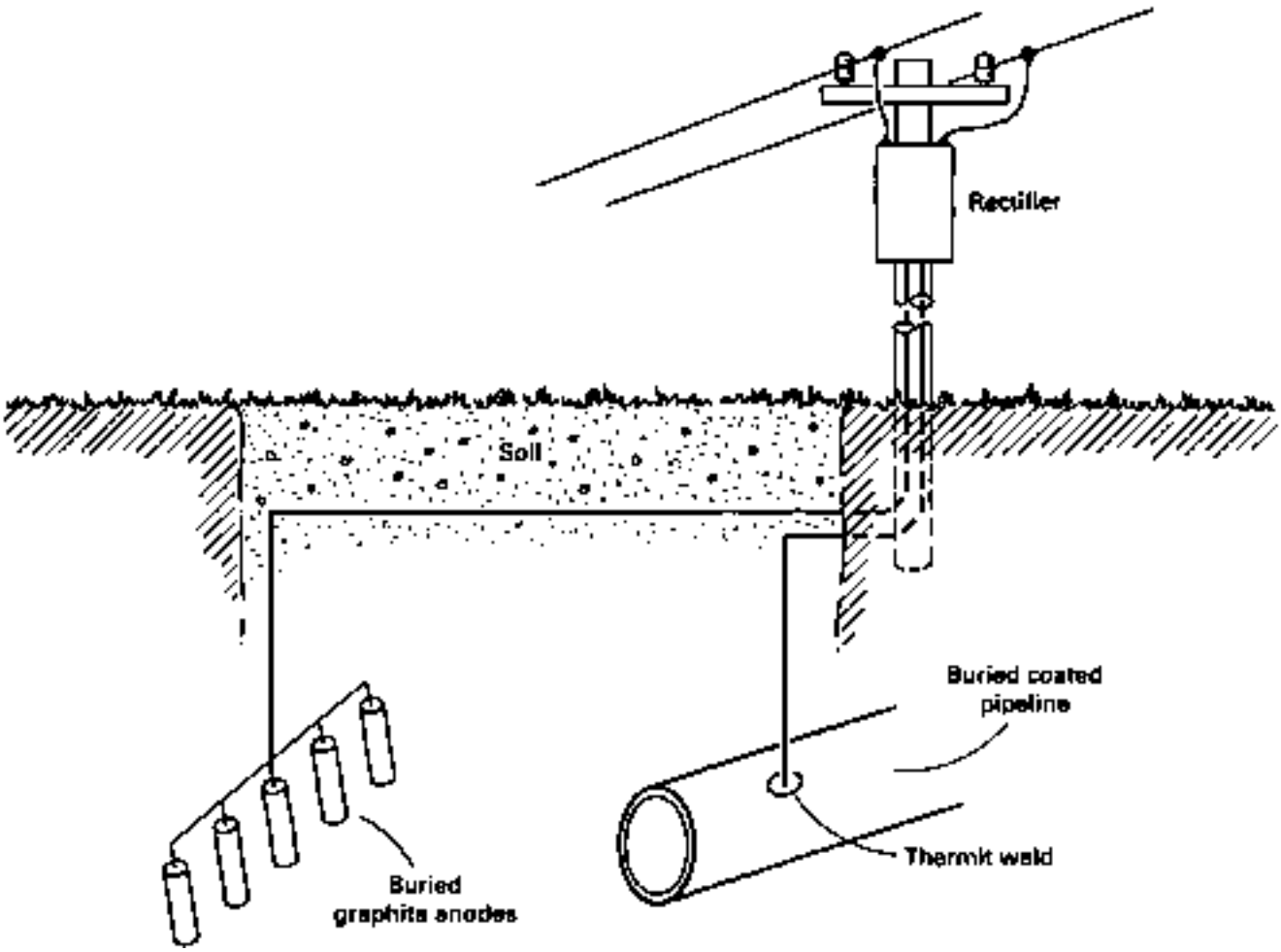


Fig. 9 Impressed-current cathodic protection of a buried pipeline using graphite anodes. Source: Ref 6

Anodic protection is based on the phenomenon of passivity. A limited number of metals can achieve passivity. Anodic protection requires that the potential of the metal be controlled. Shifting the potential of the metal to the passive range can reduce the corrosion rate of an active-passive metal. The anodic polarization of the metal results from the formation of the passive layer, which is relatively insoluble in the chemical environment. The most common uses for anodic protection include storage tanks, process reactors, heat exchangers, and transportation vessels. Anodic protection has been successfully employed in sulfuric acid environments to extend tank life.

Use of Inhibitors. Inhibitors are defined as a chemical substance or combination of substances that, when present in the environment, prevent or reduce corrosion without significant reaction with the components of the environment. Chemical treatment or inhibitors can be used for corrosion control in aqueous environments. Inhibitors are classified as anodic inhibitors, cathodic inhibitors, and mixed inhibitors. Anodic inhibitors suppress the rate of metal ions being transferred into the aqueous environments. Cathodic inhibitors impede the oxygen-reduction reaction. Mixed inhibitors hinder both reactions.

Anodic Inhibitors. There are two types of anodic inhibitors, oxidizing and nonoxidizing. Chromates and nitrites act in the absence of oxygen. Chromates are effective in protecting ferrous and nonferrous alloys. An oxidation

layer of a chromium oxide (Cr_2O_3) and a ferrous oxide (Fe_2O_3) is formed. A critical concentration of chromates must be present to protect against aggressive ions, such as chlorides and sulfates.

Similarly, nitrites also require critical concentration levels to maintain optimal corrosion protection. Nitrites are commonly used in closed recirculating systems. The pH levels are usually maintained above 7.5 to avoid extensive pitting of carbon steel materials. Borax-nitrite has been used for open recirculating cooling water systems in conjunction with the use of biocides.

Molybdates are typically combined with other inhibitors for optimal treatment. An oxidizing agent such as oxygen is required to form a protective film. Molybdates are generally most effective in the range of 5.5 to 8.5 pH. Phosphates also require the presence of oxygen in order to form a protective oxide layer. Phosphates are generally used in alkaline environments of pH greater than 8.

Cathodic inhibitors suppress the corrosion rate by restricting the availability of oxygen or by altering sites favorable for cathodic hydrogen evolution. While it is generally true that cathodic inhibitors are not as effective as anodic inhibitors, they are considered safer. Additionally, they do not cause localized attack, as anodic inhibitors do when used in insufficient amounts. There are several different kinds of cathodic inhibitors. Precipitating inhibitors produce insoluble films on the cathode in high pH conditions. By this means, the cathode is isolated from the environment. Zinc ions may be used as a means to precipitate zinc hydroxide at the cathode. The protective film is enhanced by the combination of other inhibitors of a multicomponent treatment program.

Polyphosphate-inhibitor treatments are widely used in part due to the economics of the program. The pH of the environment greatly affects the protective nature of the phosphate film. The pH should be maintained in a near-neutral range, between 6.5 and 7.5, if both steel and copper alloys are part of the system.

Phosphonates are slightly different than polyphosphates. Phosphonates form direct phosphorus-carbon bonds, while polyphosphates form a phosphorous-oxygen bond. The phosphates are sensitive to water quality and temperature.

There are numerous methods developed that use multicomponent systems. The combinations have been developed to suit a specific environment or varying environment conditions. Copper inhibitors are widely used. These inhibitors are used to prevent excessive copper deposition on steel components occurring from the presence of dissolved copper in the circulating waters. These inhibitors control the corrosion of the copper materials.

Oil and Gas Refinery Inhibitors. A number of inhibitors are designed specifically to meet the needs of the oil and gas and refinery industries. An extensive discussion regarding the use of inhibitors, the application techniques, and the monitoring of the effectiveness of inhibitor use can be found in the article “Corrosion Inhibitors for Oil and Gas Production” in *Corrosion*, Volume 13 of the *ASM Handbook* (Ref 7).

Factors affecting the inhibition systems include the corrosivity of the environment, the type of environment (oxidizing or nonoxidizing), the temperature, the pH, the metallurgy of the system components, the design of the system, and the economics of the inhibition program. Government regulations for use of the chemicals employed, regulations of the particular industry for which the inhibitor is used, potential human handling and consumption concerns, and side effects on processing fluids are all significant considerations.

Velocity may also play a major role in the effectiveness of an inhibitor. The higher the velocity of the fluid environment, the greater the amount of inhibitor reaching the surface. This increased velocity, in essence, has the same effect as increasing the inhibitor concentration. In addition, the fast-moving solution prevents the buildup of debris and deposits on the surface. This allows for better wetting of the inhibitor on the surface.

Use of Nonmetallic Coatings and Linings. The most widely used methods of corrosion control today are the application of nonmetallic coatings and linings. This includes organic coatings and liners, porcelain enamel, and chemical-setting ceramic linings. There are constant improvements in this area of protection. More surface area of metals is protected by this method than all other methods combined (Ref 8).

Surface preparation is very important for coating life and effectiveness. The Steel Structures Painting Council (SSPC) and NACE International have published surface preparation and cleaning standards. Standards published by the Swedish Standards Institute have been widely used in Europe and Asia. To ensure that the coating has been applied properly, a third-party inspection may be appropriate in some circumstances. NACE International provides training and certification of such inspectors.

This section focuses primarily on liquid applied coatings used in atmospheric and immersion service. The application of coatings is most successful when the base metal corrosion does not exceed 1.3 mm/yr (0.050 in./yr). The resin or organic binding of the coating material is the most influential factor in determining the

resistance and properties. Pigments, solvents, and additives will also influence the application properties and protective capability of the applied film.

Auto-Oxidative Cross-Linked Resins. All of the coatings in this class contain drying oils, which consist mainly of polyunsaturated fatty acids and undergo film formation by oxidation drying. Cross-linked resins include alkyd resins, modified alkyd resins, and epoxy resins.

The alkyd resins are the most common coating systems used to combat corrosion. These are not known for their exceptional chemical, alkaline, or moisture resistance; however, the ease of application, low cost, and ability to penetrate the surface, even surfaces that are poorly prepared, make alkyds one of the best choices for steel protection not directly in contact with harsh chemical environments. Some of the most common applications for alkyds are water tanks, highway bridges, structural steel, and machinery housings.

Alkyd modifications include additives that improve specific properties of the material. There are a number of modifications that are available for different applications. Reference 8 identifies several different alkyd variations and their properties.

Epoxy esters are usually prepared by reacting a drying oil with the epoxy resin at elevated temperature. Compared to the alkyd resins, epoxy resins have better adhesion, moisture, and chemical resistance and a slight increase in cost.

Thermoplastic resins are characterized by softening at elevated temperatures. The most useful thermoplastic resins are the vinyls, chlorinated rubbers, thermoplastic acrylics, and bituminous resins (coal tar and asphalt). In general, these resins have good resistance to weathering, acid or alkali exposure, and moisture. The coal tars, chlorinated rubbers, and the vinyls are excellent choices for superior moisture resistance. Vinyl-solution coatings are most often used in applications requiring toughness and water resistance. Submersion applications such as locks and dams may use vinyl-solution coatings. Vinyl paints in conjunction with zinc-rich primers are used for bridge applications.

Acrylic coatings exhibit excellent chemical resistance to weathering environments and moisture; however, they are not considered suitable for immersion service or strong chemical environments. Typical properties include excellent color retention, gloss, and ultraviolet stability. Thermoplastic acrylics are often used in the automotive industry in finishing and refinishing. One of the most familiar uses of acrylics is for water-base indoor and outdoor paints. The major benefit of acrylic-based paint is the water cleanup.

Bituminous coatings are most commonly used for roof coatings, highway or pavement sealers, underground coatings, and waterproofing compounds. Asphaltic materials are less affected by ultraviolet light and have better atmospheric weathering resistance than coal tar coatings, so the asphaltic coatings are usually used for above-grade applications. However, coal tar coatings are more resistant to moisture, acids, and alkalies than are the asphaltics.

Cross-linked thermosetting coatings refer to chemically cured coatings that harden by a chemical reaction either with a copolymer or with moisture. Examples of chemically cross-linked thermosetting coatings include epoxies, unsaturated polyesters, high-temperature curing silicones, phenolic linings, and certain urethanes. Chemically cured coatings that react with water include moisture-cured polyurethanes and most of the inorganic zinc-rich coatings. Chemically cross-linked coatings and linings are considered highly corrosion resistant to acids, alkalies, and moisture. They can be formulated to be resistant to abrasion, ultraviolet radiation, and thermal degradation. These properties are optimized when sufficient curing occurs. The curing time is dependent on weather conditions, but may take 7 days. Detailed discussion of specific cross-linked thermosetting coatings is provided in Ref 8.

Zinc-rich coatings have the unique ability to provide galvanic protection to ferrous substrates due to the metallic zinc dust pigment. Metal-to-metal contact of the coating and substrate creates a galvanic couple. Zinc is anodic to the steel and thus provides protection to the steel substrate. There are two categories, organic zinc-rich coatings and inorganic zinc-rich coatings, based on the binders used. The great advantage of this coating system is the elimination of pitting and subfilm corrosion. Zinc-rich coatings are widely used in industrial and marine environments. These coating are often used as primers or first coats, with a topcoat then applied to complete the protection system.

Porcelain Enamels. Porcelain coatings or glass coatings are applied primarily to steel, cast iron, and aluminum. They are matured at 425 °C (800 °F). The porcelain coatings offer barrier protection to the substrate, which must be free of defects. Common applications of porcelain enamels include major appliances, water heater tanks, chemical-processing vessels, heat exchangers, storage tanks, piping, and pump components. The choice of porcelain enamels is primarily for chemical resistance, corrosion protection, weather resistance, mechanical

or electrical properties, appearance, color, cleanability, and thermal shock capability. There are two types of porcelain enamels, ground-coat or cover-coat enamels. Ground-coat enamels contain oxides that promote the adherence of the coating to the substrate. Cover-coat enamels are applied to improve appearance and properties of the coating.

Chemical-Setting Ceramic Linings. Inorganic chemical-setting ceramic linings are widely used as a protective lining for construction materials. Their characteristics include resistance to high temperatures and aggressive corrosion media. Applications include floors, vessels, tanks, scrubbers, air pollution control equipment, and chimneys in a variety of industries. The linings have resistance to strong acids and solvents, temperatures up to 870 °C (1600 °F), good compressive and flexural and abrasion resistance. Monolithic linings are applied by cast or gunite (shotcreting) methods over old and new steel, concrete, and brick and mortar masonry. Inorganic monolithic linings have a certain amount of permeability, and over time acid can penetrate the lining and eventually reach the substrate. Dual-lining systems can combat this problem.

Metallic coatings are another method of corrosion protection for materials. Typical application methods include electroplating, hot dipping, cladding, thermal spray coatings, electroless process, and vapor-deposition process. Metals with greater corrosion resistance than the base material are applied.

Electroplated Coatings. Electroplating is deposition of a metal or alloy in an adherent form on an object serving as a cathode. Electrodeposition is used for cosmetic purposes, to improve conductivity and other surface properties, and for corrosion protection. Specific applications for electroplating include automobile bumpers, trim on appliances, electronic circuits, construction and architectural hardware, plumbing fittings, fasteners, gun bores, and food packaging. The selection of the coating is primarily based on the galvanic series. Coatings are difficult to apply without any imperfections. Thus, the most desirable situation is to choose a coating material that is anodic to the substrate and one that has a very low corrosion rate. Typical materials used as electrodeposited coatings include nickel, nickel-phosphorus, cadmium, zinc, chromium, tin, lead, copper, and gold.

Several factors that influence the performance of the coatings are the design of the part, the stresses in the substrate, the surface profile, and susceptibility to hydrogen absorption. Advantages of electroplating including the variety of coatings available, versatility of applications, variety of application methods, galvanic protection, color options for aluminum, precise thickness control, ductility, and formability after coating, There are some disadvantages to the electroplated coatings, which include color limitations, substrate incompatibilities, design limitations, and the difficulty to plate large objects.

Hot-dip coatings are applied by immersing a metal in a molten bath of coating metal. Typical hot-dip coated materials include zinc-coated steel (galvanized steel), aluminum-coated steel (aluminized), and aluminum-zinc alloy coated steel that are produced by batch or continuous methods. The advantages of these coatings include the ability to coat recessed areas with minimum coating thickness, resistance to mechanical damage, and good resistance to corrosion. There are, however, limitations regarding hot-dip coatings. The coating must melt at a reasonably low temperature, and the steel base metal must not undergo property alterations during the coating process. Galvanized steel has been used for more than 100 years and remains in high demand today. Aluminum-coated steel requires a more difficult process, but is used in major applications including chain links, roofing panels, and automotive exhaust components. Aluminum-zinc alloy coated steel provides a combination of galvanic protection and low corrosion rate. Applications include metal roofing, automotive components, appliances, and corrugated pipe.

Vapor-deposited coatings have been used to modify the surface properties of the materials. These coatings are most often used to improve the wear properties of a material, but they have also been used as corrosion-resistant coatings. The two major categories of these coatings are physical vapor deposition (PVD) and chemical vapor deposition (CVD). Sputtering, the process of transporting material from a source to a substrate by bombardment of a target with gas ions, is the most widely used method of application for PVD. This method is optimal for thin metal films where good adhesion and uniformity are necessary. However, the limited thickness and high cost are disadvantages of this method. Chemical vapor deposition is a process used in the semiconductor and cutting tool industry for thick, dense, and high-quality films.

Thermal Spray Coatings. Thermal spraying is a process in which fine molten metallic or nonmetallic material is sprayed onto a substrate to form a coating. Advantages of the thermal spray process are the ability to apply thick coatings, excellent paintability of the surface, no heat distortion of the substrate, no curing time required, and the capability of applying coatings in the field (Ref 9). Long-term corrosion protection of iron and steel is achieved by thermal spray coatings. The bond between the sprayed coating and the substrate is generally

mechanical. Surface preparation is critical to the performance of the coating. One application of thermal spray coatings is for the protection of steel components from heat oxidation at temperatures up to 1095 °C (2000 °F). High-temperature corrosion resistance usually requires alloys of aluminum or nickel-chromium alloys.

Zinc and aluminum coatings are commonly applied to combat corrosion. Stainless steel and Hastelloy alloys are commonly used to combat corrosion in the chemical industry. Sealers or topcoats are commonly used to extend the life of thermal spray coatings and for decorative purposes.

Conversion coatings are a form of barrier protection in which a portion of the base metal is converted into a protective compound by a chemical or electrochemical treatment. Because the continuity of the conversion coating is more certain than that of an organic coating, conversion coatings treatment often precedes the application of organic coatings to enhance overall corrosion resistance. Proper cleaning of the base metal is essential to ensure this continuous coverage.

Phosphate conversion coatings are applied to various metal substrates to enhance corrosion resistance and improve paint adhesion. Corrosion protection is not the primary function of phosphates by themselves, but they improve the subsequent corrosion protection treatment.

The three categories of phosphate coatings are iron phosphates, zinc phosphates, and heavy phosphates. Iron phosphates are used when good paint adhesion and low cost are the primary concern. Zinc phosphates provide good paint adhesion, and they hold oils and waxes well. They provide good painted corrosion resistance. Heavy phosphates (manganese phosphates) are more expensive, but have an increase in unpainted corrosion resistance.

Chromate conversion coatings are formed by a chemical or electrochemical treatment of metals in solutions containing chromium ions and usually other components. The coatings are applied to enhance bare or painted corrosion resistance, to improve the adhesion of paint or other organic finishes, and to provide the metallic surface with a decorative finish. This process is used on many metals including aluminum, zinc, steel, copper, tin, and nickel. The level of protection is dependent on the substrate metal, the type of chromate coating used and the coating weight. The mechanism of corrosion protection is that of a barrier insulation from the environment that inhibits the cathodic-corrosion reactions.

Aluminum anodizing is an electrochemical method of converting aluminum into aluminum oxide at the surface of the component. Aluminum anodizing is used extensively. Chromic, sulfuric, and hard-coat anodizing are the most common methods. Chromic is the least used. It is relatively soft, but is useful when precise dimension of the part must be maintained. Sulfuric anodizing is the most widely used due to its light to moderate wear resistance, appearance, and corrosion resistance. Chromic and sulfuric coatings are porous and are typically sealed to close the pores inherent in the process. Hard-coat is typically not sealed, as sealing softens the coating and would reduce its wear resistance.

Surface modification refers to the modification of the surface composition or structure by using energy or particle beams. Elements may be added to influence the surface characteristics of the substrate by forming alloys, metastable alloys or phases, or amorphous layers. Surface-modified layers have a greater similarity to metallurgical alloying rather than the chemically adhered and mechanical bonds created by conversion or organic coating methods. Ion implantation and laser surface processing are two methods of surface modification. Ion implantation, derived from the semiconductor industry, is a relatively specialized process because of the cost of equipment and high degree of training required of operators. The process involves electrostatically accelerating an ionized species into the substrate. It is more recognized for its improvement of wear resistance than corrosion resistance, although there are corrosion-resistance benefits in certain cases.

Laser surface modification can modify the metallurgical structure of the surface and change surface properties without adversely affecting the bulk properties. There are generally three forms of laser surface modifications: transformation hardening, surface melting, and surface alloying. Laser processing can be used to create corrosion-resistant surface layers. Usefulness of laser processing is limited because it requires a planar substrate and the substrate must be compatible with the thermal conduction requirements of the process.

References cited in this section

5. P. Elliott, Design Details to Minimize Corrosion, *Corrosion*, Volume 13, *ASM Handbook*, ASM International, 1987, p 339

6. R.H. Heidersbach, Cathodic Protection, *Corrosion*, Volume 13, *ASM Handbook*, ASM International, 1987, p 466–477

7. P.J. Stone, Corrosion Inhibitors for Oil and Gas Production, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 478

8. K.B. Tator, Organic Coatings and Linings, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 399

9. R.H. Unger, Thermal Spray Coatings, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 459

Acknowledgments

The section "Electrochemical Nature of Corrosion" was adapted from "Corrosion Failures," from the course *Principles of Failure Analysis* (Materials Engineering Institute/ASM International).

Analysis and Prevention of Corrosion-Related Failures

S.R. Freeman, Millennium Metallurgy, Ltd.

References

1. "Corrosion Costs and Preventive Strategies in the United States," FHWA-RD-01-156, Federal Highway Administration, 2002

2. *Wear and Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* ASTM

3. "Applying Statistics to Analysis of Corrosion Data," G 16, *Annual Book of ASTM Standards,* ASTM

4. "Practice for Dealing with Outlying Observations," E 178, *Annual Book of ASTM Standards,* ASTM

5. P. Elliott, Design Details to Minimize Corrosion, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 339

6. R.H. Heidersbach, Cathodic Protection, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 466–477

7. P.J. Stone, Corrosion Inhibitors for Oil and Gas Production, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 478

8. K.B. Tator, Organic Coatings and Linings, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 399

9. R.H. Unger, Thermal Spray Coatings, *Corrosion,* Volume 13, *ASM Handbook,* ASM International, 1987, p 459

Forms of Corrosion

Introduction

CORROSION is the electrochemical reaction of a material and its environment. This article addresses those forms of corrosion that contribute directly to the failure of metal parts or that render them susceptible to failure by some other mechanism. In particular, mechanisms of corrosive attack for specific forms of corrosion are described. Other articles in this Volume discuss analysis and prevention of several types of failure in which corrosion is a contributing factor. These articles include “Stress-Corrosion Cracking,” “Liquid-Metal and Solid-Metal Induced Embrittlement,” “Hydrogen Damage and Embrittlement,” “Corrosive Wear Failures,” “Biological Corrosion Failures,” and “High-Temperature Corrosion-Related Failures.”

The rate, extent, and type of corrosive attack that can be tolerated in an object vary widely, depending on the specific application and initial design. Whether the observed corrosion behavior was normal for the metal used in the given service conditions helps determine what inspection and maintenance action, if any, should be taken.

All corrosion reactions are electrochemical in nature and depend on the operation of electrochemical cells at the metal surface. This mechanism is discussed in the first section, “Galvanic Corrosion.” Galvanic corrosion applies even to generalized uniform chemical attack—described in the section “Uniform Corrosion”—in which the anodes and cathodes of the cells are numerous, small, and close together. The analysis of corrosion failures and the development of suitable corrective measures require not only a basic understanding of electrochemistry, but also the ability to apply fundamental principles of chemistry and metallurgy. It is useful to consider these principles in the context of the form that the corrosion takes. Figure 1 is a schematic of several forms of corrosion, including those described in the sections “Pitting and Crevice Corrosion,” “Intergranular Corrosion,” and “Selective Leaching” in this article. The figure also includes external conditions that produce aggressive corrosion such as velocity-affected corrosion, which is addressed in this article, and stress-corrosion cracking (SCC), which is addressed elsewhere in this Volume. These reviews of corrosion forms and mechanisms are intended to assist the reader in developing an understanding of the underlying principles of corrosion; acquiring such an understanding is the first step in recognizing and analyzing corrosion-related failures and in formulating preventive measures.

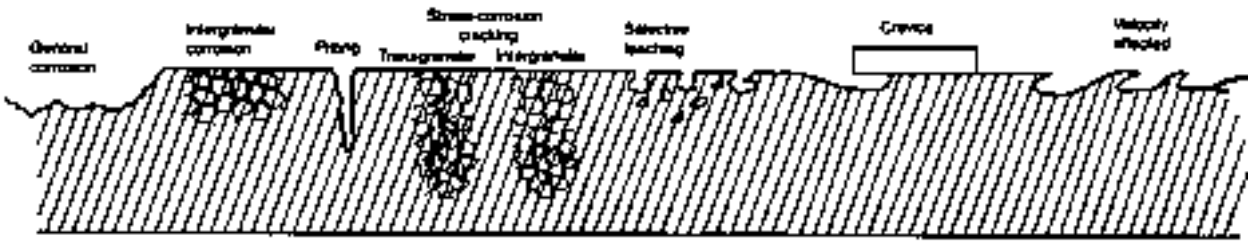


Fig. 1 Forms of corrosion. Adapted from Ref 1

More detailed discussions of the fundamentals and mechanisms of corrosion are provided in the References listed at the end of this article and in *Corrosion*, Volume 13 of *ASM Handbook.*

Forms of Corrosion

Galvanic Corrosion

Although sometimes considered a form of corrosion, galvanic corrosion is more accurately considered a type of corrosion mechanism, because galvanic action is the basis for, or can accelerate the effects of, other forms of corrosion. Uniform attack, pitting, and crevice corrosion can all be exacerbated by galvanic conditions.

Galvanic corrosion occurs when two dissimilar conducting materials (metallic or nonmetallic) are in electrical contact. It usually consists of two dissimilar conductors in electrical contact with each other and with a common conducting fluid (an electrolyte), or it may occur when two similar conductors come in contact with each other via dissimilar electrolytes. The former is the more common condition.
When two materials are in electrical contact and connected by an electrolyte, a galvanic current flows between them because of the inherent potential difference between the two. The resulting reaction is referred to as couple action, and the electrically coupled system is known as a galvanic cell. The couple consists of an anode (which liberates electrons and corrodes typically by metal dissolution and/or metal oxide formation) and a cathode (which gains electrons and typically liberates hydrogen to the electrolyte and/or reduces oxides). Both reactions at the anode and cathode must proceed simultaneously for galvanic corrosion to occur. The overall reaction results in the corrosion of the anode by metal dissolution or oxidation, and an electrical current flows from the more anodic metal through a conducting medium or electrolyte to the other more cathodic metal (liberating hydrogen and/or causing a “reduction” of oxides). This action is seen in Fig. 2 where a couple is made up of steel and mill scale.

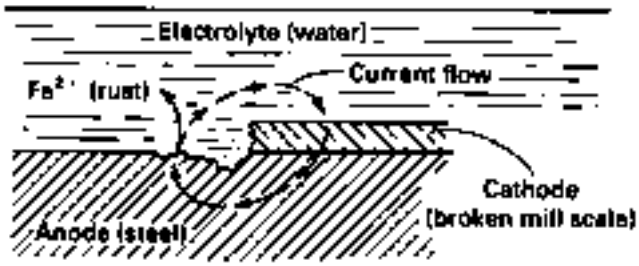


Fig. 2 Breaks in mill scale (Fe_3O_4) leading to galvanic corrosion of steel

The three essential components for galvanic corrosion are:

- Materials possessing different surface potential
- A common electrolyte
- A common electrical path

A mixed-metal system in a common electrolyte that is electrically isolated will not experience galvanic corrosion, regardless of the proximity of the metals or their relative potential or size.

Forms of Corrosion

Factors Affecting Galvanic Corrosion

The intensity of galvanic corrosion is affected by the following factors:

- The potential difference between the metals or alloys
- The nature of the environment
- The polarization behavior of the metals or alloys
- The geometric relationship of the component metals or alloys

The severity of galvanic corrosion depends on the amount of current flow. The relative potentials and polarization behavior of metals and alloys are key consideration discussed later in this article. Current flow also depends on the physical configuration, the relative area, distance, and shape, of the components.

Geometric Factors

Effect of Surface Area. The relative size of the anodic and cathodic surfaces is important. The intensity of galvanic attack is related to the relative size of the metals in electrical contact. Large cathodic areas coupled to small anodic areas will aggravate galvanic corrosion and cause severe dissolution of the more active metal. The reverse situation—large anodic areas coupled to small cathodic areas—produces very little galvanic current. This is why imperfections or holidays in protective coatings may lead to severe galvanic corrosion in the localized region of the coating imperfection. It is extremely dangerous to coat the anodic member of a couple because this may only reduce its active area, which severely accelerates the attack at these holidays in the otherwise protective coating. If inert barrier coatings are employed, both the anode and cathode must be protected.

Effect of Distance. Dissimilar metals in a galvanic couple that are in close physical proximity usually suffer greater galvanic effects than those that are farther apart. The distance affects the resistance of the current path in the solution and the external circuit. Thus, if dissimilar pipes are butt welded with the electrolyte flowing through them, the most severe corrosion will occur adjacent to the weld on the anodic member.

Effect of Shape. The geometry of the circuit elements determines the electrical potential gradient, which causes the current to flow.

Galvanic Series

Because galvanic corrosion is directly related to the electrical current caused by the natural potential difference or electromotive force (emf) between different metals, it is useful to rank metals according to their relative potentials for a given electrolytic solution. For example, the galvanic series of potentials for metals in a chloride-containing aqueous solution (i.e., seawater) is often used for purposes of general comparison. This series is shown in Fig. 3. When two metals of a galvanic cell are widely separated in the galvanic series, there is a greater flow of current than for metals with small difference in potential. The galvanic series also allows one to determine which metal or alloy in a galvanic couple is more active. Metals that are more anodic in a given galvanic cell are prone to corrode by metal dissolution or oxidation. The more cathodic material is more corrosion resistant (i.e., more noble).

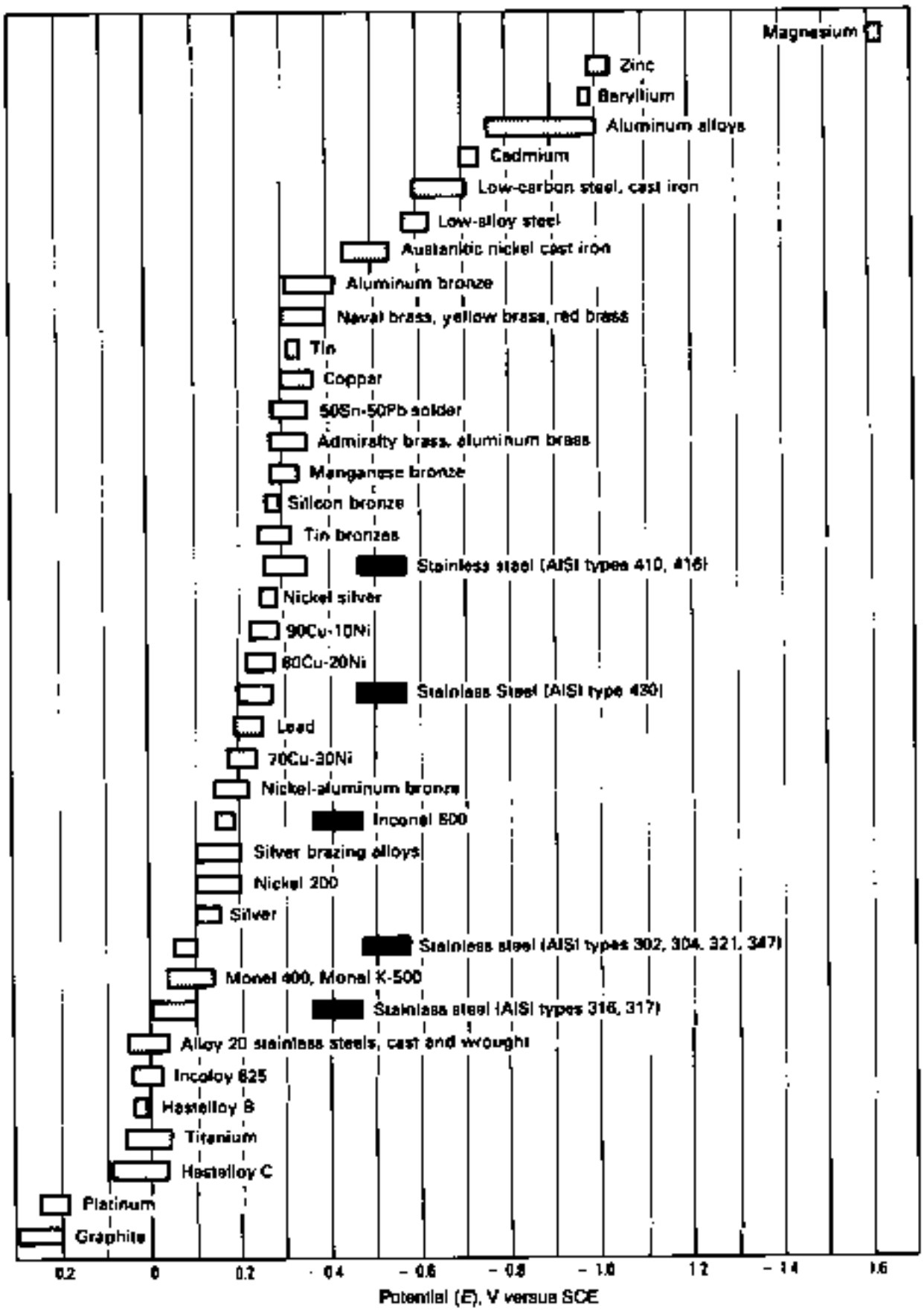

Magnesium
Zinc
Beryllium
Aluminum alloys
Cadmium
Low-carbon steel, cast iron
Low-alloy steel
Austenitic nickel cast iron
Aluminum bronze
Naval brass, yellow brass, red brass
Tin
Copper
50Sn-50Pb solder
Admiralty brass, aluminum brass
Manganese bronze
Silicon bronze
Tin bronzes
Stainless steel (AISI types 410, 416)
Nickel silver
90Cu-10Ni
80Cu-20Ni
Stainless Steel (AISI type 430)
Lead
70Cu-30Ni
Nickel-aluminum bronze
Inconel 600
Silver brazing alloys
Nickel 200
Silver
Stainless steel (AISI types 302, 304, 321, 347)
Monel 400, Monel K-500
Stainless steel (AISI types 316, 317)
Alloy 20 stainless steels, cast and wrought
Incoloy 825
Hastelloy B
Titanium
Hastelloy C
Platinum
Graphite
0.2
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Potential (E), V versus SCE

Fig. 3 Galvanic series of metals and alloys in seawater. Alloys are listed in order of the potential they exhibit in flowing seawater; those indicated by the black rectangle were tested in low-velocity or poorly aerated water and at shielded areas may become active and exhibit a potential near -0.5 V. Adapted from Ref 2

Although the measurement of potentials has limitations as previously noted, galvanic series based on seawater or other standard electrolytes are worthwhile for initial materials selection for multiple metal/alloy systems in a given environment. A properly prepared galvanic series is easy to use and can answer simple galvanic-corrosion questions regarding the positions and relative separation of metals and alloys in a galvanic series. However, the comparison of galvanic potentials is only a ranking of general susceptibility; it does not address all the conditions and factors that influence galvanic corrosion. Limitations on the use of galvanic series include:

- No information is available on the rate of corrosion
- Active-passive metals may display two, widely differing potentials
- Small changes in electrolyte can change the potentials significantly
- Potentials may be time dependent

Creating a galvanic series is a matter of measuring the corrosion potential of various materials of interest in the electrolyte of interest against a reference electrode half-cell, such as saturated calomel, as described by the procedure in ASTM G 82 (Ref 3). To prepare a valid galvanic series for a given materials and environment of interest, all the factors affecting the potential must be addressed. This includes material composition, heat treatment, surface preparation (mill scale, coatings surface finish, etc.), environmental composition (trace contaminants, dissolved gases, etc.), temperature, flow rate, solution concentration, and degree of agitation or aeration. In addition, corrosion product films and other changes in surface composition can also occur in some environments, and so exposure time is important, too.

With certain exceptions, the galvanic series in seawater is broadly applicable in other natural waters and in uncontaminated atmospheres. Bodily fluids are similar to saltwater, so as a first approximation the galvanic series is useful for implants.

However, metals behave differently in different environments. The relative positions of metals and alloys in the galvanic series can vary significantly from one environment to another. The position of alloys in the galvanic series for seawater is not necessarily valid in nonsaline solutions. For example, aluminum is anodic to zinc in an aqueous 1 *M* sodium chromate (Na_2CrO_4) solution and cathodic to iron in an aqueous 1 *M* sodium sulfate (Na_2SO_4) solution (Ref 4).

The emf series is a table that lists in order the standard electrode potentials of specified electrochemical reactions. The potentials are measured against a standard reference electrode when the metal is immersed in a solution of its own ions at unit activity (Table 1). Similar to the galvanic series, it is a list of metals and alloys arranged according to their relative potentials in a given environment. Generally, the relative positions of metals and alloys in both emf and galvanic series are the same. An exception is the position of cadmium with respect to iron and its alloys. In the emf series, cadmium is cathodic to iron, but in the galvanic series (at least in seawater), cadmium is anodic to iron. Thus, if only the emf series were used to predict the behavior of a ferrous metal system, cadmium would not be chosen as a sacrificial protective coating, yet this is the principal use for cadmium plating on steel.

Table 1 Standard emf series of metals

Metal-metal ion equilibrium (unit activity)	**Electrode potential versus SHE at 25°C (75 °F), V**
Noble or cathodic	
$Au-Au^{3+}$	1.498
$Pt-Pt^{2+}$	1.2
$Pd-Pd^{2+}$	0.987
$Ag-Ag^{+}$	0.799

$Hg-Hg^{2+}$	0.788
$Cu-Cu^{2+}$	0.337
H_2-H^+	0.000
$Pb-Pb^{2+}$	-0.126
$Sn-Sn^{2+}$	-0.136
$Ni-Ni^{2+}$	-0.250
$Co-Co^{2+}$	-0.277
$Cd-Cd^{2+}$	-0.403
$Fe-Fe^{2+}$	-0.440
$Cr-Cr^{2+}$	-0.744
$Zn-Zn^{2+}$	-0.763
$Ti-Ti^{3+}$	-1.210
$Ti-Ti^{2+}$	-1.630
$Al-Al^{2+}$	-1.662
$Mg-Mg^{2+}$	-2.363
$Na-Na^+$	-2.714
$K-K^+$	-2.925
Active or anodic	

Polarization

Electron flow occurs between metals or alloys in a galvanic couple. This current flow between the more active and more noble members causes shifts in potential, because the potentials of the metals or alloys tend to approach each other over time. This shift, where the open-circuit electrode potential changes as a result of the passage of current, is due to polarization of the electrodes. For example, during electrolysis, the potential of an anode becomes more noble, and the potential of the cathode becomes more active than their respective open-circuit (or reversible) potentials. Often polarization occurs from the formation of a film on the electrode surface. The magnitude of the shift depends on the environment, as does the initial potential. If the more noble metal or alloy is more easily polarized, its potential is shifted more toward the more active metal or alloy potential. The shift in potential of the more active metal or alloy in the direction of the cathode is therefore minimized so that accelerated galvanic corrosion is not as great as would otherwise be expected. On the other hand, when the more noble metal or alloy is not readily polarized, the potential of the more active metal shifts further toward the cathode (that is, in the direction of anodic polarization) such that appreciable accelerated galvanic corrosion occurs.

Polarization measurements on the members of a galvanic couple can provide precise information concerning their behavior. The polarization curves and the mixed potential for the galvanically coupled metals in a particular environment can be used to determine the magnitude of the galvanic-corrosion effects as well as the type of corrosion.

An important application in the use of polarization measurements in galvanic corrosion is the prediction of localized corrosion. Polarization techniques and critical potentials are used to measure the susceptibility to pitting and crevice corrosion of metals and alloys coupled in chloride solutions. In addition, this technique is valuable in predicting galvanic corrosion among three or more coupled metals or alloys.

Polarization measurements can be made by generating stepped potential or potentiodynamic polarization curves or by obtaining potentiostatic information on polarization behavior. The objective is to obtain a good indication of the amount of current required to hold each material at a given potential. Because all materials in the galvanic system must be at the same potential in systems with low solution resistivity, such as seawater, and because the sum of all currents flowing between the materials must equal 0 by Kirchoff's Law, the coupled potential of all materials and the galvanic currents flowing can be uniquely determined for the system. The corrosion rate can then be related to galvanic current by Faraday's Law. Faraday's Law establishes the proportionality between current flow and the amount of material dissolved or deposited in electrolysis. Additional information is provided in the article “Kinetics of Aqueous Corrosion,” in *Corrosion,* Vol 13 of the *ASM Handbook.*

Potentiodynamic polarization curves are generated by connecting the specimen of interest to a scanning potentiostat. This device applies whatever current is necessary between the specimen and a counterelectrode to maintain that specimen at a given potential versus a reference electrode half-cell placed near the specimen. The current required is plotted as a function of potential over a range that begins at the corrosion potential and proceeds in the direction (anodic or cathodic) required by that material. Such curves would be generated for each material of interest in the system. Additional information on the method for generating these curves is available in ASTM G 5 (Ref 5). The scan rate for potential must be chosen such that sufficient time is allowed for completion of electrical charging at the interface.

Potentiodynamic polarization is particularly effective for materials with time-independent polarization behavior. It is fast, relatively easy, and gives a reasonable, quantitative prediction of corrosion rates in many systems. However, potentiostatic techniques are preferred for time-dependent polarization. To establish polarization characteristics for time-dependent polarization, a series of specimens are used, each held to one of a series of constant potentials with a potentiostat while the current required is monitored as a function of time. After the current has stabilized or after a preselected time period has elapsed, the current at each potential is recorded. Testing of each specimen results in the generation of one potential/current data pair, which gives a point on the polarization curve for that material. The data are then interpolated to trace out the full curve. This technique is very accurate for time-dependent polarization, but is expensive and time consuming. The individual specimens can be weighed before and after testing to determine corrosion rate as a function of potential, thus enabling the errors from using Faraday's Law to be easily corrected.

The process of predicting galvanic corrosion from polarization behavior can be illustrated by the example of a steel-copper system. Steel has the more negative corrosion potential and will therefore suffer increased corrosion upon coupling to copper, but the amount of this corrosion must be predicted from polarization curves. If the polarization of each material is plotted as the absolute value of the log of current density versus potential and if the current density axis of each of these curves is multiplied by the wetted surface area of that material in the service application, then the result will be a plot of the total anodic current for steel and the total cathodic current for copper in this application as a function of potential (Fig. 4).

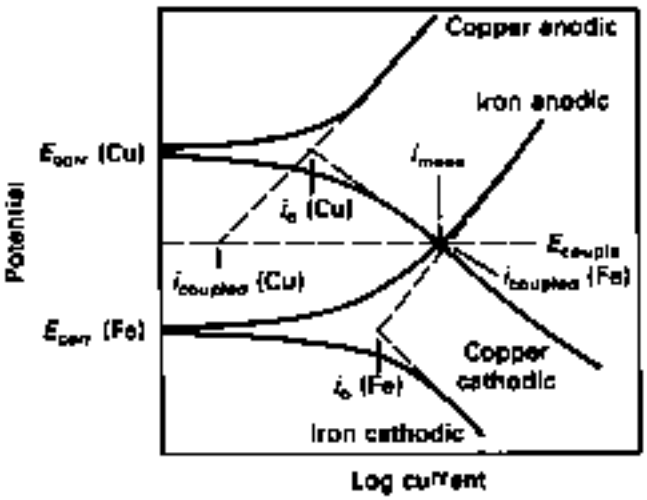


Fig. 4 Prediction of coupled potential and galvanic current from polarization diagrams. *i*, current; i_o, exchange current; E_{corr}, corrosion potential

Furthermore, when the two metals are electrically connected, the anodic current to the steel must be supplied by the copper; that is, the algebraic sum of the anodic and cathodic currents must equal 0. If the polarization curves for the two materials, normalized for surface area as above, are plotted together, this current condition is satisfied where the two curves intersect. This point of intersection allows for the prediction of the coupled potential of the materials and the galvanic current flowing between them from the intersection point. This procedure works if there is no significant electrolyte resistance between the two metals; otherwise, this resistance must be taken into account in a complex manner that is beyond the scope of this article.

References cited in this section

3. "Standard Guide for Development and Use of a Galvanic Series for Predicting Galvanic Corrosion Performance," G 82, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

4. E.H. Hollingsworth and H.Y. Hunsicker, Corrosion of Aluminum and Aluminum Alloys, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 583–609

5. "Standard Reference Test Method for Making Potentiostatic and Potentiodynamic Anodic Polarization Measurements," G 5, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

Combating Circumstances That Promote Galvanic Action

Dissimilar Metals. The combination of dissimilar metals in engineering design is quite common—for example, in heating or cooling coils in vessels, heat exchangers, or machinery. Such combinations often lead to galvanic corrosion. The obvious design considerations to minimize the adverse effects of using dissimilar metals are to:

- Select metals with the least difference in "uncoupled" corrosion potentials
- Minimize the area of the more noble metal with respect to that of the active metal
- Insert an electrically insulating gasket between the two dissimilar metals
- Seal the junction from any available electrolyte

Another possibility is coating the cathodic material for corrosion control. Where metallic coatings are used, there may be a risk of galvanic corrosion, especially along the cut edges. Rounded profiles and effective sealants or coatings are beneficial. Ineffective painting of an anode in an assembly can significantly reduce the desired service lifetime because local defects (anodes) effectively multiply the risk of localized corrosion.
It is often necessary to use different materials in close proximity. Sometimes, components that were designed in isolation can end up in direct contact in the plant (Fig. 5). In such instances, the ideals of a total design concept become especially apparent, but usually in hindsight.

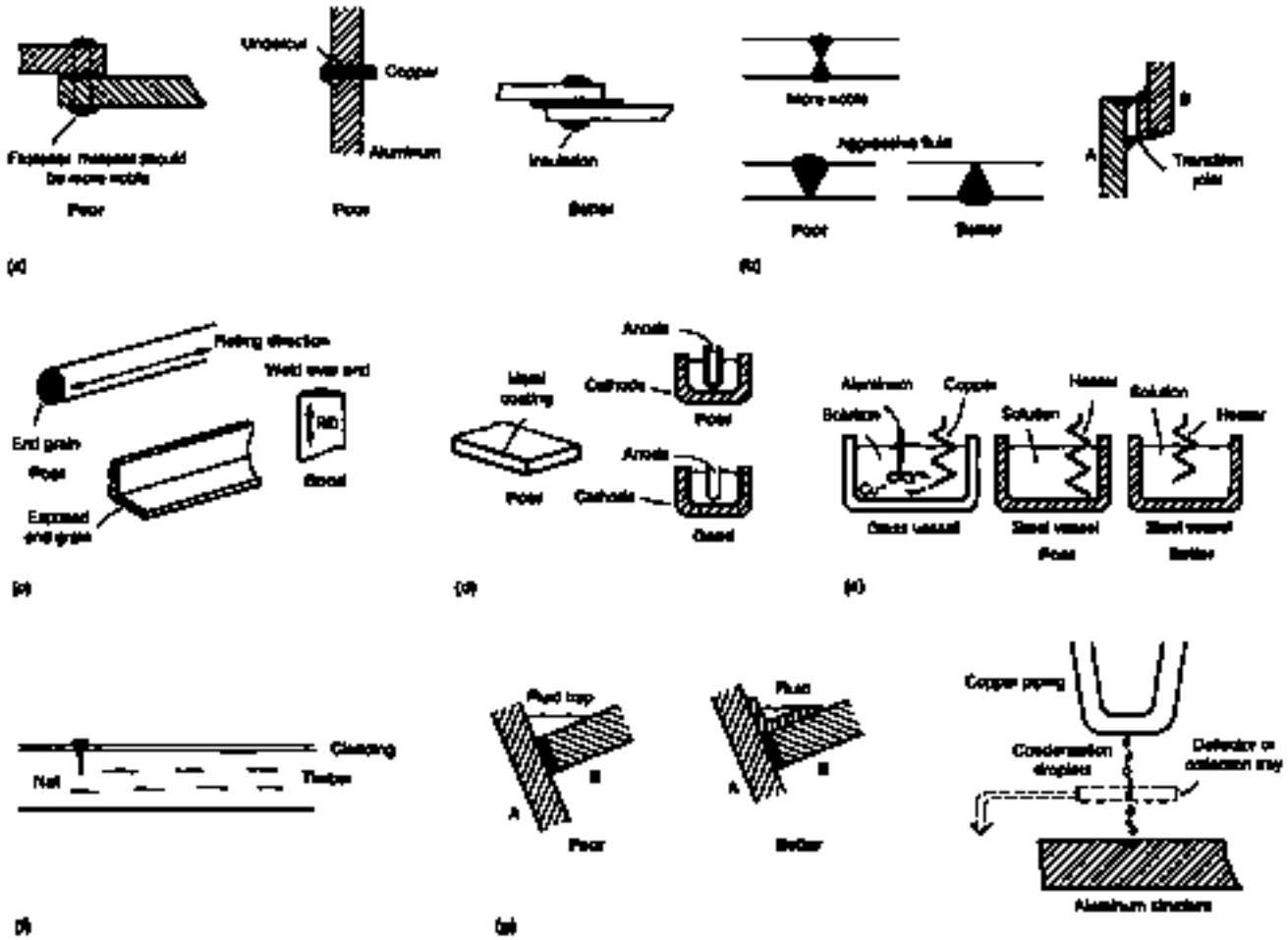


Fig. 5 Design details that can affect galvanic corrosion. (a) Fasteners should be more noble than the components being fastened; undercuts should be avoided, and insulating washers and spaces should be used to completely isolate the fastener. (b) Weld filler metals should be more noble than base metals. Transition joints can be used when a galvanic couple is anticipated at the design stage, and weld beads should be properly oriented to minimize galvanic effects. (c) Local damage can result from cuts across heavily worked areas. End grains should not be left exposed. (d) Galvanic corrosion is possible if a coated component is cut. When necessary, the cathodic component of a couple should be coated. (e) Ion transfer through a fluid can result in galvanic attack of less noble metals. In the example shown at left, copper ions from the copper heater coil could deposit on the aluminum stirrer. A nonmetallic stirrer would be better. At right, the distance from a metal container to a heater coil should be increased to minimize ion transfer. (f) Wood treated with copper preservatives can be corrosive to certain nails. Aluminum cladding can also be at risk. (g) Contact of two metals through a fluid trap can be avoided by using a drain, collection tray, or a deflector.

Where dissimilar materials are to be joined, it is advisable to use a more noble metal in a joint (Fig. 5a and b). Unfavorable area ratios should be avoided. Metal combinations should be used in which the more active metal or alloy surface is relatively large. Rivets, bolts, and other fasteners should be of a more noble metal than the material to be fastened.

Dissimilar-metal crevices, such as at threaded connections, are to be avoided, if possible. Crevices should be sealed, preferably by welding or brazing, although putties are sometimes used effectively. Replaceable sections of the more active member should be used at joints, or the corrosion allowance of this section should be increased, or both. Effective insulation can be useful if it does not lead to crevice corrosion. Some difficulties arise in the use of adhesives, which may not be sealants.

Preventive Measures. In particular cases, it is possible to reduce or eliminate galvanic-corrosion effects between widely dissimilar metals or alloys in a particular environment by altering one or more of the three key factors necessary for galvanic corrosion. Electrochemical techniques to mitigate galvanic corrosion include electrical isolation, use of transition materials, and cathodic protection.

Electrical Isolation. The joint between dissimilar metals can be isolated to break the electrical continuity. Use of nonmetallic inserts, washers, fittings, and coatings at the joint between the materials will provide sufficient electrical resistance to eliminate galvanic corrosion. The design of the insulation or maintenance must ensure that the insulator is not bridged by accumulated debris.

Transition Materials. In order to eliminate a dissimilar-metal junction, a transition piece can be introduced. The transition piece consists of the same metals or alloys as in the galvanic couple bonded together in a laminar structure. The transition piece is inserted between the members of the couple such that the similar metals mate with one another. The dissimilar-metal junction then occurs at the bond interface, excluding the electrolyte.

Cathodic Protection. Sacrificial metals, such as magnesium or zinc, may be introduced into the galvanic assembly. The most active member will corrode while providing cathodic protection to the other members in the galvanic assembly (for example, zinc anodes in cast iron waterboxes of copper alloy water-cooled heat exchangers). Cathodic protection is a common and economical method of corrosion protection that is often used for the protection of underground or underwater steel structures. The use of cathodic protection for long-term corrosion prevention for structural steels, underground pipelines, oil and gasoline tanks, offshore drilling rigs, well-head structures, steel piling, piers, bulkheads, offshore pipelines, gathering systems, drilling barges, and other underground and underwater structures is a fairly standard procedure. Magnesium, zinc, and aluminum galvanic (sacrificial) anodes are used in a wide range of cathodic protection applications.

The electric potential of an object can be changed to protect it. A direct-current potential can be applied on it through the use of a rectifier, battery, or solar cell. Conversely, stray currents can be a source of increased galvanic degradation.

Altering the Electrolyte. The use of corrosion inhibitors is effective in some cases. Elimination of cathodic depolarizers (deaeration of water by thermomechanical means plus oxygen scavengers such as sodium sulfite or hydrazine) is very effective in some aqueous systems.

Metallic Coatings. Two types of metallic coatings are used in engineering design: noble metal coatings and sacrificial metal coatings. Noble metal coatings are used as barrier coatings over a more reactive metal. Galvanic corrosion of the substrate can occur at pores, damage sites, and edges in the noble metal coating. Sacrificial metal coatings provide cathodic protection of the more-noble base metal, as in the case of galvanized steel or Alclad aluminum.

Evaluation of Galvanic Corrosion

In terms of general evaluation, the most common method of predicting galvanic corrosion is by immersion testing of the galvanic couple in the environment of interest. Although time consuming, this is the most desirable method of investigating galvanic corrosion.

In design and materials selection, screening tests are conducted to eliminate as many candidate materials as possible. These screening tests consist of one or more of the following electrochemical techniques:

- Potential measurements
- Polarization measurements
- Current measurements

Potential measurements are made to construct a galvanic series of metals and alloys, as described in the next section. As a first approximation, the galvanic series is a useful tool. However, it has several limitations. Metals and alloys that form passive films will exhibit varying potentials with time and are therefore difficult to position in the series with certainty. Also, the galvanic series does not provide information on the polarization

characteristics of the materials and so is not helpful in predicting the probable magnitude of galvanic effects. Polarization measurements are discussed in the section “Polarization” in this article.

Measurement of galvanic currents between coupled metals or alloys is based on the use of a zero-resistance milliammeter. Zero-resistance electrical continuity between the members of the galvanic couple is maintained electronically, while the resulting current is measured with the ammeter. Use of this technique should take into account certain limitations. First, when localized corrosion such as pitting or crevice corrosion is possible in the galvanic couple, long induction periods may be required before these effects are observed. Test periods must be of sufficient duration to take this effect into account. Also the measured galvanic current is not always a true measure of the actual corrosion current, because it is the algebraic sum of the currents due to anodic and cathodic reactions. When cathodic currents are appreciable at the mixed potential of the galvanic couple, the measured galvanic current will be significantly lower than the true current. Therefore, large differences between the true corrosion rate calculated by weight loss and that obtained by galvanic current measurements have been observed.

Evaluating Galvanic Corrosion as a Synergistic Corrosion Mechanism. Uniform corrosion, pitting, and crevice corrosion can all be exacerbated by galvanic conditions. In addition, any of the tests used for the more conventional forms of corrosion, such as uniform attack, pitting, or stress corrosion, can be used, with modifications, to determine galvanic-corrosion effects. The modifications can be as simple as connecting a second metal to the system or as complex as necessary to evaluate the appropriate parameters. A change in the method of data interpretation is often all that is needed to convert conventional test methods into galvanic-corrosion tests.

Example 1: Galvanic Corrosion Leading to Fatigue Failure of a Helicopter Tail Rotor (Ref 6). A tail rotor blade spar shank was constructed of 2014-T652 aluminum alloy. The spar was hollow with the cavity filled with a lead wool ballast material and sealed with a thermoset material. One blade separated in flight, but fortunately a successful emergency landing was made. The outboard section that had separated was recovered.

Examination. There was a large flat fracture with beach mark features typical of fatigue. The fatigue crack had initiated on the inner surface of the spar at the cavity (Fig. 6a and b). Fluid was found within the cavity, and the thermoset seal cap was broken. Microscopic examination revealed pitting corrosion on the inner surface of the spar cavity.

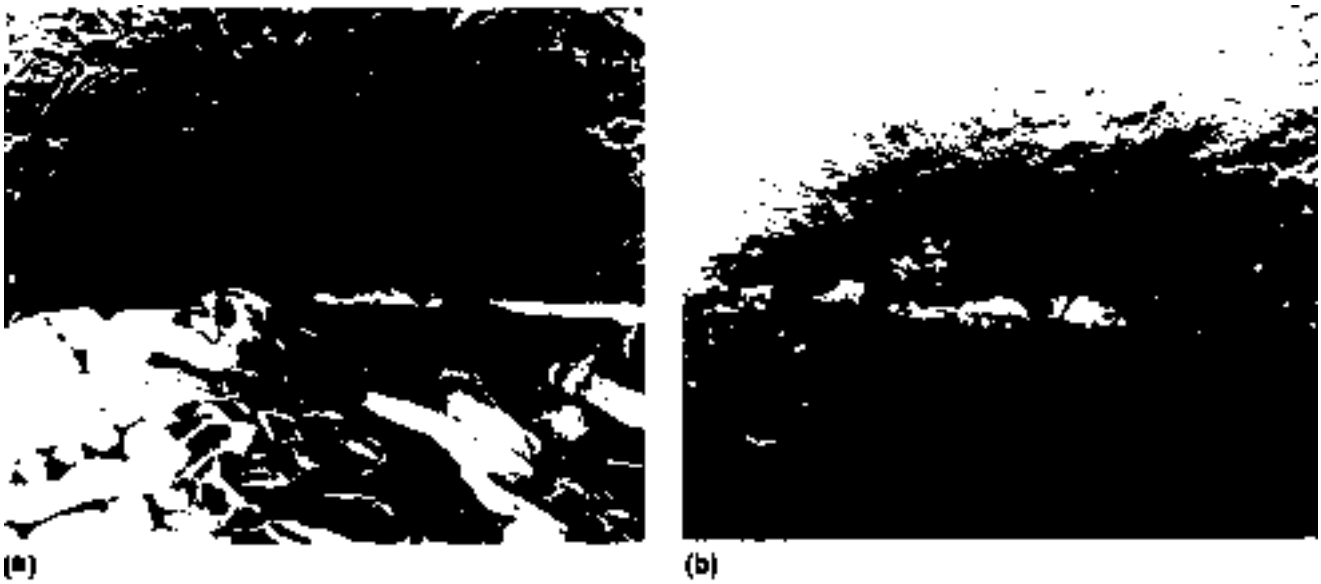


Fig. 6 Fatigue cracking of a helicopter tail rotor blade. (a) Scanning electron micrograph of the blade showing lead wool ballast in contact with the 2014-T652 aluminum spar bore cavity wall at the failure origin ~13×. (b) Greater magnification (~63×) in this same area shows the multiple pits and associated corrosion products at the failure origin. The beach marks are seen emanating from the pits, typical of fatigue failure mode. Source: Ref 6

Conclusion. The seal had failed allowing moisture into the cavity. This served as an electrolyte for galvanic corrosion between the lead and aluminum components. Corrosion pits formed at contact points of the lead wool

and the aluminum spar wall. These pits served as the point of origin for the fatigue crack leading to failure of the assembly.

Corrective Actions. To prevent the root cause of the failure, the causes of galvanic corrosion needed to be eliminated. The integrity of the seal needed to be ensured. A barrier coating between the lead and aluminum wall could be added. Alternate means of encapsulating the lead wool were also suggested. The corrective action taken was not reported.

Reference cited in this section

6. R.H. McSwain, Fatigue Fracture of a Helicopter Tail Rotor Blade Due to Field-Induced Corrosion, *Handbook of Case Histories in Failure Analysis*, Vol 2, ASM International, 1993, p 30–32

Examples of Factors Contributing to Galvanic Corrosion

Ion transfer results in the deposition of active and noncompatible deposits on a metal surface. For example, an aluminum stirrer plate used in water was extensively pitted because the water bath was heated by a copper heater coil (Fig. 5e). The pits resulted from deposition of copper ions from the heater element.

Nonmetallic Conductors. Less frequently recognized is the influence of nonmetallic conductors as cathodes in galvanic couples. Carbon brick in vessels is strongly cathodic to the common structural metals and alloys. Impervious graphite, especially in heat-exchanger applications, is cathodic to the less noble metals and alloys. Carbon-filled polymers or metal-matrix composites can act as noble metals in a galvanic couple.

Metal Ion Deposition. Ions of a more noble metal may be reduced on the surface of a more active metal—for example, copper on aluminum or steel, silver on copper. This process is also known as cementation, especially with regard to aluminum alloys. The resulting metallic deposit provides cathodic sites for further galvanic corrosion of the more active metal.

Performance of Alloy Groupings

Magnesium occupies an extremely active position in most galvanic series and is therefore highly susceptible to galvanic corrosion. Metals that combine active potentials with higher hydrogen overvoltages, such as aluminum, zinc, cadmium, and tin, are much less damaging, although not fully compatible with magnesium.

Aluminum alloys containing small percentages of copper (7000 and 2000 series and 380 die-casting alloy) may cause serious galvanic corrosion of magnesium in saline environments. Very pure aluminum is quite compatible, acting as a polarizable cathode; however, when iron content exceeds 200 ppm, cathodic activity becomes significant (apparently because of the depolarizing effect of the intermetallic compound $FeAl_3$), and galvanic attack of magnesium increases rapidly with increasing iron content. The effect of iron is diminished by the presence of magnesium in the alloy (Fig. 7).

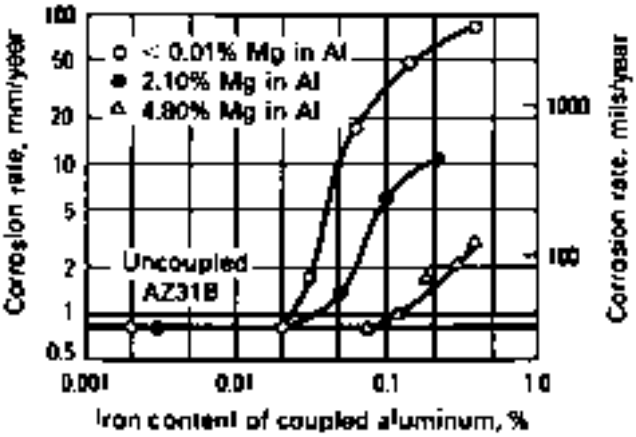


Fig. 7 Corrosion rates in 3% NaCl solution of magnesium alloy AZ31B coupled with aluminum containing varying amounts of iron and magnesium. The corrosion rate of uncoupled AZ31B is shown for comparison.

Aluminum and its alloys also occupy active positions in the galvanic series and are subject to failure by galvanic attack. In chloride-bearing solutions, aluminum alloys are susceptible to galvanically induced localized corrosion, especially in dissimilar-metal crevices. In this type of environment, severe galvanic effects are observed when aluminum alloys are coupled with more noble metals and alloys.

Cementation effects are also observed in the presence of dissolved heavy-metal ions such as copper, mercury, or lead. Some aluminum alloys are used for sacrificial anodes in seawater. An active, anodic alloy is used to clad aluminum, protecting it against pitting in some applications.

Contact of aluminum with more cathodic metals should be avoided in any environment in which aluminum by itself is subject to pitting corrosion. Where such contact is necessary, protective measures should be implemented to minimize sacrificial corrosion of the aluminum. In such an environment, aluminum is already polarized to its pitting potential, and the additional potential imposed by contact with the more cathodic metal greatly increases the corrosion current.

In the absence of chlorides or with low concentrations, as in potable water, aluminum and its alloys may be less active because of greater stability of the protective oxide film. Galvanic effects are not as severe under these conditions. Corrosion of aluminum in contact with more cathodic metals is much less severe in solutions of most nonhalide salts, in which aluminum alone normally is not polarized to its pitting potential.

In many environments, aluminum can be used in contact with chromium or stainless steels with only slight acceleration of corrosion; chromium and stainless steels are easily polarized cathodically in mild environments, so that the corrosion current is small despite the large differences in the open-circuit potentials between these metals and aluminum. Galvanic current between aluminum and another metal also can be reduced by removing oxidizing agents from the electrolyte. Thus, the corrosion rate of aluminum coupled to copper in seawater is greatly reduced wherever the seawater is deaerated.

The criterion for cathodic protection of aluminum in soils and waters has been published by the National Association of Corrosion Engineers (Ref 7). The suggested practice is to shift the potential at least -0.15 V, but not beyond the value of -1.20 V as measured against a saturated copper sulfate ($Cu/CuSO_4$) reference electrode. In some soils, potentials as low as -1.4 V have been encountered without appreciable cathodic corrosion (Ref 8).

Iron and steel are fairly active materials and require protection against galvanic corrosion by the higher alloys. They are, however, more noble than aluminum and its alloys in chloride solutions. However, in low-chloride waters, a reversal of potential can occur that causes iron or steel to become more active than aluminum. A similar reversal can occur between iron and zinc in hot waters of a specific type of chemistry.

Stainless Steels. Galvanic-corrosion behavior of stainless steels is difficult to predict because of the influence of passivity. In the common galvanic series, a noble position is assumed by stainless steels in the passive state, while a more active position is assumed in the active state (Fig. 3). These are noted by black blocks in the diagram. This dual position in galvanic series in chloride-bearing aqueous environments has been the cause of some serious design errors. More precise information on the galvanic behavior of stainless steels can be

obtained by using polarization curves, critical potentials, and the mixed potential of the galvanic couple. In chloride-bearing environments, galvanically induced localized corrosion of many stainless steels occurs in couples with copper or nickel and their alloys and with other more noble materials. However, couples of stainless and copper alloys are often used with impunity in freshwater cooling systems. Iron and steel tend to protect stainless steel in aqueous environments when galvanically coupled. The passive behavior of stainless steels makes them easy to polarize; thus, galvanic effects on other metals or alloys tend to be minimized. However, galvanic corrosion of steel can be induced by stainless steel, particularly in aqueous environments and with adverse area ratios.

Lead, Tin, and Zinc. These three materials occupy similar positions in the galvanic series, although zinc is the most active. The oxide films formed on these materials can shift their potentials to more noble values. Thus, in some environments they may occupy more noble positions than one might otherwise expect. For example, the tin coating in tin cans is anodic to steel under anaerobic conditions in the sealed container, but becomes cathodic when the can is opened and exposed to air. Zinc is an active metal. It is susceptible to galvanic corrosion and is widely used for galvanic anodes, in cathodic protection as a sacrificial coating (for example, galvanizing or electroplating), and as a pigment in certain types of coatings.

Copper and its alloys occupy an intermediate position in the galvanic series. They are not readily polarized in chloride-bearing aqueous solutions; therefore, they cause severe accelerated corrosion of more active metals, such as aluminum and its alloys and the ferrous metals. Somewhat similar to the nickel alloys, they lie between the active and passive positions for stainless steels (Fig. 3) and therefore induce localized corrosion of the active alloys.

Nickel and its alloys are not readily polarized and will therefore cause accelerated corrosion of more active materials, such as aluminum and ferrous alloys. In chloride-bearing solutions, nickel is somewhat more noble than copper, and the cupronickels lie somewhere in between. Nickel and its alloys are similar to copper alloys in their effects on stainless steels. In some environments, the cast structure of a nickel weld may be anodic to the wrought parent metals.

The combination of their passive surface with their inherent resistance places nickel-chromium alloys such as Inconel alloy 600 and Hastelloy alloy C-276 in more noble positions in the traditional galvanic series. In chloride-bearing solutions, Inconel alloy 600 is reported to occupy two positions because of existence of active and passive states in a manner similar to the stainless steels (Fig. 3). It is highlighted in black to indicate this. These alloys are readily polarized, and galvanic effects on other less noble metals and alloys therefore tend to be minimized.

Cobalt-base alloys, most of which are chromium bearing, are resistant to galvanic corrosion because of their noble position in the galvanic series. However, in environments in which their passive film is not stable, they occupy a more active position and can be adversely affected by more noble materials. The fact that they polarize readily tends to reduce their galvanic effects on less noble materials.

Reactive metals (titanium, zirconium, and tantalum) are extremely noble because of their passive films. In general, these alloys are not susceptible to galvanic corrosion, and their ease of polarization tends to minimize adverse galvanic effects on other metals or alloys. Because of the ease with which they pick up hydrogen in the atomic state, they may themselves become embrittled in galvanic couples. Tantalum repair patches in glass-lined vessels have been destroyed by contact with cooling coils or agitators made of less noble alloys. Tantalum is susceptible to attack by alkalis, such as may form in the vicinity of a cathode in neutral solutions.

Noble Metals. The term noble metal is applied to silver, gold, and platinum group metals. This designation in itself describes their position in the galvanic series and their corresponding resistance to galvanic corrosion. However, they do not polarize readily and can therefore have a marked effect in galvanic couples with other metals or alloys. This effect is observed with gold and silver coatings on copper, nickel, aluminum, and their alloys.

References cited in this section

7. Recommended Practice for Cathodic Protection of Aluminum Pipe Buried in Soil or Immersed in Water, *Mater. Protec.,* Vol 2 (No. 10), 1963 p 106

8. F.W. Hewes, Investigation of Maximum and Minimum Criteria for the Cathodic Protection of Aluminum in Soil, *Oil Week,* Vol 16 (No. 24–28), Aug–Sept 1965

Uniform Corrosion

Revised and adapted by Randy K. Kent, MDE Engineers, Inc.

Uniform corrosion, or general corrosion, is a corrosion process exhibiting uniform thinning that proceeds without appreciable localized attack. It is the most common form of corrosion and may appear initially as a single penetration, but with thorough examination of the cross section it becomes apparent that the base material has uniformly thinned.

Uniform chemical attack of metals is the simplest form of corrosion, occurring in the atmosphere, in solutions, and in soil, frequently under normal service conditions. Excessive attack can occur when the environment has changed from that initially expected. Weathering steels, magnesium alloys, zinc alloys, and copper alloys are examples of materials that typically exhibit general corrosion. Passive materials, such as stainless steels, aluminum alloys, or nickel-chromium alloys are generally subject to localized corrosion. Under specific conditions, however, each material may vary from its normal mode of corrosion.

Uniform corrosion commonly occurs on metal surfaces having homogeneous chemical composition and microstructure. Access to the metal by the attacking environment is generally unrestricted and uniform. Any gradient or changes in the attacking environment due to the corrosion reaction or stagnancy can change the degradation mode away from uniform corrosion. At the microlevel, uniform corrosion is found to be an electrochemical reaction between adjacent closely spaced microanodic and microcathodic areas. Consequently, uniform corrosion might be considered to be localized electrolytic attack occurring consistently and evenly over the surface of the metal.

This is well illustrated in binary alloys that contain phases of differing corrosion potentials. The grain in one phase becomes anodic to grain in the second phase in the microstructure, thus producing an electrolytic cell when the proper electrolyte is present.

Another illustration is ductile cast iron, where there are three compositionally different phases within the material: ferrite, pearlite, and graphite. The corrosion process will initiate as uniform corrosion with the formation of an electrolytic microcell between the graphite cathode and ferrite or pearlite anode (Ref 9, 10). The process can turn into a pitting mechanism if the microcells develop enough to produce acidic gradients within the micropits where the graphite nodules lie or are dislodged. The uniform corrosion occurring by the microelectrolytic or galvanic cell in the ductile cast iron is shown in Fig. 8. Observe the initial stages of micropits being formed by these microcells.

Fig. 8 Scanning electron micrograph of ductile cast iron graphite nodules and ferritic phase after corrosion tests. Note the loss of material at the interface of the nodule. 2000×. Source: Ref 11, 12

References cited in this section

9. R. Kent, "Anodic Polarization Measurements of Alloyed DI," NACE Int. Conf., National Association of Corrosion Engineers, 1987

10. R. Kent, "Potential Anodic Polarization Measurements and Mechanical Property Analysis of Ferritic Nodular Iron Alloyed with Nickel," University of Washington, 1986

11. F. Mansfeld and V. Bertocci, *Electrochemical Corrosion Testing*, STP 727, American Society for Testing and Materials, 1981

12. M. Henthorne, *Corrosion Causes and Control*, Carpenter Technology Corp., 1972, p. 30

Surface Conditions

All metals are affected by uniform corrosion in some environments; the rusting of steel and the tarnishing of silver are typical examples of uniform corrosion. In some metals, such as steel, uniform corrosion produces a

somewhat rough surface by the oxidation/reduction reaction, in which the end product (oxide) either dissolves in the environment and is carried away or produces a loosely adherent, porous coating now greater in thickness. Because of the porosity in the oxidation (rust), this metal is still considered active or able to continue degrading. Coating specialists have been able to formulate a phosphoric-acid-based compound that reacts with the rust to produce an oxide that makes the surface passive and seals the surface from water and other solutions.
In contrast to the active surfaces formed typically on steels, some metals form dense insulated tightly adherent passive films from uniform-corrosion processes with the metal surface remaining somewhat smooth. Examples of these include the tarnishing of silver in air, lead in sulfate-containing environments, and stainless steel in humid air creating passive films.

Classification of Uniform Corrosion

In general terms, uniform corrosion can be classified further according to the specific conditions of environmental or electrochemical attack. For example, uniform thinning can be attributed to various conditions such as:

- Atmospheric corrosion
- Aqueous corrosion
- Galvanic corrosion
- Stray-current corrosion (which is similar to galvanic corrosion, but does not rely on electrochemically induced driving forces to cause rapid attack)
- Biological corrosion (which is a microbial-assisted form of attack that can manifest itself as uniform corrosion by forming weak or cathodic oxides, or it can also produce a localized form of attack)
- Molten salt corrosion and liquid-metal corrosion (which have become more of a concern as the demand for higher-temperature heat-transfer fluids increases)
- High-temperature (gaseous) corrosion

Corrosive attack for these conditions is not necessarily restricted to just uniform thinning. Other forms of corrosive attack (such as stress-corrosion cracking, dealloying, or pitting) may also be operative.
Following are some examples of uniform-corrosion-related failures and its prevention and evaluation. Appropriate methods of controlling uniform corrosion in some situations are noted. Selection of a metal that has a suitable resistance to the environment in which the specific part is used and the application of paints and other types of protective coatings are two common methods used to control uniform corrosion. Modification of the environment by changing its composition, concentration, pH, and temperature or by adding an inhibitor is also effective.
Example 2: Uniform Corrosion of Carbon Steel Boiler Feedwater Tubes. The carbon steel tubes shown in Fig. 9 have been corroded to paper thinness, revealing a lacelike pattern of total metal loss. These steel tubes are from a boiler feedwater heater feeding a deaerator. As part of the boiler-water treatment program, it was decided to inject a chelate to control scale formation in the boiler tubes. Unfortunately, the chelate was added ahead of the preheater, where the boiler water still contained oxygen (O_2). As the chelate removed iron oxide, the O_2 formed more iron oxide, and this uniform dissolution of steel reduced the tubing to the totally corroded condition shown. Moving the chelate addition to a point after the deaerator stopped the corrosion and allowed reuse of steel in the preheater.

Fig. 9 Uniform corrosion of steel tubes in boiler feedwater containing oxygen (O_2) and a chelating water-treating chemical

Example 3: Uniform Corrosion of a Copper Pipe Coupling. The 25 mm (1 in.) copper coupling shown in Fig. 10 has been uniformly degraded around most of the circumference of the bell and partially on the spigot end. One penetration finally occurred through the thinned area on the spigot end of this pipe. The pipe was buried in noncorrosive sandy soil, but was found to incur stray currents of 2 V direct current in relation to a $Cu/CuSO_4$ half cell. Eliminating, moving, or shielding the source of stray current are obvious solutions.

Fig. 10 A 25 mm (1 in.) copper coupling from a potable water system that had degraded uniformly from stray currents. The spigot end has been penetrated near the edge of the bell. Courtesy of MDE Engineers, Inc.

Example 4: Uniform Corrosion of Copper Piping Caused by Microbiological Attack. Microbiological attack of copper piping has been well documented and was found in a closed-loop water heater system. Figure 11 shows the microbes that were cultured from the corrosion product. They were found to be sulfur-reducing bacteria. Uniform thinning occurred when the resultant oxide on the copper pipe surface eroded in low flow rates. The threshold for erosion by turbulent waters is much lower with this type of oxide and found to penetrate this same pipe, as shown in Fig. 12.

Fig. 11 Micrograph of large bacteria (SRB) that are rod shaped. Note this is a chain of two bacteria cultured from microbiologically induced corrosion product of the pipe failure shown in Fig. 12. 400×. Courtesy of MDE Engineers, Inc.

Fig. 12 Micrograph of 19 mm (0.75 in.) copper piping in a closed-loop water system with MIC and erosion of the weak oxide layer. 20×. Courtesy of MDE Engineers, Inc.

Materials Selection

Materials selection is usually straightforward for uniform corrosion, as the corrosion is relatively easy to evaluate and monitor. If a material shows only general attack with a low corrosion rate, or if only negligible contamination is present in a process fluid or on the surface, then cost, availability, and ease of fabrication may

be the dominant influence on the material of choice. An acceptable corrosion rate for a relatively low-cost material such as plain carbon steel is about 0.25 mm/year (10 mils/year) or less. At this rate and with proper design with adequate corrosion allowance, a carbon steel vessel will provide many years of low-maintenance service.

For more costly materials, such as austenitic stainless steels and copper and nickel alloys, a maximum corrosion rate of 0.1 mm/year (4 mils/year) is generally acceptable. However, a word of caution is in order. One should never assume that the higher-alloyed material or the higher-cost material is more corrosion resistant in a given environment. Proper evaluation is required. For instance, seawater corrodes plain carbon steel fairly uniformly at a rate of 0.1 to 0.2 mm/year (4 to 8 mils/year) but also severely pits certain austenitic stainless steels. The same is true for stagnant solutions; the steel becomes passive or protected, where some austenitic stainless steels will pit due to the formation of autocatalytic corrosion cells.

At times, nonmetallic coatings and linings ranging in thickness from a few tenths to several millimeters are applied to prolong the life of low-cost alloys such as plain carbon steels in environments that cause general corrosion. The thin-film coatings that are widely used include baked phenolics, catalyzed cross-linked epoxy-phenolics, and catalyzed coal tar-epoxy resins (see the section “Protective Coatings and Linings” in the article “Analysis and Prevention of Corrosion-Related Failures” in this Volume).

It is not advisable to use thin-film coatings in services where the base metal corrosion rate exceeds 0.5 mm/year (20 mils/year), because corrosion is often accelerated at holidays (pinholes or scratches) in the coating. Thick-film linings—which include glass, fiber- or flake-reinforced furan, polyester and epoxy resins, hot-applied coal tar enamels, and various elastomers such as natural rubber—may be more appropriate.

Effect of Corrosion Products

The by-products of uniform corrosion may be of greater concern in materials selection than the effect on the material itself. For food and chemical processing and biomedical applications, contamination of the process fluid by even trace amounts of corrosion products may be unacceptable. Product purity, rather than corrosion rate, is the prime consideration.

One example is storage of 93% sulfuric acid (H_2SO_4) in plain carbon steel at ambient temperature. The general corrosion rate is 0.25 mm/year (10 mils/year) or less, but traces of iron impart a color that is objectionable in many applications. Therefore, thin-film baked phenolic coatings are used on carbon steel to minimize or eliminate iron contamination. In the same way, thin-film epoxy coated carbon steel or solid or clad austenitic stainless steels are used to maintain the purity of adipic acid for various food and synthetic fiber applications.

Effect of Concentration

The effect on corrosion rate of a metal part does not follow a uniform pattern with the increase or decrease in the concentration of the corrodent. This is because of (a) ionization effects in aqueous solutions and the effects of even trace amounts of water in nonaqueous environments and (b) changes that occur in the characteristics of any film of corrosion products that may be present on the surface of the metal. Typical patterns of corrodent-concentration effects on corrosion rates are illustrated in Fig. 13 (Ref 12), which plots the corrosion rate of iron and steel as a function of the concentration of three common inorganic acids in aqueous solutions at room temperature.

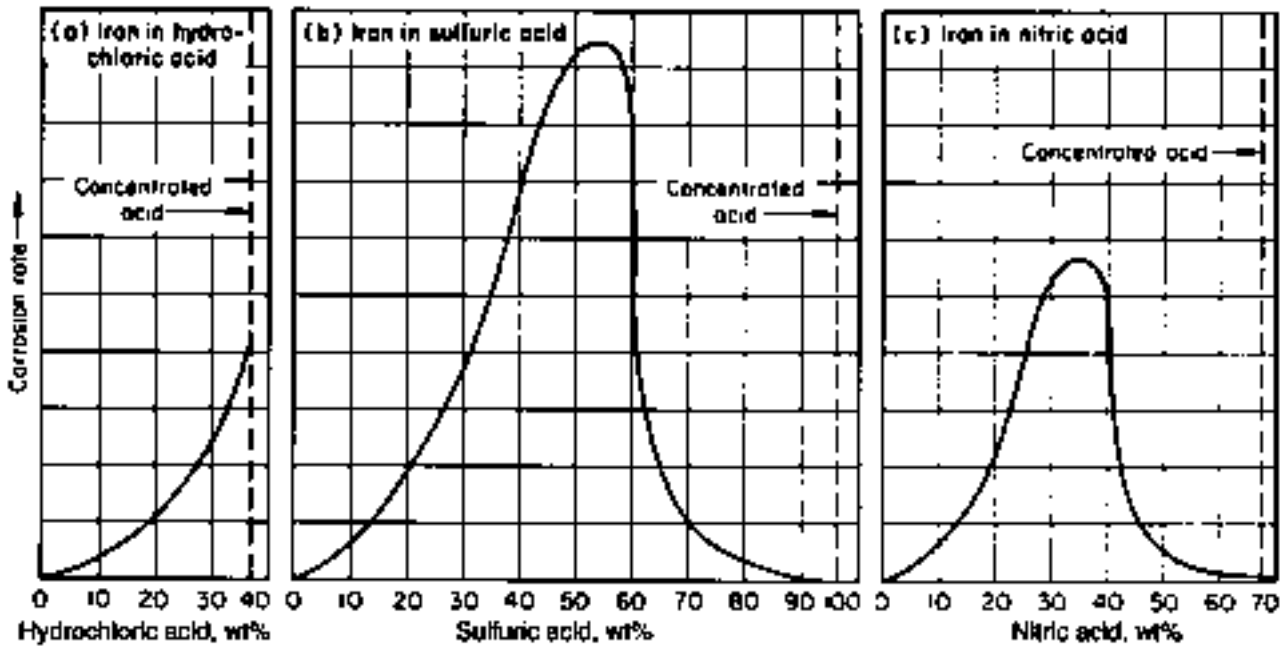


Fig. 13 Effect of acid concentration on the corrosion rate of iron completely immersed in aqueous solutions of three inorganic acids at room temperature. It should be noted that the scales for corrosion rate are not the same for all three charts. As discussed in text, the corrosion rate of iron (and steel) in nitric acid in concentrations of 70% or higher, although low compared to the maximum rate, is sufficient to make it unsafe to ship or store nitric acid in these metals. Source: Ref 12

The rate of corrosion of a given metal usually increases as the concentration of the corrodent increases, as shown in Fig. 13(a) for the corrosion of iron in hydrochloric acid. However, corrosion rate does not always increase with concentration of the corrodent; the maximum rate may occur over a preferred range of corrodent concentration, as shown in Fig. 13(b) and (c) for iron and steel in sulfuric acid and in nitric acid, respectively.

Because corrosion is electrochemical and involves anodic and cathodic reactions, process variables influence corrosion rate if they influence one or both reactions. For example, the main cathodic reaction for iron corroding in dilute inorganic acids is $2H^+ + 2e \rightarrow H_2\uparrow$. The more hydrogen ions available, the faster the rate of the cathodic reaction. In turn, this permits a high rate of anodic dissolution: ($M \rightarrow M^{+n} + ne$) where M stands for metal. This is what happens throughout the concentration range in hydrochloric acid solutions.

In both nitric acid and sulfuric acid solutions, the hydrogen-ion concentration increases with acid concentration to a point, but at the higher levels of acid concentration, it decreases again. Iron may be used to handle concentrated sulfuric acid at ambient temperatures. Care must be taken to avoid any contamination with water, because this will dilute the acid and increase the rate of attack. Also, impurities in the acid or the iron can increase the rate of attack.

The corrosion rate of iron and steel at room temperature in nitric acid decreases with increasing concentration above about 35% because of the formation of a passive oxide film on the metal surface. However, this passive condition is not completely stable.

The rate of attack for concentrations of 70% or higher, although low compared to the maximum rate shown in Fig. 13, is still greater than 1.3 mm/year (50 mils/year), making iron and steel unsuitable for use in shipping and storing nitric acid at any concentration. Nitric acid in bulk is usually stored and shipped in type 304 stainless steel (up to 95%), aluminum alloy 3003 or 1100, or commercially pure titanium (grade 2).

Metals that have passivity effects, such as Monel 400 in hydrochloric acid solutions and lead in sulfuric acid solutions, corrode at an extremely low rate at low acid concentration at room temperature. They lose their passivity at a certain limiting acid concentration above which the corrosion rate increases rapidly with increasing acid concentration.

Reference cited in this section

12. M. Henthorne, *Corrosion Causes and Control,* Carpenter Technology Corp., 1972, p. 30

Effect of Temperature

When investigating the effect of temperature on the corrosion rate of a metal in a liquid or gaseous environment, the temperature at the metal-corrodent interface must be considered. This temperature may differ substantially from that of the average of the corrodent. Laminar flow would permit temperature gradients to exist. This difference is especially important for heat-transfer surfaces, where the hot-wall effect causes a much higher corrosion rate than might be expected.

A common approximation for chemical activity, called the “10 degree rule,” states an increase of 10 °C (18 °F) in bulk temperature of the solution can increase the corrosion rate by a factor of two or more. In atmospheres, the temperature variation is more complex as temperature and humidity both affect corrosion and they are interdependent.

Hot-wall failures are fairly common in heating coils and heat-exchanger tubes. An example of this hot-wall effect is the corrosion rate of Hastelloy B (Ni-Mo, N10001) in 65% H_2SO_4. At 120 °C (250 °F), the corrosion rate would be less than 0.5 mm/year (20 mils/year). However, a Hastelloy B heating coil in the same solution might have a surface-wall temperature of 145 °C (290 °F) and a corrosion rate greater than 5 mm/year (200 mils/year), a factor of 10 greater.

In some metal-corrodent systems, there is a threshold temperature above which the corrosion rate increases dramatically. Caution is advised whenever extrapolating known corrosion rates to higher temperatures.

Temperature changes sometimes affect corrosive attack on metals indirectly. In systems where an adherent protective film on the metal may be stable in a given solution at room temperature, the protective film may be soluble in the solution at higher temperatures, with the result that corrosion can progress rapidly.

When boiling occurs in the solution, other factors also influence the rate of attack on a metal immersed in it. One factor may be simply an increase in velocity of movement of the liquid corrodent against the metal surface, which in turn increases the corrosion rate. More radical changes may occur at the metal surface, such as the substitution of steam and spray in place of liquid as the corrodent or the formation of a solid film on the metal surface—either of which may completely alter the metal-corrodent interface conditions.

There are some exceptions to the general rule that increasing temperature increases the corrosion rate. One is the reduction in rate of attack on steel in water as the temperature is increased, because the increase in temperature decreases the oxygen content of the water, especially as the boiling point of the water is approached. Other exceptions arise where a moderate increase in temperature results in the formation of a thin protective film on the surface of the metal or in passivation of the metal surface.

If thick deposits are formed on a heat-transfer surface, they have a twofold effect by changing the metal surface temperature and making crevice corrosion possible. The existence of the thick deposit may be a failure mode itself, if it limits the heat transfer. Local differences in temperature on a heat-transfer surface exposed to steam can influence corrosion rates by causing differences in the duration of exposure to condensation. For example, the inner surface of a carbon steel tube that carried saturated steam at 234 °C (454 °F) corroded to a greater depth opposite exterior heat-transfer fins than elsewhere on the inner surface (Fig. 14). The heat loss through the fins lowered the temperature of the inner surface of the tube in the area beneath the fins, resulting in the presence of condensate on this area for a greater percentage of the time and in a correspondingly greater loss of metal than elsewhere on the inner surface. All corroded areas were fairly smooth.

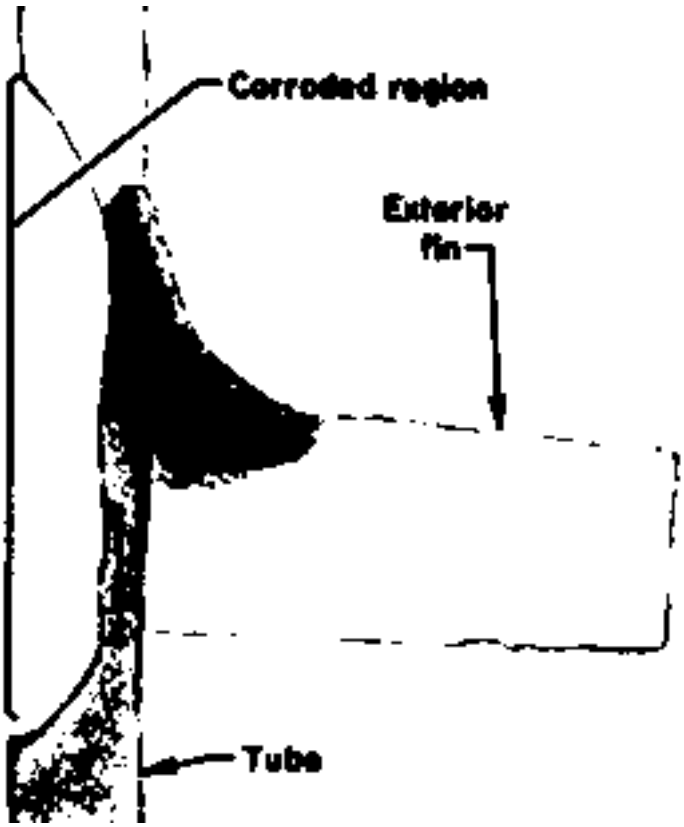


Fig. 14 Carbon steel steam tube that corroded on the inner surface more rapidly opposite the exterior heat-transfer fin than elsewhere along the tube. Etched longitudinal section. 3×

Evaluation Factors

Since uniform corrosion is one of the most common forms of corrosion; it must be considered in nearly every design. The damage may be just cosmetic at its inception, but as the thickness of the metal decreases uniformly, loss of mechanical function can result. Fortunately, uniform corrosion is usually easy to measure and predict; this facilitates proper design and the ability to take corrective action. Corrosion rate and expected service life can be calculated from periodic measurements of the general thinning produced by uniform corrosion. Because the rate of attack can change over a period of time, periodic inspection at suitable intervals is ordinarily carried out to avoid unexpected failures. Varying conditions can also accelerate some corrosion rates; therefore, generous safety factors regarding thickness allowances should be employed.
Testing Methods. Testing for uniform corrosion can be broadly classified as:

- Laboratory long-term testing
- Laboratory accelerated testing
- In-service or field testing.

Laboratory long-term testing usually consists of testing materials in simulated-service conditions on a relatively small scale. The advantage of this type of testing is that the test parameters can be controlled closely; thus, any unintentional disturbances that could occur during plant tests can be avoided.
Accelerated laboratory tests, on the other hand, are short term in nature. They are designed to compare materials under severe conditions, and the test environment may not be directly related to the service

environment. The environment and the test conditions must be carefully specified. ASTM has outlined standard methods for testing materials under accelerated conditions. ASTM G 1 (Ref 13) provides details on evaluating corrosion damage.

Service tests and plant tests involve placing test samples in actual service conditions for evaluation. The advantage of such testing is the relevance of the results to the actual service environment. However, the problems involved in tracking and interrupting normal operations are also associated with this type of corrosion testing.

Testing Goals. The purposes of measuring uniform corrosion by experimental testing are:

- Evaluation and selection of materials for a specific environment or application
- Evaluation of metals and alloys to determine their effectiveness in new environments
- Routine testing to confirm the quality of materials
- Research and development of new materials

In most cases, uniform-corrosion rates are represented as a loss of metal thickness as a function of time. This value can be directly measured from experimental data, or as is often the case, it can be calculated from mass loss data. Mass loss is a measure of the difference between the original mass of the specimen and the mass when sampled after exposure. In measuring the mass after exposure it is important to remove any corrosion product adhering to the sample (see the sections "Sample Cleaning" and "Data Acquisition" in the article "Evaluation of Uniform Corrosion" in *Corrosion,* Volume 13 of *ASM Handbook;* see also ASTM G 1, Ref 13).

Evaluations that are often included in testing for uniform corrosion include:

- Visual observations, unaided and using light optical and scanning electron microscopy
- Ion concentration increase
- Hydrogen evolution
- Loss in tensile strength (according to ASTM G 50, Ref 14)
- Electrochemical testing

Initial observations in any corrosion testing should include unaided visual or low-magnification observation of the form of corrosion. Visual evaluation can include reporting of the color and form of corrosion products and can include documentation with photographs.

In addition to the standard measurement of mass loss, the increase of metal ions in solution as the metal corrodes can be measured by analysis of samples of the corrodent removed periodically during the test (Ref 15). It may be possible to relate the increase in concentration to corrosion rate, depending on the structure of the test. In some cases, the amount of hydrogen generated in deaerated tests can be used to measure corrosion rates (Ref 16, 17). These methods are useful during in-service and field testing where the process fluid may be more available for analysis than the corroding object.

Electrochemical tests can also be used to determine corrosion rates and predict uniform-corrosion rates. Electrochemical methods include creating anodic and cathodic polarization curves, extrapolation of Tafel lines, and polarization break testing. Methods for conducting electrochemical tests are addressed in ASTM G 59 (Ref 18), G 3 (Ref 19), and G 5 (Ref 5). Reference 11 also explains some standard practices in this area, and electrochemical test methods are discussed in the article "Laboratory Testing" in *Corrosion,* Volume 13 of *ASM Handbook.*

References cited in this section

5. "Standard Reference Test Method for Making Potentiostatic and Potentiodynamic Anodic Polarization Measurements," G 5, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

11. F. Mansfeld and V. Bertocci, *Electrochemical Corrosion Testing,* STP 727, American Society for Testing and Materials, 1981

13. “Preparing, Cleaning, and Evaluating Corrosion Test Specimens,” G 1, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

14. “Standard Practice for Conducting Atmospheric Corrosion Tests on Metals,” *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

15. G. Wranglén, Corrosion and Protection of Metals, Chapman & Hall, 1985, p 238

16. R.L. Martin and E.C. French, Corrosion Monitoring in Sour Systems Using Electrochemical Hydrogen Patch Probes, *J. Pet. Technol.,* Nov 1978, p 1566–1570

17. P.A. Schweitzer, *Corrosion and Corrosion Protection Handbook,* Marcel Dekker, 1983, p 483–484

18. “Standard Test Method for Conducting Potentiodynamic Polarization Resistance Measurements,” *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

19. “Standard Practice for Conventions Applicable to Electrochemical Measurement in Corrosion Testing,” G 3, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

---

Design Considerations

When designing systems where uniform corrosion will likely occur, it is essential to establish regular inspections and preventive maintenance. The criticality of the system, ease of inspection and maintenance, and corrosion rate need to be considered when establishing the intervals. It must be understood that applied stresses, changes in loading, and varying environmental conditions could accelerate the rate of attack.

Pitting and Crevice Corrosion

Pitting and crevice corrosion are forms of localized corrosion that are significant causes of failure in metal parts. Both forms attack passivated or otherwise protected materials.

Pitting, characterized by sharply defined holes, is one of the most insidious forms of corrosion. It can cause failure by perforation while producing only a small weight loss on the metal. This perforation can be difficult to detect and its growth rapid, leading to unexpected loss of function of the component.

Crevice corrosion is pitting that occurs in slots and in gaps at metal-to-metal and metal-to-nonmetal interfaces. It is a significant contributor to component failure because such gaps often occur at critical joining surfaces.

Pitting

Causes of pitting corrosion include:

- Local inhomogeneity on the metal surface
- Local loss of passivity
- Mechanical or chemical rupture of a protective oxide coating
- Discontinuity of organic coating (holidays)
- Galvanic corrosion from a relatively distant cathode
- Formation of a metal ion or oxygen concentration cell under a solid deposit (crevice corrosion)

Every engineering metal or alloy is susceptible to pitting. Pitting occurs when one area of a metal surface becomes anodic with respect to the rest of the surface or when highly localized changes in the corrodent in contact with the metal, as in crevices, cause accelerated localized attack.

Pitting on clean surfaces ordinarily represents the start of breakdown of passivity or local breakdown of inhibitor-produced protection. When pits are few and widely separated and the metal surface has undergone little or no general corrosion, there is a high ratio of cathode-to-anode area. Penetration progresses more rapidly than when pits are numerous and close together.

The most common causes of pitting in steels are surface deposits that set up local concentration cells and dissolved halides that produce local anodes by rupture of the protective oxide film. Anodic corrosion inhibitors, such as chromates, can cause rapid pitting if present in concentrations below a minimum value that depends on the metal-environment combination, temperature, and other factors. Pitting also occurs at mechanical ruptures in protective organic coatings if the external environment is aggressive or if a galvanic cell is active.

Buried pipelines that fail because of corrosion originating on the outside surface usually fail by pitting corrosion. Another form of corrosion affecting pipelines involves a combination of pitting and erosion (erosion-corrosion). This form of corrosive attack is discussed in the section “Velocity-Affected Corrosion” in this article.

Pitting normally occurs in a stagnant environment. With corrosion-resistant alloys, such as stainless steels, the most common cause of pitting corrosion is highly localized destruction of passivity by contact with moisture that contains halide ions, particularly chlorides. Chloride-induced pitting of stainless steels usually results in undercutting, producing enlarged subsurface cavities or caverns.

Example 5: Pitting Corrosion of Stainless Steel by Potable Water in an Organic Chemical Plant Condenser (Ref 20). Several tubes in a 35 m (115 ft) type 316 stainless steel shell-and-tube condenser leaked unexpectedly in an organic chemical plant that produces vinyl acetate monomer (VAM). Leaks were discovered after five years of operation and relocation of the condenser to another unit in the same plant. Examination of tubes and tube sheets revealed pitting damage on the outside-diameter (OD) surface. Some of the pits had penetrated fully, resulting in holes. Inside-diameter (ID) surfaces were free of corrosion. Macro- and microexaminations indicated that the tubes had been properly manufactured. Pitting was attributed to stagnant water on the shell side. It was recommended that the surfaces not be kept in contact with closed stagnant water for appreciable lengths of time.

Service History. A shell-and-tube condenser is fixed vertically on top of a jacketed reactor that produces VAM (Fig. 15). The VAM vapor enters the tube side at the bottom of the condenser at about 80 °C (175 °F), is cooled and condensed by the circulating water on the shell side, and falls back to the reactor under gravity. The cooling water used is municipal supply water. The condenser had been in operation for five years in one of the units when it was shut down for nontechnical reasons. Water had been left stagnant on the shell side in the unit for three months after equipment shutoff until it was shifted to the second unit of the same plant and positioned on top of a similar reactor producing the same VAM vapors. While the shell side was being refilled with water, profuse leakage was noted. Upon opening the shell, it was revealed that almost all the tubes were pitted, with fully penetrating holes.

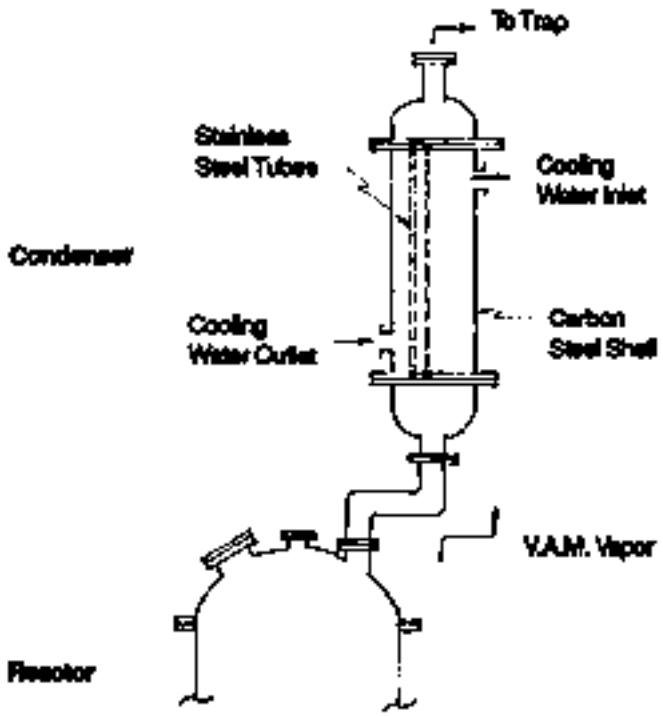


Fig. 15 Layout of a stainless steel reactor condenser that experienced pitting corrosion. VAM, vinyl acetate monomer. Source: Ref 20

Pertinent Specifications. The 117 tubes were of type 316 austenitic stainless steel manufactured by the seamless process to specification ASTM A 213. The tubes were 31.75 mm (1.25 in.) OD, had a 1.65 mm (0.065 in.) wall thickness, and were 3.0 m (9.8 ft) long. The tube sheets were also made of type 316 stainless steel, while the shell, tie rod, and baffle materials were of carbon steel. The VAM vapor was free from chlorides, with somewhat acidic condensate. The cooling water was 6.5 to 7 pH with approximately 20 ppm chlorides.

Visual Examination. Almost all the tubes had pitting damage on the OD surface. Some of the pits had penetrated fully, resulting in holes. Figure 16 shows the pits on the OD surface of two tubes. The nature of the pits indicates that they started on the OD surface and propagated both toward the interior of the tube wall and downward along the outside wall. The ID surfaces of the tubes were free from any form of corrosion. Also free from corrosion was the pipe that connects the reactor and the condenser and is exposed to the VAM vapor and the condensate.

Fig. 16 Pitting on the outside surface of type 316 stainless steel tubes, with downward propagation. Source: Ref 20

The bottom tube sheet showed deep pit marks on the shell side, while the top tube sheet was free from any form of corrosion. The baffle plates and the tie rods were completely corroded and practically disintegrated. Similarly, the shell, also of carbon steel, showed heavy corrosion on the inside, with a very rough nonuniform surface and thick, brownish rust scale.

Test Results. Chemical analysis of one of the pitted tubes indicated a composition (wt%) of 17.35% Cr, 13.46% Ni, and 2.19% Mo. These values fall within the specification for type 316 stainless steel. Macroexamination of the tested tube indicated that it was seamless. Microexamination of the tubes showed a uniform equiaxed austenitic structure with a grain size of ASTM 8 to 9, free from carbide precipitation. This indicated that the tubes had been properly manufactured.

Conclusions. The test results indicated that there was not a quality problem with the material. The discussion with plant personnel indicated that the stagnant water on the shell side (OD side of the tubes) caused the pitting. Municipal potable water with neutral pH is normally not corrosive to type 316 stainless steel. In this particular case, no corrosion problem had been experienced during operation for five years. However, even such water becomes corrosive to stainless steel if kept stagnant for extended periods. The aggressiveness increases if the stagnant water is kept closed, without access to atmospheric oxygen. Under such conditions, pitting results. Corrosion of the carbon steel contributed to the removal of oxygen in the water.

Once the pits were initiated, the liquid trapped inside the pit became more dense and more acidic than the bulk, and corrosion was accelerated, extending partly toward the inner surface of the tube and partly in a downward direction. This explains the propagation and elongation of the pits (Fig. 16).

It was recommended that the stainless steel surface not be kept in contact with closed stagnant water for extended periods, even if the water is of municipal potable quality.

Difficulty of Detection. Pits are generally small and often remain undetected. A small number of isolated pits on a generally uncorroded surface are easily overlooked. A large number of very small pits on a generally

uncorroded surface may not be detected by simple visual examination, or their potential for damage may be underestimated. When pits are accompanied by slight or moderate general corrosion, the corrosion products often mask them.

Pitting is sometimes difficult to detect in laboratory tests and in service because there may be a period of months or years, depending on the metal and the corrodent, before the pits initiate and develop to a readily visible size. Delayed pitting sometimes occurs after an unpredictable period of time in service, when some change in the environment causes local destruction of a passive film. When this occurs on stainless steels, for example, there is a substantial increase in solution potential of the active area, and pitting progresses rapidly.

Stages of Pitting. Immediately after a pit has initiated, the local environment and any surface films on the pit-initiation site are unstable. The pit may become inactive after just a few minutes if convection currents sweep away the locally high concentration of hydrogen ions, chloride ions, or other ions that initiated the local attack. Accordingly, the continued development of pits is favored in a stagnant solution.

When a pit has reached a stable stage, barring drastic changes in the environment, it penetrates the metal at an ever-increasing rate by an autocatalytic process. In the pitting of a metal by an aerated sodium chloride solution, rapid dissolution occurs within the pit, while reduction of oxygen takes place on adjacent surfaces. This process is self-propagating. The rapid dissolution of metal within the pit produces an excess of positive charges in this area, causing migration of chloride ions into the pit.

Thus, in the pit there is a high concentration of MCl_n and, as a result of hydrolysis, a high concentration of hydrogen ions. Both hydrogen and chloride ions stimulate the dissolution of most metals and alloys, and the entire process accelerates with time. Because the solubility of oxygen is virtually 0 in concentrated solutions, no reduction of oxygen occurs within a pit. Cathodic reduction of oxygen on the surface areas adjacent to pits tends to suppress corrosion on these surface areas. Thus, isolated pits cathodically protect the surrounding metal surface.

Because the dense, concentrated solution within a pit is necessary for its continuing development, pits are most stable when growing in the direction of gravity. Also, the active anions are more easily retained on the upper surfaces of a piece of metal immersed in or covered by a liquid.

This gravitational effect was seen on the tubes of Fig. 16 as discussed in example 5.

The rate of pitting is related to the aggressiveness of the corrodent at the site of pitting and the electrical conductivity of the solution containing the corrodent. For a given metal, certain specific ions increase the probability of attack from pitting and accelerate that attack once initiated. Pitting is usually associated with metal-environment combinations in which the general corrosion rate is relatively low; for a given combination, the rate of penetration into the metal by pitting can be 10 to 100 times that by general corrosion.

Depth of Penetration. The depth of pitting can be expressed by the pitting factor (Fig. 17) (Ref 21). A value of 1 would represent uniform corrosion. The maximum depth of penetration (*p*) can be measured by several methods, including metallographic examination, machining, use of micrometer, or microscope. The average penetration depth (*d*) is calculated from the weight lost by the sample. The maximum penetration depth is extremely significant if the metal is part of a barrier or tank, or is part of a pressurized system. The micrograph (Fig. 18) shows the profile of pits that have degraded an austenitic stainless steel thin-walled bellows.

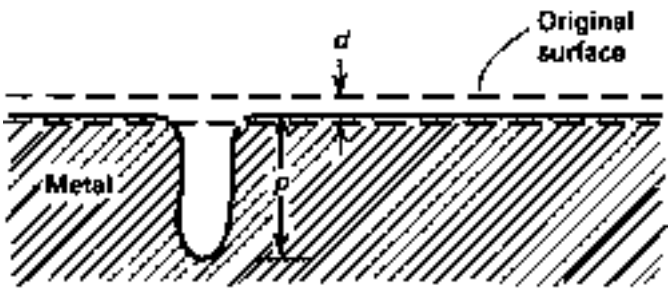


Fig. 17 Measurements used to determine the pitting factor, *p*/*d*, where *p* is maximum penetration and *d* is the average penetration depth.

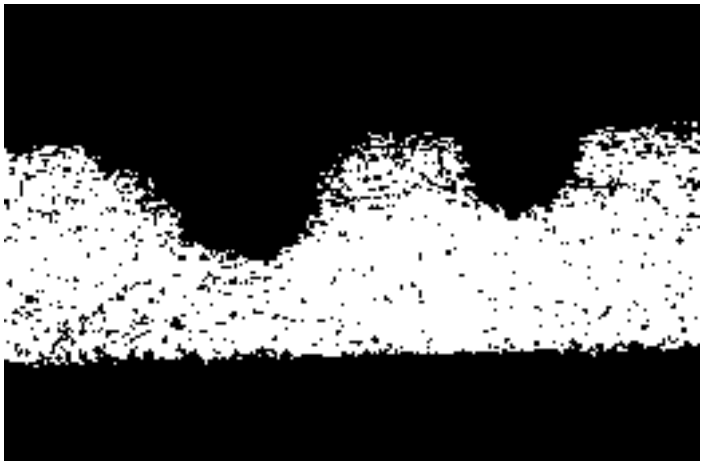


Fig. 18 Light micrograph showing the general microstructure and cross section of corrosion pits in an austenitic stainless steel thin-walled bellows. Total thickness of sheet approximately 0.5 mm (0.02 in.). 10% oxalic acid etch. 100×. Courtesy of M.D. Chaudhari, Columbus Metallurgical Service

Density and Size of Pitting. For a mechanical component, the density of pits (number per unit surface area) and size may be a more critical characteristic than maximum depth. The loss of effective cross section can decrease the strength of the component and pits can become sites of stress concentrations. ASTM G 46 points out, conversely, the loss of mechanical properties can be used to evaluate the severity of pitting corrosion (Ref 22).

Example 6: Corrosion Failure by Pitting of Type 321 Stainless Steel Aircraft Freshwater Tanks Caused by Retained Metal-Cleaning Solution. Two freshwater tanks of type 321 stainless steel were removed from aircraft service because of leakage. One tank had been in service for 321 h, the other for 10 h. The tanks were made of 0.81 mm (0.032 in.) thick type 321 stainless steel. Pitting and rusting had occurred only on the bottom of the tanks near a welded outlet, where liquid could be retained after draining. Most of the pits were about 6 to 13 mm (0.2 to 0.5 in.) from the weld bead at the outlet.

Service History. There had been departures from the specified procedure for chemical cleaning of the tanks in preparation for potable water storage. The sodium hypochlorite sterilizing solution used was at three times the prescribed strength. The tanks were drained after the required 4 h sterilizing treatment, but a small amount of solution remained in the bottom of the tanks and was not rinsed out immediately. This exposed the bottom surfaces of the tanks to hypochlorite solution that had collected near the outlet. Rinsing was delayed for 16 h for the tank that failed after 321 h of service and for 68 h for the tank that failed after 10 h of service.

Chemical Tests. Samples of the tanks were subjected to a 5% salt-spray test for three days. No corrosion was observed. Additional samples, which were subjected to one cycle of the Huey test (a test sample is boiled in 65% HNO_3 for 48 h), showed no appreciable intergranular attack.

Samples of both tanks were tested for reaction with sodium hypochlorite at three strength levels:

- The prescribed concentration of 125 ppm available chlorine
- The triple-strength solution (375 ppm available chlorine) used in sterilizing both tanks
- The full-strength solution (5% available chlorine) from the stock bottle

The samples immersed in the first solution rusted and discolored the solution in approximately 2 h. Samples in the second solution showed no corrosion for one to two days but a slight reaction after four to five days. Samples in the third solution showed no corrosion after four to five days of immersion.

Samples from the tanks, including the heat-affected zones (HAZs) near the welds at the outlet, were pickled in an aqueous solution containing 15 vol% concentrated nitric acid (HNO_3) and 3 vol% concentrated hydrofluoric acid (HF) to remove heat discoloration and scale and were then immersed in the three sodium hypochlorite solutions for several days. None of the samples showed any sign of corrosion.

Metallographic inspection of specimens from the tanks disclosed a normal microstructure with no precipitated carbides. Cross sections through pits in the tanks showed subsurface enlargement of the pits (Fig. 19).

Fig. 19 The bottom of a type 321 stainless steel aircraft freshwater storage tank that failed in service as a result of pitting. This unetched section shows subsurface enlargement and undercutting of one of the pits. Approximately 95×

Conclusions. Oxidizing solutions, such as hypochlorites, and nitrates in certain concentration ranges cause rapid pitting of stainless steels. With greater concentration, passivation is complete; at lower concentrations, chloride does not penetrate the passive film. Apparently, the concentration of sterilizing solution used was in the range that promotes pitting.

Failure of the stainless steel tanks by chloride-induced pitting resulted from using an overly strong hypochlorite solution for sterilization and neglecting to rinse the tanks promptly afterward. Failure to remove the scale from the HAZ near the welds at the outlet and failure to passivate the interior of the tanks may have accelerated the pitting, although severe pitting also occurred outside the HAZ.

Corrective Measures. Directions for sterilization and rinsing were revised. The instructions for diluting the hypochlorite solution were clarified, and the importance of using the proper strength solution was emphasized. Immediate rinsing after draining the solution was specified, using six fill-and-drain cycles and leaving the tank full of potable water as the final step.

To improve resistance to chloride-induced pitting further, it was recommended that welding be carefully controlled and adequately protected with inert gas to restrict heat effects to formation of a golden-brown to deep-purple discoloration on the metal, avoiding the formation of black scale. It was also recommended that the scale and discoloration of the welds at the outlet and HAZs be removed by abrasive blast cleaning and that the interior of the completed tanks be passivated, using either 22% HNO_3 at 60 °C (140 °F) or 50% HNO_3 at 50 °C (125 °F). Three rinse cycles of water immersion and drainage were stipulated, to be followed by final rinsing inside and out with freshwater and then oven drying.

Example 7: Fatigue Fracture of Stainless Steel Wires in an Electrostatic Precipitator at a Paper Plant (Ref 22). Several type 316L stainless steel wires in an electrostatic precipitator at a paper plant fractured in an unexpectedly short time. The wires were 2.6 mm (0.10 in.) in diameter. The chemical specification of the steel from which the wires were made was similar to a type 316L stainless steel. The wires were also specified to be $\frac{3}{4}$ hard.

Service Conditions. Wires were subjected to stress by weights hung from them and were exposed to a very corrosive gas effluent. They were periodically vibrated to free particles that adhered to them.

Analysis. Failed wires were examined using optical and scanning electron microscope (SEM), and hardness tests were conducted. Fractography clearly established that fracture was caused by fatigue originating at corrosion pits on the surface of the wire (Fig. 20). Similar pits were found on all wires at the fatigue crack origin and along the surface of the wires (Fig. 21).

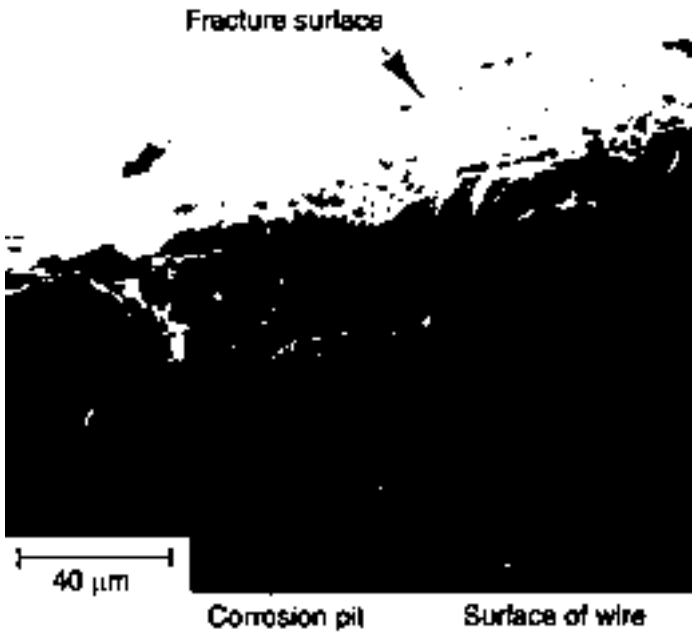


Fig. 20 Scanning electron micrograph showing a corrosion pit at the origin of a fatigue fracture in a stainless steel wire. Source: Ref 23

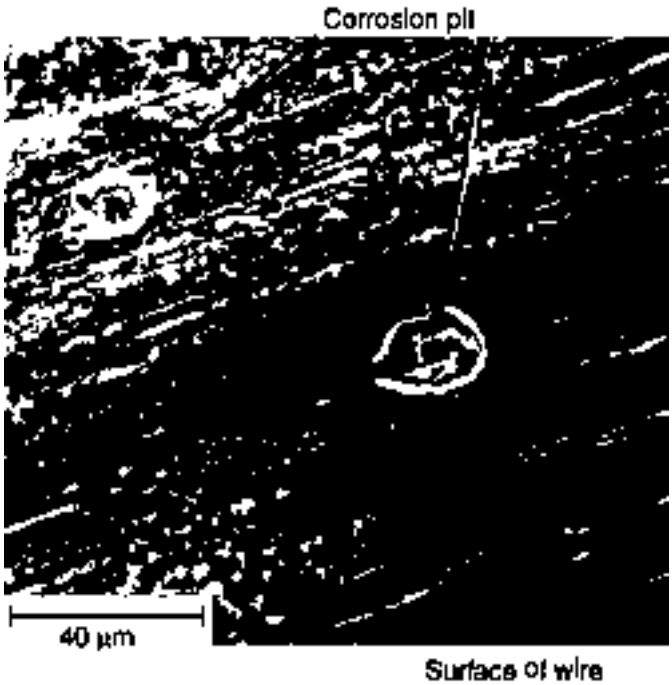


Fig. 21 Scanning electron micrograph of the surface of a stainless steel wire, showing corrosion pits. Source: Ref 23

Corrective Recommendation. The gas effluent was very corrosive, but the general attack on the surface of the wires appears slight. Thus, an austenitic stainless steel appears to be a reasonable choice. Also, specifying a steel similar to type 316L is reasonable because the molybdenum content (2 to 3%) is designed to inhibit pitting corrosion. However, pitting did occur; thus, it was recommended that a higher-molybdenum austenitic stainless steel, type 317L (3 to 4% Mo), be used (although the lower limit of molybdenum content is identical to the upper limit of type 316L). It was also recommended that the wires be used in the annealed condition, unless other aspects (e.g., static loading) required a material of high tensile strength.

Fracture Origins. Origins of corrosion-fatigue fractures are often surrounded by crack-arrest lines, or beach marks (Fig. 22). Often, there are several origins, particularly when fracture has been initiated by pitting (Fig. 23). Occasionally, no well-defined origin can be resolved.

Fig. 22 A single-origin corrosion-fatigue crack in type 403 stainless steel exposed to steam showing rubbed origin (arrow) and beach marks. Scanning electron fractograph (secondary electron image). 60×

(a)

(b)

Fig. 23 Two views of the fracture surface of a forged 17-4 PH stainless steel steam-turbine blade that failed by corrosion fatigue originating at severe corrosion pitting. (a) Light fractograph showing primary origin (arrow) and three secondary origins (along right edge below primary origin). 7.5×. (b) Scanning electron fractograph (secondary electron image) of area surrounding primary origin (in circle). 50×. White areas are corrosion pits. Black areas, one of which is indicated by arrow, are remnants of corrosion product left after the fracture surface was electrolytically cleaned.

Features at the origin of a corrosion-fatigue fracture are often indistinct because the compression portion of each stress cycle has forced mating fracture surfaces together and has formed an extremely rubbed, discolored origin (arrow, Fig. 22). Also, the area of the origin is exposed to the environment for the longest time and thus may exhibit more extensive residues of a corrosion product than the rest of the fracture surface. A corrosion product at fatigue origins can be misleading when viewed macroscopically. The oxide may be flat and tenacious, globular, or nodular. At magnifications of less than 60, globular or nodular oxide particles are easily mistaken for intergranular facets; therefore, no conclusions should be drawn until other metallographic or fractographic techniques confirm the mechanism of crack initiation. Occasionally, oxidation is so severe that no information other than location of origin can be obtained.

Other features that can be observed macroscopically are secondary cracks, pits, and fissures, all of which are often adjacent to the main origin of a particular fracture. In corrosion-fatigue failures, cracks and fissures adjacent to the primary origin are indicative of a uniform state of stress at that location. Therefore, the primary fracture simply propagated from the flaw that either induced the most severe stress concentration or was exposed to the most aggressive local environment. Sometimes, the primary crack origin is related more to heterogeneous microstructure than to the stress distribution.

Pitting of Various Metals. With carbon and low-alloy steels in relatively mild corrodents, pits are often generally distributed over the surface and change locations as they propagate. If they blend together, the individual pits become virtually indistinguishable, and the final effect is a toughened surface but a generally uniform reduction in cross section. If the initial pits on carbon steel do not combine in this way, the result is rapid penetration of the metal at the sites of the pits and little general corrosion.

Despite their good resistance to general corrosion, stainless steels are more susceptible to pitting than other metals. High-alloy stainless steels containing chromium, nickel, and molybdenum are also more resistant to pitting, but are not immune under all service conditions.

The pitting resistance equivalent number (PREN), or pitting index, is devised to quantify and compare the resistance of alloys to pitting:

$$PREN = \%Cr + 3.3(\%Mo) + 16\,(\%N) \qquad \text{(Eq 1)}$$

The higher the number the more resistant the alloy is to pitting and crevice corrosion. A variety of PREN equations have been developed for specific alloy and weld material groups (Ref 24). Equations include effects of nickel, manganese, nitrogen, tungsten, and carbon.

Pitting failures of corrosion-resistant nickel alloys, such as Hastelloy C, Hastelloy G, and Incoloy 825, are relatively uncommon in solutions that do not contain halides, although any mechanism that permits the establishment of an electrolytic cell in which a small anode is in contact with a large cathodic area offers the opportunity for pitting attack.

In active metals, such as aluminum and magnesium and their alloys, pitting also begins with the establishment of a local anode. However, a major effect on the rate of pitting penetration is that the corrosion products of these metals produce a high pH in the pit, thus accelerating the dissolution of the protective oxide film. Although the surrounding metal may be exposed to relatively pure water, the metal at the bottom of a pit corrodes at an appreciable rate, even in the absence of galvanic effects. Maintaining surface cleanness helps avoid pitting of these metals.

Pitting of aluminum and magnesium and their alloys in aqueous solutions is often accelerated by galvanic effects, which occur because these metals are anodic to most other metals, and by the effects of dissolved metallic ions and suspended particles in the solution. Pitting of these metals can be caused by even a few parts per billion of dissolved copper compounds as well as by particles of metallic iron (which produce both crevice and galvanic effects) that are embedded in or adhering to their surfaces.

Example 8: Corrosion by “Green Death” in Reboiler Bypass Duct Damper in Power- Generation Plant (Ref 24). “Green Death” is a corrosive medium used to test for pitting resistance composed of 11.9% H_2SO_4 + 1.3% HCl + 1% $FeCl_3$ + 1% $CuCl_2$. The solution is heated to boiling, 103 °C (217 °F), for the test. Such a corrosive environment occurs at power plants.

Service Environment. A continuous supply of flue gas comes in contact with a closed bypass duct. The unscrubbed combustion products condense on the cold duct. Then the closed damper conducts heat from the chimney and reheats the condensate. The environment was severe enough to cause the C-276 nickel alloy welds (N10276) on the C-276 duct floor to completely disappear in less than half a year.

Corrective Measures. 686CPT weld alloy was used to repair the C-276 plates. The plate continued to corrode at a rate greater than 1 mm/year (40 mil/year). The plate itself was then replaced by alloy 686 plate (N06686) welded with 686 CPT.

Conclusions. Though the original materials are recognized for their resistance to corrosion and heat, the 686 alloy composition has superior pitting resistance as indicated by higher PREN values according to several formulas. Tests of welded coupons corroborate this conclusion.

References cited in this section

20. K.E. Perumal, Pitting Corrosion of Stainless Steel by Potable Municipal Water in an Organic Chemical Plant, *Handbook of Case Studies in Failure Analysis,* Vol 1, ASM International, 1992, p 176–177

21. A.I. Asphahani and W.L. Silence, Pitting Corrosion, *Corrosion,* Vol 13, Metals Handbook, 9th ed., ASM International, 1987, p 112–114

22. “Examination and Evaluation of Pitting Corrosion,” G 46-94 (Reapproved 1999), *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

23. C.R. Brooks, Fatigue Fracture of Stainless Steel Wires in an Electrostatic Precipitator at a Paper Plant, *Handbook of Case Studies in Failure Analysis,* Vol 1, ASM International, 1992, p 218–222

24. S.D. Kiser, Preventing Weld Corrosion, *Adv. Mater. Process.,* Vol 160 (No. 3), March 2002

Crevice Corrosion

A crevice at a joint between two metallic surfaces, between a metallic and a nonmetallic surface, or beneath a particle of solid matter on a metallic surface provides conditions that are conducive to the development of the type of concentration-cell corrosion called crevice corrosion. Crevice corrosion can progress very rapidly (tens to hundreds of times faster than the normal rate of general corrosion in the same given solution). The classic example of this is a demonstration that a sheet of stainless steel can be cut (corroded) into two pieces simply by wrapping a rubber band around it, then immersing the sheet in seawater or dilute ferric chloride solution. The open surfaces will pit slowly, but the metal under the rubber band will be attacked rapidly for as long as the crevice between the rubber band and the steel surface exists.

In a metal-ion concentration cell, the accelerated corrosion occurs at the edge of or slightly outside of a crevice. In an oxygen-concentration cell, the accelerated corrosion usually occurs within the crevice between the mating surfaces.

Riveted and bolted joints must be considered prime sites for crevice corrosion; therefore, they require careful attention in design and assembly to minimize crevices, as well as provisions to ensure uniform aeration and

moderate but not excessive flow rates at the joints. Not only does the geometry of the joint itself promote crevice corrosion, but corners and cracks can collect debris, which exacerbates the potential for corrosion.

Replacement with welded joints can eliminate crevice corrosion, provided special care is taken in welding and subsequent finishing of the welds to provide smooth, defect-free joints. Using adhesive bonding alone or adhesive bonding prior to riveting is used to overcome crevice corrosion in aluminum (Ref 25).

Figure 24 is an example of crevice corrosion where type 316 stainless steel bubble caps contacted a type 316 tray deck. At the contact points, the type 316 stainless steel was selectively corroded by the concentrated acetic-acid solution. In this part of the distillation column, the oxidizing capacity of the acid stream is nearly exhausted, but is sufficient to prevent corrosion of the open surfaces. A change to a higher-alloy stainless steel was necessary to fully resist these corrosive conditions. Any layer of solid matter on the surface of a metal that offers the opportunity for exclusion of oxygen from the surface or for accumulation of metal ions beneath the deposit because of restricted diffusion is a probable site for crevice corrosion.

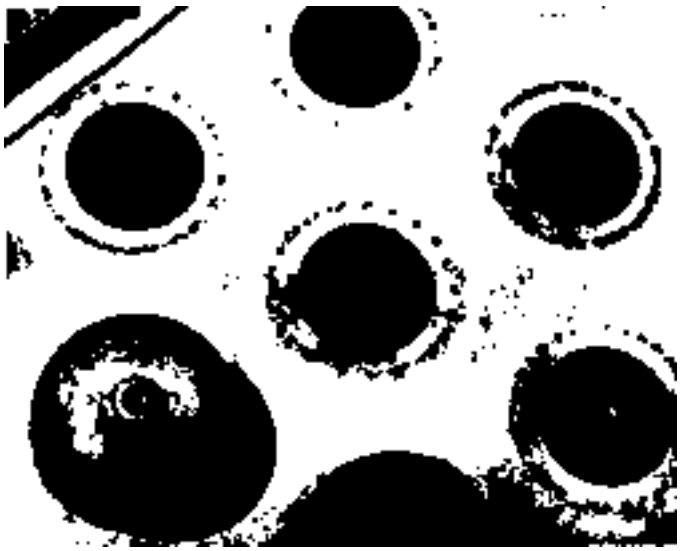


Fig. 24 Crevice corrosion pitting that has taken place where type 316 bubble caps contact a type 316 stainless steel tray deck. The oxygen-concentration cell corrosion occurred in concentrated acetic acid with minimal oxidizing capacity. $\frac{1}{8}$ actual size

Differential aeration beneath solid deposits or at cracks in mill scale is a frequent cause of crevice corrosion in boilers and heat exchangers. Suitable water treatment to provide thin protective films, together with special attention to cleaning, rinsing, and drying when boilers are shut down, minimizes the occurrence of crevice corrosion in such equipment. Breaks (holidays) in protective organic coatings or linings on vessels containing corrosive chemicals are also likely sites for the development of crevice corrosion.

Example 9: Crevice Corrosion on Stainless Steel Tube. A steel tube meeting type 304 specification (0.008 max C,18.00 to 20.00 Cr,2.00 max Mn, 8.00 to 10.50 Ni) was found to be corroded as shown in Fig. 25. The tube was part of a piping system, not yet placed in service, that was exposed to an outdoor marine environment containing chlorides. As part of the assembly, a fabric bag containing palladium oxide (PdO_2) was taped to the tube. The palladium served as a “getter.” The corrosion in this area was discovered during routine inspection. A close-up (Fig. 26) of the tube shows the corrosion products. Pits are the darkest areas. Slightly lighter areas include remnants of the fabric bag.

Fig. 25 Austenitic stainless steel tube that was corroded where a fabric bag was taped to it. Courtesy of M.D. Chaudhari, Columbus Metallurgical Service

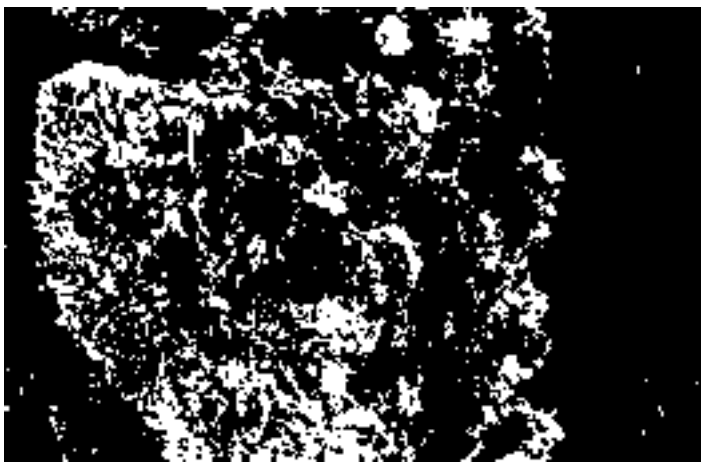


Fig. 26 Micrograph of an area about 10 mm ($\frac{3}{8}$ in.) wide from the tube shown in Fig. 25. Pits are the dark areas in the center. Lighter areas include fragments of the bag among the corrosion products. Courtesy of M.D. Chaudhari, Columbus Metallurgical Service

Analysis. Corrosion debris was analyzed using an electron-dispersive spectrometer (EDS). The profiles confirmed the presence of chlorides and palladium. Both contributed to corrosion in the crevice created by the tape on the tube, which was periodically exposed to water.

Corrective Measures. Steps were taken to prevent water from entering and being trapped in this area of the assembly.

Liquid-Level Effects. Crevice corrosion often occurs underneath deposits of solid substances that sometimes collect just above the liquid level on a metal part that is partly immersed in an electrolyte. The deposits usually remain moist or are intermittently moist and dry.

Where the liquid level fluctuates or where the liquid is agitated, the area of the metal that is intermittently wetted is called the splash zone. Splash zones are encountered in all types of tanks and equipment containing liquids; corrosion occurring in these areas is called splash-zone corrosion.

Example 10: Crevice Corrosion of a Carbon Steel Brine Tank. A failure caused by this type of crevice corrosion occurred in a carbon steel tank that contained a saturated solution of sodium chloride (brine) at room temperature. When a pump that was apparently in good condition failed to deliver brine, inspection showed that a steel suction line, partly immersed in the liquid and used to transfer small amounts of brine to other equipment, had corroded through in the intermittently wetted area. Further inspection showed that the tank had also been attacked in the intermittently wetted area. The corrosion problem was eliminated by replacing the steel suction line with a glass fiber reinforced epoxy pipe and by lining the tank with a coal-tar epoxy coating.

Example 11: Corrosion of Steel Pilings. Rust films that have little opportunity to dry do not develop protective properties. Such unfavorable conditions exist in the splash zone above the high-tide level along the seashore. The corrosivity of the moist rust films is further aggravated by the high oxygen content of the splashing seawater. Observations of old steel piling along the seashore usually reveal holes just above the water line, where the rate of corrosion is several times greater than that for continuous immersion in seawater.

Corrosion under Thermal Insulation. The application of thermal insulation to carbon steel and stainless steel equipment presents a special case of crevice corrosion. It is special because there are a number of misconceptions about the role of thermal insulation in this corrosion situation, and the phenomenon is not, strictly speaking, crevice corrosion, but does not comfortably fit in any other category. Thermal insulation is used to maintain temperature of equipment, either below ambient or above ambient; therefore, the metal normally will operate at temperatures different from the surrounding temperature. Because of this situation, the thermal insulation system—which consists of the metal outer surface, the annular space between the metal and the insulation, the insulation itself, and its weather barrier—is a dynamic, pervious combination of materials. The weather barrier or weather proofing, coating, wrapping, or cladding, while intended to protect the insulation from the ravages of weather and other mechanical damage, does not prevent the entry of humidity and moisture to the annular space between the metal surface and the inner side of the insulation.

Therefore, with moisture present on the metal surface, the insulation acts as a barrier in terms of the evaporation of the moisture and maintains it in a vapor phase in contact with the metal surface. There is also free access to oxygen as well as contaminants that might be in the insulation. This system can then become very aggressive when the metal temperature is operating in the range of 40 to 120 °C (100 to 250 °F). At these elevated temperatures, the humid annular space can corrode steel at rates as high as 0.8 mm/yr (0.03 in./yr) or, in the case of stainless steel, can initiate chloride SCC with relatively rapid failure times.

The approach to preventing corrosion under insulation has been examined by many investigators (Ref 26). The attempts to produce insulation materials of high purity and with silicate-type inhibitors added are being used. However, they offer limited protection against aggressive atmospheres that may be introduced beneath the insulation through cracks, crevices, and other openings in the insulation. The most widespread approach at this time is to use a protective coating on the metal to protect it from the water vapor existing in the annular space.

When selecting protective coatings, they should be considered as being in warm-water immersion service because of the severity of the environment. The most severe environments develop when the vessels or piping operate with cycling temperatures, that is, operate below ambient for a period of time and then go through a heating cycle that tends to evaporate the water but go through the hot metal corrosion temperature zone.

The problem of corrosion under thermal insulation has become widely recognized as one of significant concern. Because of the hidden nature of the problem, a well-insulated vessel may operate for many years without any apparent external signs of damage, while significant corrosion is taking place under the insulation.

Effects of Solid Deposits. In power-generation equipment, crevice-corrosion failures have occurred in main-station condenser tubes cooled with seawater as a result of the formation of solid deposits and the attachment of marine organisms to the tube wall. These failures have occurred particularly in condensers with stainless steel tubing.

Failures on the inside of tanks have been attributed to weld splatter that went undetected. The splatter created unnecessary and unwanted crevices. Good practices would have dictated that such weld splatter be ground off.

Crater Corrosion. Type 347 stainless steel is sometimes subject to a type of crevice corrosion known as crater corrosion, which occurs at the stop point of a weld. The failure is related to microsegregation of certain constituents in the pool of molten metal that is the last to solidify on a weld. In a manner similar to zone-melting refinement, the moving weld pool continuously sweeps selected constituents ahead of it, and the concentration of these selected constituents in the pool increases continuously until the welding is stopped and the pool solidifies. The center of the stop point is attacked rapidly in oxidizing acids, such as nitric acid, in a form of self-accelerating crevice corrosion. The final solidification pool need not extend through the thickness of the part for full perforation to occur, as described in example 12.

Example 12: Weld Craters in Stainless Steel Heat Exchanger Tubes. Beveled weld-joint V-sections were fabricated to connect inlet and outlet sections of tubes in a type 347 stainless steel heat exchanger for a nitric acid concentrator. Each V-section was permanently marked with the tube numbers by a small electric-arc pencil.

After one to two years of service, multiple leaks were observed in the heat-exchanger tubes. When the tubes were removed and examined, it was found that the general corrosion rate was normal for service of heat-exchanger tubes in a nitric acid concentrator, but that crater corrosion had perforated the tubes.

The crater corrosion occurred at two general locations. One location was at the stop point of the welds used to connect the inlet and outlet legs of the heat exchanger. The other location was at the stop points on the identifying numerals. The material was changed to type 304L stainless steel, in which the zone-melting concentration does not take place.

Crevice Corrosion of Implant Materials. A large number of stainless steels have been evaluated with a modified PREN equation for resistance to crevice corrosion in body fluids (Ref 27). To avoid allergies due to nickel, high-manganese stainless steel is under consideration, along with duplex stainless steels and other austenitic stainless steels. Titanium is widely used and cobalt-chromium alloys also exhibit a passive nature in the body.

ASTM F 746 (Ref 28) is test method for evaluating the crevice and pitting corrosion resistance of a material to rank its suitability for surgical implants. Evaluation is based on the critical electrical potential for pitting or crevice corrosion. The test is severe, so it does not follow that materials that corrode in the test will corrode when implanted. This is but one of the criteria in the material selection process. Obviously, the effect of corrosion products within the body is of paramount importance.

It has been known for some time that certain stainless steel implants had been subject to crevice corrosion. The reaction of implants to body fluids is quite like the reaction to seawater. Crevices occur at the interface where screws attach the implant, chlorides are present, and oxygen is depleted in the crevice. Evidence of crevice corrosion was observed in 90% of stainless steels in one study cited in 1985 (Ref 29). In a 1980 study (Ref 30) it was noted that in stainless steel devices where more than one component was present, such as bone plates, interfacial corrosion approached 100%.

Failure analysis is complicated by the fact that it is difficult to determine if the screw-plate interface of surgically recovered damaged implants was damaged prior to implantation, during implantation, or during removal.

References cited in this section

25. J.E. Hatch, Ed., *Aluminum: Properties and Physical Metallurgy,* American Society for Metals, 1984, p 290–291

26. *Corrosion of Metals under Thermal Insulation,* STP 880. American Society for Testing and Materials, 1985

27. A.J. Sedriks, *Corrosion of Stainless Steels,* 2nd ed., Wiley-Interscience, 1996, p 217

28. "Standard Test Method for Pitting or Crevice Corrosion of Metallic Surgical Implant Materials," F 746-87 (Reapproved 1999), *Annual Book of ASTM Standards,* Vol 13.01, American Society for Testing and Materials

29. N.D. Green, Corrosion of Surgical Implant Alloys: A Few Basic Ideas, *Corrosion and Degradation of Implant Materials,* STP 859, American Society for Testing and Materials, 1985, p 7

30. H. Skinner, et al., “Corrosion, Materials Characteristics and Local Tissue Reaction Associated with Osteosynthesis Devices, Implant Retrieval: Material and Biological Analysis,” NBS Special Publication 601, U.S. Dept. of Commerce, Jan 1981, p 425

Reducing Failures Due to Pitting and Crevice Corrosion

Steps to minimize the degrading effects of pitting and crevice corrosion include:

- Reducing the harshness of the environment
- Selecting appropriate materials
- Avoiding galvanic couples
- Avoiding rupture of any natural or applied protective coatings
- Avoiding contact with stagnant solutions or providing circulation of otherwise stagnant solutions
- Modifying component design to eliminate crevices and ensure proper drainage
- Inspecting and maintaining critical components so that pitting and crevice corrosion do not develop into more catastrophic failures

Testing can point the way to the most effective approach to avoiding failures. Testing for localized corrosion resistance is used to (Ref 31):

- Rank alloys for development and material selection
- Analyze failures
- Determine effects of changes in process parameters
- Predict penetration rates

Currently available tests aid in making more informed decisions regarding material selection and system and process design. The nature of localized corrosion makes prediction of penetration rates and service life elusive, but statistical studies show promise.

Reference cited in this section

31. R.G. Kelley, *Pitting, Corrosion Tests and Standards: Application and Interpretation,* MNL 20, R Baboian, Ed., American Society for Testing and Materials, 1995

Intergranular Corrosion

Revised by Randy K. Kent and Roy Baggerly, MDE Engineers, Inc.

Intergranular (IG) corrosion is the preferential dissolution of the anodic component, the grain-boundary phases, or the zones immediately adjacent to them, usually with slight or negligible attack on the main body of the grains. The galvanic potential of the grain-boundary areas of an alloy is anodic to that of the grain interiors due to differences of composition or structure. These differences may be due to the environmental interactions or metallurgical changes in the grain-boundary regions during manufacturing or service exposure.

Development of Intergranular Corrosion

Intergranular corrosion is usually (but not exclusively) a consequence of composition changes in the grain boundaries from elevated-temperature exposure. In general, grain boundaries can be susceptible to changes in composition, because grain boundaries are generally slightly more active chemically than the grains themselves. This is due to the areas of mismatch in the grain-boundary regions relative to the more orderly and stable crystal lattice structure within the grains. The relative disorder in the grain boundaries provides a rapid diffusion path, which is further enhanced by elevated-temperature exposure. Therefore, a variety of chemical changes may occur preferentially in the grain-boundary areas such as segregation of specific elements or compounds, enrichment of an alloying element, or the depletion of an element when precipitates nucleate and grow preferentially in this region. Impurities that segregate at grain boundaries may also promote galvanic action in a corrosive environment.

When the compositional changes in the grain boundary result in a more anodic material or galvanic potential difference, the metal becomes “sensitized” and may be susceptible to intergranular attack in a corrosive environment. A well-known example of this is the sensitization of stainless steels caused by diffusion of chromium and trace amounts of carbon around the grain boundaries during the elevated-temperature “sensitizing” exposure and precipitation of chromium carbides. Some aluminum alloys exhibit similar behavior with the precipitation of $CuAl_2$ or Mg_2Al_3, depending on the alloy.

Example 13: Intergranular Corrosion of an Aluminum Alloy Ship Hull. The 5*xxx* series of aluminum alloys are often selected for weldability and are generally very resistant to corrosion. However, if the material has prolonged exposure at slightly elevated temperatures of 66 to 180 °C (150 to 350 °F), an alloy such as 5083 can become susceptible to intergranular corrosion.

This sensitizing has occurred in ship hulls with disastrous consequences in the corrosive marine environment. This occurred in a new ship hull after only three months in service. Figure 27 is typical of many cracks that were discovered. The microstructure revealed the familiar ditch structure of sensitized aluminum alloy (Fig. 28, 29).

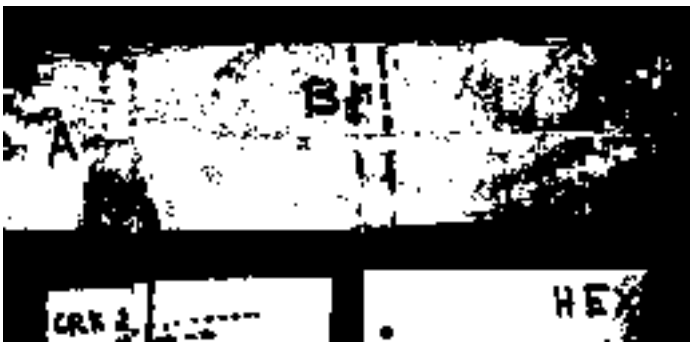


Fig. 27 Cracking in a 5083 aluminum alloy ship hull caused by sensitization. Courtesy of MDE Engineers, Inc.

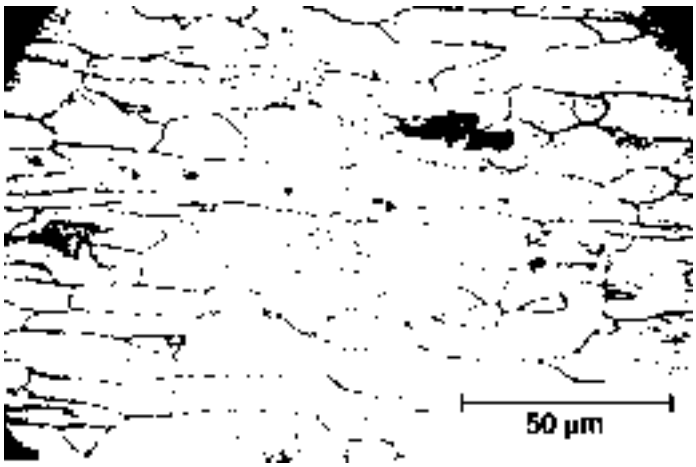


Fig. 28 Microstructure of 5083 aluminum alloy ship hull that has been sensitized. Courtesy of MDE Engineers, Inc.

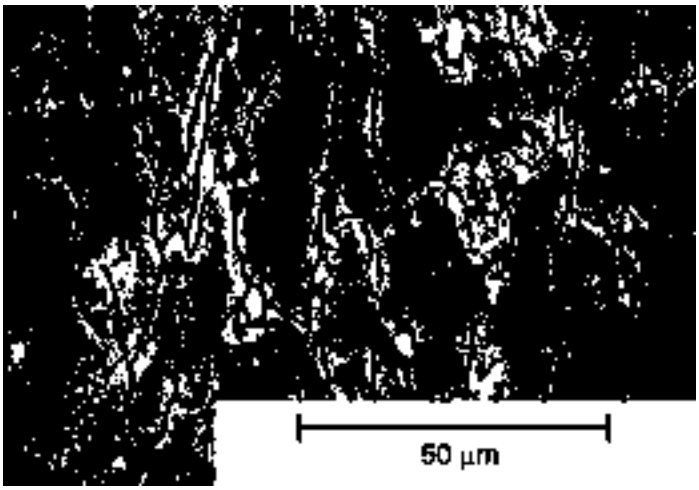


Fig. 29 Scanning electron micrograph of sensitized 5083 aluminum microstructure shown in Fig. 28. Courtesy of MDE Engineers, Inc.

A corrosion test was performed according to ASTM G 67 (Ref 32). The test confirmed the susceptibility of this particular material to intergranular corrosion as the test sample suffered a mass loss of 67 mg/cm^2 (429 mg/in.^{2}). By the standard, susceptible materials have a mass loss ranging from 25 to 75 mg/cm^2 (160 to 480 mg/in.^{2}). This relatively large mass loss occurs because corrosion around grain boundaries causes grains to fall from the sample. A material considered not susceptible to intergranular corrosion would suffer a mass loss from 1 to 15 mg/cm^2 (10 to 100 mg/in.^{2}).

During exposures to chloride solutions like seawater, the galvanic couples formed between these precipitates and the alloy matrix and led to severe intergranular attack.

Reference cited in this section

32. "Standard Practice for Determining the Susceptibility to Intergranular Corrosion of 5*xxx* Series Aluminum Alloys by Weight Loss after Exposure to Nitric Acid (NAWLT Test)," G 67, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

Alloy Susceptibility

Intergranular corrosion can occur in many alloy systems under specific circumstances; susceptibility depends on the corrosive solution and on the metallurgical factors of alloy composition, fabrication, and heat treatment parameters. Comprehensive coverage of alloy system-environment combinations is beyond the scope of this article. However, the general evaluation and some commonly occurring conditions of intergranular corrosion are briefly described in the following sections. Individual alloy families are addressed. Susceptibility of a component to intergranular corrosion can be corrected by proper heat treatment (to distribute alloying elements more uniformly), modification of the alloy, or the use of a completely different alloy.

Although grain-boundary corrosion is sometimes associated with differences in electrochemical potential (as in some aluminum alloys; see example 13 in this article), intergranular corrosion is usually not the result of active grain boundaries and a passive matrix. Usually the corroding surface is at one potential. Differences in composition produce different corrosion rates at the same potential in the passive region. Figure 30 illustrates the electrochemistry of intergranular corrosion by comparing the polarization curves for grain-boundary and matrix areas. The corrosion rates are close or the same in the active and transpassive ranges, but vary considerably with potential in the passive range.

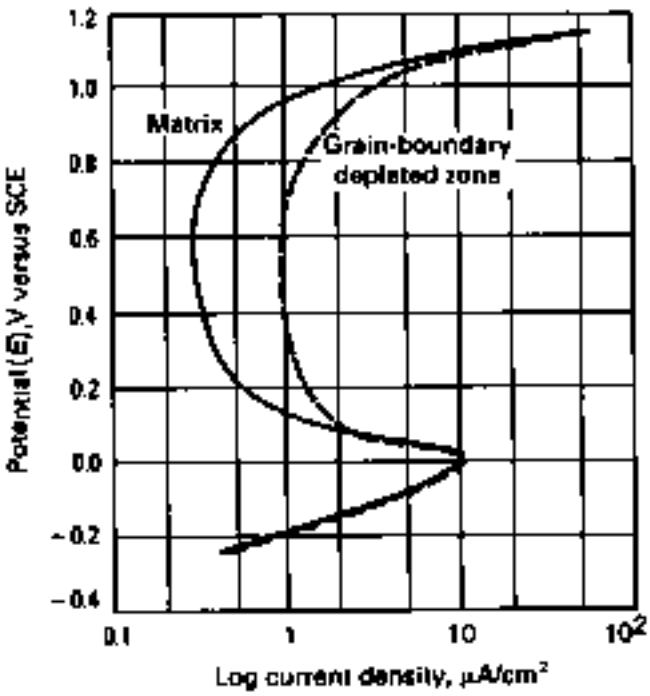


Fig. 30 Anodic polarization behavior of an active-passive alloy with grain-boundary depleted zones

Susceptibility to intergranular corrosion cannot be taken as a general indication of increased susceptibility to other forms of corrosion, such as pitting or uniform corrosion. Likewise, testing for susceptibility to intergranular attack should not be considered equivalent to evaluating the resistance of the alloy to general and localized corrosion, even though the tests used for evaluating susceptibility to intergranular attack may be severe. They are not intended to duplicate conditions for the wide range of chemical exposures present in an industrial plant.

When the attack is severe, entire grains may be dislodged because of the complete deterioration of their boundaries. When an alloy is undergoing intergranular corrosion, its rate of weight loss may accelerate with time. As the grain-boundary area dissolves, the unaffected grains are undermined and fall out; this increases the

rate of weight loss. Typical weight loss curves for an alloy undergoing intergranular corrosion are shown in Fig. 31 (Ref 33).

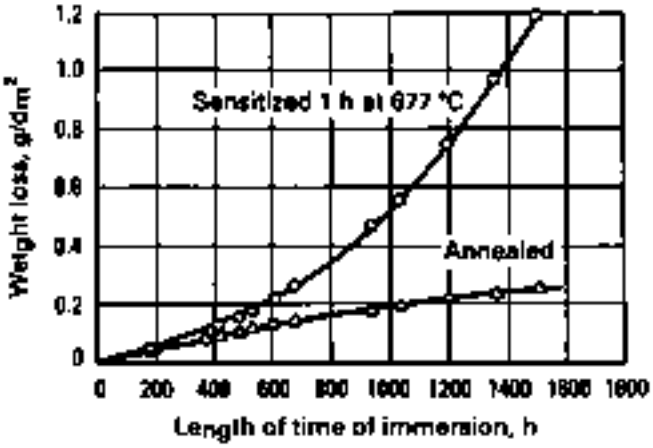


Fig. 31 Corrosion of type 304 steel in inhibited boiling 10% (H_2SO_4). Inhibitor: 0.47 g Fe^{3+}/L of solution added as $Fe_2(SO_4)_3$. Source: Ref 30

References cited in this section

30. H. Skinner, et al., "Corrosion, Materials Characteristics and Local Tissue Reaction Associated with Osteosynthesis Devices, Implant Retrieval: Material and Biological Analysis," NBS Special Publication 601, U.S. Dept. of Commerce, Jan 1981, p 425

33. M.A. Streicher, *Intergranular Corrosion of Stainless Alloys*, STP 656, R.F. Steigerwald, Ed., American Society for Testing and Materials, 1978, p 3–84

Evaluation of Intergranular Corrosion

The metallurgical changes that lead to intergranular corrosion are not always observable in the microstructure; therefore, corrosion tests may sometimes be the most sensitive indication of metallurgical changes. Corrosion tests for evaluating the susceptibility of an alloy to intergranular attack are typically classified as either simulated-service or accelerated tests. Simulated-service testing is a valuable predictor of corrosion behavior; however, accelerated tests may also be needed to evaluate the long-term susceptibility of alloys.

Over the years, specific tests have been developed and standardized for evaluating the susceptibility of various alloys to intergranular attack. Various aspects of the phenomenon of sensitization, the mechanism of intergranular corrosion, tests for sensitization, and recent advances in such testing are reviewed in Ref 34 for austenitic stainless steels. The standard practice for detecting susceptibility of austenitic stainless steels is ASTM A 262 (Ref 35), which describes five practices that are either destructive to the sample or qualitative in nature. With modification of evaluation criteria, the practices in ASTM A 262 are also used to evaluate susceptibility of ferritic stainless steels, duplex stainless steels, and nickel-base alloys to intergranular corrosion.

Electrochemical potentiokinetic reactivation (EPR) technique is described in ASTM G 108 (Ref 36). This test method overcomes the limitations of practices in ASTM A 262 by providing a nondestructive means of quantifying the degree of sensitization in austenitic stainless steels (Ref 37). This test method also has found wide acceptance in quantitative nondestructive evaluation of sensitization on intergranular corrosion and

intergranular stress-corrosion cracking (IGSCC) of ferritic stainless steels, duplex stainless steels, and nickel-base alloys.

The EPR technique detects the extent of the chromium-depleted region in stainless steels by passivating the metal and then reactivating it through a decreasing potential sweep (Ref 38). This nondestructive test provides a quantitative indication of the degree of sensitization. The EPR measurements are done by either single-loop (Ref 39) or double-loop (Ref 40, 41) procedures. Post-EPR replica metallography is also an effective means to confirm EPR results (Ref 42). The EPR test is also considered for on-site measurement with portable equipment (Ref 43).

Intergranular corrosion and SCC may occur under similar conditions for some alloys. An example of this involves the strain-hardened 5*xxx* series of aluminum-magnesium alloys. They are susceptible to intergranular attack caused by continuous grain-boundary precipitation of highly anodic Mg_2Al_3 compounds during certain manufacturing conditions, or after being subjected to elevated temperatures up to about 175 °C (350 °F). There is a parallel relationship between the susceptibility to intergranular attack (including exfoliation) and SCC for these alloys.

However, the susceptibility to intergranular corrosion does not necessarily equate with a susceptibility to IGSCC. In the presence of residual or applied stress, failure by SCC can occur before substantial intergranular attack has occurred. Conversely, some nickel-base alloys are actually more resistant to SCC when they are sensitized to intergranular corrosion (Ref 44).

References cited in this section

34. V. Kain, R.C. Prasad, and P.K. De, Detection of Sensitization and Intergranular Corrosion of Fe-Cr-Ni Alloys, *High Temp. Mater. Proc.,* Vol 16 (No. 3), July–Sept 1997, p 183–199

35. "Standard Practices for Detecting Susceptibility to Intergranular Attack in Austenitic Stainless Steels," A 262, *Annual Book of ASTM Standards,* Vol 01.03, American Society for Testing and Materials

36. "Test Method Standard Test Method for Electrochemical Reactivation (EPR) for Detecting Sensitization of AISI Type 304 and 304L Stainless Steels," G 108-94, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials, 1999

37. A.P. Majidi and M.A. Streicher, Four Non-Destructive Electrochemical Test for Detecting Sensitization in Type 304 and 304L Stainless Steels, paper 62, *Corrosion 85,* National Association of Corrosion Engineers, 1985, p 1–17; *Nucl. Technol.,* Vol 75 (No. 3), Dec 1986, p 356–369

38. A.P. Majidi and M.A. Streicher, Potentiodynamic Reactivation Method for Detecting Sensitization in AISI 304 Stainless Steels, *Corrosion,* Vol 40 (No. 1), Jan 1984, p 21–32

39. W.L. Clarke, V.M. Romero, and J.C. Danko, Detection of Sensitization in Stainless Steel Using Electrochemical Techniques, Paper No. 180 (Pamphlet), *Corrosion/77,* International Corrosion Forum, National Association of Corrosion Engineers, 1977, 32 pages

40. M. Akashi et al., Evaluation of IGSCC Susceptibility of Austenitic Stainless Steels Using Electrochemical Reactivation Method, *Corrosion Eng.,* Vol 29, 1980, p 163

41. A.P. Majidi and M.A. Streicher, The Double Loop Reactivation Method for Detecting Sensitization in AISI 304 Stainless Steels, *Corrosion,* Vol 40 (No. 11), Nov 1984, p 584–593

42. K.J. Stoner, An Evaluation of Electrochemical Potentiokinetic Reactivation Techniques for In-Service Measurements on Type 304 Stainless Steel, *Computer-Aided Microscopy and Metallography,* ASM International, 1990, p 171–182

43. M. Verneau, J. Charles, and F. Dupoiron, Applications of Electrochemical Potentiokinetic Reactivation Test to On-Site Measurements on Stainless Steels, *Application of Accelerated Corrosion Tests to*

Service Life Prediction of Materials, STP 1194, American Society for Testing and Materials, 1994, p 367–381

44. D. Van Rooyen, *Corrosion,* Vol 31, 1975, p 327–337

Intergranular Corrosion of Stainless Steels

Although stainless steels provide resistance against general corrosion and pitting, the 300- and 400-series of stainless steels may be susceptible to intergranular corrosion by sensitization. It is caused by the precipitation of chromium carbides and/or nitrides at grain boundaries during exposure to temperatures from 450 to 870 °C (840 to 1600 °F), with the maximum effect occurring near 675 °C (1250 °F). The resulting depletion in chromium adjacent to the chromium-rich carbides/nitrides provides a selective path for intergranular corrosion by specific media, such as hot oxidizing (nitric, chromic) and hot organic (acetic, formic) acids. A partial list of environments known to cause intergranular corrosion in sensitized austenitic stainless steel is given in Table 2 (Ref 45).

Table 2 Partial listing of environments known to cause intergranular corrosion in sensitized austenitic stainless steels

Environment	**Concentration, %**	**Temperature, °C (°F)**
Nitric acid	1	Boiling
	10	Boiling
	30	Boiling
	65	60 (140) to boiling
	90	Room to boiling
	98	Room to boiling
Lactic acid	50–85	Boiling
Sulfuric acid	30	Room
	95	Room
Acetic acid	99.5	Boiling
Formic acid	90	Boiling
	10 (plus Fe^{3+})	Boiling
Chromic acid	10	Boiling
Oxalic acid	10 (plus Fe^{3+})	Boiling
Phosphoric acid	60–85	Boiling
Hydrofluoric acid	2 (plus Fe^{3+})	77 (170)
Ferric chloride	5	Boiling
	25	Room
Acetic acid anhydride mixture	Unknown	100–110 (212–230)
Maleic anhydride	Unknown	60 (140)
Cornstarch slurry, pH 1.5	Unknown	49 (120)
Seawater	…	Room
Sugar liquor, pH 7	66–67	75 (167)
Phthalic anhydride (crude)	Unknown	232 (450)

Source: Ref 45

The extent of the sensitization effect is a function of time, temperature, and the alloy composition. The latest varieties of intergranular-corrosion-resistant stainless steels are described in Ref 46 along with general remedies for intergranular corrosion in various stainless steels and recently introduced procedures for intergranular corrosion testing procedures such as the copper-copper sulfate-sulfuric acid test and the EPR test. Economical evaluation of stainless steels for their resistance to intergranular corrosion is reviewed in Ref 47.

The typical appearance of intergranular corrosion of stainless steels is shown in Fig. 32 for an attack on sensitized type 304 stainless steel at 80 °C (180 °F) in water containing a low concentration of fluorides in solution. Intergranular corrosion of the type shown in Fig. 32 is more or less randomly oriented and does not have highly localized propagation, as does intergranular stress corrosion, in which cracking progresses in a direction normal to applied or residual stresses. For austenitic stainless steels, the susceptibility to intergranular corrosion is mitigated by solution heat treating at 1065 to 1120 °C (1950 to 2050 °F) and with water quenching. By this procedure, chromium carbides are dissolved and retained in solid solution. However, solution heat treatment may be difficult on many welded assemblies and is generally impracticable on large equipment or when making repairs.

Fig. 32 Sensitized 304 stainless steel exhibiting intergranular attack. 100×

Susceptible stainless steels are those that have normal carbon contents (generally >0.04%) and do not contain carbide-stabilizing elements (titanium and niobium). Examples are AISI types 302, 304, 309, 310, 316, 317, 430, and 446. Using stainless steels that contain less than 0.03% C reduces susceptibility to intergranular corrosion sufficiently for serviceability in many applications. Somewhat better performance can be obtained from stabilized types, such as types 347 or 321 stainless steel, which contain sufficient titanium and niobium (or niobium plus tantalum), respectively, to combine with all of the carbon in the steel.

Under conditions, such as multipass welding or cross welding, stabilized alloys will sensitize. They are also susceptible to a highly localized form of intergranular corrosion known as knifeline attack, which occurs in base metal at the weld fusion line. In some cases, these alloys are given stabilizing heat treatments after solution heat treatment for maximum resistance to intergranular corrosion in the as-welded condition. For example, type 321 stainless steel is stabilize-annealed at 900 °C (1650 °F) for 2 h before fabrication to avoid sensitization and knifeline attack. So treated, type 321 may still be susceptible because titanium has a tendency to form an oxide during welding: therefore, its role as a carbide stabilizer may be diminished. For this reason, type 321 is always welded with a niobium-stabilized weld filler metal, such as type 347 stainless.

In many applications involving exposure to environments that cause intergranular corrosion, low-carbon/nitrogen or stabilized alloy grades are specified. Types 304L, 316L, and 317L, with carbon limited to 0.03% (maximum) and nitrogen contents limited to 0.10% (maximum) are used instead of their higher-carbon

counterparts. These low carbon levels are readily and economically achieved with the advent of the argon-oxygen decarburization (AOD) refining process used by most alloy producers. By limiting the interstitial element content, sensitization is limited or avoided entirely during subsequent welding and other reheating operations. However, lowering the carbon/nitrogen content also lowers the maximum allowable design stresses, as noted in appropriate sections of applicable fabrication codes, such as ASME Section 8, Division 1 for unfired pressure vessels and ANSI B31.3 for process piping.

Sensitization and intergranular corrosion also occur in ferritic stainless steels. A wider range of corrosive environments can produce intergranular attack in ferritic grades than is the case for austenitic grades. The thermal processes causing intergranular corrosion in ferritic stainless steels are also different from those for austenitic stainless steels. In the case of welds, the attacked region is usually larger for ferritic grades than for austenitic grades because temperatures above 925 °C (1700 °F) are involved in causing sensitization. Ferritic grades with less than 15% Cr are not susceptible, however (Ref 48). New grades of ferritic stainless steels are also alternatives to the more common 300- and 400-series stainless steels. Ferritic stainless steel UNS S44800 with 0.025% C (maximum) plus nitrogen is replacing the higher-carbon 446 grade in applications in which resistance to intergranular corrosion is a requirement. Other grades that are potentially lower-cost alternatives to higher alloys include:

- Ferritic alloys 26Cr-1Mo (UNS S44627) and 27Cr-3Mo-2Ni (UNS S44660)
- Duplex ferritic-austenitic alloys, 26Cr-1.5Ni-4.5Mo (UNS S32900) and 26Cr-5Ni-2Cu-3.3Mo (UNS S32550)

Duplex stainless steels are resistant to intergranular corrosion when aged in the region of 480 to 700 °C (895 to 1290 °F). It has been recognized for some time that duplex grades with 20 to 40 vol% ferrite exhibit excellent resistance to intergranular corrosion. The aging of duplex stainless steels produces a variety of phases, as discussed further in Ref 48 along with an overall review of intergranular corrosion of stainless steels.

Cast Stainless Steels. Intergranular corrosion is a leading cause of failure of pump and valve castings in the chemical process industry. For cast stainless steels, when solution treatment is impractical or impossible, the low-carbon grades CF-3 and CF-3M are commonly used to preclude sensitization incurred during welding. The low carbon content (0.03% C maximum) of these alloys prevents the formation of an extensive number of chromium carbides. In addition, these alloys normally contain 3 to 30% ferrite in an austenitic matrix. By virtue of rapid carbide-precipitation kinetics at ferrite-austenite interfaces compared to austenite/austenite interfaces, carbide precipitation is confined to ferrite-austenite boundaries in alloys containing a minimum of about 3 to 5% ferrite (Ref 49, 50). If the ferrite network is discontinuous in the austenite matrix (depending on the amount, size, and distribution of ferrite pools), extensive intergranular corrosion may not be a problem in most of the environments.

High-alloy austenitic stainless steels of the CN7M type also are frequently used because they are highly resistant to sulfuric acid and other aggressive chemicals. Specifications normally require a solution anneal after weld repair to eliminate carbide precipitation in the HAZ. The introduction of the AOD secondary refining process in many stainless foundries has allowed the production of CN7M and similar alloys with very low carbon levels. Test results also have demonstrated that the heat generated by welding is insufficient to induce sensitization in the HAZ of the weld in AOD CN7M (Ref 51).

Sensitization can be induced by exposure in a furnace at 675 °C (1250 °F), but the findings indicate that AOD-refined high-alloy, fully austenitic castings behave similarly to the lower-alloyed CF3M types. A postweld solution anneal is unnecessary for material with a carbon content <0.03% after welding with a low-carbon filler.

Evaluating Intergranular Corrosion of Stainless Steels. The austenitic and ferritic stainless steels, as well as most nickel-base alloys, are generally supplied in a solution-treated condition. However, because these alloys may become sensitized from welding, improper heat treatment, or other thermal exposure, acceptance tests are implemented as a quality-control check to evaluate stainless steels and nickel-base alloys.

As noted previously, the standardized methods of testing austenitic stainless steels for susceptibility to intergranular corrosion include immersion tests described in ASTM A 262 (Ref 35) and EPR methods described in ASTM G 108 (Ref 36). There are five acid immersion tests and one etching test described in ASTM A 262. The oxalic acid etch test is used to screen samples to determine the need for further testing. Samples that have acceptable microstructures are considered to be unsusceptible to intergranular corrosion and require no further testing. Samples with microstructures indicative of carbide precipitation must be subjected to

one of the immersion tests. Electrochemical potentiokinetic reactivation methods can provide a nondestructive means of quantifying the degree of sensitization. The methods have wide acceptance in studies of the effects of sensitization on intergranular corrosion and IGSCC of stainless steels and nickel-base alloys.

The theory and application of acceptance tests for detecting the susceptibility of stainless steels and nickel-base alloys to intergranular attack are extensively reviewed in Ref 33 and Ref 52. Some standard tests include acceptance criteria, but others do not (Ref 33). A criterion is needed that can clearly separate material susceptible to intergranular attack from that resistant to attack, as listed in Table 3 (Ref 53). The evaluation tests and acceptance criteria for various stainless steels and nickel-base alloys in Table 3 have been used by the DuPont Company, the U.S. Department of Energy, and others in the chemical-processing industry.

Table 3 Appropriate evaluation tests and acceptance criteria for wrought alloys

UNS No.	**Alloy name**	**Sensitizing treatment**	**Applicable tests (ASTM standards)**	**Exposure time, h**	Criteria for passing, appearance or maximum allowable corrosion rate, mm/month (mils/month)
S43000	Type 430	None	Ferric sulfate (A 763-X)	24	**1.14 (45)**
S44600	Type 446	None	Ferric sulfate (A 763-X)	72	**0.25 (10)**
S44625	26-1	None	Ferric sulfate (A 763-X)	120	**0.05 (2) and no significant grain dropping**
S44626	26-1S	None	Cupric sulfate (A 763-Y)	120	**No significant grain dropping**
S44700	29-4	None	Ferric sulfate (A 763-X)	120	**No significant grain dropping**
S44800	29-4-2	None	Ferric sulfate (A 763-X)	120	**No significant grain dropping**
S30400	Type 304	None	Oxalic acid (A 262-A)	…	**(a)**
			Ferric sulfate (A 262-B)	120	**0.1 (4)**
S30403	Type 304L	1 h at 675 °C (1250 °F)	Oxalic acid (A 262-A)	…	**(a)**
			Nitric acid (A 262-C)	240	**0.05 (2)**
S30908	Type 309S	None	Nitric acid (A 262-C)	240	**0.025 (1)**
S31600	Type 316	None	Oxalic acid (A 262-A)	…	**(a)**
			Ferric sulfate (A 262-B)	120	**0.1 (4)**
S31603	Type 316L	1 h at 675 °C (1250 °F)	Oxalic acid (A 262-A)	…	**(a)**
			Ferric sulfate (A 262-B)	120	**0.1 (4)**
S31700	Type 317	None	Oxalic acid (A 262-A)	…	**(a)**

			Ferric sulfate (A 262-B)	120	**0.1 (4)**
S31703	Type 317L	1 h at 675 °C (1250 °F)	Oxalic acid (A 262-A)	…	**(a)**
			Ferric sulfate (A 262-B)	120	**0.1 (4)**
S32100	Type 321	1 h at 675 °C (1250 °F)	Nitric acid (A-262-C)	240	**0.05 (2)**
S34700	Type 347	1 h at 675 °C (1250 °F)	Nitric acid (A 262-C)	240	**0.05 (2)**
N08020	20Cb-3	1 h at 675 °C (1250 °F)	Ferric sulfate (G 28-A)	120	**0.05 (2)**
N08904	904L	None	Ferric sulfate (G 28-A)	120	**0.05 (2)**
N08825	Incoloy 825	1 h at 675 °C (1250 °F)	Nitric acid (A 262-C)	240	**0.075 (3)**
N06007	Hastelloy G	None	Ferric sulfate (G 28-A)	120	**0.043 (1.7) sheet, plate, and bar; 0.05 (2) pipe and tubing**
N06985	Hastelloy G-3	None	Ferric sulfate (G 28-A)	120	**0.043 (1.7) sheet, plate, and bar; 0.05 (2) pipe and tubing**
N06625	Inconel 625	None	Ferric sulfate (G 28-A)	120	**0.075 (3)**
N06690	Inconel 690	1 h at 540 °C (1000 °F)	Nitric acid (A 262-C)	240	**0.025 (1)**
N10276	Hastelloy C-276	None	Ferric sulfate (G 28-A)	24	**1 (40)**
N06455	Hastelloy C-4	None	Ferric sulfate (G 28-A)	24	**0.43 (17)**
N06110	Allcorr	None	Ferric sulfate (G 28-B)	24	**0.64 (25)**
N10001	Hastelloy B	None	20% Hydrochloric acid	24	**0.075 (3) sheet, plate, and bar; 0.1 (4) pipe and tubing**
N10665	Hastelloy B-2	None	20% Hydrochloric acid	24	**0.05 (2) sheet, plate, and bar; 0.086 (3.4) pipe and tubing**
A95005-A95657	**Aluminum Association 5*xxx* alloys**	**None**	**Concentrated nitric acid (G 67)**	**24**	**(b)**

(a) See ASTM A 262, practice A (Ref 45).

(b) See ASTM G 67, section 4.1 (Ref 32). Source: Ref 53

It also should be note that not all instances of stainless steel intergranular corrosion are associated with sensitization. Intergranular corrosion is rare in nonsensitized ferritic and austenitic stainless steels and nickel-base alloys, but one environment known to be an exception is boiling HNO_3 containing an oxidizing ion such as dichromate (Ref 54), vanadate, and/or cupric. Intergranular corrosion has also occurred in low-carbon, stabilized, and/or property solution heat treated alloys cast in resin sand molds (Ref 55). Carbon pickup on the surface of the castings from metal-resin reactions has resulted in severe intergranular corrosion in certain environments. Susceptibility goes undetected in the evaluation tests mentioned above because test samples obtained from castings generally have the carbon-rich layers removed. This problem is avoided by casting these alloys in ceramic noncarbonaceous molds.

Finally, another unique circumstance of intergranular corrosion is the following example of improper materials selection that led to intergranular corrosion of a type 303 stainless steel valve in soda-dispensing system. This is a classic case of how materials selection may involve a trade-off between manufacturing characteristics and service performance.

Example 14: Localized Corrosion of Inclusions in a Type 303 Stainless Steel Vending- Machine Valve (Ref 56). After about two years in service, a valve in contact with a carbonated soft drink in a vending machine occasionally dispensed a discolored drink with a sulfide odor, causing complaints from customers. Manufacturing specifications called for the valve body to be made of type 303 stainless steel, a free-machining steel chosen because of the substantial amount of machining necessary to make the parts. Other machine parts in contact with the drink were made from type 304 stainless steel or inert plastics. According to the laboratory at the bottling plant, the soft drink in question was one of the most strongly acidic of the commercial soft drinks, containing citric and phosphoric acids and having a pH of 2.4 to 2.5.

Investigation. The body of the valve, through which the premixed drink was discharged to the machine outlet, had an abnormal appearance on some portions of the end surface that were continuously in contact with the liquid, even during idle periods. These regions showed dark stains and severe localized corrosive attack in the stained areas. The remaining portions of the valve surface and other metal parts that had also been in contact with the liquid appeared bright and unaffected. Examination at low magnification confirmed the presence of severe highly localized attack in the stained areas.

The valve bodies were supplied by two different vendors. Chemical analysis of drillings from a corroded valve body supplied by vendor A and from an unused valve body supplied by vendor B showed that both parts met the composition requirements for type 303 stainless steel. This alloy had been specified for the valve bodies for at least nine years before complaints about its performance had come to the attention of the valve manufacturers.

Several of the valve bodies, one used and five unused, selected at random from parts on hand supplied by the two vendors, were cleaned in acetone and then continuously immersed in the highly acidic soft-drink mixture for several days. The results were as follows:

Valve	**Vendor**	**Initial condition**	Effect of teston valve
1	A	Stained and corroded	**None**
2	B	Unused	**Black stain**[(a)]
3, 4	A	Unused	**None**
5, 6	**B**	**Unused**	**None**

(a) The test also discolored the mixture and gave it a pronounced sulfide odor, making it completely unpalatable.

The valve body that was stained and corroded (from vendor A, valve body 1 in the table), the unused valve body that had stained in the immersion test (from vendor B, valve body 2 in the table), and an unused valve body from vendor A that had not been subjected to immersion testing were all sectioned through the end that would be continuously exposed to the soft drink in the vending machine and were examined metallographically at magnifications of 50 to 400×. Numerous stringer-type inclusions of manganese sulfide were observed in each of the three metallographic specimens.

Figure 33 shows a micrograph of an unetched specimen from the corroded region on the end of the used valve body (valve body 1 in the immersion test). The corroded surface is an end-grain surface, and the corrosion began at the exposed ends of manganese sulfide stringer-type inclusions, which were anodic to the surrounding metal and readily attacked by the acidic soft drink. The attack extended along the inclusion lines to a maximum depth of about 0.64 mm (0.025 in.), with the depth apparently depending on the length of the exposed sulfide stringers. A continuous line of attack that extended at least 0.58 mm (0.023 in.) below the surface of the metal is shown in Fig. 33, which also illustrates the distribution and dimensions of the sulfide stringers generally characteristic of the three specimens that were examined metallographically. No significant attack was found on the metallographic specimen taken from unused valve body 2 from vendor B, which had turned black during laboratory immersion in the soft-drink mixture.

Fig. 33 Unetched section through a type 303 stainless steel valve exposed to an acidic soft drink in a vending machine. Micrograph shows localized corrosion along manganese sulfide stringer inclusions at the end-grain surface. 100×

Conclusions. Type 303 stainless steel has only marginal corrosion resistance for this application, because of the size and distribution of sulfide stringers found in some lots of standard grades of this alloy. The sulfide stringers are anodic to the surrounding stainless steel and are preferentially corroded away. The failure of valve body 1 occurred because manganese sulfide stringers were present in significant size and concentration and were exposed at end-grain surfaces in contact with the sufficiently acidic beverage. The exposed sulfide stringers, which were anodic to the surrounding metal, were thus subject to corrosion to produce a hydrogen sulfide concentration in the immediately adjacent liquid, thus making at least the first cup of beverage dispensed discolored and unpalatable, especially after the machine stood unused overnight or over a weekend.

The inconsistent results obtained on immersing unused valve bodies 2 to 6 from the two vendors in the beverage for several days were not surprising in view of this marginal corrosion resistance, the locally varying distribution and dimensions of the sulfide stringers, and the use of only solvent cleaning before the immersion test. The complete resistance of the stained and corroded valve body (valve body 1 in the immersion tests) is not inconsistent with its past history. Any available sulfides on the corroded surface of the valve body could have already been consumed, and its passivity restored during exposure to the air after removal from service.

Recommendations. Specification of type 304 stainless steel (which contains a maximum of 0.030% S) for these and similarly exposed metal parts was recommended to avoid possible adverse effects on sales, even if the failures were infrequent. This alloy is generally satisfactory for processing and dispensing soft drinks and has been widely used for these purposes.

Example 15: Leaking Welds in a Ferritic Stainless Steel Wastewater Vaporizer. A nozzle in a wastewater vaporizer began leaking after approximately three years of service with acetic and formic acid wastewaters at 105 °C (225 °F) and 414 kPa (60 psig).

Investigation. The shell of the vessel was weld fabricated in 1972 from 6.4 mm (0.25 in.) E-Brite stainless steel plate. The shell measured 1.5 m (58 in.) in diameter and 8.5 m (28 ft) in length. Nondestructive examination included 100% radiography, dye-penetrant inspection, and hydrostatic testing of all E-Brite welds.

An internal inspection of the vessel revealed that portions of the circumferential and longitudinal seam welds, in addition to the leaking nozzle weld, displayed intergranular corrosion. At the point of leakage, there was a small intergranular crack. Figure 34 shows a typical example of a corroded weld. A transverse cross section through this weld will characteristically display intergranular corrosion with grains dropping out (Fig. 35). It was also noted that the HAZ next to the weld fusion line also experienced intergranular corrosion a couple of grains deep as a result of sensitization (Fig. 36).

Fig. 34 Top view of a longitudinal weld in 6.4 mm (0.25 in.) E-Brite ferritic stainless steel plate showing intergranular corrosion. The weld was made with matching filler metal. About 4×

Fig. 35 Intergranular corrosion of a contaminated E-Brite ferritic stainless steel weld. Electrolytically etched with 10% oxalic acid. 200×

Fig. 36 Intergranular corrosion of the inside surface heat-affected zone of E-Brite stainless steel adjacent to the weld fusion line. Electrolytically etched with 10% oxalic acid. 100×

The evidence indicated weldment contamination; therefore, effort was directed at finding the levels of carbon, nitrogen, and oxygen in the various components present before and after welding. The averaged results were as follows:

Content	Component, ppm			
	C	**N**	**C + N**	O
E-Brite base plate	6	108	114	**57**
Corroded longitudinal weld	133	328	461	**262**
Corroded circumferential weld	34	169	203	**225**
E-Brite weld wire	3	53	56	**55**
Sound longitudinal weld	10	124	134	**188**
Sound circumferential weld	**20**	**106**	**126**	**85**

These results confirmed suspicions that failure was due to excessive amounts of nitrogen, carbon, and oxygen. To characterize the condition of the vessel further, Charpy V-notch impact tests were run on the unaffected base metal, the HAZ, and the uncorroded (sound) weld metal. These tests showed the following ductile-to-brittle transition temperatures:

Specimen	Ductile-to-brittle transition temperature	
	°C	°F
Base metal	40 ± 3	**105 ± 5**
HAZ	85 ± 3	**180 ± 5**
Weld	**5 ± 3**	**40 ± 5**

Comparison of the interstitial levels of the corroded welds, sound welds, base metal, and filler wire suggested that insufficient joint preparation (carbon pickup) and faulty gas shielding were probably the main contributing factors that caused this weld corrosion failure. Discussions with the vendor uncovered two discrepancies. First, the welder was using a large, 19 mm (0.75 in.) ID ceramic nozzle with a gas lens but was flowing only 19 L/min (40 ft^3/h) of argon; this was the flow rate previously used with a 13 mm (0.5 in.) ID gas lens nozzle. Second, a manifold system was used to distribute pure argon welding gas from a large liquid argon tank to various satellite welding stations in the welding shop. The exact cause for the carbon pickup was not determined.

Conclusions. Failure of the nozzle weld was the result of intergranular corrosion caused by the pickup of interstitial elements and subsequent precipitation of chromium carbides and nitrides. Carbon pickup was believed to have been caused by inadequate joint cleaning prior to welding. The increase in the weld nitrogen level was a direct result of inadequate argon gas shielding of the molten weld puddle. Two areas of inadequate shielding were identified:

- Improper gas flow rate for a 19 mm (0.75 in.) diam gas lens nozzle
- Contamination of the manifold gas system

In order to preserve the structural integrity and corrosion performance of the new generation of ferritic stainless steels, it is important to avoid the pickup of the interstitial elements carbon, nitrogen, oxygen, and hydrogen. In this particular case, the vendor used a flow rate intended for a smaller welding torch nozzle. The metal supplier recommended a flow rate of 23 to 28 L/min (50 to 60 ft^3/min) of argon for a 19 mm (0.75 in.) gas lens nozzle. The gas lens collect body is an important and necessary part of the torch used to weld these alloys. Failure to use a gas lens will result in a flow condition that is turbulent enough to aspirate air into the gas stream, thus contaminating the weld and destroying its mechanical and corrosion properties.

The manifold gas system also contributed to this failure. When this system is first used, it is necessary to purge the contents of the manifold of any air to avoid oxidation and contamination. When that is done, the system functions satisfactorily; however, when it is shut down overnight or for repairs, air infiltrates back in, and a source of contamination is reestablished. Manifold systems are never fully purged, and leaks are common.

The contaminated welds were removed, and the vessels were rewelded and put back into service. Some rework involved the use of covered electrodes of dissimilar composition. No problems were reported.

Recommendations. First, to ensure proper joint cleaning, solvent washing and wiping with a clean lint-free cloth should be performed immediately before welding. The filler wire should be wiped with a clean cloth just prior to welding. Also, a word of caution: solvents are generally flammable and can be toxic. Ventilation should be adequate. Cleaning should continue until cloths are free of any residues.

Second, when gas tungsten arc welding, a 19 mm (0.75 in.) diam ceramic nozzle with gas lens collect body is recommended. An argon gas flow rate of 28 L/min (60 ft^3/min) is optimum. Smaller nozzles are not recommended. Argon back gas shielding is mandatory at a slight positive pressure to avoid disrupting the flow of the welding torch.

Third, the tip of the filler wire should be kept within the torch shielding gas envelope to avoid contamination and pickup of nitrogen and oxygen (they embrittle the weld). If the tip becomes contaminated, welding should be stopped, the contaminated weld area ground out, and the tip of the filler wire that has been oxidized should be snipped off before proceeding with welding.

Fourth, a manifold gas system should not be used to supply shielding and backing gas. Individual argon gas cylinders have been found to provide optimal performance. A weld button spot test should be performed to confirm the integrity of the argon cylinder and all hose connections. In this test, the weld button sample should be absolutely bright and shiny. Any cloudiness is an indication of contamination. It is necessary to check for leaks or to replace the cylinder.

Fifth, it is important to remember that corrosion resistance is not the only criterion when evaluating these new ferritic stainless steels. Welds must also be tough and ductile.

Lastly, dissimilar weld filler metals can be successfully used. To avoid premature failure, the dissimilar combination should be corrosion tested to ensure suitability for the intended service.

References cited in this section

32. "Standard Practice for Determining the Susceptibility to Intergranular Corrosion of 5*xxx* Series Aluminum Alloys by Weight Loss after Exposure to Nitric Acid (NAWLT Test)," G 67, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

33. M.A. Streicher, *Intergranular Corrosion of Stainless Alloys,* STP 656, R.F. Steigerwald, Ed., American Society for Testing and Materials, 1978, p 3–84

35. "Standard Practices for Detecting Susceptibility to Intergranular Attack in Austenitic Stainless Steels," A 262, *Annual Book of ASTM Standards,* Vol 01.03, American Society for Testing and Materials

36. "Test Method Standard Test Method for Electrochemical Reactivation (EPR) for Detecting Sensitization of AISI Type 304 and 304L Stainless Steels," G 108-94, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials, 1999

45. G. Kobrin, Material Selection, *Corrosion,* Vol 13, *Metals Handbook,* 9th ed., 1987, p 325

46. V. Kain, Intergranular Corrosion of Stainless Steels, *Conference: Workshop on Failure Analysis, Corrosion Evaluation and Metallography,* Bhabha Atomic Research Centre, Trombay, Bombay India, 1992 p 107–128

47. V. Cihal, Sr. and V. Cihal, Jr., Economical Evaluation of Stainless Steels for Their Resistance to Intergranular Corrosion, Book No. 556, *Progress in the Understanding and Prevention of Corrosion* (Barcelona, Spain), Vol II, July 1993, The Institute of Materials, 1993, p 1537–1541

48. G.F. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* 1990, p 698–736

49. T.M. Devine, Mechanism of Intergranular Corrosion and Pitting Corrosion of Austenitic and Duplex 308 Stainless Steel, *J. Electrochem. Soc.,* Vol 126 (No. 3), 1979, p 374

50. E.E. Stansbury, C.D. Lundin, and S.J. Pawel, Sensitization Behavior of Cast Stainless Steels Subjected to Simulated Weld Repair, *Proc. 38th SFSA Technical and Operating Conference,* Steel Founders' Society of America, 1983, p 223

51. C. McCaul, Evaluation of Intergranular Corrosion Susceptibility in an As-Welded High Alloy Austenitic Stainless Steel Casting, *Brit. Corros. J.,* Vol 26 (No. 4), 1991 p 239–243

52. M. Henthorne, *Localized Corrosion—Cause of Metal Failure,* STP 516, American Society for Testing and Materials, 1972, p 66–119

53. R.A. Corbett and B.J. Saldanha, Evaluation of Intergranular Corrosion, *Corrosion*, Vol 13, *Metals Handbook*, 9th ed., 1987, p 240

54. J.S. Armijo, Intergranular Corrosion of Nonsensitized Austenitic Stainless Steels, *Corrosion*, Jan 1968, p 24–30

55. W.H. Herrnstein, J.W. Cangi, and M.G. Fontana, Effect of Carbon Pickup on the Serviceability of Stainless Steel Alloy Castings, *Mater. Perform.*, Oct 1975, p 21–27

56. Corrosion Failures, *Failure Analysis*, Vol 10, *Metals Handbook*, 8th ed., American Society for Metals, 1975, p 180

Intergranular Corrosion of Nickel Alloys

The metallurgy of high-nickel alloys is more complicated than that of the austenitic stainless steels, because carbon becomes less soluble in the matrix as the nickel content rises. As in stainless steels, low carbon content is beneficial for resistance to intergranular corrosion. Stabilized grades (with additions of titanium, niobium, or other strong carbide-forming elements) are also beneficial, although they do not produce stabilized alloys in the same sense that stainless steels are stabilized. In other words, no simple ratio of Ti,Nb to C can be given that will make a particular alloy immune to intergranular corrosion. For example, titanium as a carbide-stabilizing element is used in Incoloy alloy 825 (UNS N08825), at a minimum concentration of about five times the carbon plus nitrogen content. The higher-nickel alloys 20Cb-3, Inconel 625, and Hastelloy G contain even higher concentrations of niobium—up to a maximum of about 4% in the case of alloy 625.
Nickel alloys susceptible to intergranular corrosion sensitization include:

- Inconel alloys 600 (UNS N06600) and 601 (UNS N06601)
- Incoloy alloy 800 (UNS N08800) despite the presence of titanium
- Incoloy 800H (UNS N08810)
- Nickel 200 (UNS N02200)
- Hastelloy alloys B (UNS N10001) and C (UNS N10002)

Solid-solution nickel-base alloys, such as Inconel 600, are subject to grain-boundary carbide precipitation if held at or slowly cooled through the temperature range of 540 to 760 °C (1000 to 1400 °F). If thus sensitized, they are susceptible to intergranular corrosion in hot caustic solutions, in boiling 75% HNO_3, and in high-temperature water containing low concentrations of chlorides or other salts. The ability of the thermal treatment to eliminate chromium depletion and provide good corrosion resistance is critically dependent on the grain size and carbon content of the alloy (Ref 57). Reference 57 includes a thermal treatment index, which allows calculation of the period at 700 °C (1300 °F) necessary to complete resolution of carbon. Resistance to intergranular attack also has a close dependency on the segregated impurity concentration and the morphology and distribution of intergranular carbides (Ref 58). Evaluation of alloy 600 has also been done in examining the potential impact of grain-boundary design and control as a practical application in the general field of corrosion prevention and control (Ref 59).
Hastelloy B and C are susceptible after being heated at 500 to 705 °C (930 to 1300 °F) but are made immune to intergranular corrosion by heat treatment at 1150 to 1175 °C (2100 to 2150 °F) for Hastelloy B and at 1210 to 1240 °C (2210 to 2260 °F) for Hastelloy C, followed by rapid cooling in air or water.
Precipitation-hardenable nickel alloys corrode intergranularly in some environments. lnconel X-750 is susceptible to intergranular corrosion in hot caustic solutions, in boiling 75% HNO_3, and in high-temperature water containing low concentrations of chlorides or other salts. It can be made resistant to intergranular

corrosion in these corrodents, but not necessarily in others, by heat treating the cold-worked alloy at 900 °C (1650 °F).

Some specialty alloys have low interstitial element content plus the addition of stabilizing elements for resistance to intergranular corrosion. These alloys include the higher-nickel Hastelloy alloy G-3 (UNS N06985), which contains 0.015% C (maximum) and niobium plus tantalum up to 0.5%. Alloy G-3 has good weldability and resistance to intergranular corrosion in the welded condition.

The intergranular corrosion behavior of high-nickel alloys was reviewed in considerable detail in Ref 60 and 61. Isothermal heat treatments produced susceptibility to intergranular corrosion in all the alloys examined in these studies. Fortunately, the problems produced by the isothermal heat treatment are not usually manifested in welding. When welded by qualified procedures, most modern high-nickel alloys are resistant to intergranular corrosion.

Commercially pure Nickel 200 with a maximum carbon content of 0.15% will precipitate elemental carbon or graphite in the grain boundaries when heated in the range of 315 to 760 °C (600 to 1400 °F). This results in embrittlement and susceptibility to intergranular corrosion. Where embrittlement and intergranular corrosion must be avoided, Nickel 201 (UNS N02201) with a maximum carbon content of 0.02% is specified.

References cited in this section

57. D.R. Johns and F.R. Beckitt, Factors Influencing the Thermal Stabilisation of Alloy 600 Tubing Against Intergranular Corrosion, *Corros. Sci.,* Vol 30 (No. 2–3), 1990, p 223–237

58. C.M. Younes, F.H. Morrissey, G.C. Allen, and P. McIntyre, Effect of Heat Treatment on the Grain Boundary Chemistry and the Resistance to Intergranular Corrosion of Alloy 600 and Alloy 690, *Second International Conference on Corrosion-Deformation Interactions,* CDI '96, Institute of Materials, 1997, p 421–434

59. P. Lin, G. Palumbo, U. Erb, and K.T. Aust, Influence of Grain Boundary Character Distribution on Sensitization and Intergranular Corrosion of Alloy 600, *Scr. Metall. Mater.,* Vol 33 (No. 9), 1 Nov 1995, p 1387–1392

60. M.H. Brown, *Corrosion,* Vol 25, 1969, p 438

61. M.H. Brown and R.W. Kirchner, *Corrosion,* Vol 29, 1973, p 470–474

Intergranular Corrosion of Aluminum Alloys

Various types of intergranular corrosion are shown in Fig. 37. The precipitated phases in high-strength aluminum alloys make them susceptible to intergranular corrosion. The effect is most pronounced for alloys such as 2014 containing precipitated $CuAl_2$ and somewhat less for those containing $FeAl_3$ (1100), Mg_2Si (2024), $MgZn_2$ (7075), and $MnAl_6$ (5*xxx*) along grain boundaries or slip lines. Solution heat treatment makes these alloys almost immune to intergranular corrosion, but substantially reduces their strength. Some magnesium alloys are similarly attacked unless solution heat treated.

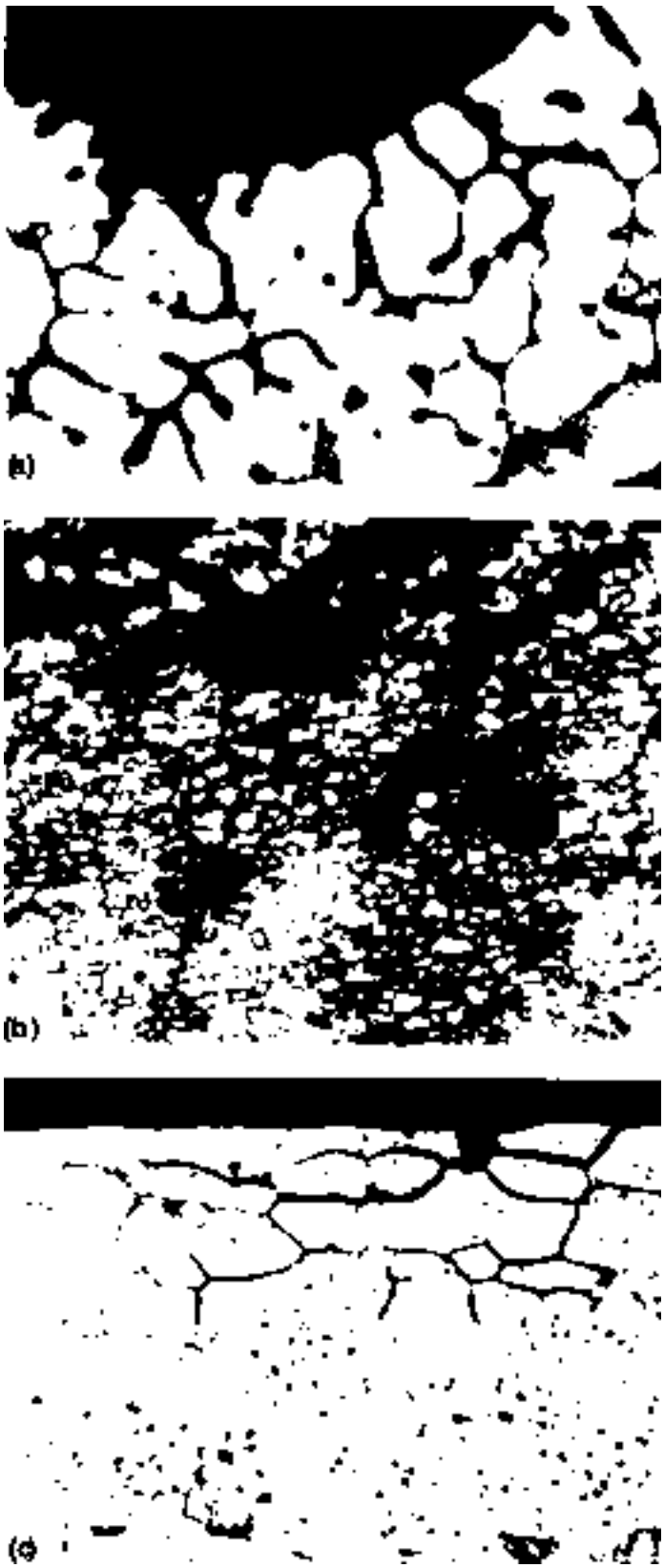


Fig. 37 Various types of intergranular corrosion. (a) Interdendritic corrosion in a cast structure. (b) Interfragmentary corrosion in a wrought, unrecrystallized structure. (c) Intergranular corrosion in a recrystallized wrought structure. All etched with Keller's reagent. 500×

Although many types of intergranular corrosion are not associated with a potential differences between the grain-boundary region and the adjacent grain bodies (as previously noted in discussions with Fig. 27), intergranular corrosion of aluminum alloys may occur from potential differences between the grain-boundary

region and the adjacent grain bodies (Ref 62). The location of the anodic path varies with the different alloy systems.

In 2*xxx* series alloys, it is a narrow band on either side of the boundary that is depleted in copper; in 5*xxx* series alloys, it is the anodic constituent Mg_2Al_3 when that constituent forms a continuous path along a grain boundary; in copper-free 7*xxx* series alloys, it is generally considered to be the anodic zinc- and magnesium-bearing constituents on the grain boundary; and in the copper-bearing 7*xxx* series alloys, it appears to be the copper-depleted bands along the grain boundaries (Ref 63, 64). The 6*xxx* series alloys generally resist this type of corrosion, although slight intergranular attack has been observed in aggressive environments. The electrochemical mechanism for intergranular corrosion proposed by E.H. Dix has been verified (Ref 65) and related to the pitting potentials of aluminum (Ref 66).

Evaluation of intergranular attack is more complex than evaluation of pitting. Visual observations are generally not reliable. For 5*xxx* series alloys, a weight-loss test method is ASTM G 67 (Ref 32). Electrochemical techniques provide some evidence of the susceptibility of a particular alloy or microstructure to intergranular corrosion, but such techniques should be accompanied by a metallographic examination of carefully prepared sections.

References cited in this section

32. "Standard Practice for Determining the Susceptibility to Intergranular Corrosion of 5*xxx* Series Aluminum Alloys by Weight Loss after Exposure to Nitric Acid (NAWLT Test)," G 67, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

62. E.H. Dix, Acceleration of the Rate of Corrosion by High Constant Stresses, *Trans. AIME,* Vol 137, 1940, p 11

63. W.L. Fink and L.A. Willey, Quenching of 75S Aluminum Alloy, *Met. Technol.,* Vol 14 (No. 8), 1947, p 5

64. M.S. Hunter, G.R. Frank, and D.L., Robinson, *Fundamental Aspects of Stress-Corrosion Cracking* (Proc. Conf.), R.W. Staehle, Ed., National Association of Corrosion Engineers, 1969, p 497

65. H. Kaesche, Pitting Corrosion of Aluminum and Intergranular Corrosion of Aluminum Alloys, *Localized Corrosion,* B.F. Brown, J. Kruger, and R.W. Staehle, Ed., National Association of Corrosion Engineers, 1974, p 516

66. J.R. Galvele and S.M. de Micheli, Mechanism of Intergranular Corrosion of Al-Cu Alloys, *Corros. Sci.,* Vol 10, 1970, p 795

Intergranular Corrosion of Copper Alloys

Intergranular corrosion is an infrequently encountered form of attack that occurs most often in applications involving high-pressure steam. This type of corrosion penetrates the metal along grain boundaries, often to a depth of several grains, which distinguishes it from surface roughening. Mechanical stress is apparently not a factor in intergranular corrosion. The alloys that appear to be the most susceptible to this form of attack are Muntz metal, admiralty metal, aluminum brasses, and silicon bronzes.

Copper alloy C26000 (cartridge brass, 70% Cu) corrodes intergranularly in dilute aqueous solutions of sulfuric acid, iron sulfate, bismuth trichloride, and other electrolytes. Figure 38 shows the intergranular corrosion attack on an inhibited admiralty brass caused by hot water containing 0.1 to 0.2% H_2SO_4.

Fig. 38 Intergranular attack of Admiralty B brass in a hot water containing a small amount of sulfuric acid. 150×

Intergranular Corrosion of Zinc

Zinc Anodes. The zinc industry has long been aware of the susceptibility of zinc alloys containing aluminum to intergranular corrosion when exposed to elevated temperatures. When zinc alloy anodes are exposed to service conditions involving elevated temperatures, the use of ASTM type II in ASTM B 41 is recommended.

Intergranular Corrosion of Galvanized Steel. It has been known since 1923 that zinc die casting alloys are susceptible to intergranular attack in an air-water environment (Ref 67). The adverse effect of intergranular corrosion of hot-dip galvanized steel was first observed in 1963 and was investigated at Inland Steel Company in 1972 (Ref 68). The observed effect associated with intergranular corrosion was termed delayed adhesion failure. Delayed adhesion failure is a deterioration in coating adhesion due to selective corrosion at grain boundaries. It was found that the small amount of lead normally added to commercial galvanizing spelters was a critical factor in the susceptibility of the zinc coating to intergranular attack. By using lead-free zinc spelter (<0.01% Pb), the damaging effect of intergranular corrosion was essentially eliminated.

For the continuous hot-dip galvanizing process, the main reason to add 0.07 to 0.15% Pb to zinc spelter was to produce a spangled coating and to lower the surface tension of the zinc bath in order to provide the necessary fluid properties to produce a ripple-free coating. It was found that antimony additions to the zinc spelter achieved the beneficial effects of lead additions without causing intergranular corrosion. For galvanized coatings produced by electroplating process, no intergranular corrosion has been observed.

One study (Ref 69) describes the influence of aging at 100 °C on the development of intergranular cracking in a galvanized coating (from a melt containing Zn-0.2Al-0.035Pb (wt%). Most other work indicates that high humidity is necessary for intergranular attack, but this work indicated that significant development of intergranular cracking could occur in the absence of a high-humidity environment. They attributed the increased cracking to diffusion and accumulation of oxygen at the grain boundaries. This study examines the development of intergranular cracking in galvanized samples exposed to environments with different levels of humidity.

References cited in this section

67. H.F. Brauer and W.M. Peirce, The Effect of Impurities on the Oxidation and Swelling of Zinc-Aluminum Alloys, *Trans. AIME,* Vol 60, 1923, p 796

68. H.H. Lee, Galvanized Steel With Improved Resistance to Intergranular Corrosion, *Proc. Galvanized Committee,* Vol 69, 1977, p 17

69. D.J. Willis, A. De Liseo, and B. Gleeson, Intergranular Corrosion in Continuously Galvanized Steel, *Galvatech '95. The Use and Manufacture of Zinc and Zinc Alloy Coated Sheet Steel Products into the 21st Century,* Iron and Steel Society/AIME, 1995, p 145–153

Selective Leaching

S.R. Freeman, Millennium Metallurgy, Ltd.

Selective leaching is defined as the removal of one element or phase from a solid alloy by a corrosion process (Ref 70). Selective leaching is also known as dealloying, and when referring to the noble metals, it is also called parting (Ref 71). The phenomenon of selective leaching was first reported by Calvert and Johnson in 1866 (Ref 72). Since that time, there has been no general consensus regarding the exact corrosion mechanism. The removal of one element results in an altered matrix usually consisting of a porous mass. There are several different alloy families that undergo selective leaching, the best known being brass alloy corrosion due to dezincification. Table 4 illustrates the various combinations of alloys and environments in which dealloying can occur (Ref 56).

Table 4 Combinations of alloys and environments subject to selective leaching and elements removed by leaching

Alloy	**Environment**	Element removed
Brasses	Many waters, especially under stagnant conditions	**Zinc (dezincification)**
Gray iron	Soils, many waters	**Iron (graphitic corrosion)**
Aluminum bronzes	Hydrofluoric acid, acids containing chloride ions	**Aluminum**
Silicon bronzes	Not reported	**Silicon**
Copper nickels	High heat flux and low water velocity (in refinery condenser tubes)	**Nickel (denickelification)**
Monels	Hydrofluoric and other acids	**Copper in some acids, and nickel in others**

Alloys of gold or platinum with nickel copper, or silver	Nitric, chromic, and sulfuric acids	**Nickel, copper, or silver (parting)**
High-nickel alloys	Molten salts	**Chromium, iron, molybdenum, and tungsten**
Cobalt-tungsten-chromium alloys	Not reported	**Cobalt**
Medium-carbon and high-carbon steels	Oxidizing atmospheres, hydrogen at high temperatures	**Carbon (decarburization)**
Iron-chromium alloys	High-temperature oxidizing atmospheres	**Chromium, which forms a protective film**
Nickel-molybdenum alloys	**Oxygen at high temperature**	**Molybdenum**

References cited in this section

56. Corrosion Failures, *Failure Analysis,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 180

70. H. Uhlig and R.W. Revie, *Corrosion and Corrosion Control,* John Wiley & Sons, 1985, p 14, 331–334

71. M.G. Fontana and N.D. Greene, *Corrosion Engineering,* McGraw-Hill, 1978, p 67–71

72. E.D. Verink, Jr., and R.H. Herdersbach, Jr., Evaluation of the Tendency for Dealloying in Metal Systems, *Localized Corrosion—Causes of Metal Failure,* STP 516, American Society for Testing and Materials, 1972, p 303–322

Dealloying Mechanisms

There are numerous theories regarding the mechanisms of selective leaching, but two predominant mechanisms prevail. The first theory states that the two elements or phases in the alloy dissolve and then one redeposits on the surface. The second theory is that one element or phase selectively dissolves from the alloy, leaving the more noble elements or phases in a porous mass (Ref 56).
The selective leaching process has specific names based on the elements that are removed from particular alloy families.

Reference cited in this section

56. Corrosion Failures, *Failure Analysis,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 180

Dezincification

The most common process of selective leaching occurs in brass alloys, and most frequently in alloys containing more than 15% Zn. This is called dezincification. Of the two dealloying mechanism, dezincification was most commonly accepted to occur by the brass corroding, the zinc ions staying in solution, and the copper replating (Ref 71).

Figure 39 illustrates the application of each of the dezincification theories to 70Cu-30Zn alloy (UNS C26000) (Ref 72). The diagram shows that higher potentials and low pH values favor selective removal of zinc. At negative potentials and acidic conditions, copper and zinc dissolve. At zero to slightly positive potentials and acidic conditions, a region exists in which copper is expected to redeposit. Thus, the research results show that both dezincification mechanisms can occur independently or in conjunction, depending on the given environment conditions. This is the widely held current belief.

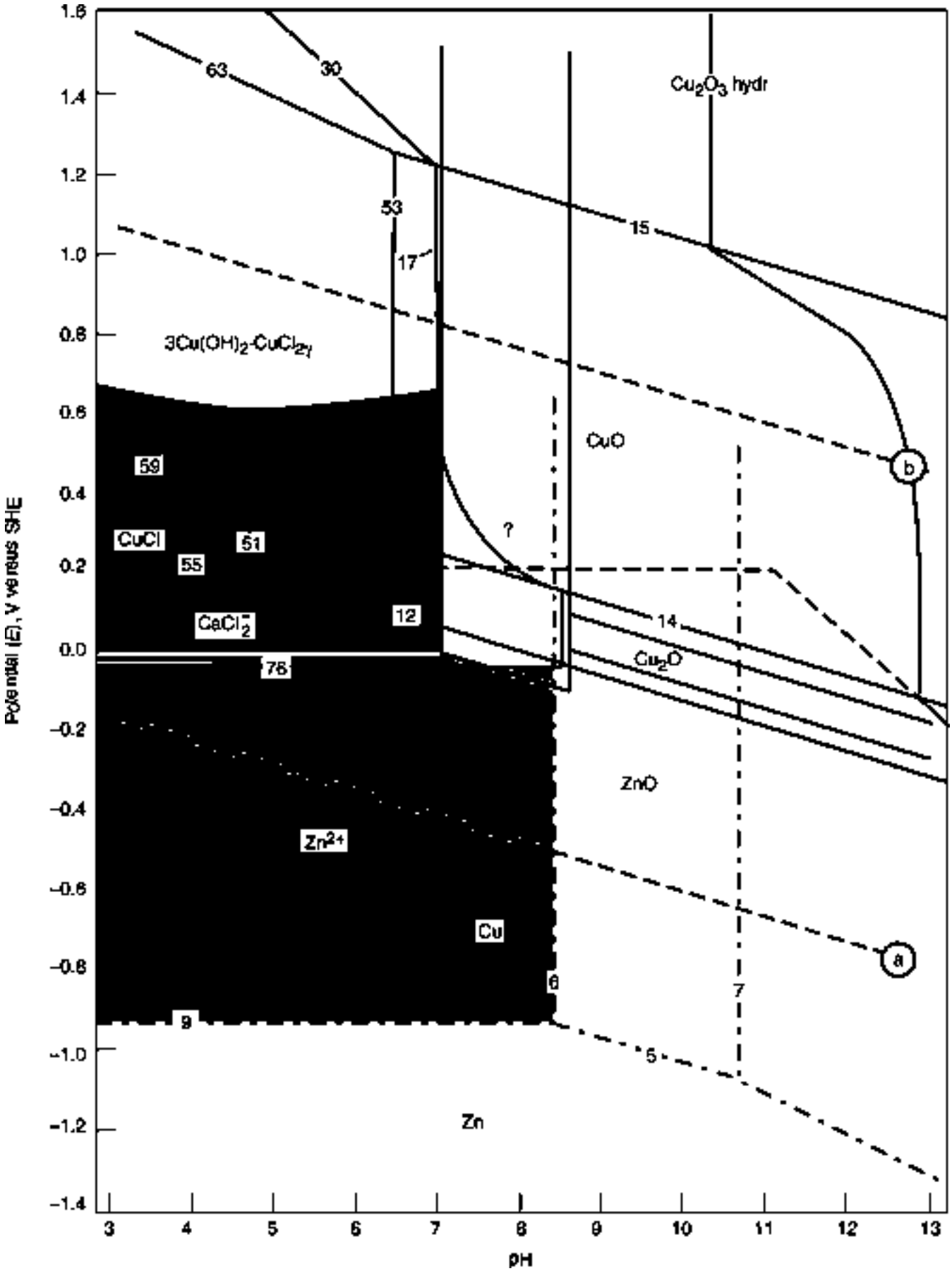


Fig. 39 Superimposed *E*-pH diagram of a 70Cu-30Zn alloy in 0.1 *M* NaCl. Lightly shaded area indicates the domain in which selective removal of zinc is expected in solutions free of copper ions. Intermediate shaded area indicates the domain in which both copper and

zinc dissolves. Dark shaded area indicates the region in which copper is expected to deposit. Source: Ref 72

Two types of damage can be characterized; one type of dezincification is uniform, and the second is plug-type. Uniform or layer-type dezincification results in a relatively uniform zone of dezincified material, with the underlying material remaining unaffected. Brasses with high zinc content in an acidic environment are prone to uniform dezincification (Ref 71). Plug-type dezincification results in localized penetrations of dezincified areas that progress through the wall thickness of the material. Overall dimensions of the material do not change. The dezincified areas are weakened or in some cases perforated. Plug-type corrosion is most likely to occur in basic or neutral environments and at elevated temperatures.

Prevention of dezincification can be achieved most readily by proper alloy selection. Alloys containing greater than 85% Cu are considered resistant to dezincification. Tin is added to act as an inhibitor. In addition inhibited copper-zinc alloys containing 0.020% to 0.6% arsenic, antimony or phosphorus are also considered resistant to dezincification. This is likely due to the formation of a redeposited film of protective elements. Other possible remedies for dezincification include the use of cathodic protection, liners, or coatings.

Example 16: Dezincification of a Chromium-Plated Copper Alloy Tube. A 12.7 mm (0.5 in.) diam tube was removed from a potable water supply due to leaks. The tube wall thickness was 0.711 mm (0.028 in.) with a thin layer of chromium plate on the OD surface. The tube had been in service for approximately 33 years.

Investigation. The tube section had a through-wall perforation (Fig. 40a). After cross-sectioning the tube longitudinally, the ID surface showed intermittent green deposits (Fig. 40b). Deposit samples were removed for EDS, and cross sections were removed for metallurgical evaluations. The samples were examined initially in the unetched condition. The metallurgical examination of the tube sample showed porous material typical of plug-type dezincification initiating from the inside surface (Fig. 40c). Where the dezincification had progressed through the tube wall, the chromium plate had exfoliated from the base material and cracked. (Fig. 40d). The EDS results of the deposits removed from the inside diameter surface showed the pipe material was a leaded brass, most likely a Muntz metal, 60% Cu and 40% Zn (now UNS C28000). The deposits showed varying levels of zinc, indicative of the dezincification corrosion mechanism. The copper content remained consistent.

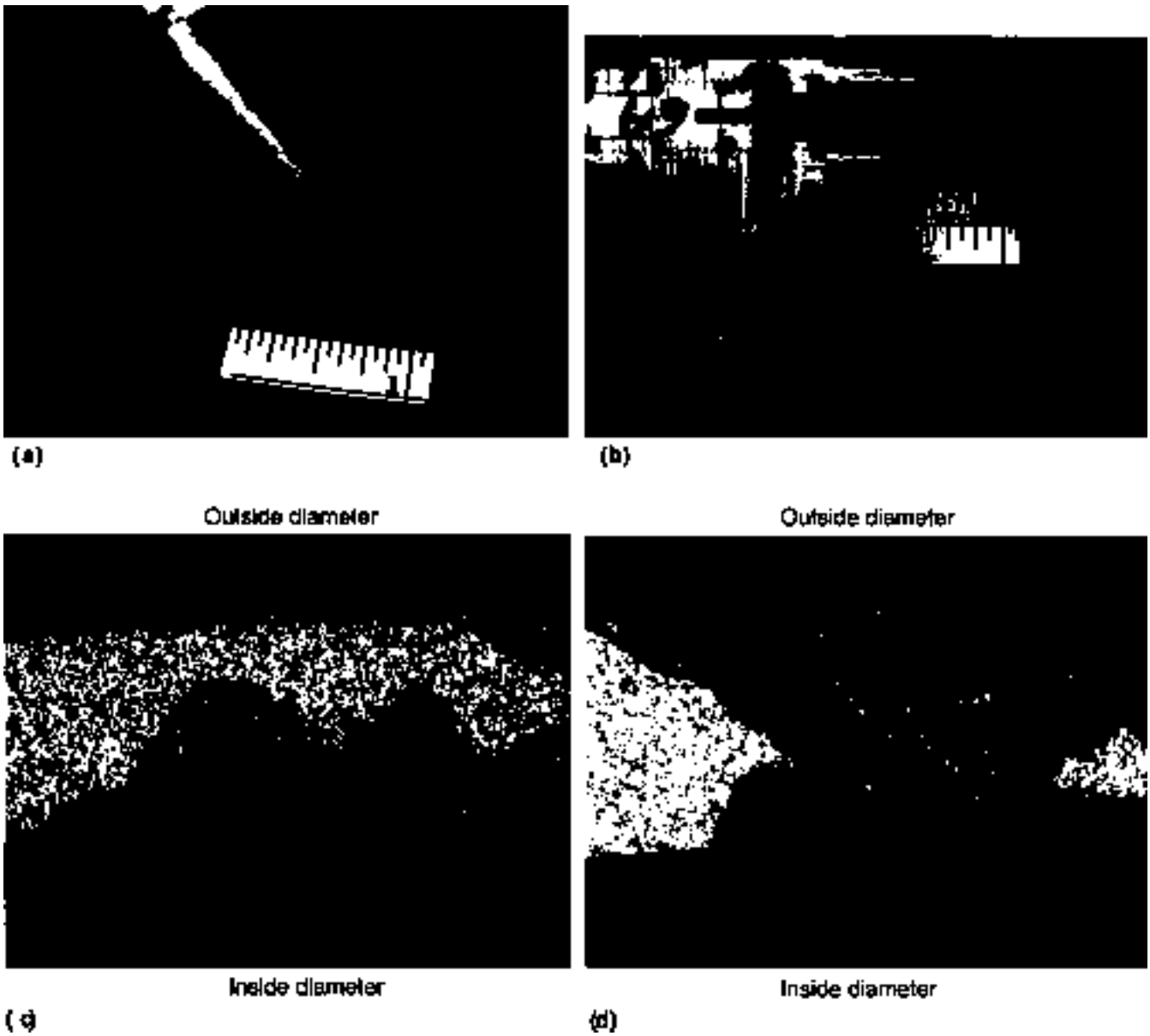


Fig. 40 Views of a through-wall perforation of a chromium-plated α brass (70Cu-30Zn) tube removed from a potable water system due to dezincification. (a) Macroview of tube. (b) Inside diameter surface of the tube shown in (a), depicting localized green deposits at the areas of dezincification. (c) Micrograph of plug-type dezincification from a cross section cut from an area of the tube shown in (b). The dezincification initiated on the inside diameter surface. Original photograph shows an area 1.52 mm (0.060 in.) across. (d) Micrograph of plug-type dezincification that penetrated the tube wall from a cross section cut from an area of the tube shown in (b). Chromium plating was lost from the outside diameter surface. Original photograph shows an area 1.52 mm (0.060 in.) across. Courtesy of S.R. Freeman, Millennium Metallurgy, Ltd.

Recommendations. The piping required replacement with a more corrosion-resistant material.

Corrosion By-Products. Copper alloys with a copper content greater than 85% are resistant to dezincification. Copper alloys such as red brass (UNS C23000), inhibited Admiralty brass (UNS C44300), and arsenical aluminum brass (UNS C68700) are less susceptible to corrosion. Suitability for use in potable water systems is the prime consideration in this case, however.

This points out that not only are the degrading effects of selective leaching on the function of the component important, but also the effects of corrosion by-products must be evaluated. Leaching of copper itself in drinking water is a matter of public health debate and regulation, beyond the scope of this article.

If the replacement must have a chromium decorative coating, ease of plating is also a factor. Here again, the high-zinc brass is susceptible to dezincification through the somewhat porous chromium plating in an aggressive exterior environment.

References cited in this section

71. M.G. Fontana and N.D. Greene, *Corrosion Engineering*, McGraw-Hill, 1978, p 67–71

72. E.D. Verink, Jr., and R.H. Herdersbach, Jr., Evaluation of the Tendency for Dealloying in Metal Systems, *Localized Corrosion—Causes of Metal Failure*, STP 516, American Society for Testing and Materials, 1972, p 303–322

Graphitic Corrosion

Graphitic corrosion is a form of dealloying that occurs in cast iron material. This corrosion mechanism is usually found in gray cast irons and is associated with the presence of graphite flakes. The graphite is cathodic to the iron matrix. Exposure to an electrolyte results in selective leaching of the iron matrix, leaving behind a porous mass of graphite flakes. Graphitic corrosion is most often a long-term mechanism, resulting from exposure of 50 years or more. Pipelines made of cast iron, especially those buried in clay-based soils and soils containing sulfates are susceptible (Ref 73).

In cases where graphitic corrosion has caused extensive wall loss, a reduction in component strength will occur. Thus, it is not unusual for cracking to accompany graphitic corrosion. In some cases, graphitic corrosion has been observed on the fracture surfaces, indicating the long-term existence of the crack. Figure 41 shows an example of a section of gray cast iron sewer pipe removed from the mud floor of a New England bay. Graphitic corrosion was apparent across approximately 25% of the total wall thickness.

Fig. 41 Graphitic corrosion of a gray cast iron sewer pipe section removed from the mud bottom of a seawater bay. Graphitic corrosion on the outside diameter surface is apparent for approximately 25% of the wall thickness. Courtesy of S.R. Freeman, Millennium Metallurgy, Ltd.

It is generally thought that ductile and malleable iron are not affected by graphitic corrosion, so ductile cast iron is routinely used for underground water and sewer pipes. However, a case of ductile cast iron subject to graphitic corrosion and microbial-influenced corrosion was reported about a buried ductile cast iron pipe (Ref 74). The porous graphite structure was accompanied by massive tubercles indicating that the biodeposits played a significant role in the formation of a graphite mass of corrosion residue. Ductile cast iron graphitic corrosion has, otherwise, gone undocumented and is considered rare.

Graphitization is not related to graphitic corrosion, but the term is sometimes confused with the graphitic corrosion mechanisms. Graphitization is the microstructural change that occurs in carbon steel materials (steels with 0.5% Mo or less) after exposure to temperatures in the range of 399 to 593 °C (750 to 1100 °F) for an extended period of time. The pearlite transforms to nodular graphite and ferrite. At grain boundaries, the formation of chains of graphite can result in brittle fracture or cracking.

Example 17: Failure of a Cast Iron Water Pipe due to Graphitic Corrosion. A 25.4 cm (10 in.) diam gray cast iron water main pipe was buried in the soil beneath a concrete slab. The installation was believed to have been completed in the early 20th century. A leak from the pipe resulted in flooding of a warehouse. Once removed, the pipe revealed through-wall perforations and cracking along its axis. The perforations and the crack were at the 6 o'clock position.

Investigation. Visual examination of the pipe showed two large holes (Fig. 42a). The areas surrounding the holes showed delaminations and corrosion deposits. Radiography was performed to evaluate the material quality. The radiographic results showed various film densities indicating corrosion or variations in wall thickness. Secondary cracking was also observed beyond the more apparent primary crack. Three cross sections were removed: one from the perforated area, one from the crack area, and one from an area considered undamaged. The samples were metallurgically prepared and examined in the unetched condition. A macrograph of a cross section through the crack depicted graphitic corrosion primarily on the OD surface (Fig. 42b). A porous mass and significant wall loss was observed on the OD surface (Fig. 42c). This is indicative of graphitic corrosion. Graphitic corrosion occurs as the ferritic phase is leached from the matrix leaving a porous mass of graphite flakes (Fig. 42d). The structure of the material was typical of gray cast iron with random orientation of the graphite flakes. Indications of graphitic corrosion were also observed to a lesser degree on the ID surface of the pipe. More than 50% of the wall thickness showed evidence of graphitic corrosion. Wall-thickness readings were fairly uniform measuring approximately 11 mm ($\frac{7}{16}$ in.). Wall loss due to graphitic corrosion is difficult to detect since a porous mass of material remains in place.

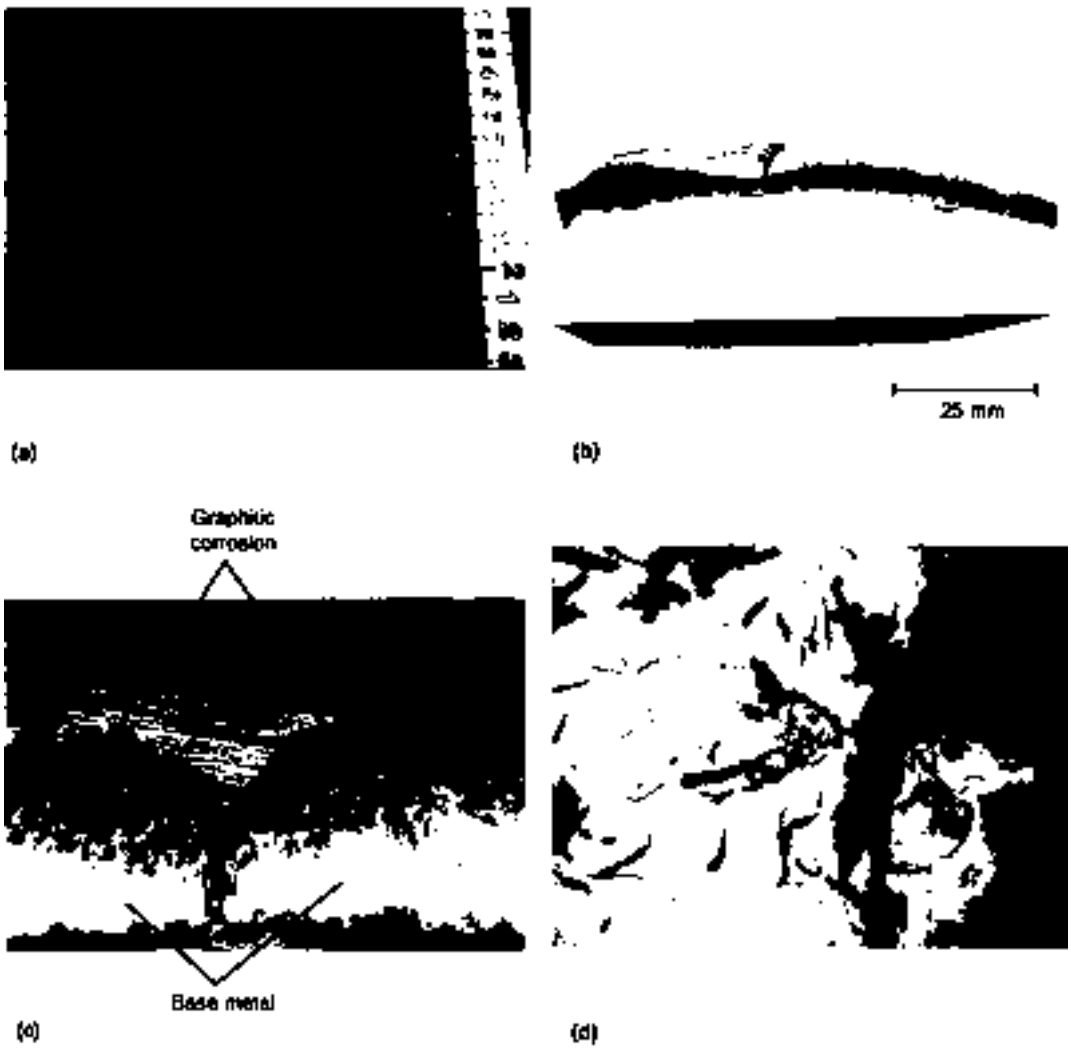


Fig. 42 A 25 cm (10 in.) diam gray cast iron pipe that failed due to graphitic corrosion. The pipe was part of a water supply to a fire protection system. The external surface was covered with soil and the inside surface in contact with water. The pipe had been experienced cracking and through-wall perforations. (a) View showing a hole and cracking in the pipe. (b) Macrograph of a cross section through the crack observed in (a) showing graphitic corrosion primarily on the outside diameter (top). (c) Micrograph of the cross section shown in (b). Significant wall loss and graphitic corrosion was observed on the outside diameter surface. Original micrograph image shows an area 16.5 mm (0.65 in.) wide. (d) Micrograph of crack surface in (c) depicting graphitic corrosion surrounding the graphite flakes. Original micrograph image shows an area 1.14 mm (0.045 in.) wide. Courtesy of S.R. Freeman, Millennium Metallurgy, Ltd.

Tensile testing was also performed. A sample was removed from the area adjacent to the crack. A second sample was removed from an area 90° from the crack. The sample near the crack showed a 40% reduction in load capacity. The pipe leak occurred as result of years of exposure to ground water in the soil resulting in graphitic corrosion. Soils containing sulfates are particularly aggressive.

Recommendations. The extensive damage due to graphitic corrosion required pipe replacement. The wall thickness had been sufficiently reduced that the pipe could no longer support the required load. Water mains are designed for more than 100 years life. Ductile iron or coated and lined steel pipe, generally not susceptible to graphitic corrosion, would be suitable replacement materials. Cathodic protection is also a possibility.

References cited in this section

73. S.R. Freeman, Graphitic Corrosion—Don't Forget about Buried Cast Iron Pipes, *Mater. Perform.*, Aug 1999, p 68–69

74. K. Kasahara and F. Kajtyama, The Role of Bacteria in the Graphitic Corrosion of Buried Ductile Cast Iron Pipes, *Microbial Corrosion*, Second EFC Workshop, March 1991

Dealuminification

Dealloying of aluminum bronze and nickel-aluminum bronze castings has been well known since the early 1960s. It has been determined over the years that the mechanism of dealloying or selective leaching in aluminum bronze castings occurs by corrosion of the duplex structures ($\alpha + \gamma$-2) (Ref 75). This type of dealloying is rarely detectable visually. Nickel-aluminum bronze castings containing 4% or more Ni show diminished selective leaching. When dealuminification does occur, the mechanism involves the dealloying of nonequilibrium microconstituents, such as retained β and γ-2. Figure 43 shows dealuminification of a cast aluminum bronze furnace electrode pressure ring exposed to recirculating cooling waters. A temper-anneal heat treatment has been shown to change the microstructure to all $\alpha + \kappa$, further reducing the susceptibility to selective leaching. However, large nickel-aluminum bronze castings may still be susceptible to dealloying along the $\alpha + \kappa$ eutectoid network. It appears that heat treatment is effective on smaller castings having fine and uniformly distributed microstructures, but larger castings have coarser and more segregated microstructures and are less responsive.

Fig. 43 Dealuminification of a cast aluminum bronze furnace electrode pressure ring exposed to recirculating cooling water (pH = 7.8 to 8.3, conductivity = 1000 to 1100 µS). The preferentially attacked γ phase left behind a residue of copper (darkened regions in eutectoid and along grain boundaries). The α particles within the eutectoid (light gray areas) are unattacked. Etched with $FeCl_3$. Magnification: 260×. Courtesy of R.D. Port. Permission granted by Nalco Chemical Co., 1987

Reference cited in this section

75. R.J. Ferrara and T.E. Caton, Review of Dealloying of Cast Aluminum Bronze and Nickel-Aluminum Bronze Alloys in Sea Water Service, *Mater. Perform.*, Feb 1982, p 30–34

Denickelification

Denickelification involves the removal of nickel from copper nickel alloys. Only a few instances of this phenomenon have been reported. The conditions likely to induce this selective leaching process include low-flow conditions, temperatures above 100 °C (212 °F) and high local heat flux (Ref 76). Figure 44 shows the residual copper layer remaining from a UNS C71500 feedwater pressure tube that underwent denickelification.

A recent report of selective leaching of nickel in type 304L stainless steel material occurred in lithium hydride at temperatures greater than 715 to 850° C (Ref 77).

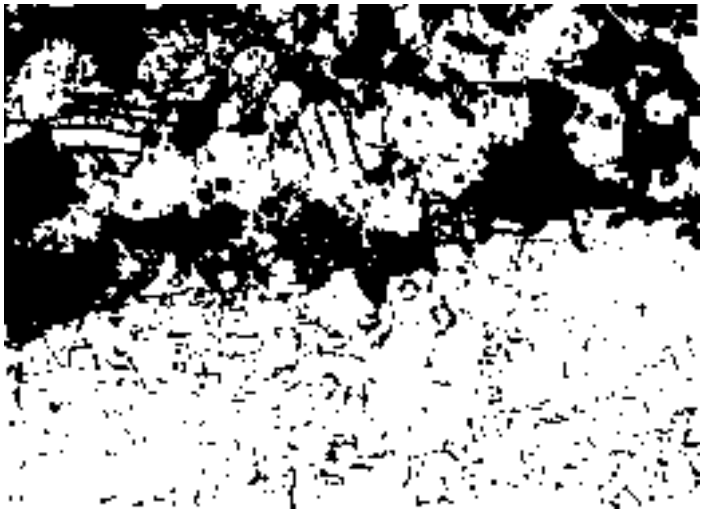


Fig. 44 Residual copper layer from a UNS C71500 feedwater pressure tube that underwent denickelification. The tube was subject to 205 °C (400 °F) steam on the external surface and boiling water on the internal surface 175 °C (350 °F), at pH 8.6 to 9.2). Courtesy of James J. Dillion. Permission granted by Nalco Chemical Co., 1987

References cited in this section

76. R. Steigerwald, Metallurgically Influenced Corrosion, *Corrosion,* Vol 13, ASM Handbook, 1987, p 130–133

77. S.J. Pavel, Type 304L Stainless Steel Corrosion in Molten Lithium Hydride as a Function of Temperature, *J. Nucl. Mater.,* Vol 200 (No. 2), April 1993, p 184–199

Destannification and Desiliconification

Specific instances of dealloying tin in cast tin bronzes have been observed in hot brine or steam. Isolated cases of desiliconification in silicon bronzes have occurred in high-temperature steam with acidic species (Ref 76).

Reference cited in this section

76. R. Steigerwald, Metallurgically Influenced Corrosion, *Corrosion,* Vol 13, ASM Handbook, 1987, p 130–133

Dealloying of Noble Metals

Dealloying may also occur in copper-gold and silver-gold alloys. A study of failure mechanisms in Cu_3Au single crystals indicated that dealloying or selective leaching of the copper occurs, leaving a gold sponge (Ref 78). Testing was performed in a ferric chloride solution and was done primarily to distinguish between true transgranular stress-corrosion cracking (T-SCC) and chemical embrittlement or selective leaching in Cu_3Au single crystals. Cu_3Au samples 1 to 2 mm (0.04 to 0.08 in.) thick were exposed to stress-free immersion in aqueous $FeCl_3$ for 30 days. Other than a slight discoloration, the samples did not show indications of attack. There was no obvious volume change. Weight loss corresponded with total loss of copper, however. Dealloying had occurred leaving a single-crystal gold sponge with the dimensions and crystal orientation of the original copper-gold crystal.

Evaluations in the cited work showed the gold sponge failed in a brittle manner by macrocleavage under small applied load in air. Figures 45(a) and (b) show the facet-step fracture surface. However, close examination of the fracture surface (Fig. 45c) revealed the failure occurred by ductile overload of the small gold ligaments of the sponge.

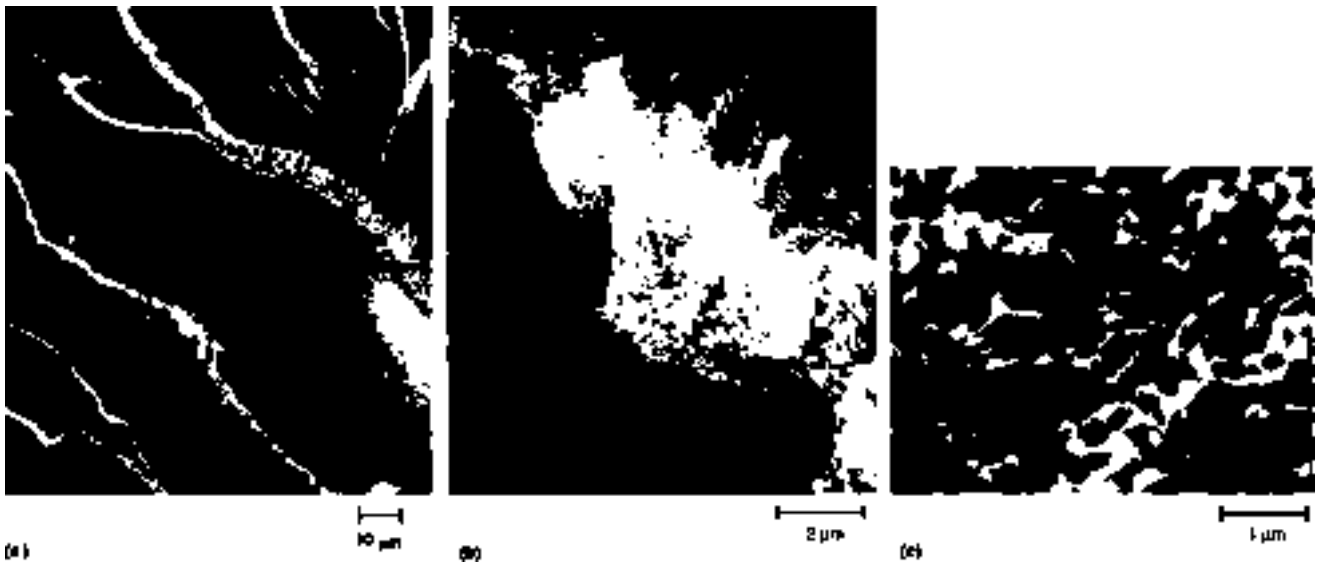


Fig. 45 Fracture surface of a Cu-25Au (at.%) single crystal upon bending in air following 30 day stress-free immersion in aqueous $FeCl_3$. (a) Scanning electron micrograph of fracture surface of gold sponge. (b) Scanning electron micrograph of the boxed area showing facet-step structure in gold sponge. (c) Scanning electron micrograph of gold sponge structure revealed on the surface of the sample shown in (b). Average interpore spacing is ~200 nm. Evidence of ductile fracture can be seen on the surface (necking of the gold ligaments).

For true (T-SCC), samples failed following 1 to 20 h exposure to the corrodent depending on the applied electrical potential. There was no evidence of that cracks had occurred in the plane of localized dealloying. Crack growth was over two orders of magnitude greater than sponge formation.

It is believed that gold sponge formation may assist in nucleating T-SCC cracks at a free surface.

Reference cited in this section

78. T.B. Cassagne, W.F. Flanagan, and B.D. Lichter, On the Failure Mechanism of Chemically Embrittled Cu_3Au Single Crystals, *Metall. Trans.*, April 1986, p 703–710

Velocity-Affected Corrosion

Revised by Colin Sandwith, University of Washington, and Randy K. Kent, MDE Engineers, Inc.

The attack on metal immersed in water may vary greatly, depending on the relative velocity between the water and the metal surface. For metals that show passive behavior or form other protective films in water, as most metals do, attack will occur where the changes in water velocity are most pronounced. Water corrosivity can be dramatically increased by dissolved gases, acids, salts, strong bases, entrained abrasives, high temperature, fluctuating pressure, cavitation, or impingement.
Relative difference in velocity between the metal and the aqueous corrosive influence any of the common varieties of iron or steel, including low-carbon or high-carbon steel, low-alloy steel, wrought iron, and cast iron. These corrode in slow-moving freshwater or seawater at almost the same rate, which is about 0.13 mm/year (0.005 in./year). At higher temperatures with equal values of dissolved gas concentration, the rate tends to increase, but remains relatively low. Therefore, steel can be used for boilers in contact with deaerated water. Commercially pure aluminum typically corrodes less in aerated or deaerated freshwater than iron and carbon steels, making it a suitable material for handling distilled water.

Forms of Corrosion

Low-Velocity Effects

Slow-moving and stagnant waters can prevent, damage, or remove passive films. The low velocity allows loosely adherent solid corrosion products to form on metal surfaces and allows debris to collect, which facilitates further corrosion damage. In closed systems, if a corrosion inhibitor is used, its effectiveness is often reduced where the water is stagnant or quiet.
In designing for corrosion control, stagnant zones should be eliminated by the following methods:

- Circulation of stagnant liquids or relative movement of metallic surfaces
- Allowing free drainage of water
- Filtering suspended solids
- Providing an N_2 blanket (carbon steels) or free access to O_2 (stainless steels). Weakly passivating metals such as carbon steels are attacked by high O_2, whereas strongly passivating metals such as stainless steels are protected by uniform O_2 distributions.
- Maintaining concentration of dissolved passivating chemicals such as O_2 by infusion or injection
- Installing baffling to eliminate stagnant liquid zones
- Providing gas-vent lines
- Maintaining clean and smooth surfaces, avoiding crevices
- Using cathodic protection
- Using proper inhibitors where applicable

Example 18: Failure of Stainless Steel Piping in Stagnant Seawater. Stagnant seawater can be quite destructive to some alloy systems. An austenitic stainless steel piping used in the fire-sprinkler system in a large saltwater passenger and car ferry failed by rapid leaking. The original author had failure analysis experience with a similar, seawater-filled fire main piping system in a high-performance hydrofoil ship, two decades earlier.
Investigation. The following design and parameters were revealed in investigation of the failure:

- *Piping Material:* Type 316/316L stainless steel, schedule 40, 64 mm (2.5 in.) diam and larger pipe was used.
- *Pitting Damage:* Observed through pipe wall directed from interior to exterior. Most pits were leaking in the HAZ at the weld joints. Some pitting was found midlength at the bottom of the pipe.
- *Environment:* Stagnant seawater at ambient temperatures. Pipe was in service for four weeks when three leaks appeared.

Figure 46 shows a pit near the end of a ballpoint pen lying in the trough of a sectioned open AISI 316L pipe. A close-up of the same pit is shown Fig. 47. The stagnant (quiet, zero flow rate) seawater created a hole approximately 1.5 mm ($\frac{1}{16}$ in.) diam through the pipe wall in about four weeks of exposure to the stagnant and pressurized seawater inside the pipe. The location of the pit is at the bottom of the pipe in the center of a long straight horizontal pipe run. The pipe was butt welded by electric arc processes at some joints and butt welded to flanged fittings at other joints. Some air may have been trapped in the top of this and other pipe runs.

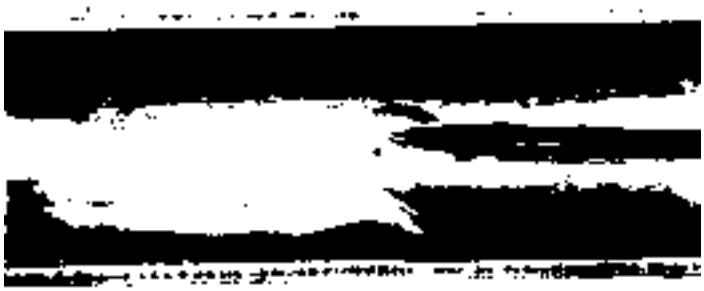


Fig. 46 Photograph of the inside of a stainless steel pipe with corrosion pits. The pipe was from a fire-sprinkler system for a car and passenger ferry boat.

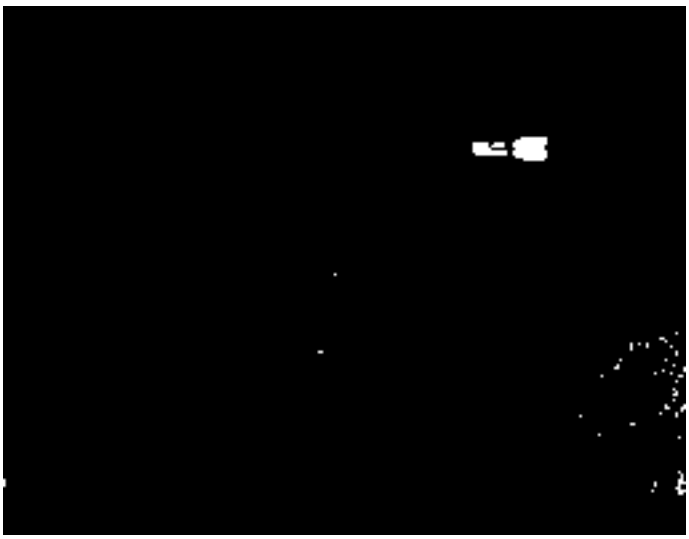


Fig. 47 Close-up of the corrosion pit in the center of Fig. 46

Discussion. These conditions were ideal for three modes of highly accelerated pitting of austenitic stainless steel. The three modes can be distinguished by how and where the pit starts and how the hydrochloric acid (HCl) collects. The morphology or surface conditions, contours, waviness, and relative surface area of each mode strongly influences pit shape and depth as well as route and growth rate of corroded path.

The first mode is shown in Fig. 46 and 47, where it is characterized by pitting in the bottom of the long run of pipe. Cl ions, which are plentiful in seawater, attack and damage the passive film and activate the surface of the metal at that point. As the corrosion starts at the tiny active sites, hydrolysis and precipitates of the less-soluble Fe^{2+} hydroxides produce increased concentrations of hydrochloric acid by the following reaction steps:

$$Fe \rightarrow Fe^{2+} + 2e^-$$

$$Fe^{2+} + 2e^- + 3H^+ + 2OH^- + Cl^- \rightarrow Fe(OH)_2 \downarrow + H_2 \uparrow + H^+ + Cl^-$$

or simply,

$$Fe^{2+} + 2Cl^- + 2H_2O \rightarrow Fe(OH)_2 \downarrow + 2HCl$$

The metal hydroxide precipitates, H_2 escapes, and the HCl (acid) product remains in solution as shown by the right side of the equation. The acid is concentrated with pH lowering to near 2 by gravity in puddles or pits, which further accelerates the corrosion at the site.

In this first mode, the bottom center of the horizontal pipe run is activated when the Cl ions pick the weakest site in this massive bottom surface area and cause that location to become activated. The morphology is typical of pipe, neither very rough nor smooth. The relative surface area of the cathode to anode is large because of the expanse of the pipe involved. Corrosion by pitting starts, the dense acid puddles by gravity over a relatively wide area, and the remainder of the pipe section acts as a cathode. The pit is saucer shaped, relatively large, and uniform, but grows rapidly due to decreasing pH and develops by autocatalytic corrosion (Ref 79).

Mode two pits develop in the weld or HAZ, even without the presence of sensitization, because of the inevitable surface micro-nonuniformities, roughness, and imperfections, as well as some chemical differences from welding products. The rough and irregular morphology of welds that have not been ground flat provides sites that are easily activated for pitting and hard to completely passivate. The pipe and irregular welded surface, especially if porous and if slag and spatter are present, comprise a relatively massive cathode with a small anodic pit. Pitting starts and acid collects in the tiny pores of the damaged weld, and the process of accelerated pitting results. The pits tend to be long, thin, and capillarylike, routed in a path though the weld or HAZ, just as mold grows through bread. In sensitized stainless steel, the route is clearly intergranular, but still circuitous. Active pit sites can be located at any position on a weld. If an acid reservoir can be established, pits will grow in the direction preferred by gravity and will eventually penetrate the weld or pipe wall. The pitting rate can be very fast. Because only a microscopic amount of metal must be corroded to produce a leak, component failure is likewise fast.

The third mode develops by crevices. Areas around mating flanges, gaskets, O-ring seals, and machined faying surfaces all contain crevices. If O-rings are used, the flanges would be machined and polished. Other gasketed flanges are also smoother than the surrounding metal. The polished surface has the smallest overall surface area and hence is easiest to passivate, polarize, and keep clean, making it the most resistant to activation. If a crevice is present, it will produce a relatively long active site. The area under rubber or in a tight joint is easily activated by Cl ion attack. The absence of sufficient oxidizers does not allow the passive film to be replenished and maintained. The dense HCl collects by gravity in the lower regions of the crevice, and the cycle of accelerated corrosion is established.

The speed at which a leak develops depends on the direction of the crevice in relation to gravitational force. In the case of a horizontal seal edge, the corrosion damage can be very long and uniform, with a good deal of metal corrosion prior to any leaking. On the other hand, if the plane of the seal edge is vertical or just off of vertical by several degrees, the corrosion can be more like pitting than crevice corrosion. Less metal must corrode to produce a leak. Crevice corrosion and pitting are similar in the mechanisms and driving forces and identical regarding the corrosion reactions. Macroscopic geometry determines the category.

These austenitic stainless steels can be active or passive in adjacent areas on the same part or pipe. Because of the relatively large surface area of the cathode and hence the low current density of the cathode, it tends to *not* polarize. Therefore, the cathode to anode potential difference remains great and thus the high corrosion rate at the anode is maintained. However, cathodic protection (CP) can help retard or prevent the onset of pitting and reduce the rate of growth of established shallow pits. The application of CP reduces corrosion by forcing the passive and active areas to exhibit a more anodic potential, i.e., protective potential. For example, the criteria for cathodically protecting steels in most waters and soils is accomplished by changing the steel's potential by about 200 mV in the anodic direction.

The electromotive voltage or galvanic potential between these different metal surfaces, can easily reach 500 mV open circuit. The passive area is protected and becomes passive by developing an invisible complex chromium oxide film. This passive oxide film must be maintained by the presence of a dissolved oxidizer such as O_2, $FeCl_2$, or Fe^{3+} ions.

Recommendations. Proper material selection for piping, flanges, and weld rods can make the greatest improvement in resistance. Even low carbon steel has better resistance to pitting in stagnant seawater piping systems. Seal design is important. The thicker the O-ring and the greater the quantity of the proper lubricant, the better for corrosion resistance. Proper filtering to prevent entrained abrasive is critical. Surface abrasion damages passivation. Corrosive agents can sometimes be countered with inhibitors. Flow rates above several feet per second (~1.5 m/s, or 5 ft/s) can reduce, but should not be expected to prevent, such pitting corrosion damage unless no locations can become active. To be effective, the flow must deliver the repairing/maintaining oxidizer and wash away any acids. Timely breakdown inspections are one way of possible early detection of damage, as shown by internal rust pits, carbuncles, and rust deposits.

Reference cited in this section

79. M.G. Fontana, *Corrosion Engineering*, 3rd ed., McGraw-Hill, 1986

High-Velocity Effects

Swift-moving water may carry dissolved metal ions away from corroding areas before the dissolved ions can be precipitated as protective layers. Gritty suspended solids in water scour metal surfaces and continually expose fresh metal to corrosive attack.

In freshwater, as water velocity increases, it is expected that corrosion of steel first increases, then decreases, and then increases again. The latter occurs because erosive action serves to break down the passive state.

The corrosion of steel by seawater increases as the water velocity increases. The effect of water velocity at moderate levels is shown in Fig. 48, which illustrates that the rate of corrosive attack is a direct function of the velocity until some critical velocity is reached, beyond which there is little further increase in corrosion. At much higher velocities, corrosion rates may be substantially higher.

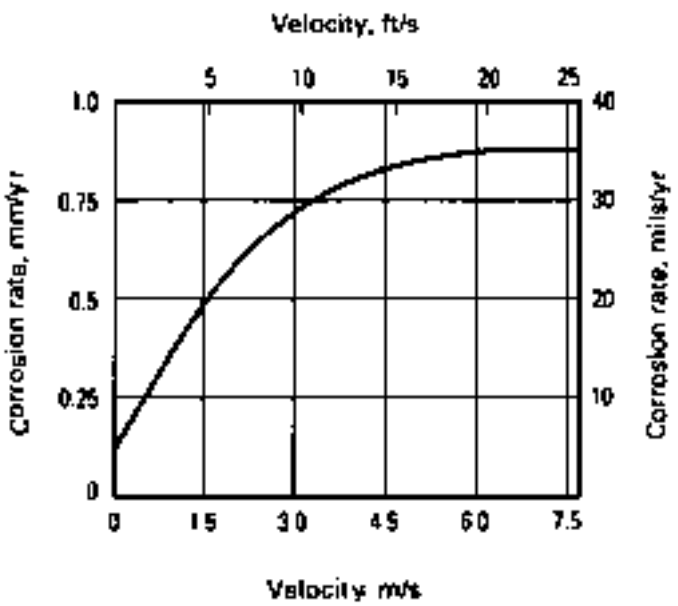


Fig. 48 Effect of velocity of seawater at atmospheric temperature on the corrosion rate of steel

The effect of changes in water velocity on the corrosion resistance of stainless steels, copper alloys, and nickel alloys shows much variation from alloy to alloy at intermediate velocities. Type 316 austenitic stainless steel may pit severely in typical seawater especially when stagnant (as indicated by the example above) and at velocities of less than 1 or 1.5 m/s (4 or 5 ft/s), but is usually very corrosion resistant at higher velocities. Aluminum brass alloy (C68700) has satisfactory corrosion resistance if the seawater velocity is less than 2.4 m/s (8 ft/s).

Copper alloy C71500 (copper nickel, 30% Ni) has excellent resistance to swift-moving seawater and to many types of freshwater. Because copper alloy C71500 is also subject to biofouling, velocities should be kept above 1.5 m/s (5 ft/s) since higher velocities inhibit attachment.

In seawater at high velocity, metals fall into two distinctly different groups: those that are velocity limited (carbon steels and copper alloys) and those that are not velocity limited (stainless steels and many nickel alloys).

Metals that are not velocity limited are subject to virtually no metal loss from velocity effects or turbulence short of cavitation conditions. The barrier films that form on these metals seem to perform best at high velocities with the full surface exposed and clean. It is in crevices and under deposits that form from slow-moving or stagnant seawater that local breakdown of the film and pitting begin.

Erosion-corrosion is a form of corrosion, typically affecting pipelines, that involves a combination of pitting and erosion. When movement of a corrosive over a metal surface increases the rate of attack due to mechanical wear and corrosion, especially abrasive particles, the attack is called erosion-corrosion. It is encountered when particles in a liquid or gas impinge on a metal surface, causing the removal of protective surface films, such as air-formed protective oxide films or adherent corrosion products, and exposing new reactive surfaces that are anodic to noneroded neighboring areas on the surface. This results in rapid localized corrosion of the exposed areas in the form of smooth-bottomed shallow recesses. As temperature increases, the protective film may become more soluble and/or less resistant to scouring. Hence, the same flow rates, but at higher temperature, can cause an increase in corrosion rates or new forms and modes of corrosion.

Nearly all turbulent corrosive media can cause erosion-corrosion. The attack may exhibit a directional pattern related to the path taken by the corrosive as it moves over the surface of the metal.

Figure 49 shows the interior of a 50 mm (2 in.) copper river waterline that has suffered pitting and general erosion due to excessive velocity of the water. The brackish river water contained some suspended solids that caused the polishing of the copper pipe surface. The horseshoe-shaped pits (facing up stream) are typical of the damage caused by localized turbulence. The copper pipe was replaced with fiberglass-reinforced plastic piping.

Fig. 49 Erosion pitting caused by turbulent river water flowing through copper pipe. The typical horseshoe-shaped pits point upstream. Approximately 0.5× actual size

Impingement corrosion is a severe form of erosion-corrosion. It occurs frequently in turns or elbows of tubes and pipes, and on surfaces of impellers or turbines. It occurs as deep, clean, horseshoe-shaped pits with the deep, or undercut, end pointing in the direction of flow. Impingement-corrosion attack can also occur as the result of partial blockage of a tube. A stone, a piece of wood, or some other object can cause the main flow to

deflect against the wall of the tube. The impinging stream can rapidly perforate tube walls. Water that carries sand, silt, or mud will have an additional severely erosive effect on tubes.

The energy transfer is the rate of change of the momentum of the flowing medium. As this energy transfer takes place over a smaller and smaller area per unit time, the energy or power density of the process becomes damaging to the substrate.

A beneficial extension of this effect at sufficiently high power densities is abrasive machining and cutting. At slower rates and lower power densities, it is called impingement corrosion. Steam erosion is another form of impingement corrosion, occurring when high-velocity wet steam impacts a metal surface. The resulting attack usually produces a roughened surface showing a large number of small cones with the points facing in the direction of flow.

Example 19: Impingement-Corrosion Failure of a Ferritic Malleable Iron Elbow. Leakage was detected in a malleable iron elbow after only three months in service. Life expectancy for the elbow was 12 to 24 months. The 21 mm (0.82 in.) ID 90° elbow connected segments of 19 mm (0.75 in.) pipe with a wall thickness of 2.9 mm (0.11 in.). The piping alternately supplied steam and cooling water to a tire-curing press. The supply line and elbow were subjected to 14 heating and cooling cycles per hour for at least 16 h/day, or a minimum of 224 cycles/day. Steam pressure was 1035 kPa (150 psi), and water pressure was 895 kPa (130 psi). Based on pump capacity, the water-flow rate was estimated at 1325 L/min (350 gal/min). Water-inlet temperature was 10 to 15 °C (50 to 60 °F); water-outlet temperature was 50 to 60 °C (120 to 140 °F). The water had a pH of 6.9.

The elbow was cast from ASTM A 47, grade 35018, malleable iron and had a hardness of 76 to 78 HRB. ASTM A 47 (Ref 80) covers applications for normal ambient conditions up to 400 °C (750 °F). Composition of the iron was:

Element	Content, %
Carbon	**1.95**
Manganese	**0.60**
Silicon	**1.00**
Sulfur	**0.15**
Phosphorus	**0.05**
Copper	**0.17**
Chromium	**0.03**
Nickel	**0.02**
Molybdenum	**0.001**

Investigation. Specimens were cut from two areas on the elbow, one just below the point of leakage (region A, Fig. 50a) and another further downstream (region B, Fig. 50a). The deepest penetration in the first specimen was at the top, just below the point of leakage, where the wall thickness had been reduced to 1.6 mm (0.062 in.). Maximum wall thickness was 9.5 mm (0.37 in.).

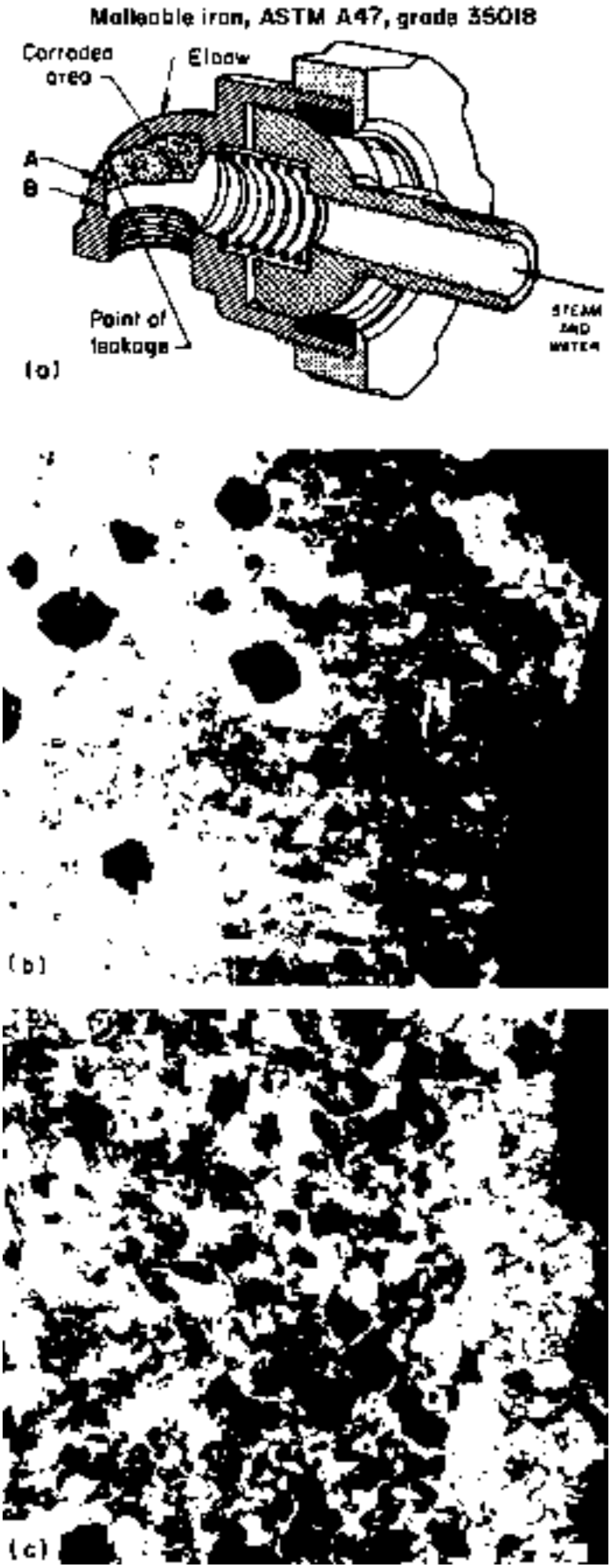


Fig. 50 Malleable iron elbow in which impingement corrosion caused leakage and failure at the bend. (a) Section through the elbow showing extent of corrosion and point of leakage. Regions A and B are locations of specimens shown in micrographs (b) and (c), respectively. (b) Micrograph of a nital-etched specimen from region A (just below the

failure area) showing ferritic surface (light areas) corroded away, exposing the subsurface pearlitic zone (dark areas). 67×. (c) Micrograph of a nital-etched specimen from region B (an uncorroded area of the elbow) showing a typical ferrite zone at the surface, a subsurface pearlitic zone (with twice the thickness of the ferrite zone), and a ferritic malleable microstructure in the interior. 67×

Metallographic examination of the first specimen showed that moderate but irregular attack had occurred (Fig. 50b). A small area of ferrite remained at the top, but the surface ferritic zone (light areas, Fig. 50b) had been eroded and corroded away on the remainder of the surface, exposing the pearlitic zone (dark areas, Fig. 50b). The interior of the specimen showed a ferritic malleable microstructure.

The second specimen, which showed no signs of attack, had a typical ferrite zone at the surface, then a subsurface pearlitic zone with about twice the thickness of the ferrite zone, and a ferritic malleable microstructure in the interior of the specimen (Fig. 50c).

Conclusions. Examination of the micrographs (Fig. 50b and c) indicated that the elbows had been given the usual annealing and normalizing treatment for ferritizing malleable iron. This resulted in lower resistance to erosion and corrosion than pearlitic malleable iron.

Recommendations. It was recommended that replacement elbows be heat treated to produce a pearlitic malleable microstructure, which has longer life under the given conditions of service. (Additional information on the heat treatment and properties of ferritic and pearlitic malleable irons is provided in the article "Malleable Iron" in *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Volume 1 of *ASM Handbook.*)

In piping systems, erosion-corrosion can be reduced by increasing the diameter of the pipe, thus decreasing velocity and turbulence. The streamlining of bends is useful in minimizing the effects of impingement. Inlet pipes should not be directed onto the vessel walls if this can be avoided. Flared tubing can be used to reduce problems at the inlet tubes in a tube bundle.

Cavitation erosion is the one of the most severe form of erosion-corrosion. It occurs principally when relative motion between a metal surface and a liquid environment and thermodynamic conditions in the liquid causes vapor bubbles to develop and then rapidly collapse. When the bubbles collapse, they impose hammerlike blows, which produce stresses on the order of 400 MPa (60 ksi), simultaneously with the initiation of tearing action that appears to pull away portions of the surface. The tearing action can remove any protective oxide film that exists on the surface of a metal, exposing active metal to the corrosive influence of the liquid environment. It has been shown that corrosion is not essential to cavitation erosion, however, as discussed in the article "Liquid-Impact Erosion" in this Volume.

Cavitation erosion occurs on rotors, pumps, on the trailing faces of propellers and of water-turbine blades, and on the water-cooled side of diesel-engine cylinders. Here high velocities give rise to extremely low-pressure areas, vapor bubbles develop by a phase change in thermodynamically unstable liquids, and then the bubbles collapse cataclysmically as higher pressures and/or lower temperature make the vapor state thermodynamically unstable. The massive momentum change of liquid rushing in to fill the void at collapse of the vapor bubble destroys the protective film on the metal surface or disrupts and often plastically deforms the metal itself.

Damage can be reduced by operating rotary pumps at the highest possible head of pressure to avoid formation of momentarily stable vapor bubbles. For turbine blades, aeration of water serves to cushion the damage caused by the collapse of bubbles. Neoprene or similar elastomer coatings on metals are somewhat reasonably resistant to damage from this cause. To reduce cavitation damage to diesel-engine cylinder liners, the addition of 2000 ppm sodium chromate to the cooling water has proved effective, as has the use of Ni-Resist (high-nickel cast iron) liners. Table 5 rates some metals frequently used in seawater in four groups on the basis of their resistance to cavitation erosion in seawater.

Table 5 Ratings of some metals for resistance to cavitation erosion in seawater

Group No. and characteristics	Alloys
Group I	
Most resistant. Subject to little or no damage. Useful under extremely severe conditions	**Stellite hardfacing alloys** **Titanium alloys**

	Austenitic and precipitation-hardening stainless steels **Nickel-chromium alloys such as Inconel alloys 625 and 718** **Nickel-molybdenum-chromium alloys such as Hastelloy C**
Group II	
These metals are commonly used where a high order of resistance to cavitation damage is required. They are subject to some metal loss under the most severe conditions of cavitation.	**Nickel-copper-aluminum alloy Monel K-500** **Nickel-copper alloy Monel 400** **Copper alloy C95500 (nickel-aluminum bronze, cast)** **Copper alloy C95700 (nickel-aluminum-manganese bronze, cast)**
Group III	
These metals have some degree of cavitation resistance. They are generally limited to low-speed low-performance applications.	**Copper alloy C71500 (copper-nickel, 30% Ni)** **Copper alloys C92200 and C92300 (leaded tin bronzes M and G, cast)** **Manganese bronze, cast** **Austenitic nickel cast irons**
Group IV	
These metals normally are not used in applications where cavitation damage may occur, except in cathodically inhibited solutions or when protected by elastomeric coatings.	**Carbon and low-alloy steels** **Cast irons** **Aluminum and aluminum alloys**

Applies to normal cavitation-erosion intensities, at which corrosion resistance has a substantial influence on the resistance to damage. Source: Ref 81

References cited in this section

80. "Standard Specification for Ferritic Malleable Iron Castings," A 47, *Annual Book of ASTM Standards,* Vol 01.02

81. A.H. Tuthill and C.M. Schillmoler, "Guidelines for Selection of Marine Materials," Ocean Science and Engineering Conference, Marine Technology Society, Washington, June 1965

Acknowledgments

Sections of this article have been based in part on the corresponding sections of “Corrosion Failures” by W.G. Ashbaugh in *Failure Analysis and Prevention,* Volume 11 of *ASM Handbook,* 1986, p 172–202.
Additionally, the section “Galvanic Corrosion” is based in part on content from the following *ASM Handbook* articles:

- R. Baboian and S. Pohlman, Galvanic Corrosion, *Corrosion,* Vol 13, *ASM Handbook,* 1987, p 83–87
- F.P. Ford and P.L. Andresen, Design for Corrosion Resistance, *Materials Selection and Design,* Vol 20, *ASM Handbook,* 1997, p 545–572

The section “Uniform Corrosion” is based in part on “Evaluation of Uniform Corrosion” by C.A. Natalie, *Corrosion,* Volume 13 of the *ASM Handbook,* 1987, p 229–230.

Forms of Corrosion

References

1. R.W. Revie, Ed., *Uhlig's Corrosion Handbook,* John Wiley & Sons, 2000, p 32

2. R.S. Treseder, Ed., *The NACE Corrosion Engineer's Reference Book,* National Association of Corrosion Engineers, 1980

3. “Standard Guide for Development and Use of a Galvanic Series for Predicting Galvanic Corrosion Performance,” G 82, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

4. E.H. Hollingsworth and H.Y. Hunsicker, Corrosion of Aluminum and Aluminum Alloys, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 583–609

5. “Standard Reference Test Method for Making Potentiostatic and Potentiodynamic Anodic Polarization Measurements,” G 5, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

6. R.H. McSwain, Fatigue Fracture of a Helicopter Tail Rotor Blade Due to Field-Induced Corrosion, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 30–32

7. Recommended Practice for Cathodic Protection of Aluminum Pipe Buried in Soil or Immersed in Water, *Mater. Protec.,* Vol 2 (No. 10), 1963 p 106

8. F.W. Hewes, Investigation of Maximum and Minimum Criteria for the Cathodic Protection of Aluminum in Soil, *Oil Week,* Vol 16 (No. 24–28), Aug–Sept 1965

9. R. Kent, “Anodic Polarization Measurements of Alloyed DI,” NACE Int. Conf., National Association of Corrosion Engineers, 1987

10. R. Kent, “Potential Anodic Polarization Measurements and Mechanical Property Analysis of Ferritic Nodular Iron Alloyed with Nickel,” University of Washington, 1986

11. F. Mansfeld and V. Bertocci, *Electrochemical Corrosion Testing,* STP 727, American Society for Testing and Materials, 1981

12. M. Henthorne, *Corrosion Causes and Control,* Carpenter Technology Corp., 1972, p. 30

13. “Preparing, Cleaning, and Evaluating Corrosion Test Specimens,” G 1, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

14. “Standard Practice for Conducting Atmospheric Corrosion Tests on Metals,” *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

15. G. Wranglén, Corrosion and Protection of Metals, Chapman & Hall, 1985, p 238

16. R.L. Martin and E.C. French, Corrosion Monitoring in Sour Systems Using Electrochemical Hydrogen Patch Probes, *J. Pet. Technol.,* Nov 1978, p 1566–1570

17. P.A. Schweitzer, *Corrosion and Corrosion Protection Handbook,* Marcel Dekker, 1983, p 483–484

18. “Standard Test Method for Conducting Potentiodynamic Polarization Resistance Measurements,” *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

19. “Standard Practice for Conventions Applicable to Electrochemical Measurement in Corrosion Testing,” G 3, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

20. K.E. Perumal, Pitting Corrosion of Stainless Steel by Potable Municipal Water in an Organic Chemical Plant, *Handbook of Case Studies in Failure Analysis,* Vol 1, ASM International, 1992, p 176–177

21. A.I. Asphahani and W.L. Silence, Pitting Corrosion, *Corrosion,* Vol 13, Metals Handbook, 9th ed., ASM International, 1987, p 112–114

22. “Examination and Evaluation of Pitting Corrosion,” G 46-94 (Reapproved 1999), *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

23. C.R. Brooks, Fatigue Fracture of Stainless Steel Wires in an Electrostatic Precipitator at a Paper Plant, *Handbook of Case Studies in Failure Analysis,* Vol 1, ASM International, 1992, p 218–222

24. S.D. Kiser, Preventing Weld Corrosion, *Adv. Mater. Process.,* Vol 160 (No. 3), March 2002

25. J.E. Hatch, Ed., *Aluminum: Properties and Physical Metallurgy,* American Society for Metals, 1984, p 290–291

26. *Corrosion of Metals under Thermal Insulation,* STP 880. American Society for Testing and Materials, 1985

27. A.J. Sedriks, *Corrosion of Stainless Steels,* 2nd ed., Wiley-Interscience, 1996, p 217

28. "Standard Test Method for Pitting or Crevice Corrosion of Metallic Surgical Implant Materials," F 746-87 (Reapproved 1999), *Annual Book of ASTM Standards,* Vol 13.01, American Society for Testing and Materials

29. N.D. Green, Corrosion of Surgical Implant Alloys: A Few Basic Ideas, *Corrosion and Degradation of Implant Materials,* STP 859, American Society for Testing and Materials, 1985, p 7

30. H. Skinner, et al., "Corrosion, Materials Characteristics and Local Tissue Reaction Associated with Osteosynthesis Devices, Implant Retrieval: Material and Biological Analysis," NBS Special Publication 601, U.S. Dept. of Commerce, Jan 1981, p 425

31. R.G. Kelley, *Pitting, Corrosion Tests and Standards: Application and Interpretation,* MNL 20, R Baboian, Ed., American Society for Testing and Materials, 1995

32. "Standard Practice for Determining the Susceptibility to Intergranular Corrosion of 5*xxx* Series Aluminum Alloys by Weight Loss after Exposure to Nitric Acid (NAWLT Test)," G 67, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials

33. M.A. Streicher, *Intergranular Corrosion of Stainless Alloys,* STP 656, R.F. Steigerwald, Ed., American Society for Testing and Materials, 1978, p 3–84

34. V. Kain, R.C. Prasad, and P.K. De, Detection of Sensitization and Intergranular Corrosion of Fe-Cr-Ni Alloys, *High Temp. Mater. Proc.,* Vol 16 (No. 3), July–Sept 1997, p 183–199

35. "Standard Practices for Detecting Susceptibility to Intergranular Attack in Austenitic Stainless Steels," A 262, *Annual Book of ASTM Standards,* Vol 01.03, American Society for Testing and Materials

36. "Test Method Standard Test Method for Electrochemical Reactivation (EPR) for Detecting Sensitization of AISI Type 304 and 304L Stainless Steels," G 108-94, *Wear and Erosion; Metal Corrosion,* Vol 03.02, *Annual Book of ASTM Standards,* American Society for Testing and Materials, 1999

37. A.P. Majidi and M.A. Streicher, Four Non-Destructive Electrochemical Test for Detecting Sensitization in Type 304 and 304L Stainless Steels, paper 62, *Corrosion 85,* National Association of Corrosion Engineers, 1985, p 1–17; *Nucl. Technol.,* Vol 75 (No. 3), Dec 1986, p 356–369

38. A.P. Majidi and M.A. Streicher, Potentiodynamic Reactivation Method for Detecting Sensitization in AISI 304 Stainless Steels, *Corrosion,* Vol 40 (No. 1), Jan 1984, p 21–32

39. W.L. Clarke, V.M. Romero, and J.C. Danko, Detection of Sensitization in Stainless Steel Using Electrochemical Techniques, Paper No. 180 (Pamphlet), *Corrosion/77,* International Corrosion Forum, National Association of Corrosion Engineers, 1977, 32 pages

40. M. Akashi et al., Evaluation of IGSCC Susceptibility of Austenitic Stainless Steels Using Electrochemical Reactivation Method, *Corrosion Eng.,* Vol 29, 1980, p 163

41. A.P. Majidi and M.A. Streicher, The Double Loop Reactivation Method for Detecting Sensitization in AISI 304 Stainless Steels, *Corrosion,* Vol 40 (No. 11), Nov 1984, p 584–593

42. K.J. Stoner, An Evaluation of Electrochemical Potentiokinetic Reactivation Techniques for In-Service Measurements on Type 304 Stainless Steel, *Computer-Aided Microscopy and Metallography,* ASM International, 1990, p 171–182

43. M. Verneau, J. Charles, and F. Dupoiron, Applications of Electrochemical Potentiokinetic Reactivation Test to On-Site Measurements on Stainless Steels, *Application of Accelerated Corrosion Tests to Service Life Prediction of Materials,* STP 1194, American Society for Testing and Materials, 1994, p 367–381

44. D. Van Rooyen, *Corrosion,* Vol 31, 1975, p 327–337

45. G. Kobrin, Material Selection, *Corrosion,* Vol 13, *Metals Handbook,* 9th ed., 1987, p 325

46. V. Kain, Intergranular Corrosion of Stainless Steels, *Conference: Workshop on Failure Analysis, Corrosion Evaluation and Metallography,* Bhabha Atomic Research Centre, Trombay, Bombay India, 1992 p 107–128

47. V. Cihal, Sr. and V. Cihal, Jr., Economical Evaluation of Stainless Steels for Their Resistance to Intergranular Corrosion, Book No. 556, *Progress in the Understanding and Prevention of Corrosion* (Barcelona, Spain), Vol II, July 1993, The Institute of Materials, 1993, p 1537–1541

48. G.F. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* 1990, p 698–736

49. T.M. Devine, Mechanism of Intergranular Corrosion and Pitting Corrosion of Austenitic and Duplex 308 Stainless Steel, *J. Electrochem. Soc.,* Vol 126 (No. 3), 1979, p 374

50. E.E. Stansbury, C.D. Lundin, and S.J. Pawel, Sensitization Behavior of Cast Stainless Steels Subjected to Simulated Weld Repair, *Proc. 38th SFSA Technical and Operating Conference,* Steel Founders' Society of America, 1983, p 223

51. C. McCaul, Evaluation of Intergranular Corrosion Susceptibility in an As-Welded High Alloy Austenitic Stainless Steel Casting, *Brit. Corros. J.,* Vol 26 (No. 4), 1991 p 239–243

52. M. Henthorne, *Localized Corrosion—Cause of Metal Failure,* STP 516, American Society for Testing and Materials, 1972, p 66–119

53. R.A. Corbett and B.J. Saldanha, Evaluation of Intergranular Corrosion, *Corrosion,* Vol 13, *Metals Handbook,* 9th ed., 1987, p 240

54. J.S. Armijo, Intergranular Corrosion of Nonsensitized Austenitic Stainless Steels, *Corrosion,* Jan 1968, p 24–30

55. W.H. Herrnstein, J.W. Cangi, and M.G. Fontana, Effect of Carbon Pickup on the Serviceability of Stainless Steel Alloy Castings, *Mater. Perform.,* Oct 1975, p 21–27

56. Corrosion Failures, *Failure Analysis,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 180

57. D.R. Johns and F.R. Beckitt, Factors Influencing the Thermal Stabilisation of Alloy 600 Tubing Against Intergranular Corrosion, *Corros. Sci.,* Vol 30 (No. 2–3), 1990, p 223–237

58. C.M. Younes, F.H. Morrissey, G.C. Allen, and P. McIntyre, Effect of Heat Treatment on the Grain Boundary Chemistry and the Resistance to Intergranular Corrosion of Alloy 600 and Alloy 690, *Second International Conference on Corrosion-Deformation Interactions,* CDI '96, Institute of Materials, 1997, p 421–434

59. P. Lin, G. Palumbo, U. Erb, and K.T. Aust, Influence of Grain Boundary Character Distribution on Sensitization and Intergranular Corrosion of Alloy 600, *Scr. Metall. Mater.,* Vol 33 (No. 9), 1 Nov 1995, p 1387–1392

60. M.H. Brown, *Corrosion,* Vol 25, 1969, p 438

61. M.H. Brown and R.W. Kirchner, *Corrosion,* Vol 29, 1973, p 470–474

62. E.H. Dix, Acceleration of the Rate of Corrosion by High Constant Stresses, *Trans. AIME,* Vol 137, 1940, p 11

63. W.L. Fink and L.A. Willey, Quenching of 75S Aluminum Alloy, *Met. Technol.,* Vol 14 (No. 8), 1947, p 5

64. M.S. Hunter, G.R. Frank, and D.L., Robinson, *Fundamental Aspects of Stress-Corrosion Cracking* (Proc. Conf.), R.W. Staehle, Ed., National Association of Corrosion Engineers, 1969, p 497

65. H. Kaesche, Pitting Corrosion of Aluminum and Intergranular Corrosion of Aluminum Alloys, *Localized Corrosion,* B.F. Brown, J. Kruger, and R.W. Staehle, Ed., National Association of Corrosion Engineers, 1974, p 516

66. J.R. Galvele and S.M. de Micheli, Mechanism of Intergranular Corrosion of Al-Cu Alloys, *Corros. Sci.,* Vol 10, 1970, p 795

67. H.F. Brauer and W.M. Peirce, The Effect of Impurities on the Oxidation and Swelling of Zinc-Aluminum Alloys, *Trans. AIME,* Vol 60, 1923, p 796

68. H.H. Lee, Galvanized Steel With Improved Resistance to Intergranular Corrosion, *Proc. Galvanized Committee,* Vol 69, 1977, p 17

69. D.J. Willis, A. De Liseo, and B. Gleeson, Intergranular Corrosion in Continuously Galvanized Steel, *Galvatech '95. The Use and Manufacture of Zinc and Zinc Alloy Coated Sheet Steel Products into the 21st Century,* Iron and Steel Society/AIME, 1995, p 145–153

70. H. Uhlig and R.W. Revie, *Corrosion and Corrosion Control,* John Wiley & Sons, 1985, p 14, 331–334

71. M.G. Fontana and N.D. Greene, *Corrosion Engineering,* McGraw-Hill, 1978, p 67–71

72. E.D. Verink, Jr., and R.H. Herdersbach, Jr., Evaluation of the Tendency for Dealloying in Metal Systems, *Localized Corrosion—Causes of Metal Failure,* STP 516, American Society for Testing and Materials, 1972, p 303–322

73. S.R. Freeman, Graphitic Corrosion—Don't Forget about Buried Cast Iron Pipes, *Mater. Perform.,* Aug 1999, p 68–69

74. K. Kasahara and F. Kajtyama, The Role of Bacteria in the Graphitic Corrosion of Buried Ductile Cast Iron Pipes, *Microbial Corrosion,* Second EFC Workshop, March 1991

75. R.J. Ferrara and T.E. Caton, Review of Dealloying of Cast Aluminum Bronze and Nickel-Aluminum Bronze Alloys in Sea Water Service, *Mater. Perform.,* Feb 1982, p 30–34

76. R. Steigerwald, Metallurgically Influenced Corrosion, *Corrosion,* Vol 13, ASM Handbook, 1987, p 130–133

77. S.J. Pavel, Type 304L Stainless Steel Corrosion in Molten Lithium Hydride as a Function of Temperature, *J. Nucl. Mater.*, Vol 200 (No. 2), April 1993, p 184–199

78. T.B. Cassagne, W.F. Flanagan, and B.D. Lichter, On the Failure Mechanism of Chemically Embrittled Cu_3Au Single Crystals, *Metall. Trans.*, April 1986, p 703–710

79. M.G. Fontana, *Corrosion Engineering*, 3rd ed., McGraw-Hill, 1986

80. "Standard Specification for Ferritic Malleable Iron Castings," A 47, *Annual Book of ASTM Standards*, Vol 01.02

81. A.H. Tuthill and C.M. Schillmoler, "Guidelines for Selection of Marine Materials," Ocean Science and Engineering Conference, Marine Technology Society, Washington, June 1965

Effect of Environment on the Performance of Plastics

Donald E. Duvall, Engineering Systems Inc.

Introduction

THE MECHANICAL PROPERTIES of polymeric materials are often segregated into short-term and long-term properties. The category of short-term properties includes such things as tensile and impact strengths. Long-term properties include creep, stress relaxation, and creep (stress) rupture. Both categories of properties are affected by exposure to external chemical environments.

With any polymeric material, chemical exposure may have one or more different effects. Some chemicals act as plasticizers, changing the polymer from one that is hard, stiff, and brittle to one which is softer, more flexible, and tougher. Often these chemicals can dissolve the polymer if they are present in large enough quantity and if the polymer is not crosslinked. Other chemicals can induce environmental stress cracking (ESC), an effect in which brittle fracture of a polymer will occur at a level of stress well below that required to cause failure in the absence of the ESC reagent. Finally, there are some chemicals that cause actual degradation of the polymer, breaking the macromolecular chains, reducing molecular weight, and diminishing polymer properties as a result. Each of these effects is examined in subsequent paragraphs.

Effect of Environment on the Performance of Plastics

Donald E. Duvall, Engineering Systems Inc.

Plasticization, Solvation and Swelling

Certain interactions between liquid chemicals and polymers can be understood through the use of solubility parameters. The Hildebrand solubility parameter is the square root of the cohesive energy density, the later being the energy required to vaporize one mol of a liquid (Ref 1). Solubility parameter values for both low molecular weight liquids and polymeric materials are tabulated in many references (Ref 2, 3). When linear or branched thermoplastic polymers are exposed to large enough quantities of solvents having solubility

parameters within approximately ±2 H of that of the polymer, dissolution of the polymer will occur. In smaller quantities, these solvents will be adsorbed by the polymer. With polymer-solvent combinations having solubility parameter differences outside this range, some adsorption of the solvent by the polymer may still occur. When large differences between solvent and polymer solubility parameters exist, the solvent will have no apparent effect. Amorphous polymers absorb chemicals more readily than crystalline polymers, and the rate varies inversely with the degree of crystallinity. Cross-linked polymers will not dissolve but will swell significantly when exposed to chemicals having similar solubility-parameter values.

The impact of these interactions on the mechanical properties and failure of an affected polymer are many. One effect can be plasticization of the polymer by the adsorbed chemical. Plasticization of a polymer can result in the polymer being transformed from a rigid, glassy material to a soft flexible material (Ref 4, 5). Great advantage is taken of this effect in the polyvinyl chloride (PVC) industry. The PVC can be altered from the rigid material from which plastic pressure pipes are made to the extremely soft, flexible material from which medical tubing is made through the judicious use of plasticizers (Ref 6). However, the plasticization effect reduces both tensile strength and stiffness of the affected plastic and also accelerates the creep rate of the material if it is under stress. If an application of a particular plastic requires a certain minimum level of strength or stiffness, unintentional plasticization by exposure to a chemical that the plastic can adsorb could accelerate failure by reducing those properties (Ref 7).

Plasticizer loss from an intentionally plasticized polymer may also have an adverse effect on polymer performance. Plasticizer migration from PVC is a well known phenomenon (Ref 8) that results not only in embrittlement and/or a loss of flexibility of the PVC part but also loss of the plasticizer chemical(s) into the environment. Plasticizer migration from flexible PVC products also results in some small amount of shrinkage of the products. This may be problematic if close dimensional tolerances are necessary.

The effects mentioned previously will occur in both amorphous and crystalline plastics, but they may not be as visibly evident in crystalline ones. Crystalline plastics often do not appear to be as affected by interactions with solvents, since diffusion of solvent into the crystalline regions is much more limited (Ref 9). Adsorption of solvents into the amorphous regions of a crystalline polymer will create the discussed effects within those regions of the polymer morphology; this can result in changes in polymer mechanical properties. For example, nylon plastics will absorb moisture from the air. An extremely dry nylon may be rather brittle, while that same nylon exposed to 50% relative humidity for several days can be quite tough. However, the short-term tensile strength and modulus of the hydrated nylon will be somewhat reduced, as will the long term (creep rupture) strength of the material. Creep deformation of the hydrated nylon will proceed more rapidly than that in the dry material at the same level of stress. Hydrocarbon liquids will have similar effects on polyethylene.

In crosslinked rubber products, swelling will result when exposure to chemical solvents occurs (Ref 10). The swelling can be rather extreme when the solubility parameter difference between rubber and solvent is small. Highly swollen rubbers will exhibit a severe loss of strength and stiffness, a high creep rate, and a high rate of stress relaxation or compression set. This last phenomenon can have an extremely adverse effect on rubber gaskets and o-rings by allowing the compressive stress in such a seal to decrease to a level so low that leakage of the seal can occur. Even with polymer/solvent pairs for which the solubility parameter difference is somewhat greater and swelling seems not as severe, these effects can occur. Rubber compounds are sometimes intentionally plasticized like PVC in order to achieve a desired level of compressibility, flexibility, and so forth while using a specific polymer as the base for the compound. Loss of these intentional plasticizers into the environment will have the same effects as with PVC.

With respect to identifying adsorbed chemicals in plastics, since the chemical is adsorbed, it can usually be extracted in some way and identified. In some cases, simply heating the plastic will drive off the chemical, which can be collected and fed into a gas chromatograph (GC) or a GC/mass spectrometry (GC/MS) analysis. Sometimes extraction with a second chemical solvent is necessary, one that is a better solvent for the plasticizer than the polymer is. The extract can then be tested for the presence of other chemicals to identify plasticizers. When the quantity of absorbed chemical is great, an infrared spectrum of the plasticized plastic can be obtained by Fourier transform infrared (spectroscopy) (FTIR). An FTIR spectrum of the unplasticized plastic, or a reference spectrum from a spectral library of such, can then be subtracted from the subject spectrum by the on-board computer of the instrument. The resultant subtraction spectrum will often be that of the absorbed chemical, which can then be identified by its infrared spectrum.

References cited in this section

1. M.P. Stevens, *Polymer Chemistry: An Introduction,* 2nd ed., Oxford University Press, 1990, p 43

2. E.A. Grulke, Solubility Parameter Values, Chap. VII, *Polymer Handbook,* 4th ed., J. Brandup, E.H. Immergut, and E.A. Grulke, Ed., John Wiley & Sons, New York, 1999

3. *Physical Properties of Polymers Handbook,* J.E. Mark, Ed., AIP Press, Woodbury, NY, 1996

4. E. Miller, Properties Modification by Use of Additives, *Engineered Materials Handbook,* Vol. 2, *Engineering Plastics,* ASM International, 1988, p 493–507

5. A. Kumar and R. Gupta, *Fundamentals of Polymers,* Section 2.3, McGraw-Hill, 1998.

6. J.K. Sears and J.R. Darby, Solvation and Plasticization, Chap. 9, *Encyclopedia of PVC,* Vol 1, 2nd ed., Marcel Dekker, Inc., 1986.

7. M. Ezrin, *Plastics Failure Guide: Cause and Prevention,* Hanser Publishers, New York, 1996, p 22–23

8. J.-M. Vergnaud, *Polym.-Plast. Technol. Eng.,* Vol 20 (No. 1), 1983, p 1–22

9. F.W. Billmeyer, *Textbook of Polymer Science,* 3rd ed., Wiley-Interscience, New York, 1984, p 344

10. S.L. Rosen, *Fundamental Principals of Polymeric Materials,* Section 3.4, 2nd ed., Wiley-Interscience, New York, 1993

Effect of Environment on the Performance of Plastics

Donald E. Duvall, Engineering Systems Inc.

Environmental Stress Cracking

Environmental stress cracking (ESC) of plastics has been defined as "… the failure in surface initiated brittle fracture of a specimen or part under polyaxial stress in contact with a medium in the absence of which fracture does not occur under the same conditions of stress" (Ref 11). Virtually all plastics are stress cracked by some chemical environments. The biggest problem with this is that each plastic has its own set of stress cracking reagents, and those chemicals that stress crack one type of plastic will have no effect on others. Thus, the potential stress cracking effect of a specific chemical on a specific plastic must be known from prior work or elucidated by direct experiment in order to know whether or not a problem exists.
Several physical characteristics are typical of environmental stress crack failures:

- Failure is always nonductile, even in plastics that would normally exhibit a ductile yielding failure mechanism.
- The brittle fracture is surface initiated.
- The surface at which cracking initiated was in contact with a chemical reagent.
- The plastic was mechanically stressed in some way; internal (residual) stresses or externally applied stresses both qualify.

Chemicals that induce ESC usually have no other apparent effect on the plastic in question; that is, no swelling (or dissolution in large quantities of the chemical) and no physical or chemical changes in the polymer that

might be detected by analytical methods. In the absence of a mechanical stress, the ESC chemical has no discernible effect. Conversely, the magnitude of stress that will cause ESC will not cause fracture if imposed in the absence of the stress cracking reagent. Thus conventional chemical resistance tests run on plastics, in which unstressed tensile bars are soaked in a chemical and withdrawn periodically for testing, give no indication of the possibility of ESC for any polymer-reagent system. It is only in the presence of both mechanical stress and chemical environment that ESC occurs.

Failures from ESC may occur early or late in the life of a product. In some cases, ESC will occur as soon as a part is loaded, if the reagent is already present on the surface of a previously unstressed part. Stress cracking reagents also impact the creep rupture properties of plastics by shortening the time for brittle fracture to occur over that which exists in the absence of the reagent. Sometimes ESC reagents will create brittle fracture at a low stress level in a polymer-like polyethylene that normally fails in a highly ductile manner. These effects are summarized in the schematic creep rupture curves of Fig. 1. High density polyethylene exhibits ductile failure (elongations to break of several hundred percent) at stresses near to its reported yield stress. At lower stresses and longer failure times, a different molecular mechanism controls failure, and brittle fracture occurs at elongations of less than five percent. Presence of an ESC reagent on the surface of a plastic can dramatically shorten the time for failure to occur at a given stress level and change the failure mechanism from highly ductile to macroscopically brittle.

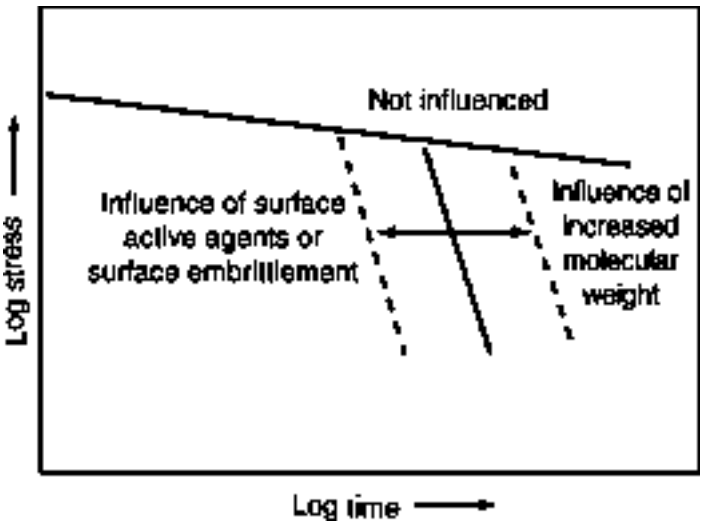


Fig. 1 Effect of environmental stress cracking agents on creep rupture performance

One condition in which the possibility of ESC should be considered is when there is an apparent stain or other deposit residue on the surface of a fractured part. Even though stress cracking chemicals are not adsorbed into the plastic to any significant extent, surface residues are often left behind that can be identified. Once the chemical that left the deposit is known, reference to prior work reported in the literature may tell whether or not it is an ESC agent for the particular plastic in question. If no previously identified problems with that chemical can be found, laboratory stress cracking tests of the polymer/chemical combination can be conducted to assess the likelihood of ESC.

Reference cited in this section

11. J.B. Howard, *SPE J.*, Vol 15, 1959, p 397

Donald E. Duvall, Engineering Systems Inc.

Polymer Degradation by Chemical Reaction

Another effect that chemical environments can have on plastics is to actually degrade the polymer; that is, to break down the polymer chains into lower molecular weight compounds that no longer have the desirable strength or toughness properties of the original. Certain polymer types are more susceptible than others to specific degradation mechanisms, but all polymers can be degraded by at least one mechanism. The most common degradation mechanisms are discussed in the following paragraphs.

Hydrolysis. Polymers created by stepwise reactions, for example, polyesters and nylons, form water as a reaction product along with the polymer. Under certain circumstances of exposure to aqueous environments, the polymerization reaction can essentially be reversed and the polymer broken down. Normally, these hydrolytic degradation reactions occur at extremely slow rates, and certainly nylon and polyester fabrics can be repeatedly washed in water without adverse effect. However, at conditions of either low (>4) or high (<10) pH, the rate of hydrolysis may become perceptible and result in molecular weight reduction and mechanical property diminution.

These polymer types can also degrade during processing (i.e., extrusion or injection molding) if there is moisture in the material. Even at neutral pHs, the elevated temperatures used for polymer molding or extrusion (175 to 250 °C, or 350 to 500 °F, or more) will cause hydrolytic degradation if there is moisture in the resin. Because of this, polymer resin manufacturers advise drying of the material just prior to processing to reduce the moisture content to a low enough level that hydrolysis will not occur while the resin is heated in the manufacturing equipment. It is often the case with plastics that are susceptible to hydrolytic degradation that a reduced polymer molecular weight is found during the failure analysis. The challenge for the analyst then becomes deciding whether the degradation occurred during fabrication of the part or on exposure to an aqueous service environment.

Thermal Degradation. High molecular weight polymers will also break down upon exposure to elevated temperatures. Sufficient thermal energy can be input to a polymeric material to break the covalent chemical bonds that hold polymer molecules together. This bond breakage (chain scission), if it occurs in the polymer backbone, will reduce molecular weight. Chain scission of side chain branches may also alter the polymer structure sufficiently to change appearance or mechanical properties enough to create a premature failure.

As with hydrolysis, thermal degradation can occur both in processing and in an end-use environment. In molding or extrusion operations, the molten plastic is exposed not only to elevated temperatures but also to mechanical shearing. The combination of the two may reduce molecular weight to the extent that performance properties will suffer. Thermally degraded plastics also tend to discolor and sometimes a plastic product is deemed a failure because it no longer has the desired cosmetic appearance due to thermal degradation. These same changes may be observed in end-use environments, albeit at much slower rates due to the lower temperatures at which plastics are usually used.

Thermal degradation of polymers is a chain reaction that begins when an atom (usually a hydrogen atom) is abstracted from the polymer chain, leaving behind an unpaired electron from the broken covalent bond at an atom on the chain (Ref 12). The free radical thus formed may react in several different ways, one of which results in chain scission and molecular weight reduction. Fortunately for plastics usage, there are chemical additives compounded into polymers that will react with the unpaired electron, interrupt the chain reaction, and postpone or at least greatly retard thermal decomposition. If thermal degradation is believed to be a contributing factor to failure, the type and amount of these additives should be checked to be certain that failure was not due to degradation in an unprotected polymer material.

Oxidation. Many polymers, especially the olefins and others with long olefinic segments in the polymer chain backbone, will oxidize when exposed to oxygen-containing environments. As with thermal degradation, oxidation usually commences by formation of a free radical on the polymer chain. An oxygen atom from the environment will then react with the unpaired electron to form a hydroperoxy radical. This will then degrade by

one of several reactions, some of which result in chain scission and property loss (for details of oxidation chemistry, see Ref 12).

As with thermal degradation, chemical additive stabilizers and antioxidants can be added to the polymer that will break the chain reaction in a variety of ways, preserving polymer properties at least until the additives have been consumed. If oxidative degradation is a possible contributing factor to a premature failure, it becomes necessary to determine what allowed it to occur. It may be that the stabilizers were not present originally in the proper types or amounts. If the polymer resin did contain antioxidants, then it must be determined what caused them to become ineffective. In some cases stabilizers can bloom to the surface of a plastic part and be removed by ablation, dissolution, or evaporation into the environment (Ref 13). In other cases, the additives may simply have been consumed doing the job for which they were intended, and premature oxidation occurred because the service environment was at a higher temperature than the design engineer anticipated. All these can lead to an oxidized polymer with reduced mechanical properties, unacceptable appearance, or other deficiencies.

Photodegradation. With many polymeric materials, ultraviolet (UV) radiation can be the source of energy that will abstract an atom from the polymer backbone and start the degradation process. It is well known that prolonged outdoor exposure of plastics will initially cause color changes that may be undesirable. Oxidation initiated by UV radiation will result in eventual loss of properties as well. Once again, there are chemical additives that will retard these processes, but eventually they will be consumed and degradation will proceed. The plastics design engineer must be certain that the UV radiation stabilizers are present in the proper types and amounts to yield a product that will operate for its intended life without undergoing an inordinate amount of degradation from the exposure. There are both accelerated indoor and outdoor test methods that are used to assess the level of stability to UV exposure of a plastic material.

References cited in this section

12. N. Grassie and G. Scott, *Polymer Degradation and Stabilization,* Cambridge University Press, 1985

13. U.W. Gedde et al., *Polym. Eng. Sci.* Vol 34 (No. 24), 1994, p 1773

Donald E. Duvall, Engineering Systems Inc.

Surface Embrittlement

An adverse effect of polymer degradation on plastic part performance does not require changes in the bulk of the material in that part. In many cases, it is only necessary to cause degradation in a thin surface layer of the part in order for performance to be compromised. This phenomenon has been observed in both long-term and short-term properties of many polymeric materials.

In fact, polymer degradation is often limited, at least initially, to the surfaces of exposed plastic products. Oxidative degradation initiated by either purely thermal means or by UV radiation occurs initially at the surfaces, because that is where oxygen concentration is the greatest. In order for oxidation to occur deeper in a specimen, it must diffuse in. Because oxygen reacts very rapidly with free radical species, oxidation below the surface of a polymer part is diffusion limited and occurs very slowly compared to surface oxidation. Hydrolytic degradation also occurs first and most rapidly at surfaces since that is where the concentration of water is the greatest. Since often all that is needed for premature failure to occur is to generate a sufficient level of degradation at the surface, even if the bulk material within the plastic part is unaffected, the fact that degradation is initially limited to the surfaces creates problems for product performance.

The impact of surface degradation on short-term properties has been demonstrated by many authors. Numerous studies have shown that considerable reductions in tensile strength, impact strength, and toughness have been observed for oxidation degradation extending only a short distance into a specimen (Ref 14, 15). Figure 2

illustrates this effect in polyethylene. Choi (Ref 17, 18) demonstrated that the creep rupture behavior of a polyethylene pipe resin could be compromised by a certain level of oxidative degradation occurring only in the first 50 μm below the surface of a 2.5 mm (0.1 in.) thick specimen. Figure 3 (Ref 17) illustrates how surface degradation of a plain strain tension specimen alters the ductile brittle transition in polyethylene creep rupture.

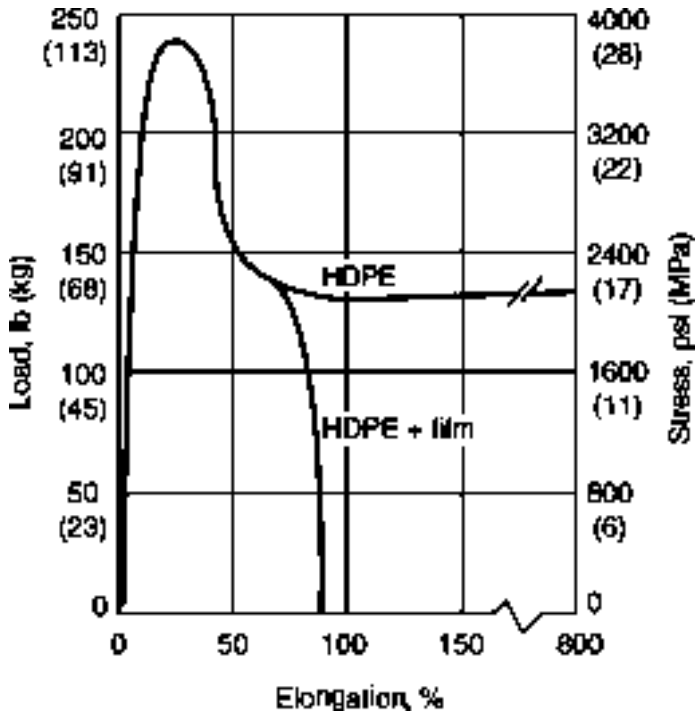


Fig. 2 Effect of thin brittle film on stress-strain behavior of high density polyethylene. Source: Ref 16

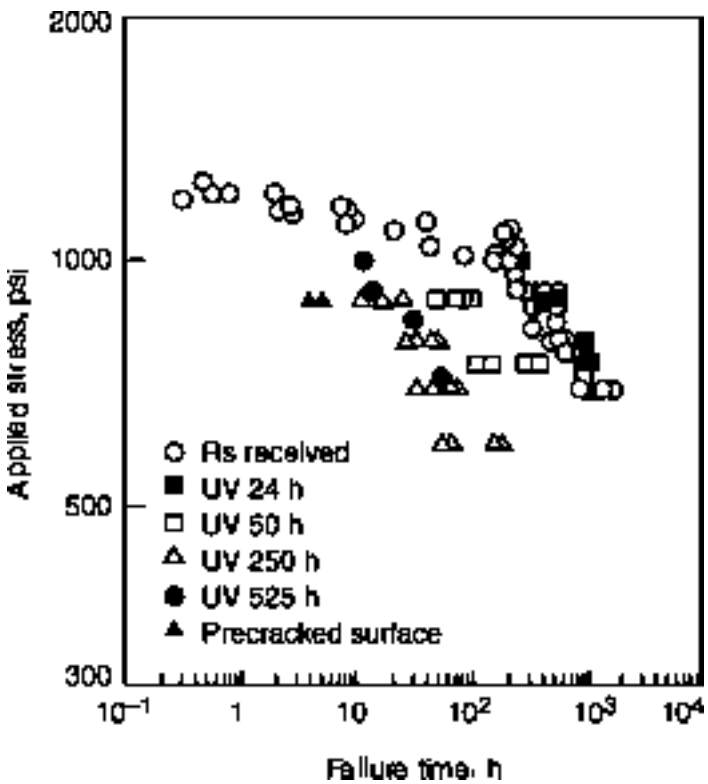


Fig. 3 Effect of surface embrittlement from varied UV exposure times on creep rupture behavior of polyethylene at 80°C (175 °F). Source: Ref 17

The existence of this surface embrittlement phenomenon requires that the evaluation of surface-initiated brittle fracture in an otherwise ductile polymer include characterization of material taken only from the surface. The results can be compared to similar testing of the core of the failed part or to testing of a control sample of the material to determine whether or not degradation of material at the surface is a contributing factor.

References cited in this section

14. P.K. So and L.J. Broutman, *Polym. Eng. Sci.*, Vol 22 (No. 14), 1982, p 888

15. J.C.M. de Bruijn, "The Failure Behavior of High Density Polyethylene Products with an Embrittled Surface Layer Due to UV Exposure," Delft University of Technology, Delft, Netherlands, 1991.

16. L.J. Broutman, "Surface Embrittlement of Polyethylene," Final Report to the Gas Research Institute, GRI-81-0030, Chicago, Nov 1981

17. S.-W. Choi, "Surface Embrittlement of Polyethylene," Ph.D. dissertation, Illinois Institute of Technology, Chicago, 1992

18. S.W. Choi and L.J. Broutman, Proceedings of the 11th Plastic Fuel Gas Pipe Symposium, San Francisco, 3–5 Oct 1989, p 296–320

Effect of Environment on the Performance of Plastics

Donald E. Duvall, Engineering Systems Inc.

Temperature Effects

One final effect of environment on polymer performance is that of temperature. Polymeric materials will exhibit a transition between two very different types of mechanical behavior as the environmental temperature passes through what is called the "glass transition temperature" (T_g) (Ref 19). Figure 4 shows how the modulus of an amorphous polymer changes as temperature is increased or decreased through this critical region. At temperatures well below T_g, this polymer behaves like a glassy material, with a relatively high modulus and low energy to break. At temperatures well above T_g, the same polymer has two or three orders of magnitude lower modulus and will either flow like a very viscous liquid or fail in tension at high extension, depending on whether or not the polymer is cross-linked. In the temperature range over which the glass transition is occurring, a mixed mechanical property behavior will occur, with reduced modulus and increased ductility versus the glassy state but not as extreme as the rubbery state.

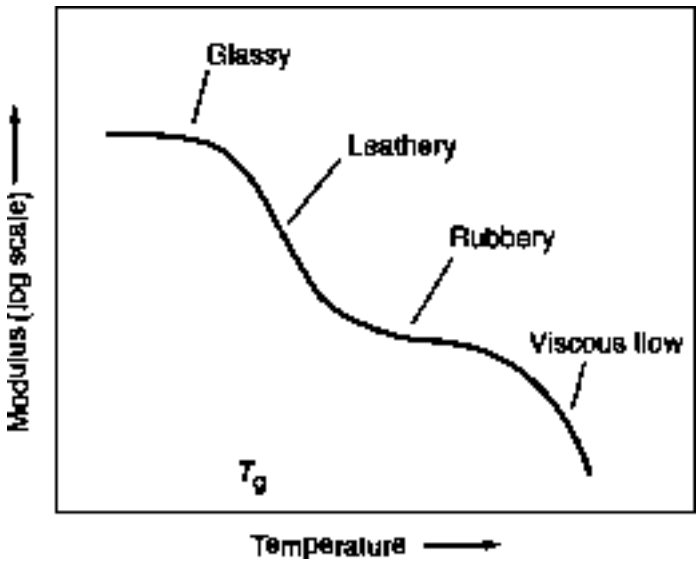


Fig. 4 Modulus versus temperature for a typical linear polymer. Source: Ref 19

This phenomenon also manifests itself in semicrystalline polymers, where the amorphous component of the material also exhibits a T_g. The T_g is always lower than the melting temperature (T_m) of a semicrystalline polymer. The extent to which mechanical properties are altered as temperature changes around T_g depends on the relative amounts of crystalline and amorphous material that exist in the polymer in question. Of course, when temperature is above both T_g and T_m, the material exhibits either rubbery or viscous fluid behavior depending on whether or not it is cross-linked.

The actual temperature range over which the glass transition phenomenon occurs will vary somewhat as the rate of deformation of the polymer changes. High-speed (high strain rate) deformation favors nonductile failure while low-speed (low strain rate) deformation favors more ductile failure. The actual numerical value of T_g will vary with the rate of testing used to make the measurement. Thus, when assessing temperature effects on failure mode, it is necessary to know how the environmental temperature compares to the polymer T_g. It is also necessary to factor in loading rate, especially when the environmental temperature is near T_g.

Reference cited in this section

19. A. Eisenberg, The Glassy State and the Glass Transition, Chap 2, *Physical Properties of Polymers*, 2nd ed., American Chemical Society, 1992

Effect of Environment on the Performance of Plastics

Donald E. Duvall, Engineering Systems Inc.

Conclusions

Environmental effects on polymeric materials cover a broad range of different behaviors. Plasticization, solvation, and swelling, involving diffusion of the chemicals into the polymer, do not alter the molecular structure irreversibly. Their effects can be reversed if the chemical is removed from the material, for example, by evaporation. Environmental stress cracking occurs without significant absorption of an environmental reagent by the polymer but will promote brittle fracture in normally ductile plastics at levels of stress or strain well below those that would usually cause failure at all. Actual degradation of a polymer, which reduces molecular weight and therefore mechanical properties, does not need to be particularly pervasive in order to be problematic. Degradation of a thin surface layer of material on a plastic part can facilitate premature failure or brittle failure under conditions where ductile failure would normally occur. All of this is further exacerbated by

the effects of changing temperature and strain rate. All of these conditions may need to be considered in determining cause of failure in a product manufactured from a polymeric material.

References

1. M.P. Stevens, *Polymer Chemistry: An Introduction,* 2nd ed., Oxford University Press, 1990, p 43

2. E.A. Grulke, Solubility Parameter Values, Chap. VII, *Polymer Handbook,* 4th ed., J. Brandup, E.H. Immergut, and E.A. Grulke, Ed., John Wiley & Sons, New York, 1999

3. *Physical Properties of Polymers Handbook,* J.E. Mark, Ed., AIP Press, Woodbury, NY, 1996

4. E. Miller, Properties Modification by Use of Additives, *Engineered Materials Handbook,* Vol. 2, *Engineering Plastics,* ASM International, 1988, p 493–507

5. A. Kumar and R. Gupta, *Fundamentals of Polymers,* Section 2.3, McGraw-Hill, 1998.

6. J.K. Sears and J.R. Darby, Solvation and Plasticization, Chap. 9, *Encyclopedia of PVC,* Vol 1, 2nd ed., Marcel Dekker, Inc., 1986.

7. M. Ezrin, *Plastics Failure Guide: Cause and Prevention,* Hanser Publishers, New York, 1996, p 22–23

8. J.-M. Vergnaud, *Polym.-Plast. Technol. Eng.,* Vol 20 (No. 1), 1983, p 1–22

9. F.W. Billmeyer, *Textbook of Polymer Science,* 3rd ed., Wiley-Interscience, New York, 1984, p 344

10. S.L. Rosen, *Fundamental Principals of Polymeric Materials,* Section 3.4, 2nd ed., Wiley-Interscience, New York, 1993

11. J.B. Howard, *SPE J.,* Vol 15, 1959, p 397

12. N. Grassie and G. Scott, *Polymer Degradation and Stabilization,* Cambridge University Press, 1985

13. U.W. Gedde et al., *Polym. Eng. Sci.* Vol 34 (No. 24), 1994, p 1773

14. P.K. So and L.J. Broutman, *Polym. Eng. Sci.,* Vol 22 (No. 14), 1982, p 888

15. J.C.M. de Bruijn, “The Failure Behavior of High Density Polyethylene Products with an Embrittled Surface Layer Due to UV Exposure,” Delft University of Technology, Delft, Netherlands, 1991.

16. L.J. Broutman, “Surface Embrittlement of Polyethylene,” Final Report to the Gas Research Institute, GRI-81-0030, Chicago, Nov 1981

17. S.-W. Choi, “Surface Embrittlement of Polyethylene,” Ph.D. dissertation, Illinois Institute of Technology, Chicago, 1992

18. S.W. Choi and L.J. Broutman, Proceedings of the 11th Plastic Fuel Gas Pipe Symposium, San Francisco, 3–5 Oct 1989, p 296–320

19. A. Eisenberg, The Glassy State and the Glass Transition, Chap 2, *Physical Properties of Polymers*, 2nd ed., American Chemical Society, 1992

Corrosion Failures of Industrial Refractories and Technical Ceramics*

Michel Rigaud, École Polytechnique

Introduction

*REFRACTORY MATERIALS are used in various industrial applications to line the interior of a multitude of vessels (mainly made of steels). The two primary objectives are to minimize heat losses and to prevent various fluids and solids from coming in direct contact with the outer skin (again, most of the time, made of steels).

Unlike thin organic linings and coatings, refractory linings may consist of one or many times two layers or more with thicknesses varying from at least 100 mm (4 in.) up to 500 mm (20 in.) or more. Refractories are either shaped materials (bricks and various other formats, including lances, tubes, nozzles, and so forth) or unshaped or monolithic materials (delivered to the users in bags, sacks, or other containers for bulk materials).

Technical ceramics recovered a very wide diversity of materials. This article mainly discusses structural ceramics used in gas turbine components, the automotive and aerospace industries, or as heat exchangers in various segments of the chemical and power generation industries.

The most common cause of premature failures of refractories is corrosion. However, corrosion is infrequently the cause of premature failure for technical ceramics. Corrosion of refractories occurs mainly when refractory materials are in contact with metallurgical slags or glasses (solids, S, + liquid, L, type of reaction), but atmosphere corrosion also occurs under specific circumstances (under reducing or oxido-reducing conditions), as, for example, under attack from carbon monoxide or alkalis and other condensable gases (S + gas, G, type of reaction).

Corrosion of technical ceramics is attributed to either S + G or S + L types of reactions (G for high-temperature combustible gases, in turbine-, diesel-, or other-fuel-engines; L mainly for molten salts or even aqueous solutions under special circumstances).

This article is divided into three sections. The first section, “Basic Principles,” covers fundamental aspects of chemical corrosion (as opposed to electrochemical corrosion). The second section is a compilation of corrosion resistance characteristics of specific classes of refractories and structural ceramics, and the third section is devoted to the prevention strategies that minimize corrosion failures of both classes of materials.

For the present purpose, refractories are defined as mainly coarse-grain ceramics made of either natural and/or synthetic raw materials, with an overall open porosity greater than 10%. Refractories are used as containment materials (containing heat and mass). Refractories for heat containment only are called insulating materials. Insulating materials are not covered specifically in the context of corrosion failures. The accepted definition for industrial refractories is, according to the ASTM (Ref 1), “Nonmetallic materials having those chemical and physical properties that make them applicable for structures, or as components of systems, that are exposed to environments above 538 °C (1000 °F).” Nowadays, refractories are made of not only oxides and nonoxide compounds (carbides, nitrides, and carbons), but also of metallic additions. With the new generations of “black refractories,” which contain 7% or more carbon materials (mainly natural graphite), adding one or two metallic additions to act as “antioxidants” or to protect carbon is mandatory.

The technical ceramics, on the other hand, constitute an ill-defined class of materials. They are often called “engineering ceramics” or “advanced ceramics,” subdivided into “structural ceramics” (which exhibit superior

mechanical properties as compared to traditional ceramics) and “functional ceramics” (which exhibit superior electrical, magnetic, or optical properties, again as compared to traditional ceramics). For the present purpose, technical ceramics are defined as fine-grain (as opposed to coarse) ceramics made of high-purity powders, oxides, and/or non oxides (same as for refractories), with very little batch-to-batch variability and as little as possible liquid-solid sintering. Technical ceramics nevertheless possess a fired density as close as possible to theoretical density (with small pores and very little apparent porosity—a major difference when compared to refractories).
The refractory materials to be considered can be classified as:

- *Basic refractories:* magnesia- or doloma-based, with or without graphite, with some chromia and alumina additions in the case of MgO, and magnesia and minor amounts of zirconia additions, in the case of doloma (CaO · MgO)
- *Alumina and alumino-silicate refractories:* high-alumina, with or without minor addition of graphite and carbon, magnesia, or silicon carbides; high-alumina with clays and then the fireclays and silica refractories
- *Others:* includes zircon-zirconia and silicon carbide refractories

The technical ceramics to be considered are high-purity materials of alumina or zirconia (stabilized), or silicon-carbide, silicon-nitride, or boron-nitride.
Refractories and structural ceramics are generally thought to be inert and corrosion resistant (as compared to metallic alloys), and they are, relatively so, at room temperature, under dry atmosphere, and over long time intervals. With increasing temperatures and specific chemical, mechanical, and physical gradients, the propensity to degradation increases rapidly.
Degradation, deterioration, decomposition, and wear are all words used to describe corrosion of these materials. Corrosion of refractories and ceramics is indeed a complex phenomenon to describe, and no single word properly fits. Corrosion involves a combination of different mechanisms, such as dissolution and invasive penetration (where diffusion, grain boundary, and stress corrosion may all be present), and oxidation-reduction reactions (where absorption, desorption, and mass transport phenomena all come into play).

Footnote

* *Adapted with permission from Corrosion of Refractories and Ceramics by M. Rigaud in *Uhlig's Corrosion Handbook*, 2nd Edition, John Wiley & Sons, 2000

Reference cited in this section

1. “Standard Terminology Relating to Refractories,” ASTM C 71, *Annual Book of ASTM Standards*, ASTM

Corrosion Failures of Industrial Refractories and Technical Ceramics*

Michel Rigaud, École Polytechnique

Basic Principles

It is generally accepted that corrosion of refractories and ceramics are the result of some specific chemical reactions rather than electrochemical reactions. It is then appropriate to estimate first the driving force or the propensity of a given reaction to occur, using the appropriate thermodynamic calculations. This should be done in two steps:

1. Verify the available thermodynamic data for the thermal stability of each constituent.
2. Make the appropriate thermodynamic calculations to estimate the free enthalpy variations $(\Delta G)_r$, for all possible reduction or oxidation (redox) reactions that may occur between constituents themselves and between constituents and the environment (gases and/or liquids).

To select the best material for a specific application, however, thermodynamic calculations are not enough. This is only the first part of the story. Identification of the mechanisms, which will control the degree of advancement of the reaction, is also duly required. Kinetic models and kinetic data then are needed.
In this context, the principles of penetration, dissolution, and spalling are considered.

Thermodynamic Calculations

To make the appropriate calculations, first evaluate the thermal stability of each constituent. Then, consider the melting and dissolution behavior under equilibrium conditions. Finally, assess the redox potentials, taking into account the partial pressure values of the initial constituents and of all the compounds that can be formed at different temperatures and encountered in the system considered.
Thermal Stability. Several polymorphic transitions may occur, and these can change the microstructural integrity of the solids as they are being heated. Some are reversible, and some are irreversible, with some well-known disruptive transformations, as for silica or zirconia. Other transitions are the decomposition of mixed oxides (e.g., $ZrSiO_4 \rightarrow ZrO_2 + SiO_2$), the devitrification of glassy phases, and the crystallization of high-temperature phases for amorphous or poorly crystallized ones (e.g., the graphitization of carbon). For such modifications, large volume variations may occur, creating sufficiently large compressive, tensile, or shear forces to cause microcracks, hence porosity, an important aspect of the corrosion of refractories.
Other volume changes or debonding may occur due to thermal mismatch when solids are heated. All noncubic lattice refractory compounds, for example, Cr_2O_3 (hexagonal); Fe_2O_3 (trigonal); and ZrO_2 (monoclinic), are susceptible to disruptive intercrystalline debonding, since they exhibit thermal expansion anisotropy. Another cause of debonding is the juxtaposition of two constituents of different properties, such as different thermal expansion coefficients (e.g., alumina-mullite, mullite-silica, and magnesia-chromite). In fact, in most refractories, phase boundary microcracking can be expected.
Melting Behavior and the Use of Phase Diagrams. For many years, compositions of refractories were limited to individual minerals as mined. With time, it was recognized that nature did not proportion the oxide contents of the minerals in the most suitable ratio for optimum refractory performance. As pure oxides became available and affordable in terms of cost (first alumina, then magnesia), improvements in performance (in particular, corrosion resistance) became possible. The use of phase diagrams was then recognized as a very good research tool to accelerate such improvements.
Phase diagrams are used to design refractories and to understand the role of slags (of the same nature as the liquid phase formed in the refractories). The early book by Muan and Osborn (Ref 2) is a reading requirement for new researchers in this field, as are the reviews by Kraner (Ref 3) on phase diagrams for fired refractories and by Alper et al. (Ref 4), on fusion-cast refractories. Phase equilibria in systems containing a gaseous component have also been examined; in particular for the effects of oxygen partial pressure on phase relations in oxide systems. The relative importance of these issues has greatly increased with the advent of the so-called carbon-bonded refractories, magnesia-dolomite and alumina-based systems with carbon and graphite. Also, the roles of oxycarbides, oxynitrides, and silicon-aluminum oxynitrides, as well as metals and alloys in refractories, have been documented. These aspects are treated in the following paragraphs and sections.
In general, the corrosion resistance of a refractory is high if formation of low-melting eutectics and of a large amount of liquid can be avoided. Phase diagrams can be used to predict the formation of these phases. Of course, the better the system is defined, the easier it is to make an accurate prediction. There are, nevertheless, many limitations; for example, phase diagrams are readily available for no more than three constituents, while in practical cases of magnesia or dolomite basic refractories, one needs to characterize the CaO-MgO-FeO-Fe_2O_3-Cr_2O_3-SiO_2 system. At any rate, for the basic refractories, a particularly important feature is the CaO/SiO_2 ratio and its effect on the quantities of low-temperature phases to be expected.
Oxidation and Reduction Behavior. As for melting or solid-liquid reactions, the thermodynamic calculations of gas-solid reactions are a powerful tool for describing the stability of refractory and ceramic materials (in particular, the carbon-containing refractories that are now very widely used). Gas-solid reactions are important

not only for dealing with the direct oxidation of carbon in air but also for evaluating the reduction of aggregates (MgO, Al_2O_3, SiO_2, etc.) by carbon (indirect oxidation of carbon) and the role of the antioxidants, being more reducing than carbon in most systems. A powerful use of thermodynamics is to assess oxidation-reduction behavior of multicomponent systems using elaborate computer codes, such as Chem Sage or Solgasmix (Ref 5, 6). Graphical displays such as Ellingham diagrams and volatility and predominance diagrams are often used to describe the simplest systems of the type $S_1 + G_1 \rightarrow S_2 + G_2$ (Ref 7). For the Ellingham diagrams, all the condensed phases of the reactions are assumed to be pure phases, and therefore, at unit activity, but deviations from unit activity are common. Corrections must be applied. Predominance diagrams and volatility diagrams are graphical representations in which the gaseous products are considered at various nonstandard conditions, and these diagrams are used for complex systems involving several gases (e.g., metal-oxygen-carbon-nitrogen, under a wide range of partial pressures). These diagrams, the equivalent to Pourbaix diagrams for studying corrosion of a solid in aqueous media, are used to predict the direction of reactions and the phases present.

From thermodynamic calculations, it is possible to explain why and how redox variations are harmful to any refractory containing iron oxides as impurities; how the reduction of magnesia to magnesium gas and the reoxidation to MgO can be used with great advantage for MgO-carbon bricks (a destruction-reconstruction mechanism that leads to the dense-zone magnesia formation theory); how nonoxide refractories (and nonoxide structural ceramics) without exception are subject to high-temperature oxidation; and finally, how, in the presence of alkalis, refractory chromites (Cr^{3+}) can be oxidized in part to toxic chromates (Cr^{6+}). The alkali chromites ($Na_2Cr_2O_7$ or Na_2CrO_4) are water leachable and therefore present an environmental contamination risk in disposal sites.

Kinetic Considerations

Penetration, dissolution, and spalling are the most important phenomena that control the kinetics of the corrosion of industrial refractories and technical ceramics. They apply as well to the S + L reactions as to the solid S + G reactions.

Liquids can be either slags or fluxes, molten salts, or molten metals, each presenting their own peculiarities. Slags are characterized by their basicity/acidity ratio (CaO/SiO_2 ratio either greater or lower than 2 to 1) and fluidity or viscosity (very much a function of temperature and overheat, measured by $\Delta T = T_{service} - T_{melting}$ of the lowest melting eutectic compounds in the system). Molten salts are known for their low melting temperature, high fluidity, and high fugacity (high volatility of alkalis). Molten metals are less reactive toward refractories but are not inert.

The solid refractory material, S, has already been defined as a multiphase material having a texture with intricate surface properties, most often used under a thermal gradient and usually containing some inherent amount of liquid (ℓ) at the hot face, where ℓ may be equivalent to L, in terms of affinity.

Various hot gases, reactive or nonreactive, with or without dusts, are represented by G. The reactivity of G can be determined by starting with the acid-base series shown hereafter, noting that the strongest acids and bases are volatile, and by using thermodynamic data in predominance and volatility diagrams:

Order listed	Acids and bases
Gases	
From most acidic to most basic	**N_2O_5 (G)**
	SO_3 (G)
	SO_2 (G)
	$CO_2 \cdot ZrO_2$
	$B_2O_3 \cdot ZrO_2$
	$V_2O_5 \cdot ZrO_2$
Solids	
From most acidic to neutral	**SiO_2**
	TiO_2
	ZrO_2
	Fe_2O_3
	Cr_2O_3
	Al_2O_3

From neutral to most basic	**FeO**
	NiO
	MnO
	MgO
	CaO
	Na_2O (S or G)
	K_2O (S or G)

Penetration. It is useful to distinguish between physical penetration and chemical invasion. Physical penetration, without dissolution at all, occurs when a strictly nonwetting liquid is forced into the pores of a solid by gravity or external forces. Chemical invasion occurs when dissolution and penetration are tied together. Both physical and chemical penetration are favored by effective liquid-solid wetting and by low-viscosity liquid. Silicates, particularly silicate glasses, are usually viscous; simple oxidic compounds and basic slags are less viscous. Halides and elemental molten metals are, in general, the most fluid liquids.

Penetration is the result of an interplay between capillary forces (surface tension), hydrostatic pressure, viscosity, and gravity. Mercury penetration in a capillary glass tube is the best example of physical penetration. The rate of penetration in a horizontal pore $d\ell/dt$, with a pore radius, r, is given by the following expression:

$$\frac{d\mathbf{l}}{dt} = \frac{r_{\mathbf{l}t}\cos}{4h\mathbf{l}}$$

where $\gamma_{\ell g}$ is the surface tension of mercury in air, $\boldsymbol{\theta}$ the wetting angle of mercury on the glass wall, η the viscosity of the penetrating liquid, Hg, over a length, ℓ, at time, t. This expression is valid only at constant temperature, whereas $\gamma_{\ell g}$ and η are liquid properties greatly influenced by temperature. Such an equation has often been used to describe the penetration of liquids in refractories without distinction between physical and chemical invasion. However, in the case of chemical invasion, the penetration-dissolution causes changes in composition (which, in turn, affect the value of $\gamma_{\ell g}$ and η) and changes in porosity geometry.

When the pore size distribution is narrow (i.e., pores of the same size), penetration and filling of the porosity by capillarity produces a relatively uniform front moving gradually from the hot face and remaining parallel to it. When pore size distribution is wide (i.e., very large and very small pores) or when open joints, cracks, or gaps between bricks in a refractory wall are accessible at the hot face, rapid and irregular liquid intrusions occur. There are many penetration paths in a refractory, and the texture of the material is of primary importance. It is important to distinguish between interconnected versus isolated porosity, between open and total porosity, and between pore sizes and unbounded boundaries between grains (aggregates, and/or matrix) due to thermal mismatches during heating.

For a given temperature gradient, the pertinent eutectic temperature of the penetrating liquid determines its maximum liquid penetration depth.

Dissolution. The simplest case of pure dissolution is represented by the reaction $S_1 + L_1 \rightarrow L_2$, where L_2 is a solution $L_1 + S_1$, S_1 having a continuous surface. A general equation useful to describe such a dissolution process is:

$$j = \left\{ k\left(1 + k\frac{d}{D}\right)^{-1} \right\}(C_{sat} - C_{\infty}) \qquad \text{(Eq 1)}$$

where j is the rate of dissolution per unit area at a given temperature, T; K is the surface reaction rate constant; δ is the thickness of the boundary layer in the liquid phase; C_{sat} is the concentration of the dissolving solid, in the liquid, at the interface; C_{∞} is the concentration of the dissolving phase in the bulk of the liquid; and D is the effective diffusion coefficient in the solution for the exchange of solute and solvent.

The parameter δ is further defined by the expression:

$$d = \frac{C_{sat} - C_{\infty}}{(dc/dy)} \qquad \text{(Eq 2)}$$

where (dc/dy) is the concentration at the interface.

The dissolution rate, j, may be visualized as the ratio of a potential difference (C_{sat} - C_{∞}) divided by a resistance term:

$$K^{-1}\left(1+\frac{kd}{D}\right)$$

Three different cases will be briefly treated.

The first case is when $K >>> D/\delta$, that is, when the chemical reaction takes place so rapidly at the solid-solvent interface that the solution is quickly saturated and remains so during the dissolution process. In this case, the dissolution rate, j, is controlled by mass transport. Equation 1 reduces to Eq 3:

$$j = \frac{D}{d}\left(C_{sat} - C_{\infty}\right) \qquad \text{(Eq 3)}$$

The transport process is enhanced by convection due to density differences between bulk and saturated solution (natural convection) and/or by the hydrodynamics of the system under forced convection. Expressions for the boundary layer, δ, have been derived from first principles, for both natural (or free) and for forced convection, for a variety of simple geometries, and Eq 3 has been validated many times. This process is often called direct dissolution.

The second case is when $K <<< D/\delta$, that is, when the dissolution rate is phase-boundary controlled as opposed to mass-transport controlled. Equation 1 reduces to Eq 4:

$$j = K(C_{\text{sat}} - C_{\infty}) \qquad \text{(Eq 4)}$$

The phase-boundary reaction rate is then fixed by the movement of ions across the interface and, hence, is governed by molecular diffusion. The effective diffusion length over which mass is transported is proportional to $(Dt)^{1/2}$ and, therefore, the change in thickness of the specimen, proportional to the mass dissolved, varies with $t^{1/2}$. This process is often called indirect dissolution.

The third case is when K **;** D/δ, both phase boundary and mass-transport are controlling; in this case of mixed control, the potential difference $(C_{\text{sat}} - C_{\infty})$ can be seen as divided into two parts. $(C_{\text{sat}} - C^*)$ is the part that drives the phase boundary condition, and $(C^* - C_0)$ is the part that drives the transport process, so as to keep the dissolution rates for each process equal.

The dissolution rate is extremely temperature sensitive, largely determined by the exponential temperature dependence of diffusion, and can be expressed by the following equation;

$$j_c = A\exp\left[-B\left(\frac{1}{T}-\frac{1}{T_1}\right)\right] \qquad \text{(Eq 5)}$$

where A is the dissolution rate at temperature T_1 and B is a model constant.

In most practical dissolution problems, the reaction of a solid in a solvent leads to a multicomponent system; the chemical composition is no longer defined by a single concentration. There is not a single saturation composition at a given temperature; instead of pure dissolution, the reaction is of the type $S_1 + L_1 \rightarrow L_2 + S_2$ with the formation of S_2 causing the interface composition to change.

For porous solids such as refractories with open porosity and with matrix materials being fine and highly reactive, both dissolution and penetration occur; hence, most slagging situations involve chemical attack of the matrix or low-melting constituent phases, which disrupts the structure and allows the coarse-grained aggregates to be carried away by the slag movement. When penetration is more important than dissolution, structural spalling, another mechanism of degradation, needs to be considered.

Structural or Chemical Spalling. While *spalling* is a general term for the cracking or fracture caused by stresses produced inside a refractory, *chemical* or *structural spalling* is a direct consequence of corrosion penetration. It should not be confused with pure thermal or pure mechanical spalling. Structural spalling is a net result of a change in the texture of the refractory, leading to cracking at a plane of mismatch, at the interface of an altered structure and the unaffected material.

When slag penetration does not cause direct loss of material, the slag partially or completely encases a volume of refractory, reducing its apparent porosity (by sometimes more than one half) and causing differential expansion with the associated development of stresses. As a result, there is a degradation in material strength and stiffness, the appearance of microcracks and, eventually, totally disruptive cracks parallel to the hot face of the lining. This is structural spalling.

Quantification of the mass of spalled materials from the hot face has been attempted by Chen and Buyukozturk (Ref 8), who combined slag dissolution and spalling effects into one expression in terms of the residual

thickness, X, and the rate of thickness decrease. They proposed expressions as a function of the hot-face temperature, T_H, and the maximum depth of slag penetration D_P, referred to the location of the hot face at time, t_i, during the (i - 1)th and ith spalling.

The same authors have developed a very interesting notion to approximate the value of D_P, using a critical temperature criterion; they postulated that, for a given system, slag penetration occurs only when temperature reaches a critical value, T_c, which could be related to the temperature above which a certain percentage of liquid phase may still exist in the refractory (always under a thermal gradient).

Corrosion by Gases and Dusts versus Liquids. For corrosion by gases and dusts, the same qualitative relationships are applicable as for gas and liquid penetration into a refractory, except that wetting is not a factor. In gas corrosion, the driving force is the pressure gradient. Gas penetration is less rapid, in terms of mass, than that of wetting liquids; reactions do not necessarily commence at the hot face. However, the depth of penetration of gases can be much greater than that of liquids.

Moving down a thermal gradient, gases may condense locally to form a liquid, which then may initiate a dissolution-corrosion process and migrate still further into the solid. Alternatively, gases may react chemically to form new compounds (liquid or solid). Once condensation has occurred, the basic criteria for liquid corrosion apply. The condensed liquid fills the pores and debonds the ceramic material, creating thermal-expansion mismatches, weakening, softening, swelling, and slabbing.

Sulfur dioxide, usually from combustion or smelting operations, provides good examples of the condensation-corrosion patterns experienced in practice. Condensation leads to the formation of sulfate in magnesia at 1100 °C (2000 °F), in lime at 1400 °C (2550 °F), and in alumina at 600 °C (1100 °F). When alkali vapors (Na_2O) are present with SO_2, condensation of Na_2SO_4 overlaps with SO_2 attack. Volatile chlorides also play a role.

When dusts are carried out in hot gases, they may lead to deposition or abrasion. Deposition leads to scaling or caking with either beneficial or life-shortening consequences. Dusts entrained in gases that condense within refractories or ceramics may facilitate the gradual adhesion of solid particles to the working surface. The resulting buildup can be either a hard scale or lightly sintered cake acting as a protective barrier. In less obvious cases, dust and condensing gases fuse together to create an invasive liquid. In other instances, scaling and caking of refractory walls may not be acceptable to a particular process, and periodic descaling operations can lead to degradation of refractories.

Although no deposition may occur, abrasion can take place when refractory or ceramic linings are bombarded by dusts. The extent of abrasive wear depends on the impinging particle size, shape, hardness, mass, velocity, angle of impact, the fluid dynamics of the system, and the viscosity of the eluant gas, among other parameters. Abrasion exacerbates corrosion.

Exposure of a refractory or a ceramic either to a liquid or to its saturated vapor is thermodynamically the same. The differences lie in the transport mechanism. In approximate order of decreasing thermodynamic power, the oxidizing gases are NO_x, Cl_2, O_2, HCl, CO_2, H_2O, and SO_2, and the reducing gases are reactive metal vapors, NH_3, H_2, C_xH_y, CO, common metal vapors, and H_2S.

References cited in this section

2. A. Muan and E.F. Osborn, *Phase Equilibria Among Oxides in Steelmaking,* Addison-Wesley, Reading, MA, 1965

3. H.M. Kraner, *Phase Diagrams, Materials Science and Technology,* Vol 6-II, A.M. Alper, Ed., Academic, 1970, p 67–115

4. A.M. Alper, R.C. Doman, R.N. McNally, and H.C. Yeh, *Phase Diagrams Materials Science and Technology,* Vol 6-II, A.M. Alper, Ed., Academic, 1970, p 117–146

5. G. Ericksson and K. Hack, *Metall. Trans. B,* 1990, p 1013

6. G. Ericksson, Chem. Scr., 8, 1975, p 100

7. V.L.K. Lou and A.H. Heuer, *High-Temperature Corrosion of Technical Ceramics,* R.J. Fordham, Ed., Elsevier Applied Science, 1990, p 33–52

8. E.S. Chen and O. Buyukozturk, *Am. Ceram. Soc. Bull.*, Vol 64 (No. 7), 1985, p 995

Michel Rigaud, École Polytechnique

Corrosion of Specific Classes of Refractories and Technical Ceramics

A simple, all-encompassing, general theory of corrosion of refractories and structural ceramics does not exist. The interplay between dissolution, penetration, and texture is not sufficient to take into account all the interactions possible with so many parameters involved. This complexity may provide the reason for the shortage of reliable comparative information on the resistance of ceramics to chemical corrosion. Carniglia and Barna (Ref 9), have provided useful compilations on refractories, and McCauley (Ref 10) and Lay (Ref 11), on technical ceramics.

Industrial Refractories

To assess the corrosion resistance of different types of refractories, a rule of thumb is to consider the temperature limits above which corrosion is considered to be excessive, as established considering the expected durability. From one application to another, this durability does vary greatly, as does the mode of degradation: corrosion-dissolution, penetration and structural spalling, or both. In Table 1, two cases are distinguished: slagging resistance and hot gas corrosion resistance. The first case is represented by a reaction of the type S + L, where L can be either a slag (from ironmaking, steelmaking, copper smelting, etc.) or a flux (from other nonferrous metallurgy), either basic ($CaO/SiO_2 > 2$) or acidic ($Ca/SiO_2 < 2$, on a molar ratio), where S is a given refractory with a given texture. The second case is represented by a reaction of the type S + G, where G can be various hot gases, with or without dusts, either oxidizing (oxygen, air, or CO_2) or reducing (CO, H_2). In Table 1, six types of refractories, from basic magnesia to acid silica and neutral zircon-zirconia are considered.

Table 1 Corrosion resistance of refractories based on the maximum temperature criterion

Type of refractory	Temperature limit in indicated atmosphere, °C (°F)			
	Slags and fluxes		Hot gases	
	Basic	**Acid**	**Oxidizing**	Reducing
Group 1				
Magnesia (M)	1700 (3090)	NR	>2000 (>3630)	**1700 (3090)**
Doloma (D)	1700 (3090)	NR	>1800 (>3270)	**1700 (3090)**
(M), (D), or (M + D) + graphite	1800 (3270)	1700 (3090)	800 (1470)	**1700 (3090)**
Magnesia-chrome (MK)	1700 (3090)	1600 (2910)	1800 (3270)	**NR**
Group 2				
High-alumina (A)	1600 (2910)	1600 (2910)	1900 (3450)	**1900 (3450)**
(A) + graphite	1700 (3090)	1700 (3090)	600 (1110)	**1700 (3090)**
Group 3				
Clays + (A) (65% Al_2O_3)	1300 (2370)	1300 (2370)	1450 (2640)	**NR**
Clays + (A) + C	1550 (2820)	1550 (2820)	600 (1110)	**1700**

				(3090)
Group 4				
Super-duty fireclays (F)	1200 (2190)	1250 (2280)	1400 (2550)	**NR**
Medium-low duty (F)	NR	1100 (2010)	1200 (2190)	**NR**
Group 5				
High-silica	NR	1400 (2550)	1750 (3180)	**1600 (2910)**
Group 6				
Zircon-zirconia	1450–1600 (2640–2910)	1500–1700 (2730–3090)	>2000 (>3630)	**1800 (3270)**
Silicon carbide	**<1400 (<2550)**	**1500 (2730)**	**1200 (2190)**	**1600 (2910)**

NR, not recommended. Temperature limit not reached for “normal” durability

Basic Refractories. Basic bricks are commonly used for the refractories in various industries; for example, in metallurgical industries, in steelmaking, and so forth. Magnesia is the most commonly used basic refractory, followed by doloma. Magnesia bricks are made of various magnesia raw materials (sinters from natural magnesite, sinters from seawater magnesia, or grains of fused magnesia). Many other types of magnesia bricks are used, in combination with chrome (MK), alumina (NA), and doloma (MD). The most corrosion resistant magnesia bricks are made of fused magnesia grains. Magnesia-chrome bricks were the best bricks available before magnesia-carbon bonded bricks containing flake graphite additions were developed for their slagging resistance. Dolomitic refractories are made of burnt natural dolomite, usually low in iron oxide (0.3%) and alumina (0.2%). As for magnesia, very efficient doloma-carbon bricks have been developed. As shown in Table 1, doloma performance may match that of magnesia for specific applications.

High-Alumina Refractories. High-alumina refractories are those that contain more than 65% Al_2O_3, up to 92–95% Al_2O_3 composed of α-alumina (pure Al_2O_3, 99.3+ %), and mullite $(Al_2O_3)_3 \cdot (SiO_2)_2$ or calcined bauxites. For high corrosion resistance, electrofused products of alumina or mullite should be considered. As for any type of refractory, there are many different classes of products: fired and unfired, with many different bond systems, a commonly used one being a phosphate chemical bond system for common purposes. Corrosion-dissolution resistance is improved with increasing alumina content, but slag penetration resistance of such bricks then decreases. Hence, the structural spalling resistance also decreases. Improvements have been achieved with addition of chrome or silicon carbide, and, more importantly, of carbon.

High-Alumina Refractories with Clay. High-alumina bricks with clay refractories have alumina content between 45 and 65%; they may contain calcined kaolin, andalusite, or mullite mixed with calcined fire clays (chamotte) and binders. Corrosion resistance varies greatly with mineral composition and relatively low refractoriness (measured by the lowest temperature at which at a liquid phase is present in the material). The maximum temperature limit, as indicated in Table 1, may not apply for specific applications.

Fireclay Refractories. Super-duty, medium-duty, and low-duty fireclays are products with alumina content varying between 40 and 45%, 30 and 40%, and 25 and 30%, respectively. They are used in a wide variety of kilns and furnaces because of their low price but are limited to low-temperature applications (1200 to 1250 °C, or 2190 to 2280 °F).

Silica Refractories. Silica bricks contain typically >93% SiO_2 with minor amounts of lime (CaO <3.5%), alumina (Al_2O_3 <2%), and iron oxide (Fe_2O_3 <1.5%). They have high corrosion resistance to acid slags but are sensitive to high temperatures and to thermal shock, even at medium temperature levels. Bricks with reduced porosity and lower alumina content may be used to prevent rapid degradation in the presence of alkalis, when hot gas corrosion prevails. However, fused amorphous silica products exhibit very good thermal shock properties, as long as recrystallization can be prevented.

Zircon-Zirconia and Silicon Carbide Refractories. Among the several available zircon-zirconia refractories, zircon bricks containing mainly zircon sand ($ZrSiO_4$) have a good wear resistance and poor wettability by molten metals at processing temperatures up to 1450 °C (2640 °F).

Zircon is decomposed to zirconia and silica at 1540 °C (2800 °F). The most common zircon-zirconia bricks are the very special electrofused, A-Z-S bricks, containing alumina-zirconia-silica and made especially for glass-melting furnaces. Refractory materials with high zirconia content require the use of stabilized zirconia (CaO or MgO in solid solution) to avoid cracking on heating, due to monoclinic to tetragonal transformation in the

temperature range of 1000 to 1200 °C (1830 to 2190 °F), with a large volumetric change. Properly prepared zirconia shapes have very high corrosion resistance and are not wetted by metals including steels.
Silicon carbide refractories are the nonoxide refractories most often used with carbon blocks, usually under reducing conditions. Many different types of silicon carbide refractories are available. They have the following advantages over oxide refractories: high thermal conductivity, superior spalling resistance, superior abrasion resistance (and, hence, high corrosion resistance against nonoxidizing slag, as in some nonferrous metallurgical applications). Properties and costs of SiC refractories vary greatly, depending on the sintering aid used, but they do tolerate many more "impurities" than do structural SiC ceramics, which are discussed in the following sections.

Technical Ceramics

The fundamental principles of corrosion of industrial refractories also apply to corrosion of structural ceramics, in both the S + L and S + G cases. Nevertheless, as porosity of structural ceramics is much lower, dissolution is more important than penetration. For the S + L case, the corrosion resistance of structural ceramics to molten salts determines their usefulness as components in the chemical industry, such as pumps and heat exchangers. The S + G type of reaction is important for structural ceramics used as kiln furniture in industrial furnaces and as components in gas turbine and diesel engines.
Corrosion by Molten Salts. Comparative information is presented in Table 2 on the corrosion resistance to fused salts, alkalis, and low-temperature oxides for two different classes of structural ceramics, oxides and nonoxides. The data in Table 2 show that both dense oxides and nonoxides are susceptible to attack at low and intermediate temperatures. The normally protective SiO_2 layer that forms on SiC and Si_3N_4 may explain why such ceramics have a poor corrosion resistance in basic salts.

Table 2 Corrosion resistance of structural ceramics to fused salts, alkalis, and low-melting oxides

Ceramics at purity (>99.5%)	Resistance to indicated material at indicated temperature[(a)], °C (°F)							
	NaCl	**NaCl + KCl**	**KNO_3**	**Na_2CO_3**	**Na_2SO_4**	**KOH**	**Na_2O**	V_2O_5
Alumina	A, 1000 (1830)	A, 800 (1470)	A, 400 (750)	A, 900 (1650)	A, 1000 (1830)	B, 500 (930)	B, 500 (930)	**C, 800 (1470)**
Zirconia (stabilized)	…	C, 800 (1470)	…	C, 900 (1650)	A, 1000 (1830)	B, 500 (930)	C, 500 (930)	**C, 800 (1470)**
SiC (reaction bonded)	B, 900 (1650)	C, 800 (1470)	A, 400 (750)	C, 900 (1650)	C, 1000 (1830), air	C, 500 (930), air	C, 600 (1110)	**C, 800 (1470)**
Si_3N_4 (RBSN)	…	…	A, 400 (750)	C, 750 (1380), air	C, 1000 (1830), air	C, 500 (930)	C, 500 (930)	**C, 800 (1470)**
Si_3N_4 (HPSN)	A, 800 (1470)	B, 900 (1650)	…	C, 900 (1650), air	B, 1000 (1830), air	C, 500 (930)	B, 500 (930)	**C, 800 (1470)**
BN (HP)	**…**	**…**	**…**	**C, 900 (1650), air**	**C, 1000 (1830), air**	**C, 500 (930)**	**C, 500 (930)**	**…**

(a) A, resistant to attack up to temperature indicated; B, some reaction at temperature indicated; C, appreciable attack at temperature indicated
Values have been compiled from the combination of references (Ref 10, 11, 12) and refer to the chemical inertness of pure crystalline materials, with purity >99.5%. For this reason, the data in Table 2 are of limited generality. For example, it is to be expected that ceramics with less purity will be attacked more rapidly. The code A, B, C, should be used only for comparing the materials considered. It is to be assumed that the information is based on short-term observations at ambient pressure. McCauley (Ref 10) provides a more complete coverage of the subject, including attack by glasses, aqueous solutions, and molten metals, but no further generalizations can be made from the specific studies that they reviewed because of limitations in the scope of the studies.

Corrosion by Hot Gases. The S + G reactions discussed subsequently refer to oxidation at high temperatures in air and oxygen, gaseous corrosion in the presence of condensed deposits, and reduction or oxidation by hot gases, CO (G), H_2 (G), H_2S (G), and H_2O (G). High-temperature oxidation in air or oxygen is the most important type of corrosion for carbide- and nitride-based ceramics. Although the corrosion of structural ceramics in hot gases is similar in some aspects to the oxidation of metals and the same theories are used to describe the processes, there are some important differences:

- Ceramics have a much higher oxidation resistance than do metals.
- Because oxidation occurs at much higher temperatures, the corrosion products formed are liquids and gases, as well as solids; this can lead to either weight gains or losses, causing practical experimental difficulties of interpretation.
- The textures of ceramics are not as homogeneous as those of metals, and so it is very difficult to determine the controlling mechanism of the oxidation process without considering both the manufacturing characteristics and the intrinsic nature of the raw materials used.

Gaseous corrosion in the presence of condensed phase deposits, in combustion applications, is the second most important type of corrosion, with sodium sulfate and vanadate being the most corrosive deposits. Nonoxide ceramics are particularly sensitive to this type of environment.
In Table 3, the corrosion resistance of dense structural ceramics to hot gases are presented using the same code as in Table 2. The oxidation resistance of nonoxide ceramics is clearly less than that of alumina and zirconia.

Table 3 Corrosion resistance of structural ceramics to hot gases

Ceramics at purity (>99.5%)	Resistance to indicated gas at indicated temperature[(a)], °C (°F)					
	Air	**Steam**	**CO**	**H_2**	**H_2S**	F (gas)
Alumina	A, 1700 (3090)	A, 1700 (3090)	A, 1700 (3090)	A, 1700 (3090)	…	**…**
Zirconia (stabilized)	A, 2400 (4350)	C, 1800 (3270)	…	…	…	**…**
SiC (reaction bonded)	A, 1200 (2190)	B, 300 (570)	A, >1000 (>1830)	A, >1000 (>1830)	A, 1000 (1830)	**A, >800 (>1470)**
Si_3N_4 (reaction bonded)	B, 1200 (2190)	A, 220 (430)	A, >800 (>1470)	…	A, 1000 (1830)	**…**
Si_3N_4 (hot pressed)	B, 1250 (2280)	A, 250 (480)	A, >900 (>1650)	A, >800 (>1470)	A, 1000 (1830)	**A, >1000 (>1830)**
BN (hot pressed)	**C, 1200 (2190)**	**C, 250 (480)**	**A, 2000 (3630)**	**A, >800 (>1470)**	**…**	**…**

(a) A, resistant to attack up to temperature indicated; B, some reaction at temperature indicated; C, appreciable attack at temperature indicated

Alumina. Alumina is considered to be one of the most versatile materials that resist corrosive attack by acids, alkalis, molten metals, molten salts, as well as oxidizing and reducing gases, even at high temperatures. It is widely used in many high-temperature laboratory and pilot-plant scale applications, such as for crucibles, tubes, and special shapes.

Zirconia and Thoria. Thoria is the most thermodynamically stable oxide and the one with the highest refractoriness, with a melting point of about 3300 °C (5970 °F). Unfortunately, it is not a readily available material, being used essentially in the atomic energy industrial sector. Zirconia is also a very stable pure oxide, but it has to be used only when stabilized, fully or partially, with CaO, MgO, or Y_2O_3 additions. Such additions make the solid solutions • ZrO_2-X• more vulnerable to acid attacks and to alkalis.

Silicon Carbide. Silicon carbide has very good resistance to acidic solutions (even HF) and molten metals but is prone to attack under severe alkaline (basic) conditions, in particular in air. The SiC ceramics, produced by reaction sintering and containing free silicon, are more susceptible to oxidation and chemical corrosion than is single-phase SiC obtained by hot pressing. The SiC is very susceptible to attack by alkali-containing phases, such as sodium or potassium sulfate. On oxidation, SiC forms a layer of SiO_2, but the possibility of forming

SiO (G) increases with temperature and when the partial pressure of oxygen is decreased. The transition between passive oxidation (SiO_2 protective layer) and active oxidation [SiO (G), without protection] occurs at low pressures and high temperatures. SiC ceramics are more sensitive to hot corrosion than is Si_3N_4 (Ref 13).

Silicon Nitride. Thermodynamic calculations of equilibria in the system Si_3N_4-O_2 reveal that several reactions resulting in the formation of solids SiO_2, Si_2N_2O, SiO, and gaseous N_2, N_2O, NO, SiO, and SiN are possible. Reactions leading to the formation of SiO_2 (S) are characterized by the most negative change in free enthalpy, but at high temperatures and low-oxygen partial pressure, the possibility of reactions leading to the formation of SiC increases (significant at 1300 °C, or 2370 °F, and higher). The notion of passive and active Si_3N_4 oxidation has been well documented, but the controlling factor can be either diffusion of oxygen to the Si_3N_4/SiO_2 interface, diffusion of nitrogen to the SiO_2/air interface, or diffusion of impurity ions and sintering aids from the inner layers to the surface of the ceramic. A detailed survey of the subject is presented in Ref 14.

For hot corrosion to complement the data listed in Table 2, only one significant result obtained on reaction-bonding silicon nitride (RBSN) will be recalled. Although Si_3N_4 exhibits high corrosion resistance in the stream of combustion products of pure fuel (10^{-5}% Na and V, 0.5%) below 1400 °C (2550 °F), it corrodes rapidly at 900 °C (1650 °F) when fuel contains 0.005% Na, 0.005% V, and 3% S. In liquid media, Si_3N_4 resists attack by acids up to ~100 to 200 °C (~210 to 390 °F). The Si_3N_4 exhibits higher resistance in alkaline solutions than in alkali melts and is rapidly attacked by (Na_2SO_4, Na_2SO_4 - NaCl, Na_2SO_4 + V_2O_5) melts but has good resistance to chloride melts NaCl + KCl up to 600 °C (1110 °F), and up to 1100 °C (2010 °F) in NaF.

Boron Nitride. Boron nitride (BN) ceramics are usually prepared by hot pressing and normally contain a small proportion of boric oxide, a useful impurity that helps pressing but is detrimental for its oxidation behavior at high temperatures. The BN ceramics are reported to be attacked by strong acids but are relatively inert towards alkalis. The nonwetting behavior of BN, similar in some ways to pure graphite, may account for this behavior.

Boron nitride starts to oxidize at about 700 °C (1290 °F). Oxidation mechanisms differ from Si_3N_4 and SiC, because in this case the oxide layer firstly formed, B_2O_3, is liquid and tends to evaporate to BO (G), BO_2 (G), and B_2O_3 (G) as temperature increases. Under reducing conditions, pure BN may be usable up to 1800 °C (3270 °F). The most valuable characteristic of BN is its resistance to wetting by metals and alloys.

References cited in this section

9. S.C. Carniglia and G.L. Barna, *Handbook of Industrial Refractories Technology*, Noyes Publication, Park Ridge, NJ, 1992

10. R.A. McCauley, *Corrosion of Ceramics*, Marcel Dekker, New York, 1995

11. L.A. Lay, *Corrosion Resistance of Technical Ceramics*, HMSO, London, 1983

12. R. Morrell, *Handbook of Properties of Technical and Engineering Ceramics*, Part I, HMSO, London, 1985

13. D.S. Fox and J.L. Smialek, “Burner Rig Hot Corrosion of Silicon Carbide and Silicon Nitride,” *J. Am. Ceram. Soc.*, Vol 73, 1990 p 303–311

14. Y.G. Gogotsi and V.A. Lavrenko, *Corrosion of High-Performance Ceramics*, Springer-Verlag, Berlin, 1992, p 110–156

Corrosion Failures of Industrial Refractories and Technical Ceramics*

Michel Rigaud, École Polytechnique

Strategies for Analysis and Prevention of Failures

Minimizing corrosion of refractories and technical ceramics under severe conditions at high temperature is a very important task that requires the following three sets of problems to be addressed simultaneously:

- Material selection considering both intrinsic and extrinsic characteristics, such as the design and configuration of the component, and overall learning from past experience, when appropriate
- Installation methods and maintenance procedures, considering that each material is a part of a bigger identity and taking a systematic approach
- Process control to minimize variability and extreme values, defining the key variables in the operation

Material Selection

For refractories and for technical ceramics as well as for any other type of material, the selection problem involves one of two situations:

- Selection of the materials to meet the needs of a new process or a new design
- Substitution for an existing design to improve reliability and performance

To tackle the first situation, the essential qualifications for a new case are the thermal, chemical, and mechanical constraints to be met. Definition of those constraints is not always satisfactorily done, and, soon after, the case falls into the second type of situation, where the essential step is to determine if the material to be replaced has failed to meet the user's expectancy due to a design deficiency (wrong shapes selection for instance), a manufacturing deficiency (in the making of the material up to its final shape), or a faulty installation. The latter can involve wrong procedure, for instance inadequate expansion allowances, for refractories as well as ceramics, or too much mixing-water addition in monolithic refractories, or improper zoning, and so forth or, at last, an unexpected variation in the process, causing higher than expected constraints.
The criteria to choose from to meet the thermal qualifications are the refractoriness in relation to the maximum service temperature, the microstructural stability, and dimensional integrity under changing thermal environments in relation to the thermal stress and thermal shock resistances. The main criteria to meet the chemical qualifications is the corrosion resistance. Durability for refractories and technical ceramics is a matter of both phase composition and texture. The texture (nature and dimension of the grains, coarse or fine, mono-size or multiple sizes) and the porosity as well as the bonding criteria (type of sintering) for maximum thermal stress resistance are, on the whole, opposed to those for maximum corrosion resistance. The main mechanical criterion that affects material selection is the load-bearing capability at the service temperature. This capability may be embodied in the hot strength (hot modulus of rupture for refractories) or in the creep resistance (for technical ceramics).
To deal with failures by corrosion only, the most significant improvements for refractory materials have been obtained through texture control, recalling that total porosity of 12% or so (at service temperature) is close to the minimum required to maintain good thermal shock resistance and acceptable insulating properties.
While maintaining the porosity at this value, further improvements are possible through reducing the pore-size distribution so that the larger pores are of a smaller size, and through reducing the permeability of the refractory by using fused grains instead of sintered grains to reduce the openness of the porosity.
For structural ceramics, if high strength is more important than thermal shock resistance, porosity can be controlled to very low values (<0.5% in many cases). For these materials, various sintering and densification techniques have been developed, such as reaction-bonded sintering and hot pressing, using very pure, fully densified powder of very small grain size. Porosity control for corrosion resistance in structural ceramics is, nevertheless, important. For example, to achieve full density, Si_3N_4 and SiC can be impregnated, either with an organosilicon compound that is subsequently decomposed to produce a pore-filling oxide, or with a material that is subsequently exposed to a nitriding or a carbiding treatment.
Factors Governing Penetration. Once porosity has been adjusted to its optimal value (fixed for a refractory, nil for a structural ceramic), penetration can be minimized by: (1) adjusting the composition of the penetrant fluid; (2) adjusting the wetting/nonwetting characteristics of the same fluid; (3) adjusting the thermal gradient in the material, if at all feasible; and (4) glazing or coating the working face, or in some specific cases, the back or cold face, when oxidation of nonoxide constituents is important.

Composition Adjustment of the Penetrant Fluid. The composition of the penetrant fluid changes during penetration because dissolution and penetration occur simultaneously. If either the melting temperature or the viscosity of the liquid penetrant increases, the depth of penetration can be minimized. This desirable result can be achieved by carefully selecting the matrix additives; for example, a small amount of microsilica can be added to a cement-free alumina-spinel used in the steelmaking applications.

Changing the Wetting Characteristics. In steelmaking applications, wetting characteristics can be changed by adding nonoxides to traditional oxide compositions and using nonfired instead of fired products; for example, by using carbon-bonded refractories and graphite. Insertion of graphite in magnesia-based materials limits penetration and improves corrosion resistance by a factor of 10. Graphite is not readily wetted by $CaO\text{-}SiO_2$ slags and is, therefore, a barrier to penetration. Of course, graphite can be oxidized and should not be used in every situation. Impregnation of pitches and tars modifies the bonding phases in refractories and influences penetration phenomena very significantly. Numerous other examples are available in the technical literature, for example, the use of specific additives such as $BaSO_4$ and CaF_2 in alumina-silica based castables to hold liquid aluminum alloys.

Adjusting the Thermal Gradient. To minimize corrosion, it is important that the penetrating liquid solidify as closely as possible to the hot face so that the rate of penetration is minimized. For a given hot-face temperature, the thermal gradient should be as steep as possible. This can be achieved by vigorously cooling the cold face, by using highly conductive material, and by increasing the thermal conductivity of the lining by conductive backup material.

Use of Glazing or Coating. Sacrificial coatings can be used in some specific situations in which penetration is an important factor. For black (carbon-containing) refractories and for structural ceramics sensitive to oxidation (nonoxide ceramics), the protection of the back face to avoid oxidation has been identified as an important factor. Glazing the hot face using compounds of low melting temperature has to be analyzed very carefully, because glazing can lead to less penetration but more dissolution. Preoxidation treatment to form a protective oxide layer on Si_3N_4 or SiC materials can be beneficial. Uses of multiple layers of new materials, such as “Sialons” (silicon-aluminum oxynitride) and “Simons” (silicon-magnesium oxynitride), have also been studied.

Factors Governing Dissolution. Penetration and dissolution act together and interact, so that the factors controlling them are, to a certain extent, inseparable. Dissolution is controlled by both the chemical reactivity and the specific area of the constituents: the smaller the grains, the faster they dissolve. In refractories, however, the smallest grains, by design, bond with the coarser grains (the aggregate), and since this bonding occurs by liquid-solid sintering, the matrix materials (the bonding phases) have the lowest melting temperatures and the highest reactivity. In general, the dissolution of the bonding phases (or the matrix) leads to debonding between the coarser grains, the dissolution of which is inherently slower and usually apparent only at the refractory hot face.

To minimize dissolution of the matrix, the composition can be altered to raise its melting temperature by increasing the chemical purity of the matrix, so as to decrease the importance of segregated impurities. Practical examples include the use of tabular alumina instead of bauxite grains and of very high-purity instead of low-purity magnesia grains. Alternatively, to raise the melting temperature of the matrix, finely powdered matrix additives can be included in the refractory mix. Because the matrix is a flux in sintering, these additions require higher sintering temperatures to optimize the bonding and the mechanical properties of the resulting products. The logic of reducing or eliminating additives of low melting temperature for improved corrosion resistance is compelling, but the consequence is that the sintering temperature must be raised.

A thorough knowledge of the different binder systems (for nonfired refractory products) and the different bonding systems (after firing) goes a long way to explain the different dissolution characteristics of the many refractory materials and ceramics of a given type. The texture and, in particular, the porosity are key parameters to be considered. Dissolution of the coarse grains becomes a life-limiting factor only when penetration and matrix invasion are sufficiently minimized.

Factors Governing Oxidation. Oxidation of carbon-bonded refractories is considered first. Carbon and graphite oxidation is very detrimental because there, material losses lead to porosity. The low specific gravity of carbon causes a 5% material loss to lead to 7 to 10% porosity. To minimize the rate of carbon and graphite oxidation, new additives, called antioxidants or oxygen inhibitors, have been successfully used. Some general reviews on the role of antioxidants, in such cases, are available, but this topic is outside the scope of this chapter. Among the many additives (Ref 15) now available, the metallic aluminum-based alloys are the most commonly used, sometimes in conjunction with various nonmetals such as silicon carbide and boron compounds. These

additives are used to delay carbon debonding and graphite oxidation. In addition, they often act as pore-blockers, and many of them have a net positive effect on the mechanical strength of the refractories.
Oxidation of nonoxide structural ceramics can be improved by using protective coatings. The use of a chemically vapor deposited (CVD) layer of the same composition as the substrate is a very promising application of coatings to regard the hot corrosion of structural ceramics (Ref 16). Other methods investigated are cathode sputtering and thin multilayer coating.
As an example, a potential high-temperature oxidation protection scheme could involve a coating system consisting of a refractory oxide outer layer for erosion protection, a silica glass inner layer for an oxygen diffusion barrier and crack sealant, another refractory oxide layer for isolation from the carbon surface, and a refractory carbide inner layer for a carbon diffusion barrier. This system, as other current or potential protection technologies, involves interfaces between metal oxides and carbides or oxynitrides. The oxidation resistance of nonoxide ceramics is a dominant problem that restricts the wider usage of these materials.

Installation Methods and Maintenance Procedures

The problems of installation and maintenance are much more acute in corrosion failures of refractories than technical ceramics, simply because of the sizes and volumes considered in each case. Refractory problems occur often in large units involving two (or even three) layers of linings of different thicknesses and different qualities, delimiting large zones, with variable constraints from one zone to another.
The first problem is then one of continuity in the lining; when using bricks of various shapes, the gaps between bricks act as very large pores at the scale of the lining. The use of mortar in joints is then the next issue. In many applications, it can be demonstrated that the best mortar is no mortar and that joint attack can be reduced to a minimum by dry-laying accurately sized bricks. The standard of size tolerances is extremely variable among manufacturers around the world, and the success of one product against another has been, in some instances, directly related to differences in brick size tolerances and to the type of mortar used to compensate for such poor brick tolerances.
For construction, at least for basic bricks, it is preferable to lay them as headers (bricks laid with the length perpendicular to the face of the wall) rather than as stretchers (length parallel to the face of the wall).
Nowadays, there is a great tendency to replace bricks (due in part to the shortage of skilled brick layers) by monolithics. Among the monolithics, there is a tendency to replace bricks by refractory castables, either cast and vibrated in place, or pumped, gunned, or finally shotcreted, the latest technique to be used. In every case, the recommended procedures for installing monolithics are very critical and have direct consequences on the corrosion resistance, in particular when too much mixing water is used, creating exaggerated porosity (see the section “Material Selection”). The errors in installation methods can be many, from excess water to poor mixing, causing segregation, poor flow characteristics, setting that is too fast or too slow, poor drying, poor curing, wrong selection of the first heat-up rate, and poor anchoring systems when anchors have to be used. This list is not exhaustive. In general, errors in installation methods on such unfired materials before use (castables, ramming mixes, plastics) are less forgiving than errors in bricklaying.
To detect installation faults, profiling of the lining at the shutdown of the unit to reline is a must. During tear-out, it is important to record residual thicknesses and locations of holes (if any) in the lining. The profile taken should be used to assist in material selection, definition of zoning, and control of installation procedures, and, in specific cases, of materials-maintenance procedures.
Poor installation procedures, poor commissioning controls, and poor production control can ruin the best efforts in terms of material selection. More detailed information to illustrate such issues can be found for a number of specific cases, such as Ref 17 as regards pneumatic steelmaking. When a premature failure occurs, it is important to be able to differentiate between slagging with clear evidences of partial melting (rounded grains) from creep, abrasion, or spalling due to thermal or mechanical loads rather than structural spalling due to penetration. Although slagging is the most common cause of industrial refractories failures, many mechanisms other than chemical attack can act simultaneously or separately to enhance the degradation rate of the materials.

Process Control

Inherent variabilities in any given process or radical changes in the input or output parameters have many times unforeseen effects on corrosion failures of refractories and technical ceramics. On the slagging issue only,

viscosity, velocity, actual time contact, temperature, partial pressures, total pressure, heat transfer, and mass transfer have a direct effect on failure rates.

As an example, Ref 18 includes an enlightening section on the importance of the scheduling of production in basic oxygen furnace steelmaking plants, as well as another section on the effects of operating variables on lining life.

"Operational consistency and control of the melting process are key ingredients in meeting the demands of modern-day steelmaking. Key variable standard operating procedures (KVSOP) should be developed for all aspects of the operation and be frequently audited to make certain that they are being routinely and correctly implemented," according to Ref 18 (page 12). It is also well recognized that operators who have the control over the operating variables can influence refractories failures and extend the lifetimes of linings. Often, however, several variables are changed in the same time period, and separating the individual influences becomes difficult. Of course, the same philosophy is applicable to technical ceramics: knowledge on how they are being used, the variables describing the environment in which they have to perform, and the minimization of the amount of variations or of the cyclic nature of the constraints help in avoiding premature failures.

As a final consideration, corrosion resistance of refractories and technical ceramics should be considered as a measure of a complex process involving many intrinsic and extrinsic parameters. Pragmatically, the definition of all parameters and their evolution with time is as important as the definition of the material and of its installation and commissioning into service.

References cited in this section

15. M. Rigaud, *8th CIMTEC Proceedings, Ceramics: Charting the Future,* P. Vincenzini, Ed., Techna Srl., Faenza, Vol 3A, 1995, p 399–414

16. J. Desmaison, *High-Temperature Corrosion of Technical Ceramics,* R.J. Fordham, Ed., Elsevier Applied Science, 1990, p 93–108

17. K. Lowrance, Refractory Lining Construction, *Pneumatic Steelmaking, Volume 3 on Refractories,* M.A. Rigaud and R.A. Landy, Ed., Iron and Steel Society, 1996, p 137–150

18. R. Jones, Scheduling of Production, *Pneumatic Steelmaking, Volume 3 on Refractories,* M.A. Rigaud and R.A. Landy, Ed., Iron and Steel Society, 1996, p 11–20

Corrosion Failures of Industrial Refractories and Technical Ceramics*

Michel Rigaud, École Polytechnique

References

1. "Standard Terminology Relating to Refractories," ASTM C 71, *Annual Book of ASTM Standards,* ASTM

2. A. Muan and E.F. Osborn, *Phase Equilibria Among Oxides in Steelmaking,* Addison-Wesley, Reading, MA, 1965

3. H.M. Kraner, *Phase Diagrams, Materials Science and Technology,* Vol 6-II, A.M. Alper, Ed., Academic, 1970, p 67–115

4. A.M. Alper, R.C. Doman, R.N. McNally, and H.C. Yeh, *Phase Diagrams Materials Science and Technology,* Vol 6-II, A.M. Alper, Ed., Academic, 1970, p 117–146

5. G. Ericksson and K. Hack, *Metall. Trans. B,* 1990, p 1013

6. G. Ericksson, Chem. Scr., 8, 1975, p 100

7. V.L.K. Lou and A.H. Heuer, *High-Temperature Corrosion of Technical Ceramics,* R.J. Fordham, Ed., Elsevier Applied Science, 1990, p 33–52

8. E.S. Chen and O. Buyukozturk, *Am. Ceram. Soc. Bull.,* Vol 64 (No. 7), 1985, p 995

9. S.C. Carniglia and G.L. Barna, *Handbook of Industrial Refractories Technology,* Noyes Publication, Park Ridge, NJ, 1992

10. R.A. McCauley, *Corrosion of Ceramics,* Marcel Dekker, New York, 1995

11. L.A. Lay, *Corrosion Resistance of Technical Ceramics,* HMSO, London, 1983

12. R. Morrell, *Handbook of Properties of Technical and Engineering Ceramics,* Part I, HMSO, London, 1985

13. D.S. Fox and J.L. Smialek, "Burner Rig Hot Corrosion of Silicon Carbide and Silicon Nitride," *J. Am. Ceram. Soc.,* Vol 73, 1990 p 303–311

14. Y.G. Gogotsi and V.A. Lavrenko, *Corrosion of High-Performance Ceramics,* Springer-Verlag, Berlin, 1992, p 110–156

15. M. Rigaud, *8th CIMTEC Proceedings, Ceramics: Charting the Future,* P. Vincenzini, Ed., Techna Srl., Faenza, Vol 3A, 1995, p 399–414

16. J. Desmaison, *High-Temperature Corrosion of Technical Ceramics,* R.J. Fordham, Ed., Elsevier Applied Science, 1990, p 93–108

17. K. Lowrance, Refractory Lining Construction, *Pneumatic Steelmaking, Volume 3 on Refractories,* M.A. Rigaud and R.A. Landy, Ed., Iron and Steel Society, 1996, p 137–150

18. R. Jones, Scheduling of Production, *Pneumatic Steelmaking, Volume 3 on Refractories,* M.A. Rigaud and R.A. Landy, Ed., Iron and Steel Society, 1996, p 11–20

Hydrogen Damage and Embrittlement

Introduction

HYDROGEN DAMAGE is a term used to designate a number of processes in metals by which the load-carrying capacity of the metal is reduced due to the presence of hydrogen. The detrimental effects of hydrogen on metals are often enhanced when the hydrogen is present in combination with residual or applied tensile stresses. Many types of hydrogen damage are not really due to corrosion processes at all. However, because hydrogen damage has many similarities regardless of the source of the hydrogen, damage processes caused by corrosion and manufacturing problems are both described in this article.

Many metals and alloys are susceptible to hydrogen damage; however, specific forms of hydrogen damage are generally more common within particular alloy systems. Hydrogen damage in one form or another can severely restrict the use of certain materials.

This article is divided into three parts. The first part is an introduction to the general forms of hydrogen damage, and the second part is an overview of the different types of hydrogen damage in all the major commercial alloy systems. These two sections cover the broader topic of hydrogen damage, which can be quite complex and technical in nature.

The third part of this article focuses on failure analysis where hydrogen embrittlement of a steel component is suspected. Many of the types of parts that were commonly diagnosed with hydrogen embrittlement in the 1980s and 1990s were low-cost commodity products, such as self-tapping screws and small spring clips. The financial resources available to perform a failure analysis in such cases are often limited, if nobody was injured. Therefore, the third section gives practical advice for the failure analysis practitioner or for someone who is contemplating procurement of a cost-effective failure analysis of commodity-grade components suspected of hydrogen embrittlement. Because of the wide economic impact in commodity-grade components, it is worthwhile to look at failure analysis for this specific type of hydrogen damage. The third section of the article is intended to be able to be understood and used by engineering managers, quality engineers, design engineers, and manufacturing engineers and technicians. Some prevention strategies for design and manufacturing problem-induced hydrogen embrittlement are also provided.

Hydrogen Damage and Embrittlement

Overview of Hydrogen Damage Processes

Hydrogen is known to cause various problems in many metals and is readily available during the production, processing, and service of metals from operations such as:

- Refining (precipitates upon solidification from supersaturated concentrations)
- Acid cleaning (pickling)
- Electroplating
- Contact with water or other hydrogen-containing liquids or gases

Depending on the type of hydrogen-metal interaction, hydrogen damage of metal manifests itself in one of several ways. The types of interactions between hydrogen and metals include the formation of solid solutions of hydrogen in metals, molecular hydrogen, gaseous products from reactions between hydrogen and alloying or segregated impurity elements, or brittle intermetallic hydride compounds.

Specific types of hydrogen damage, some of which occur only in specific alloys under specific conditions (Table 1), are classified in this article as follows:

- Hydrogen embrittlement
- Hydrogen-induced blistering
- Cracking from precipitation of internal hydrogen
- Hydrogen attack
- Cracking from hydride formation

Table 1 Specific types of hydrogen damage

Specific damage process/mechanism	**Most common in**	Comments
Hydrogen embrittlement	High-strength steels and heavily cold-worked low-carbon steels, ambient temperatures	**Needs sustained stress. Not relevant for impact-induced cracking**
Hydrogen-induced blistering	Low-strength steels at ambient temperatures	**Hydrogen forms gas-filled blisters**
Cracking from precipitation of internal gaseous hydrogen	Heavy steel sections	**Induced by high-temperature exposure followed by rapid cooling**
Hydrogen attack	Steels subject to combined high-temperature and high-pressure hydrogen. Also affects copper	**Irreversible chemical reactions of hydrogen with matrix or alloy elements form high-pressure pockets of gases other than molecular hydrogen.**
Cracking from hydride formation	**Transition, rare earth, alkaline-earth metals, and their alloys (includes titanium, tantalum, zirconium, uranium, and thorium)**	**Brittle hydrides often form preferentially where the stress is highest.**

The first three types are usually observed at ambient temperatures and are closely related to one another. Hydrogen damage usually manifests itself as hydrogen embrittlement in high-strength steels and as hydrogen-induced blistering in low-strength steels. The solubility and diffusivity of hydrogen in steel sharply decrease with lowering temperatures; therefore, when a heavy section of steel containing hydrogen at elevated temperature is rapidly cooled to ambient temperature, the hydrogen remaining in the steel precipitates out in the gaseous state. The pressure of the hydrogen gas is often great enough to produce internal cracks. Hydrogen attack is an elevated-temperature phenomenon, in which hydrogen reacts with metal substrates or alloy additions. A number of transition and rare earth metals form hydrides, and the formation of metal hydrides can result in cracking.

Hydrogen damage is a complex phenomenon that takes on several forms, and many mechanisms have been proposed, which include (Ref 1):

- Hydrogen pressure theory
- Reduced surface energy (surface adsorption) theory
- Decohesion theory
- Enhanced plastic flow theory
- Decarburization (hydrogen attack)
- Hydride formation

The pressure theory assumes that embrittlement is caused by the pressure exerted by gaseous hydrogen on the crack front. For example, when steel is saturated with hydrogen at elevated temperature and then is cooled, gaseous hydrogen precipitates in microvoids, and an extremely high pressure of the gas can be developed. Flakes in heavy forgings and underbead cracks in weldments can be explained by the pressure theory. Hydrogen charged into steel during aqueous corrosion or cathodic charging can also produce a very high pressure at the surfaces of closed voids or pores.

The reduced surface energy theory states that the absorption of hydrogen decreases the surface free energy of the metal and enhances crack propagation. This theory may explain the crack propagation of high-strength steels in low-pressure gaseous hydrogen. The decohesion theory (Ref 2) holds that dissolved hydrogen tends to migrate toward regions of hydrostatic (often termed as triaxial) tensile stress and weakens the material by lowering the bond strength between metal atoms. This theory is the one that is most applicable to the type of hydrogen embrittlement generally found in newly processed parts that crack during shipping or are found cracked at some point during the assembly process. Grain boundaries of steel are apparently particularly sensitive to this type of atomic bond weakening.

The enhanced plastic flow theory is based on fractographic evidence, where the grain facets of intergranular fracture after hydrogen stress cracking reveal dimples at high magnifications (1000 to 5000×). This, however, is complicated by the influence of grain-boundary precipitates, which may act as sites for nucleation of both voids and hydrogen during its migration in grain boundaries. The formation of methane gas bubbles at grain boundaries is another mechanism (see, for example, Fig. 4 in the article "Intergranular Fracture" in this Volume). For titanium alloys, the repeated formation and rupture of the brittle hydride phase at the crack tip is the dominant mechanism. Probably no one mechanism is applicable to all metals, and several mechanisms may operate simultaneously to embrittle a material. Whatever the mechanism, the end result is an adverse effect on the mechanical properties of the material.

Fig. 1 Cadmium-plated AISI 8740 steel nut that failed by hydrogen embrittlement. Failure occurred seven days after installation on an aircraft wing structure. See also Fig. 2. 5×. Courtesy of Lockheed-Georgia Company

References cited in this section

1. U.K. Chatterjee, S.K. Bose, and S.K. Roy, *Environmental Degradation of Metals*, Marcel Deckker, 2001

2. J.P. Hirth, Theories of Hydrogen Induced Cracking, *Hydrogen Embrittlement and Stress Corrosion Cracking*, R. Gibala and R.F. Hehemann, Ed., American Society for Metals, 1984, p 29–41

Hydrogen Embrittlement

Hydrogen embrittles several metals and alloys, and the effects can range from a slight decrease in the percent reduction of area at fracture to a brittle macroscopic fracture at a relatively low applied stress (often well below

the yield strength). Even small amounts of hydrogen can have a deleterious effect, particularly for high-strength steels with tensile strengths of 1240 MPa (180 ksi) or more. A few parts per million of hydrogen dissolved in steel can cause hairline cracking and loss of tensile ductility. Even when the quantity of gas in solution is too small to reduce tension-test ductility, hydrogen-induced delayed fracture may occur.

Although hydrogen embrittlement has been evaluated extensively, it takes on several forms. The various forms or manifestations of hydrogen embrittlement and the terminology associated with it can be very confusing, because the term has been used without clarified definitions. One classification breaks hydrogen embrittlement into three types (Ref 3, 4):

- Internal reversible hydrogen embrittlement
- Hydrogen environment embrittlement
- Hydrogen reaction embrittlement

This division is roughly equivalent to the classification of hydrogen embrittlement in steels that is defined in *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Volume 1, *ASM Handbook,* as follows (Ref 5):

- Hydrogen environmental embrittlement
- Hydrogen stress cracking
- Loss in tensile ductility

This latter division focuses on the mechanical effects of HSC and loss of ductility from reversible hydrogen embrittlement, as opposed to the other categories that are based more on the nature of hydrogen. Either approach could be considered, but this article considers the former based on Ref 4.

Internal Reversible Hydrogen Embrittlement

Internal reversible hydrogen embrittlement has also been termed slow strain rate embrittlement and delayed failure. The mechanical effects, as described subsequently, may range from loss of ductility to cracking. To be fully reversible, of course, embrittlement must occur without the hydrogen undergoing any type of chemical reaction or causing microscopic cracks to be initiated.

Internal reversible hydrogen embrittlement can occur after a very small average concentration of hydrogen has been absorbed from the environment. However, local concentrations of hydrogen may be substantially greater than average bulk values, because the absorbed hydrogen diffuses into grains or (as is more likely) preferentially into the grain boundary. Degradation is enhanced by slow strain rates, which thus suggests time-dependent diffusion as a controlling factor. Moreover, embrittlement is usually most severe at room temperature during either sustained stresses or slow strain rate tension testing. Higher temperatures cause hydrogen to diffuse away from areas of localized concentration. The effect of hydrogen is also strongly influenced by other variables, such as:

- The strength (or hardness) level of the alloy
- The microstructure
- The magnitude of the applied stress
- The presence of a localized triaxial tensile stress
- The amount of prior cold work
- The degree of segregation of tensile low melting-point elements such as phosphorus, sulfur, nitrogen, tin, or antimony at the grain boundaries

This is the classical type of hydrogen embrittlement. Its unique aspect is the delayed nature of the fractures; that is, after a specimen is charged with hydrogen, fracture does not occur instantly but only after the passage of a certain amount of time. This time-dependent nature (incubation period) for embrittlement by hydrogen is also why it is most noticeable at low strain rates. Internal reversible hydrogen embrittlement has also been observed in a wide variety of other materials, including nickel-base alloys and austenitic stainless steels, provided they are severely charged with hydrogen. However, most attention has been given to the delayed cracking under

stress of high-strength steel components exposed to hydrogen from processing treatments such as electroplating, pickling, and other sources. In steels, a few parts per million of dissolved hydrogen can cause hairline cracking and loss of tensile ductility. While hydrogen embrittlement causes strength reduction along with reduced ductility, the hardness remains the same.

The loss of tensile ductility, one of the earliest documented forms of hydrogen damage, occurs in steels and other alloys exposed to hydrogen. It generally affects lower-strength alloys. Fracture does not necessarily occur, as it does with higher-strength alloys; these alloys experience only a decrease in tensile elongation and reduction in area, depending on strain rate and hydrogen content. This type of embrittlement in steel is reversible; that is, ductility can be stored by relieving the applied stress and aging at room temperature, provided microscopic cracks have not yet initiated.

Hydrogen stress cracking is characterized by the brittle fracture of an alloy due to hydrogen exposure from either some time prior to manufacturing or from environmental hydrogen. It is also commonly considered a subcritical crack growth mechanism that often produces time-delayed fractures in production parts (especially high-strength steels) even with no externally applied stress. The phenomenon is characterized by the simultaneous interactions of sustained static or very slowly changing loads on susceptible materials that have internal dissolved atomic hydrogen.

Hydrogen stress cracking involves the absorption of hydrogen, followed by hydrogen diffusion to regions of high tensile stress, particularly those associated with notches. Because it is a diffusion-controlled process, more slowly applied stresses and, in particular, sustained or static stresses are more likely to allow the hydrogen embrittlement problem to manifest. This gives hydrogen more time to diffuse to the potential crack plane(s). Thus, there is a delay time, or incubation time, before cracking occurs. In this regard, hydrogen stress cracking has also been referred to as static fatigue, sustained-load cracking, or hydrogen-induced delayed fracture.

This type of hydrogen damage occurs most often in high-strength steels (tensile strength in excess of about 1240 MPa, or 180 ksi), although it can also occur in relatively soft steels that have been heavily cold formed. The factors that influence the likelihood of cracking include the hardness or strength level, stress level, the duration of the sustained load, and the concentration of hydrogen. The cracking is most prevalent at ambient temperature, and it can occur under load when the quantity of gas in solution is too small to reduce ductility. The cracking tendency decreases with increasing temperature, and the hydrogen-embrittlement phenomenon virtually disappears in steels above 200 °C (390 °F).

Cracks typically initiate internally, usually below the root of a notch at a region of maximum localized tensile triaxial stress. Tensile residual stresses may also promote the initiation of hydrogen stress cracking. Crack branching is much less prevalent in hydrogen embrittlement than in stress-corrosion cracking (SCC). The crack path is often intergranular, especially for high-strength steels. However, hydrogen-induced fractures are not always intergranular (Ref 5). In high-strength steels, cracking is often intergranular, but transgranular cleavage is also observed. For ductile low-strength steels, the fracture mode is ductile, with decreases in the dimple size.

Cracking Threshold. For an alloy exhibiting a specific strength level and microstructure, there is a stress-intensity value, K_I, below which, for all practical purposes, hydrogen embrittlement cracking does not occur. This threshold crack tip stress-intensity factor (K_{th}) is an experimentally determined quantity that depends on:

- The nominal strength of the unembrittled part
- The amount of hydrogen present in the steel
- The location of the hydrogen in the microstructure
- The presence of other embrittling elements or microstructural phases

The actual value of the tensile stress that creates a crack in a particular component depends on these factors. For example, prior temper embrittlement of high-strength alloy steels enhances hydrogen stress cracking. The threshold stress intensity for crack growth is lowered, and intergranular fractures occur. The influence of temper embrittlement on hydrogen stress cracking is particularly pronounced as the yield strength is decreased (Ref 5). Very low-impurity-content steels do show high threshold stress-intensity values and freedom from intergranular fracture (Ref 6).

Hydrogen from Preservice Pickling and Electroplating. As noted, hydrogen stress cracking can occur from hydrogen pickup from manufacturing. Properly performed pickling alone is not usually a direct cause of hydrogen damage (unless internal voids or other imperfections lead to local formation of atomic hydrogen), because much of the absorbed hydrogen diffuses out of the metal shortly after it is absorbed. Heating at 150 to

200 °C (300 to 390 °F) hastens the removal of hydrogen. Also, the addition of suitable inhibitors to the pickling solution eliminates or minimizes attack on the metal and the consequent generation of nascent hydrogen. Salt baths operated at approximately 210 °C (410 °F) can be used for descaling titanium alloys, superalloys, and refractory metals to avoid the possibility of hydrogen charging associated with pickling.

Plating solutions and plating conditions selected to produce a high-cathode efficiency minimize the amount of hydrogen generated on the metal surface. Because the metallic coatings plated on metal often prevent the hydrogen from leaving the base metal, elevated-temperature baking after plating is generally required to allow the hydrogen to move to microstructural positions in the part interior that are less damaging to the atomic bonds of the iron matrix.

In one hydrogen-damage failure caused by improper baking after electroplating, several cadmium-plated American Iron and Steel Institute (AISI) 8740 steel nuts cracked in seven days after installation on an aircraft wing structure (Fig. 1). Examination of the fracture surface (Fig. 2) showed predominantly intergranular fracture typical of hydrogen embrittlement. The nuts had not been baked for a sufficient time; thus, hydrogen picked up during the plating process was not deactivated.

Fig. 2 Fracture surface of failed cadmium-plated nut in Fig. 1. (a) Macrograph of fracture surface; 15×. (b) Scanning electron micrograph of the area in the box in (a) showing typical intergranular fracture; 3950×. Courtesy of Lockheed-Georgia Company

Procedures for prevention of hydrogen damage in electroplating and preparation for electroplating are described in ASTM A 143, “Standard Practice for Safeguarding against Embrittlement of Hot Dip Galvanized Structural Steel Products and Procedure for Detecting Embrittlement.” Other relevant standards include:

- ASTM F 1940, “Standard Test Method for Process Control Verification to Prevent Hydrogen Embrittlement in Plated or Coated Fasteners”
- ASTM F 519, “Standard Test Method for Mechanical Hydrogen Embrittlement Evaluation of Plating Processes and Service Environments”
- ASTM G 142, “Standard Test Method for Determination of Susceptibility of Metals to Embrittlement in Hydrogen-Containing Environments at High Pressure, High Temperature, or Both”
- ASTM F 1459, “Standard Test Method for Determination of the Susceptibility of Metallic Materials to Gaseous Hydrogen Embrittlement”

Hydrogen Environmental Embrittlement

Hydrogen embrittlement, in classic form, is reversible, in that the damage should be eliminated by removing the hydrogen. However, hydrogen can also cause permanent damage from exposure to environmental hydrogen during service. Gaseous environments containing hydrogen are also damaging. Hairline cracking usually follows prior-austenite grain boundaries and seems to occur when the damaging effect of dissolved hydrogen is superimposed on the stresses that accompany the austenite-to-martensite transformation. Hairline cracking is readily recognized by metallography and is most common near the center of fairly bulky components, where constraint of plastic deformation is high. Affected areas are recognized on fracture surfaces by their brittle appearance and high reflectivity, which usually contrasts with the matte appearance of surrounding regions of ductile fracture. This has led to such areas being described as flakes or fisheyes. These and other forms of

hydrogen damage are described in subsequent sections on blistering, precipitation of internal hydrogen, hydrogen attack, and hydride formation.
However, in terms of the various types of in-service environments, hydrogen embrittlement can also occur. This type of situation may be complicated by more factors than the preservice type of delayed HSC. In particular, hydrogen from corrosion reactions may be a factor, or SCC mechanisms may be involved. In particular, the distinction between SCC and hydrogen embrittlement can be difficult. However, the distinction is well defined in terms of electrochemical potentials (Fig. 3) (Ref 7), where hydrogen embrittlement occurs in a lower region associated with hydrogen formation. Current mechanisms of corrosion-assisted, sustained-load cracking are variations of two basic theories: crack advance by anodic dissolution or hydrogen embrittlement. The controlling factors in these two are as follows:

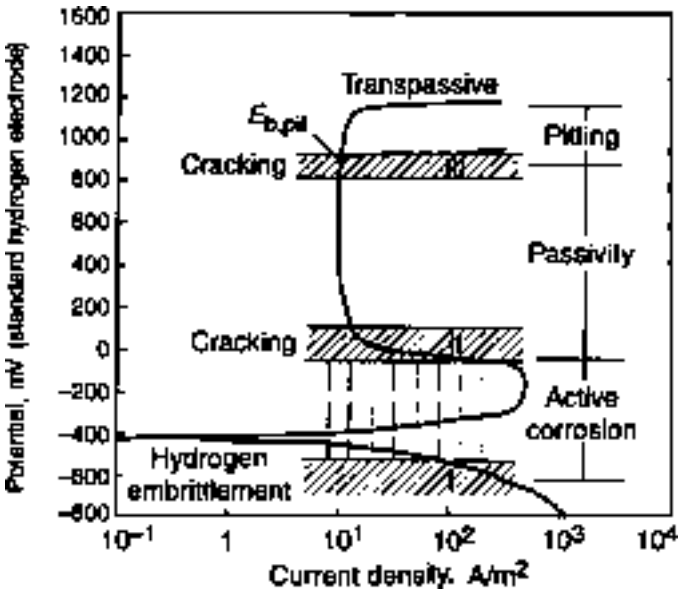


Fig. 3 Potential ranges of environmentally assisted cracking by (I) hydrogen embrittlement, (II) cracking of unstable passive film, and (III) cracking initiated at pits near the pitting potential. Vertical dashed lines define potential range over which nonpassivating films may crack under stress Source: Ref 7

Anodic dissolution (SCC) is characterized by:

- Grain-boundary precipitate size, spacing, and/or volume fraction
- Grain-boundary precipitate-free zone (PFZ) width, solute profile, or deformation mode
- Matrix precipitate size/distribution and deformation mode
- Oxide rupture and repassivation kinetics

Hydrogen embrittlement is characterized by:

- Hydrogen absorption leading to grain-boundary or transgranular decohesion
- Internal void formation via gas pressurization
- Enhanced plasticity (adsorption and absorption arguments exist)

Hydrogen embrittlement and SCC are also sometimes distinguished in terms of application and the source of the hydrogen atoms. If the hydrogen is coming from an intentional manufacturing process, and that hydrogen facilitates cracking at lower than expected stresses, whether the hydrogen production itself is intentional or not, the term hydrogen embrittlement is used. If the service conditions (moisture, for example, or chloride ions) allow the removal of atoms from their intended location in the component (corrosion), and if the hydrogen from this reaction facilitates cracking, then SCC is used.
Regardless of the definition, environmental factors causing the cracking must be properly evaluated, and obviously, in-service conditions may involve more environmental factors than hydrogen charging during

manufacturing. Therefore, the electrochemical definition of hydrogen embrittlement (Fig. 3) has more specificity for complex environmental conditions in service.

In-Service Cracking from Hydrogen Charging in an Aqueous Environment. When metal corrodes in a low-pH solution, the cathodic partial reaction is reduction of the hydrogen ion. Although most of the reduced hydrogen reacts to form H_2 and leaves the metal surface as gaseous hydrogen, a part of the reduced hydrogen enters into the metal as atomic hydrogen. The presence of certain chemical substances that prevent the recombination of hydrogen to form molecular hydrogen enhances the absorption of nascent (atomic) hydrogen into the metal. These substances are called cathodic poisons, and they include phosphorus, arsenic, antimony, sulfur, selenium, tellurium, and cyanide ion. Among the cathodic poisons, sulfides are among the most common. Environments containing hydrogen sulfide can cause severe embrittlement of steels and some other high-strength alloys. On the other hand, corrosion inhibitors lower the corrosion rate and thus the amount of hydrogen charged into the metal.

In-Service Atmospheric and Aqueous Corrosion. Most high-strength steels are susceptible to hydrogen embrittlement when they are stressed and exposed to fresh or sea water and even during atmospheric exposure. The susceptibility of steels to hydrogen embrittlement generally increases with increasing tensile strength.

Steels having a tensile strength greater than approximately 1000 MPa (150 ksi) are susceptible to embrittlement. Above a tensile-strength level of 1250 MPa (180 ksi), most alloy steels, such as AISI 4130 and 4340, and precipitation-hardening stainless steels are susceptible to hydrogen embrittlement cracking in marine atmospheres when the residual or applied tensile stresses are sufficiently high. The cracking occurs after some time has passed. Steels with tensile strengths less than 690 MPa (100 ksi) appear to be much more resistant to hydrogen embrittlement cracking, and the structures made with such steels have been used in service without serious problems in various environments, as long as they do not contain hydrogen sulfide.

As noted, hydrogen-embrittlement cracking of high-strength steel in aqueous environments may be sometimes erroneously called SCC. Cracking of high-strength steel in aqueous environments is caused by hydrogen absorbed into the steel during corrosion. It is not caused by an anodic dissolution mechanism, such as the chloride SCC of austenitic stainless steels or caustic cracking of low-alloy steels. The analyst should bear in mind, however, that sometimes the boundary between SCC and hydrogen embrittlement may not be very clear, depending on the understanding of the physics, chemistry, and metallurgy of a particular situation. The more important concern for most failure analysts or their clients or internal customers is whether the investigation can provide insight on prevention strategies. In both SCC and hydrogen embrittlement, the presence of hydrogen is clearly a significant factor contributing to the ease of crack formation.

When working on failure analysis cases involving hydrogen damage, it is important to realize that the electrochemical conditions at the tip of a pit or an advancing crack are not the same as those of the bulk solution. This is because the solution chemistry in the region of the pit or crack tip may be quite different from that of the bulk solution. In a laboratory study, when wedge-opening loading specimens were exposed to sodium chloride solutions, the pH value of the solution at the crack tip was measured to be approximately 3.5, regardless of that of the bulk solution. At the crack tip, where diffusion is limited, the pH value of the solution is lowered by the acidic hydrolysis reaction, and in such a low-pH solution, the cathodic partial reaction is the reduction of the hydrogen ion. As a result, nascent hydrogen is generated at the tip of the pit or crack and absorbed into the metal. It has also been shown that hydrogen can be generated at the crack tip even when an anodic potential is applied to the bulk metal.

In-Service Failures in Environments Containing Hydrogen Sulfide (Sulfide Stress Cracking). High-strength steel pipes used in drilling and completion of oil and gas wells may exhibit delayed cracking in environments containing hydrogen sulfide. This type of fracture is referred to as sulfide stress cracking. The basic cause of sulfide stress cracking is embrittlement resulting from hydrogen absorbed into steel during corrosion in sour environments. The presence of hydrogen sulfide in the environment promotes hydrogen absorption into steel, thereby making the environment more severe and thus more likely to cause hydrogen embrittlement. Although hydrogen sulfide gas, like gaseous hydrogen, can cause embrittlement, water ordinarily must be present for sulfide stress cracking to occur.

The susceptibility to sulfide stress cracking increases with increasing hydrogen sulfide concentration or partial pressure and decreases with increasing pH. The ability of the environment to cause sulfide stress cracking decreases markedly above pH 8 and below 100 Pa (~0.001 atm) partial pressure of hydrogen sulfide. The cracking tendency is most pronounced at ambient temperature and decreases with increasing temperature. For a given strength level, tempered martensitic steels have better sulfide stress cracking resistance than normalized-

and-tempered steels, which in turn are more resistant than normalized steels. Untempered martensite demonstrates poor resistance to sulfide stress cracking. It is generally agreed that a uniform microstructure of fully tempered martensite is desirable for sulfide stress cracking resistance.

Researchers in sulfide stress cracking are not in agreement as to the effect of alloying elements on the sulfide stress cracking resistance of carbon and low-alloy steels, with the exception of one element. Nickel is detrimental to sulfide stress cracking resistance. Steels containing more than 1% Ni are not recommended for service in sour environments.

Welds and associated heat-affected zones (HAZs) are often less resistant to sulfide stress cracking than the adjacent base metal. The high hardness and residual stresses resulting from welding are believed to increase the susceptibility. In a laboratory study of the sulfide stress cracking resistance of submerged arc weldments in a hydrogen-sulfide-saturated aqueous solution of 0.5% acetic acid and 5% sodium chloride, no failures were observed for the welds with hardness values below 191 HB (91 HRB); all of the welds with hardness values of 225 HB (20 HRC) or higher cracked. Steels with hardness values less than 22 HRC are considered acceptable for sour service. Metallic materials resistant to sulfide stress cracking for oil-field equipment are covered in National Association of Corrosion Engineers standard (MR-01) “Sulfide-Stress-Cracking-Resistant Metallic Material for Oil-Field Equipment.”

In-Service Cathodic-Protection-Related Damage. Marine structures and underground pipelines are cathodically protected either with sacrificial anodes or with impressed current. Although the cathodic protection of such structures usually does not cause HE, high-strength steel can be embrittled by cathodic protection.

In-Service Cracking from Gaseous Hydrogen. Steel vessels and other equipment that contains hydrogen gas at high pressures at ambient temperature are susceptible to fracture by hydrogen embrittlement. Many high-strength steels are subject to severe embrittlement in tension tests conducted in high-pressure hydrogen gas, and the cracking tendency increases with increasing pressure of hydrogen. As has been noted many times in this article for many types of hydrogen damage, the cracking susceptibility of steels in gaseous hydrogen generally increases with increasing hardness.

Generally, high-strength steels and high-strength nickel alloys display severe degradation of tensile properties in hydrogen gas. Austenitic stainless steels, aluminum alloys, and alloy A286 show very little embrittlement in this environment; most other engineering metals and alloys are affected to a lesser degree. Seamless steel vessels have been used for the containment of hydrogen gas at moderately high pressure without any reported problems. When austenitic stainless steels, such as AISI 304L and 309S, are exposed to high-pressure hydrogen gas (69 MPa, or 10 ksi), the macroscopic tensile ductility is markedly reduced, although the microscopic fracture may be dimpled microvoid coalescence. The loss in tensile reduction of area can be correlated with a change in dimple size (Ref 8).

Hydrogen gas at atmospheric pressure can cause embrittlement in martensitic high-strength steels. Subcritical crack growth is observed with high-strength steel at very low stresses in pure hydrogen gas, and oxygen in small quantities stops the crack growth. Oxygen is believed to form an oxide barrier at the crack tip by preferential adsorption, thereby protecting the crack tip from the damaging effects of the hydrogen.

Fracture Characteristics

Crack Origin. When a hydrogen-damage failure is caused by hydrogen embrittlement or hydrogen-induced blistering, the crack typically originates in the interior of the part or very close to the metal surface. In metals of relatively low strength and hardness and in high-strength metals that are not subjected to significant levels of applied or residual tensile stresses, cracking almost always originates in the interior of the metal. In high-strength metals subjected to applied or residual tensile stresses, especially if a severe stress raiser such as a sharp notch is present at the surface, cracking is likely to originate near the surface.

Crack Morphology. A hydrogen-assisted crack is usually a single crack that shows no significant branching (Fig. 4). The crack path can be either intergranular or transgranular, and the crack path sometimes changes from one to the other as it propagates. Intergranular fracture becomes more likely at higher strength and with impurities. When transgranular fracture from hydrogen occurs in less ductile ((body-centered cubic, or bcc) alloys, the mechanisms may be brittle cleavage or quasi-cleavage. In more ductile low-strength steels or face-centered cubic (fcc) alloys, fracture may be by dimpled microvoid coalescence. Therefore, the results of metallographic and electron fractographic examinations should be carefully interpreted for failure analysis. Examination at low magnification should precede examination at higher magnifications. Examination of the

fracture surface at high magnifications (1000× or more) may yield clues on the potential mechanisms of intergranular fracture in the grain boundaries.

Deposits on the Fracture Surface. Corrosion products are ordinarily not present on the surface of hydrogen-damaged fractures, unless the surfaces have been exposed to corrosive environments. Foreign matter, if present, is in trace amounts only, unless the fracture surfaces had been exposed to contaminants. Energy-dispersive x-ray analysis allows the determination of the elements on the fracture surface.

Laboratory cracking tests are used to verify that the material is susceptible to hydrogen damage under a given environmental condition or to distinguish between hydrogen damage and SCC. One method of distinguishing between hydrogen embrittlement and SCC is to note the effect of small impressed electric currents on time-to-failure in constant-load test. If cracking has occurred by a stress-corrosion mechanism, application of a small anodic current shortens the time to failure. When hydrogen embrittlement is the cracking mechanism, a cathodic current accelerates cracking. Electrochemical methods are also used to determine whether hydrogen can be charged into metal in given solutions.

Example 1: Hydrogen Embrittlement of a Type 431 Stainless Steel Mushroom-Head Closure. The type 431 stainless steel mushroom-head closure shown in two views in Fig. 4 fractured in service at a hydrogen pressure of 3000 atm. Fracture occurred at room temperature after miscellaneous chemical service that included exposures to hydrogen at temperatures from ambient to 350 °C (662 °F).

The fracture surface displayed areas of both transgranular cleavage, as shown in the stereo pair of Fig. 5, and intergranular separation, as shown in the stereo pair of Fig. 6, which is characteristic of hydrogen embrittlement of high-stress steels. Debris from the final reaction to which the part was exposed has resulted in artifacts on the transmission electron microscopy (TEM) replicas, but there appears to be no evidence of grain-boundary corrosion in Fig. 6. This is an important finding, in that it differentiated this intergranular fracture, which was caused by hydrogen embrittlement, from similar ones produced by SCC. Only by meticulous fractographic examination can clues be gained that make it possible to distinguish between these two fracture mechanisms.

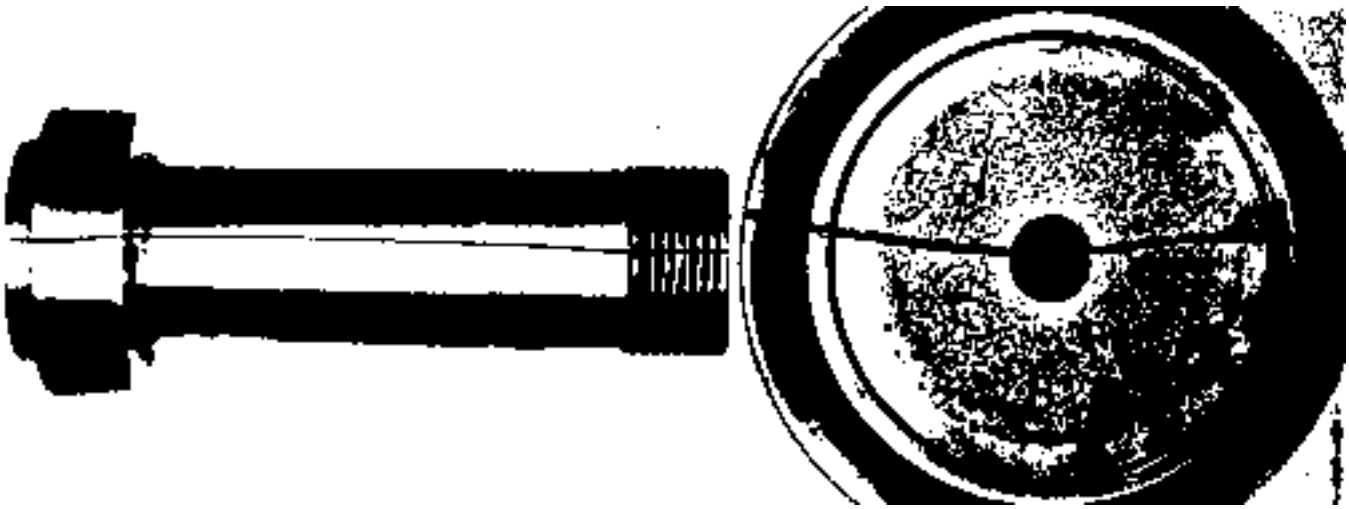


Fig. 4 Two views of a fracture from hydrogen embrittlement of a type 431 stainless steel mushroom-head closure (Example 1). This is not typical; HE cracking on cylinders is usually circumferential.

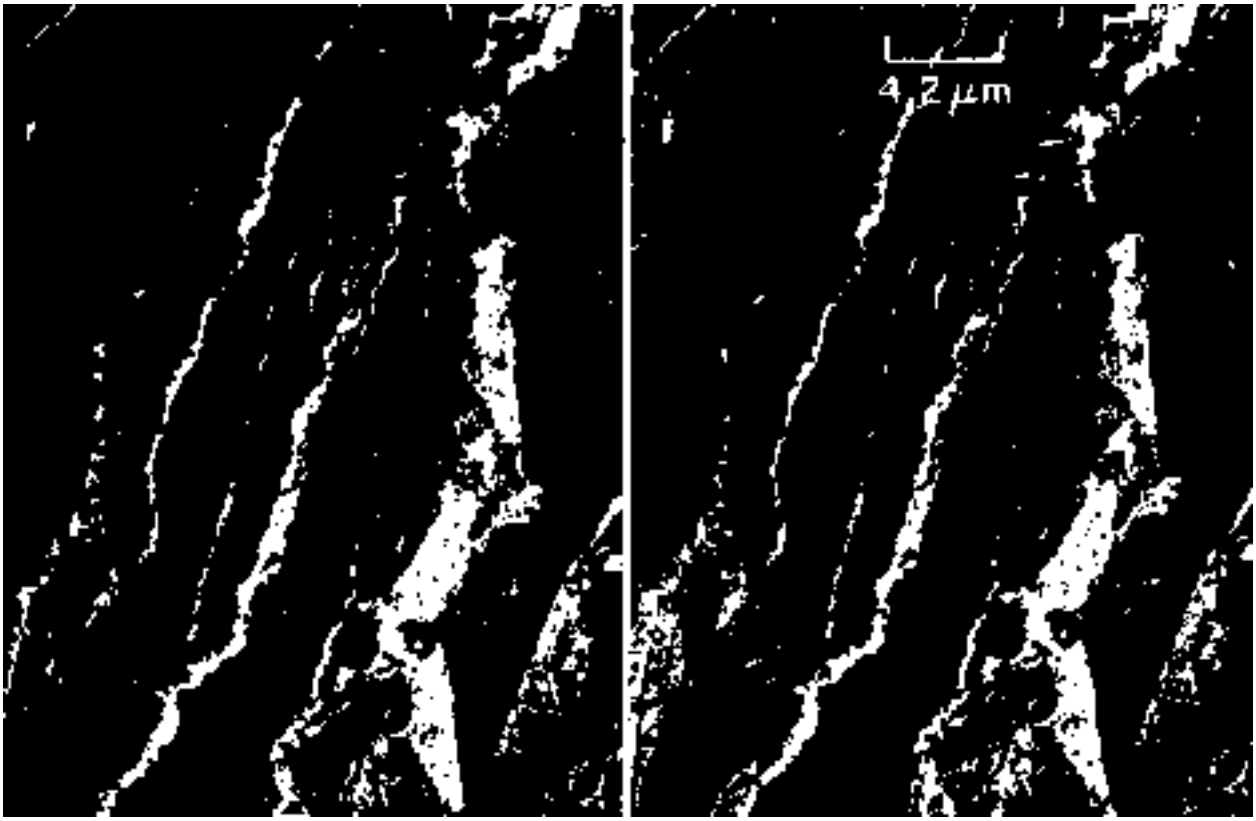


Fig. 5 Transgranular cleavage in an area of the surface of the hydrogen embrittlement fracture of the type 431 stainless steel mushroom-head closure shown in Fig. 4. See also Fig. 6. When viewed in three-dimension, this stereo pair shows a massive ridge running from top to bottom at the left, with pronounced cleavage steps on its right-hand slope. The corrosion debris remaining from service and extracted by the replica may be seen projecting from the surface. TEM; 2400×

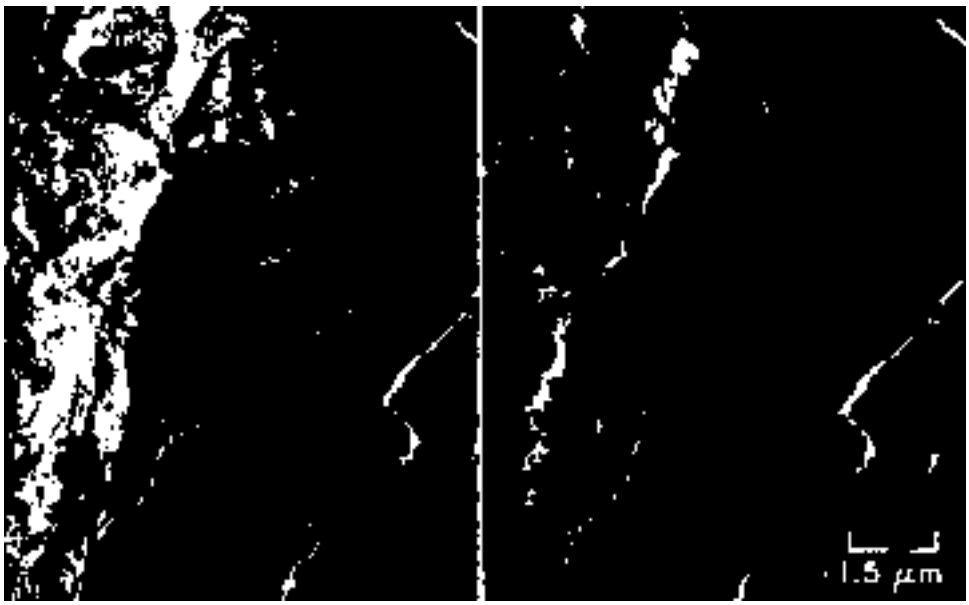


Fig. 6 Intergranular separation in an area of the surface of the hydrogen embrittlement fracture of the type 431 stainless steel mushroom-head closure of Fig. 4 and 5. When viewed in three-dimension, the somewhat rounded separated-grain surfaces show cavities akin to dimples in appearance but more probably sites of previous intergranular

inclusions. The corrosion debris may be seen projecting from the fracture surfaces. TEM; 6600×

References cited in this section

3. L. Raymond, Ed., *Hydrogen Embrittlement Testing,* STP 543, American Society for Testing and Materials, 1972

4. L. Raymond, Evaluation of Hydrogen Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 283–290.

5. G.F. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, p 689–736

6. R. Viswanathan and R.I. Jaffie, Clean Steels to Control Hydrogen Embrittlement, *Current Solutions to Hydrogen Problems in Steels,* American Society for Metals, 1982, p 275–278

7. E.E. Stansbury and R.A. Buchanan, *Fundamentals of Electrochemical Corrosion,* ASM International, 2000, p 371

8. A.W. Thompson, The Mechanism of Hydrogen Participation in Ductile Fracture, *Effect of Hydrogen on Behavior of Materials,* American Institute of Mining, Metallurgical, and Petroleum Engineers, 1976, p 467–479

Hydrogen Reaction Embrittlement

Although the sources of hydrogen may be any of those mentioned previously, this type of embrittlement is quite distinct from hydrogen environment embrittlement. Once hydrogen is absorbed, it may react near the surface or diffuse substantial distances before it reacts. Hydrogen can react with itself, with the matrix, or with a foreign element in the matrix. The chemical reactions that comprise this type of embrittlement or attack are well known and are encountered frequently. The new phases formed by these reactions are usually quite stable, and embrittlement is not reversible during room-temperature aging treatments.

Atomic hydrogen (H) can react with the matrix or with an alloying element to form a hydride (MHx). Hydride phase formation can be either spontaneous or strain induced. Atomic hydrogen can react with itself to form molecular hydrogen (H_2). This problem is frequently encountered after steel processing and welding; it has been termed flaking or fisheyes. Atomic hydrogen can also react with a foreign element in the matrix to form a gas. A principal example is the reaction with carbon in low-alloy steels to form methane (CH_4) bubbles. Another example is the reaction of atomic hydrogen with oxygen in copper to form steam (H_2O), resulting in blistering and a porous metal component.

Hydrogen-Induced Blistering. Blistering is observed in low-strength steels and other metals. It is caused by atomic hydrogen diffusing to internal defects or inclusions, where it precipitates as molecular (diatomic) hydrogen. This generates substantial pressure, which produces blisters. Shatter cracks, flakes, and fisheyes are terms that describe cracks or the surface appearance of a fracture in castings, forgings, wrought alloys, or weldments. These defects are caused by the presence of excessive hydrogen in the liquid melt prior to solidification. Microperforation may also occur, mainly when steels are exposed to very high-pressure hydrogen near room temperature.

Hydrogen-induced blistering is most prevalent in low-strength alloys, and it is observed in metals that have been exposed to hydrogen-charging conditions, for example, acid pickling or corrosion in environments containing hydrogen sulfide. When hydrogen is absorbed into metal and diffuses inward, it can precipitate as molecular hydrogen at internal voids, laminations, or inclusion/matrix interfaces, and it can build up pressure great enough to produce internal cracks. If these cracks are just below the surface, the hydrogen-gas pressure in the cracks can lift up and bulge out the exterior layer of the metal, so that it resembles a blister (Fig. 7). The equilibrium pressure of the molecular hydrogen in the void, which is in contact with the atomic hydrogen in the surrounding metal, is great enough to rupture any metal or alloy. The absorbed hydrogen may come from various sources.

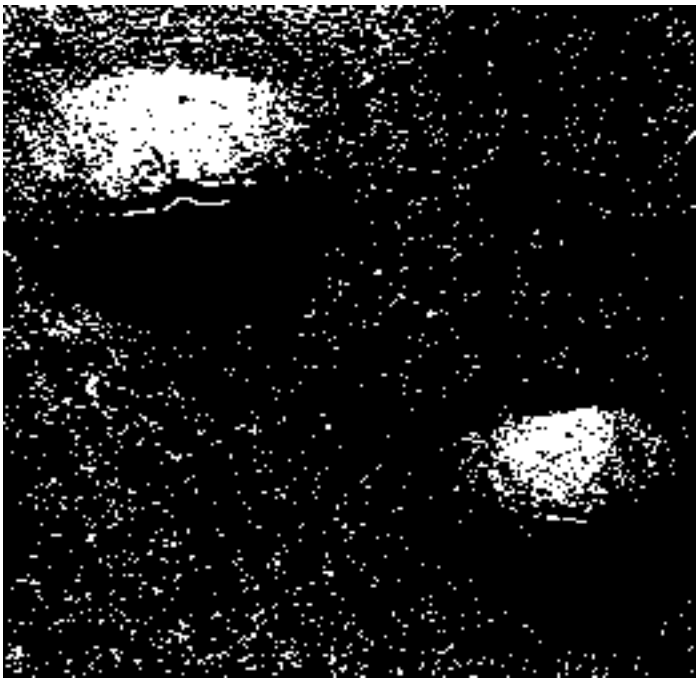


Fig. 7 Hydrogen-induced blistering in a 9.5 mm (3/8 in.) thick carbon steel plate (ASTM A 285, grade C) that had been in service one year in a refinery vessel. 1.5×

Corrosion-generated hydrogen causes blistering of steel in oil-well equipment and in petroleum-storage and refinery equipment. In the refinery, hydrogen-induced blistering has been found most frequently in vessels handling sour (hydrogen-sulfide-containing) light hydrocarbons and in alkylation units where hydrofluoric acid is used as a catalyst. Storage vessels with sour gasoline and propane are highly prone to blistering, but sour crude storage tanks are less prone to blistering, apparently because the oil film of the heavier hydrocarbons acts as a corrosion inhibitor. In storage vessels, blistering is generally at the bottom or in the vapor space where water is present. Gas-plant vessels in catalytic hydrocarbon-cracking units are particularly prone to blistering, because the cracking reaction generates cyanides. The presence of cyanides, hydrogen sulfide, and water enhances hydrogen absorption into steel. Hydrogen-induced blistering also occurs on steel plates used as cathodes in industrial electrolysis.

Line pipe transmitting wet sour gas can develop hydrogen-induced cracking in the pipe wall. When line-pipe steel is exposed to sour brine in the absence of applied stress, a number of cracks parallel to the longitudinal axis of the pipe may develop through the wall, and the tip of one crack may link up with another in stepwise fashion (Fig. 8). This type of cracking is called stepwise cracking, and it can significantly reduce the effective wall thickness of the pipe. It is believed that a single, straight longitudinal crack is less harmful than stepwise cracks. The presence of hydrogen sulfide in the corrodent greatly promotes the absorption of hydrogen into steel. Thus, the fugacity of hydrogen generated during sulfide corrosion is extremely high—often on the order of 10^3 MPa (10,000 atm).

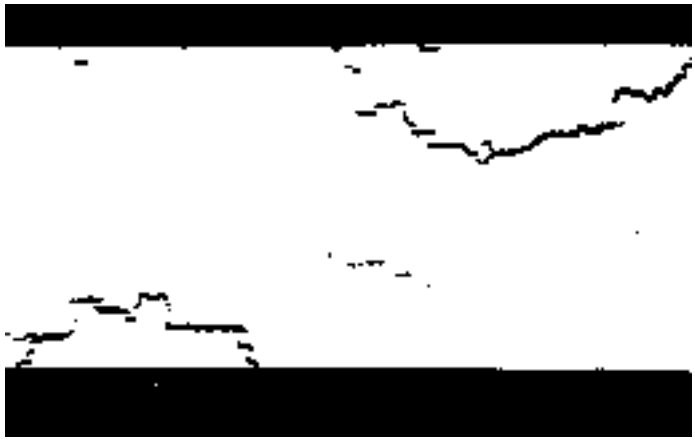


Fig. 8 Stepwise cracking of a low-strength pipeline steel exposed to hydrogen sulfide (H_2S). 6×

Hydrogen-induced cracks in line-pipe steels are always associated with certain metallurgical features, such as inclusions, large precipitate particles, or martensite bands. Elongated inclusions, such as type II manganese sulfides and glassy silicates, are particularly detrimental to hydrogen-induced cracking resistance. The oxygen content in molten steel markedly affects the shape of sulfide inclusions in rolled steel. Manganese sulfides in semikilled steel deform to a lesser extent during hot rolling than those in silicon-aluminum killed steels. Semikilled steels usually contain ellipsoidal manganese sulfides (type I MnS) and generally have better hydrogen-induced cracking resistance than fully killed steels having elongated manganese sulfides (type II MnS). Although the hydrogen-induced cracking resistance of steel increases with decreased sulfur content, lowering the sulfur level alone does not provide immunity to hydrogen-induced cracking. Most modern steels are produced in the killed form, because the usually more-economical continuous casting processes only work for killed steels.

The presence of martensite banding in line-pipe steel increases the susceptibility to hydrogen-induced cracking. The formation of martensite can be controlled by selecting suitable steel compositions and the proper thermomechanical treatment. Addition of approximately 0.3% Cu is known to improve the hydrogen-induced cracking resistance of steel, but the beneficial effect of copper prevails only in environments with relatively high pH (approximately 5 or more) and diminishes in low-pH (3.5 or less) solutions.

Flakes. Heavy steel forgings often contain a number of hairline cracks in the center part, and such cracks are called flakes. On fracture surfaces, flakes appear as small areas of elliptical bright cracks. Flakes are formed during cooling after the first forging or rolling and not during cooling after solidification. Flakes are caused by localized hydrogen embrittlement resulting from internal hydrogen.

The primary source of hydrogen in steelmaking is moisture in the atmosphere and in the additives, and the hydrogen content of molten steel after refining can be as high as 5 to 8 ppm. Hydrogen dissolves more in the fcc crystal structure than in the bcc crystal structure, and the solubility decreases exponentially with decreasing temperature. The lattice solubility of hydrogen in steel is much smaller than 0.1 ppm at room temperature. Therefore, on cooling, hydrogen precipitates as molecular hydrogen at inclusions or micropores, and because such regions are already embrittled by hydrogen, flakes are readily formed by the gaseous hydrogen pressure. It is generally known that flakes form at temperatures below 200 °C (390 °F). Flakes are generally oriented within the forging grain or segregated bands. Flaking sensitivity increases with increasing hydrogen content.

Fisheyes are another example of localized hydrogen embrittlement. The term is used to describe small, shiny spots occasionally observed on the fracture surface of tension specimens from steel forgings or plates having a high hydrogen content. Fisheyes are associated with reduced tensile ductility. Fractographic examination usually reveals microscopic stress raisers, such as pores or nonmetallic inclusions, within the fisheye. Baking or prolonged room-temperature aging of tension specimens frequently eliminates fisheyes and restores tensile ductility.

When this type of hydrogen damage occurs in welding, it is called underbead cracking. This form of cracking develops in the HAZ of the base metal and runs roughly parallel to the fusion line. Because cracking caused by hydrogen may occur hours or days after welding, it is also known as delayed cracking. The factors controlling this type of cracking are dissolved hydrogen, tensile stress, and low-ductility microstructure, such as martensite.

Hydrogen can be transferred to the molten weld pool from the arc atmosphere. Sources of hydrogen include the shielding gas, flux, or surface contamination. As the weld metal cools, it becomes supersaturated with hydrogen, which diffuses into the HAZ, which is in a state of high triaxial tensile stress. As the austenite transforms into martensite on rapid cooling, the hydrogen is retained in this region, and, as a result, the metal is embrittled by its presence. The stresses generated by external restraint and by volume changes due to the transformation can easily produce cracks in this region. More information on hydrogen damage in welds is available in the article “Failures Related to Welding” in this Volume.

Hydrogen attack occurs at high temperatures rather than at or near room temperature. Hydrogen reacts with carbon or carbides within the steel to form methane gas. This may simply decarburize the steel, lowering its strength, or may produce cracks or fissures. This problem does not occur below 200 °C (390 °F). Steel exposed to high-temperature, high-pressure hydrogen appears to be unaffected for days or months and then suddenly loses its strength and ductility. However, it is important to note that hydrogen attack is different from hydrogen embrittlement. Hydrogen attack is irreversible damage, and it occurs at elevated temperatures, whereas hydrogen embrittlement is often reversible and occurs at temperatures below 200 °C (390 °F).

During hydrogen attack of steel, methane bubbles form along grain boundaries; these bubbles subsequently grow and merge to form fissures. Failure by hydrogen attack is characterized by decarburization and fissuring at grain boundaries (Fig. 9) or by bubbles in the metal matrix. This type of hydrogen damage is most commonly experienced in steels that are subjected to elevated temperatures in petrochemical-plant equipment that often handles hydrogen and hydrogen-hydrocarbon streams at pressures as high as 21 MPa (3 ksi) and temperatures up to 540 °C (1000 °F).

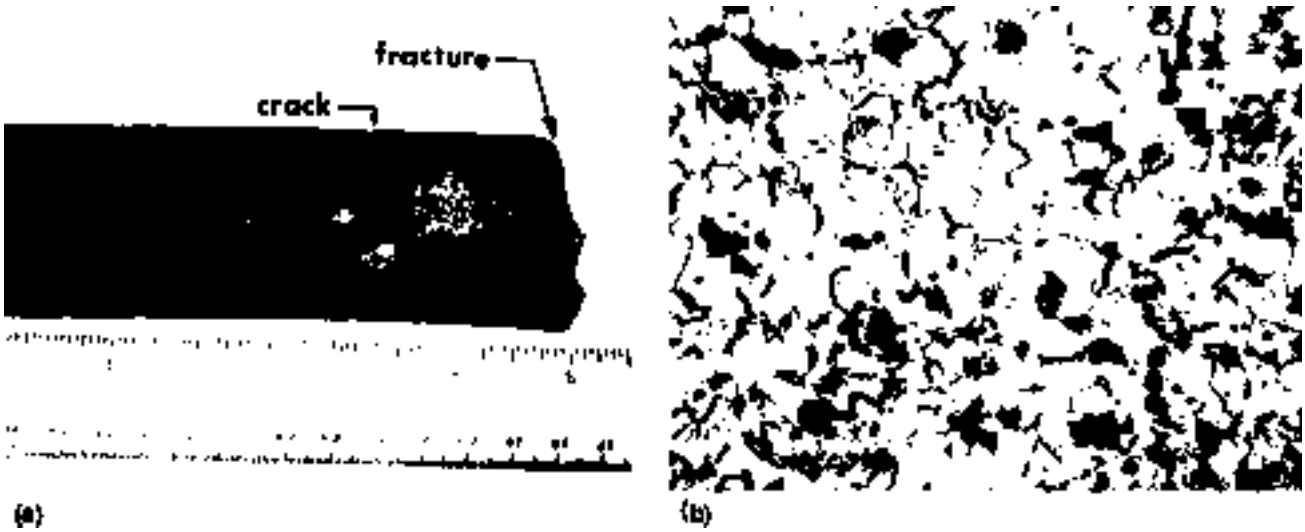


Fig. 9 Section of ASTM A 106 carbon steel pipe with wall severely damaged by hydrogen attack. The pipe failed after 15 months of service in hydrogen-rich gas at 34.5 MPa (5000 psig) and 320 °C (610 °F). (a) Overall view of failed pipe section. (b) Microstructure of hydrogen-attacked pipe near the midwall. Hydrogen attack produced grain-boundary fissures that are radially aligned.

The severity of hydrogen attack depends on temperature, hydrogen partial pressure, stress level, exposure time, and steel composition. Additions of chromium and molybdenum to the steel composition improve the resistance to hydrogen attack. On the basis of industrial experience, the American Petroleum Institute prepared a materials selection guide for hydrogen service at elevated temperatures and pressures (Ref 9). This is known as the Nelson diagram, and it is very useful for proper selection of materials for high-temperature hydrogen service.

Moisture in hydrogen enhances decarburization of steel. Hydrogen gas with a dewpoint of -45 °C (-50 °F) does not decarburize hypoeutectoid steels. Also, the decarburization effect of hydrogen on these steels at temperatures below 700 °C (1290 °F) is negligible. Although hydrogen can severely decarburize steel under certain conditions of temperature and dewpoint, decarburization is not limited to hydrogen atmospheres only, and it may occur in a variety of hydrogen-free heat treatment atmospheres, such as oxygen, air, and carbon dioxide, and in molten salt baths.

Damage by hydrogen embrittlement or blistering may also occur in steel after rapid cooling after prolonged exposure to high-temperature, high-pressure hydrogen. The excess hydrogen cannot then escape during cooling and causes cracking or blistering.

Hydrogen attack can also occur in copper. At elevated temperature, hydrogen absorbed in tough pitch copper internally reduces Cu_2O particles present in the copper, forming pockets of steam. Because water vapor cannot diffuse away, the steam pressure in the pockets can be very high, causing cavities or blisters. Similar behavior can occur in silver.

Cracking from Hydride Formation. A number of transition, rare earth, alkaline-earth metals, and the alloys of these metals are subject to embrittlement and cracking due to hydride formation. Among these metals, the commercially important ones are titanium, tantalum, zirconium, uranium, thorium, and their alloys. The presence of hydrides in these metals can cause significant increases in laboratory-measured strength and large losses in ductility and toughness. As in other types of alloys, excess hydrogen is readily picked up during melting or welding, and hydride formation takes place during subsequent cooling. The use of vacuum melting and the modification of compositions can reduce susceptibility to hydride formation. Hydrogen can often be removed by annealing in vacuum. Welding generally requires the use of inert gas shielding to minimize hydrogen pickup.

The hydride particles often have the form of platelets and show preferred orientation within the parent lattice, depending primarily on the metal or alloy composition. The large volume change associated with hydride formation leads to a strong interaction between the hydride-formation process and externally applied stresses. Applied stresses can cause preferential alignment of hydrides or realignment. In most cases, the hydride phase has a much lower ductility than the matrix.

Reference cited in this section

9. “Steel for Hydrogen Service at Elevated Temperatures and Pressures in Petroleum Refineries and Petrochemical Practice,” API 941, American Petroleum Institute, July 1970

Susceptibility of Various Metals

Most metals and alloys are susceptible to hydrogen damage, and many are susceptible to more than one type of hydrogen damage. Carbon and low-alloy steels can fail by several types of hydrogen damage. Failure by hydrogen embrittlement is often encountered with high-strength steels, especially when the tensile strength is above 1034 MPa (150 ksi). Low-strength steels are considered resistant to hydrogen embrittlement, but they are susceptible to hydrogen-induced blistering. In high-temperature hydrogen environments, carbon and low-alloy steels are also subject to hydrogen attack. However, steels are not susceptible to cracking caused by hydride formation.

Hydrogen Damage and Embrittlement

Stainless Steels

The susceptibility of the different types of stainless steels (austenitic, ferritic, martensitic, and precipitation hardening) to failure by hydrogen damage varies widely.

Austenitic Stainless Steels. Although austenitic stainless steels are very susceptible to chloride SCC, they are more resistant to hydrogen embrittlement than bcc steels. This may be due to the larger size of the interstitial site in the center of the unit cell. Another factor in the resistance of austenitic stainless steels to hydrogen

damage may be the low hydrogen diffusivities in these steels. Hydrogen diffusivity in austenite is several orders of magnitude lower than that in ferrite. However, because the outgassing rate is lower, low-energy trapping of hydrogen can occur in fcc iron-base alloys at imperfections generated by deformation (e.g., dislocations and ε or α martensite) (Ref 2). There is a general correlation between hydrogen resistance and austenite stability.

Ferritic Stainless Steels. In the annealed condition, ferritic stainless steels are very resistant to hydrogen damage, because of their low hardness. However, in the cold-worked or as-welded condition, ferritic stainless steels are susceptible to hydrogen embrittlement.

Martensitic and precipitation-hardening stainless steels have high strength and are subject to hydrogen embrittlement. The susceptibility to cracking increases with increasing yield strength, and almost any corrosive environment can cause failure by hydrogen embrittlement in these types of stainless steels. Even a mild environment, such as fresh water at room temperature, may cause cracking in especially susceptible alloys. In one case, for example, a precision bolt of 17-4 PH (type 630) stainless steel cracked from hydrogen embrittlement when exposed for approximately nine months in a warm seacoast atmosphere while in contact with 7075 aluminum alloy (Ref 10).

References cited in this section

2. J.P. Hirth, Theories of Hydrogen Induced Cracking, *Hydrogen Embrittlement and Stress Corrosion Cracking,* R. Gibala and R.F. Hehemann, Ed., American Society for Metals, 1984, p 29–41

10. *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 235

Nickel-Base Alloys

Hydrogen embrittlement of nickel-base alloys is exemplified by three forms: brittle (usually intergranular) delayed fracture, a loss in reduction of area while often retaining a microvoid coalescent fracture, or a reduction in properties such as fatigue strength. Although cleavage-type cracks have been reported in nickel-base alloys, they are not the predominant mode of fracture. Table 2 lists alloys and environments in which hydrogen embrittlement is known to occur.

Table 2 Nickel-base alloys exhibiting hydrogen embrittlement

Alloy	Environment
600	**Cathodic charge/tensile test**
C-276, C-4, MP35N, 625, 718	**…**
Pure nickel	**Cathodic charge H_2SO_4**
K500	**Sour oil and gas environment**
X750	**5% NaCl + 5% CH_3COOH + H_2S**
718, A286, 625	**Hydrogen gas (34 MPa, or 5000 psi) charged at room temperature**
Pure nickel, 600	**Hydrogen gas charged at 723 K, tested at room temperature**
K500	**Cathodic charge, seawater**

Annealed Nickel Alloys. With few exceptions, fully annealed nickel-base alloys are essentially immune to hydrogen-induced delayed failure, but alloy B-2 is susceptible to hydrogen embrittlement in H_2S-containing environments even in the annealed condition. Some nickel alloys do not show significant susceptibility to hydrogen embrittlement. These alloys include alloys 825, G, and G-3, and alloy 20Cb3, as well as austenitic nickel-base alloys with correspondingly high iron contents.

Large reductions in the ductility of nickel-base alloys due to hydrogen embrittlement have been observed at strain rates near 10^{-4} to 10^{-6} s^{-1}. Only when the strength level of the alloys is increased by cold working or by heat treatment is the incidence of hydrogen embrittlement significant. Also, only the low-iron-containing alloys (generally less than 10% Fe) are significantly affected by hydrogen-induced delayed fracture.

The fracture mode may remain microvoid coalescence, but annealed nickel has been shown to be susceptible to intergranular hydrogen embrittlement. This phenomenon of intergranular fracture appears to be related to the sulfur content/segregation in nickel.

Nickel-Base Superalloys. In general, hydrogen embrittlement has not been a serious problem with heat-resistant alloys. However, laboratory tests have shown that certain nickel-base alloys are susceptible to embrittlement when exposed to highly oxidizing environments or to pure hydrogen at a pressure of 34 MPa (5 ksi) and a temperature of 680 °C (1250 °F). The susceptibility of precipitation-hardenable alloys, such as alloys X-750, alloy K-500, and alloy IN-718, has been reported from laboratory tests and from field exposures with less susceptibility at lower strength levels. Some of the solid-solution alloys, such as alloy C-276 and alloy 625, are susceptible to hydrogen embrittlement under conditions of hydrogen charging when cold worked to high strength levels. Aging at temperatures (540 °C, or 1000 °F) at which ordering and/or grain-boundary segregation can occur greatly increases this susceptibility to hydrogen embrittlement.

Aluminum and Aluminum Alloys

Hydrogen damage occurs occasionally in aluminum and aluminum alloys, but it is not usually a serious problem. When a high-strength aluminum alloy is cathodically charged, its ductility is reduced. Hydrogen embrittlement of aluminum alloys can result in intergranular or transgranular cracking. Dry hydrogen gas does not cause significant hydrogen embrittlement in aluminum alloys, but SCC of Al-Zn-Mg alloys in moist gases probably involves hydrogen embrittlement. Problems with hydrogen in aluminum alloys arise mostly from formation of gas-filled voids during solidification. These voids can affect both cast and wrought products. In ingots for wrought products, the presence of hydrogen gas in voids inhibits healing on subsequent working and is responsible for such defects as bright flakes in thick sections and blisters on the surface of annealed or heat treated material. The principal effect of bright flakes is a reduction in short-transverse ductility.

Copper and Copper Alloys

Copper and its alloys are not susceptible to attack by hydrogen, unless they contain copper oxide. Tough pitch coppers, such as C11000, contain small quantities of Cu_2O. Deoxidized coppers with low residual deoxidizer contents—C12000, for example—may contain Cu_2O but contain less than tough pitch coppers. These deoxidized coppers are not immune to hydrogen embrittlement. Deoxidized coppers with high residual deoxidizer contents, however, are not susceptible to hydrogen embrittlement, because the oxygen is tied up in complex oxides that do not react appreciably with hydrogen.

Titanium and Titanium Alloys

Hydrogen damage in titanium alloys results from embrittlement caused by absorbed hydrogen. Hydrogen may be supplied by a number of sources, including water vapor, pickling acids, and hydrocarbons. The amount of absorption depends primarily on the titanium oxide film on the metal surface, and an adherent unbroken film can significantly retard hydrogen absorption.

Titanium and its alloys become embrittled by hydrogen at concentrations that produce a hydride phase in the matrix. The exact level of hydrogen at which a separate hydride phase is formed depends on the composition of the alloy and the previous metallurgical history. In commercial unalloyed material, this hydride phase is normally found at levels of 150 ppm of hydrogen; however, hydride formation has been observed at levels as low as 40 or 50 ppm of hydrogen. Figure 10 is an example of hydrided titanium.

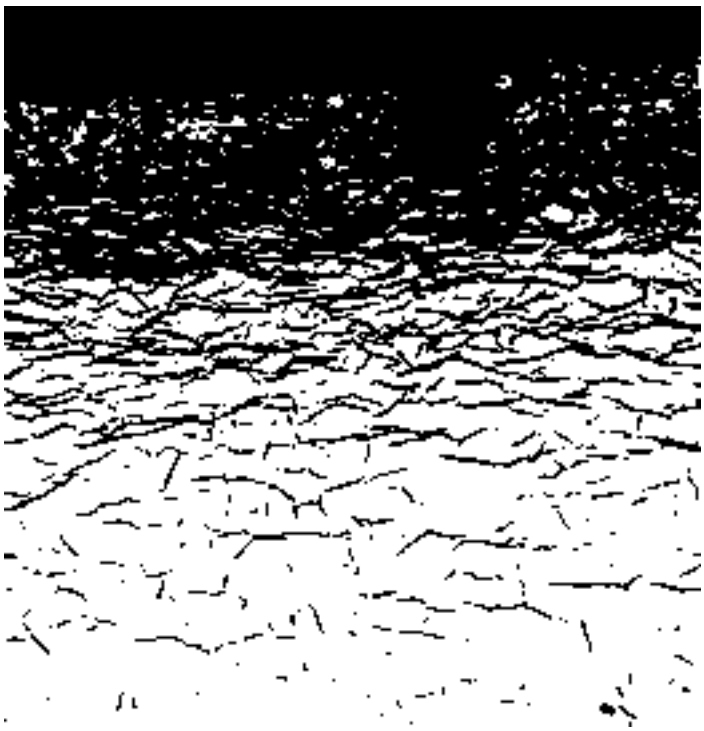


Fig. 10 Severely hydrided unalloyed titanium

At temperatures near the boiling point of water, the diffusion rate of hydrogen into the metal is relatively slow, and the thickness of the layer of titanium hydride formed on the surface rarely exceeds approximately 0.4 mm (0.015 in.), because spalling takes place when the hydride layer reaches thicknesses in this range.

Hydride particles form much more rapidly at temperatures above approximately 250 °C (480 °F), because of the decrease in hydrogen solubility within the titanium lattice. Under these conditions, surface spalling does not occur, and the formation of hydride particles through the entire thickness of the metal results in complete embrittlement and high susceptibility to failure. This type of embrittlement is often seen in material that has absorbed excess hydrogen at elevated temperatures, such as during heat treatment or welding, and subsequently has formed hydride particles during cooling.

There have been instances of localized formation of hydrides in environments where titanium has otherwise given good performance. Investigations of such instances suggest that the localized formation of hydrides is the result of impurities in the metal (particularly the iron content) and the amount of surface contamination introduced during fabrication.

There is a strong link between surface-iron contamination and formation of hydrides of titanium. Severe hydride formation has been noted in high-pressure dry gaseous hydrogen around particles of iron present on the surface. Anodizing in a 10% ammonium sulfate solution removes surface contamination and leads to thickening of the normal oxide film.

In chemical-plant service, where temperatures are such that hydrogen can diffuse into the metal if the protective oxide film is destroyed, severe embrittlement may occur. For example, in highly reducing acids where the titanium oxide film is unstable, hydrides can form rapidly. Hydrogen pickup has also been noted under high-velocity conditions where the protective film erodes away as rapidly as it forms.

Hydrogen contents of 100 to 200 ppm may cause severe losses in tensile ductility and notched tensile strength in titanium alloys and may cause brittle delayed failure under sustained loading conditions. The sensitivity to hydrogen embrittlement from formation of hydrides varies with alloy composition and is reduced substantially by alloying with aluminum.

Care should be taken to minimize hydrogen pickup during fabrication. Welding operations generally require inert gas shielding to minimize hydrogen pickup. Hydrogen can be removed from titanium by annealing in vacuum.

Transition and Refractory Metals

Tantalum, zirconium, uranium, thorium, and alloys of these metals, when exposed to hydrogen, can sustain severe damage due to hydride formation.

Tantalum absorbs hydrogen at temperatures above 250 °C (480 °F), resulting in the formation of tantalum hydride. Hydrogen absorption also can occur if tantalum is coupled to a more active metal in a galvanic cell, particularly in hydrochloric acid. As little as 100 ppm of hydrogen in tantalum cause severe embrittlement by hydride formation.

Zirconium and its alloys are highly susceptible to embrittlement by hydride formation. Absorption of gaseous hydrogen into zirconium and its reaction with zirconium to produce the hydride ZrH_2 occur at an extremely fast rate at 800 °C (1470 °F). This hydride is very brittle, and it may be crushed into a powder. The hydrogen may then be pumped off, leaving a powder of metallic zirconium.

Zirconium alloys can pick up substantial amounts of hydrogen during exposure to high-pressure steam. Zirconium with 2.5% Nb corrodes in water at 300 °C (570 °F) at a rate of approximately 25 μm per year and picks up hydrogen at the rate of 2 to 4 mg/cm^2 (13 to 26 mg/in.^{2}) per day. Hydrogen can be removed from zirconium and its alloys by vacuum annealing.

Uranium, uranium alloys, and thorium are susceptible to hydrogen embrittlement. Hydrides in thorium have been identified as ThH_2 and Th_4H_{15}. The hydride in uranium is UH_3. Uranium absorbs hydrogen from three sources: fused salt bath annealing, etching and electroplating operations, and corrosion. A hydrogen pickup of 1 to 2 ppm as a result of salt bath annealing noticeably reduces elongation. Hydrogen absorbed during etching and electroplating is concentrated at the surface and does not degrade tensile properties. Corrosion of uranium in water or water vapor reduces tensile ductility. Uranium is embrittled in water over the pH range of 5 to 10.

Analysis of Hydrogen Embrittlement in Commodity-Grade Steels

Debbie Aliya, Aliya Analytical

Hydrogen embrittlement refers to the degradation of mechanical properties due to the presence of dissolved atomic hydrogen, and it often results in ambient-temperature and delayed cracking of steel parts that have recently been assembled. Because of the wide economic impact in commodity-grade components and the frequent difficulty in diagnosing HE failures, this section discusses preservice and early-service fractures of commodity-grade steel components suspected of hydrogen embrittlement. Preservice and early-service failures of commodity steels strongly suggest the delayed nature of hydrogen embrittlement, but in many cases, it may difficult to conclusively confirm hydrogen as the root cause; other factors besides hydrogen may be contributing.

The sources of the difficulty in analyzing parts that are suspected of hydrogen embrittlement failure are many. One reason is because the effects of hydrogen embrittlement can be quite varied. In many cases, the loss of strength due to the presence of hydrogen is only one of many factors that allow an unexpected fracture. In other cases, the loss of strength is spectacular, leaving only 10% of the expected tensile strength compared to an unembrittled component. However, even in these severe cases, it may be difficult to prove with any degree of certainty that the most significant cause of the fracture was the presence of hydrogen.

When the compromised mechanical properties in the subject part(s) are due to hydrogen embrittlement, there may be a tendency to focus on the reduction of hydrogen as a solution. In fact, the hydrogen concentration is only one of many factors that combine to create a realistic picture of "the physical root cause" of the unexpected hydrogen embrittlement fracture.

Preservice and Early-Service Failures

Hydrogen embrittlement can be a likely potential problem in preservice and early-service fractures of commodity-grade steel components. In certain applications, where steels are processed so that their yield strength is a high percentage of their tensile strength (such as springs, spring clips, and externally threaded high-strength or self-tapping fasteners), hydrogen embrittlement can cause a lot of consequential economic damage. This is especially a problem when the cracks are found after assembly, and labor and other components of the assembly must be scrapped along with the cracked parts. Parts with very hard surface layers, intentionally produced by carburizing or carbonitriding, also increase the likelihood of hydrogen embrittlement problems.

People employed in industries that use high volumes of small components also are sometimes faced with the prospect of creating lot-containment plans when only a few of a large batch of components were diagnosed with hydrogen embrittlement. This happens when any further failure is unacceptable for either safety or economic reasons. Significant expenditures for prevention of hydrogen embrittlement may be worthwhile economically if the damage of the often-inexpensive embrittled components causes a more complex and costly assembly to be scrapped.

Hydrogen embrittlement problems are detected during shipping, inspection, and shortly after placement in service. If the part broke when it was subject to an impact condition, it would not usually be considered a hydrogen embrittlement problem, although it is possible that a preexisting hydrogen embrittlement condition allowed the impact event to separate the fragments more easily. This small example illustrates the difficulty encountered by anyone who tries to make up a simplified testing and data interpretation guideline for any type of failure.

In these types of situations, the first step is to review the history of the failed part(s) and determine if conditions necessary for delayed hydrogen-induced cracking may be present. Hydrogen embrittlement in steels is most prevalent in components that:

- Have been in contact with atomic hydrogen at some time during their processing
- Are subject to sustained tensile stresses (applied and/or residual)
- Have high hardness from strengthening by heat treatment, cold working, or surface treatment

Other major factors include the hardness range specified and the consistency of actual hardness values; design stress level and actual stress consistency; dimensional consistency; composition, including intentional alloying elements and residuals; and microstructure. The significance of dimensional consistency should not be overlooked, because in some designs, small changes in dimensions can be associated with large changes in localized stress levels.

These factors are briefly reviewed in a little more detail in the following sections. This is followed by discussions on the factors that may influence the threshold of delayed hydrogen stress cracking.

Hydrogen Sources. Many small commodity-grade components are produced with various secondary manufacturing operations that provide ready sources for the atomic form of hydrogen. The most common causes are toward the top of the list:

- Electroplating
- Acid cleaning
- Phosphate conversion coating
- Caustic cleaning
- Heat treating where hydrogen is used as a protective atmosphere and parts are not tempered after exposure
- Moisture pickup during melting and refining of molten steel

Although caustic cleaning has been associated with embrittlement of steels, electroplating processes can be particularly damaging, because many of the baths produce heavy concentrations of atomic hydrogen at the interface of the base material and plated layer. While the hydrogen atoms can diffuse readily into the iron matrix of the steel, the zinc or cadmium plating acts as a barrier to any hydrogen atoms that may have otherwise randomly diffused away from the structural portion of the component. This creates a situation where hydrogen builds up in the steel.

Note that the chemical reaction that creates the hydrogen during acid cleaning may be very similar to that which corrodes the base metal in SCC. However, the focus here is on pre- or early-service failures. Of course, many corrosion processes can generate hydrogen that can cause embrittlement, but pre- or early-service failures are obviously limited to the context of hydrogen charging during processing of the part. If the local hydrogen concentration and stress-level combination is more than the material can resist (over the threshold level), the crack starts to grow. However, the distinction between hydrogen embrittlement and stress corrosion still should be considered carefully in preservice situations. For example, evidence of hydrogen embrittlement due to an excessive time in the acid cleaning bath should be described as hydrogen embrittlement from the manufacturing process. However, evidence of hydrogen embrittlement due to corrosion of the assemblies that sat on a shipping dock unprotected from ice-melting salt might more correctly be described as SCC.

Sustained tensile stresses are more likely to cause hydrogen embrittlement related cracking, because tensile stresses actually change the shape of the interatomic spaces. This makes it easier for the hydrogen atom to “fit” in. Thus, there is a synergistic effect for the efficiency with which hydrogen atoms cause embrittlement. The regions of a part that have high tensile stress are more likely to crack, not just because many cracks initiate in the presence of a tensile stress, but the tensile stress itself causes the hydrogen atoms to diffuse toward the areas of highest tensile stress. Grain boundaries stressed in tension seem to be a particularly attractive location for the hydrogen atoms. The actual manner in which the presence of the hydrogen in the high-tensile-stress grain boundaries reduces the material strength or otherwise embrittles the material is not entirely known or necessarily explained by one mechanism, such as the decohesive theory previously noted (Ref 1, 2).

Many people who are involved for the first time in a hydrogen embrittlement problem may wonder why there is a sustained tensile stress before the part is put into service. In this regard, residual stresses from heat treating or assembly stresses are often much higher than most people realize and can produce residual tensile stresses in components. Thus, a practical understanding of residual stresses is important, along with techniques in the measurement of residual stresses (e.g., see “X-Ray Diffraction Residual Stress Measurement in Failure Analysis” in this Volume, or the article “Residual Stress Measurements” in *Mechanical Testing and Evaluation,*) Volume 8, in *ASM Handbook.*

As an example of a part with a sustained tensile stress due to assembly, consider externally threaded fasteners. When a bolt is threaded through holes in two pieces of sheet metal, it must be tightened to hold the components securely in place. While most threaded fastener installation specifications still (in 2002) call out a minimum torque, it is the tensile clamping stress that actually provides the fastening effect. Thus, the whole body of the fastener between the underside of the head and the nut is in tension before the whole assembly is ever placed into service. Normally occurring stress concentration factors mean that the highest stress is at the root of the external thread. Because cracks can initiate wherever the local stress exceeds the local strength, thread roots are an area that has a higher chance than the surrounding material to initiate a crack. Again, the presence of the hydrogen allows cracking to occur at stress levels much lower than those required to propagate a crack in the absence of the hydrogen.

As another example, consider a U-shaped spring clip that is inserted into a notch in a plastic molded part. In order for the clip to stay in place during its service life, the assembly is designed so that the clip is held in a position that is slightly pinched closed and from which it “wants” to spring back open. This configuration creates a tensile stress at the outer surface of the U-portion of the clip. The presence of any tool marks or scratches can increase the local stress level to an even greater degree.

References cited in this section

1. U.K. Chatterjee, S.K. Bose, and S.K. Roy, *Environmental Degradation of Metals*, Marcel Deckker, 2001

2. J.P. Hirth, Theories of Hydrogen Induced Cracking, *Hydrogen Embrittlement and Stress Corrosion Cracking*, R. Gibala and R.F. Hehemann, Ed., American Society for Metals, 1984, p 29–41

Factors Affecting Delayed Hydrogen Stress Cracking

As previously noted in the section “Internal Reversible Hydrogen Embrittlement,” there is a particular value of the stress or threshold that causes a particular component to crack due to hydrogen embrittlement. This threshold is dependent on a number of factors, including:

- The nominal strength of the unembrittled part
- The amount of hydrogen present in the steel
- The location of the hydrogen in the microstructure
- The presence of other embrittling elements or microstructural phases

The latter item is particularly important, because other embrittling elements or phases in the grain boundaries can cause intergranular fracture or assist hydrogen concentration in the grain boundaries. It should also be noted that hydrogen embrittlement may be present in a component without causing a crack. If the combined levels of stress, hydrogen concentration, and susceptibility are not present, evidence of the embrittlement may not appear.

The Nominal Strength of the Unembrittled Part. While higher-strength parts have less ability to deform in general, the effect of hydrogen is strongly nonlinear in this respect. Many older references (1970s and 1980s) use 40 HRC or tensile strength of 1240 MPa (180 ksi) as the lower susceptibility limit for hydrogen embrittlement. This is an important point, because components that have a local equivalent hardness value of 45 HRC may lose up to 90% of the strength value that they would have without the hydrogen. This shows that a lower hardness and thus, lower-strength part (by standard static load test rating systems), may perform when a higher-strength part fractures! Design engineers who are not familiar with hydrogen embrittlement often have trouble understanding that the way to prevent recurrence of the fracture is to lower the specified strength. While components of much lower strength levels than 40 HRC have also been found to have hydrogen embrittlement, these situations are less common. The residual stress levels required to produce hydrogen embrittlement in steels of hardness values in the Rockwell B range are very high and are uncommon even after reasonably severe forming operations. Heat treated medium-carbon steels seem to have a greatly increased susceptibility to hydrogen embrittlement related cracks when their hardness level is above 43 HRC. Differences in purity levels and other microstructure features, as well as production volumes, may affect both the susceptibility of the steels to HE and the analyst's ability to detect that susceptibility due to the presence of actual cracks.

The Amount of Hydrogen Present in the Steel. As more hydrogen is present, the embrittling effect increases. Some materials exhibit a threshold combination of stress and hydrogen concentration. Below this value, there is no danger of hydrogen embrittlement cracking (Ref 11). Because most industrial processes for high-volume

parts have significant variations in actual processing conditions, it may be extremely difficult to know what the actual range of hydrogen concentration is. It is obviously even more difficult to know which particular part had the most damaging combination of stress, hardness, and hydrogen concentration. Also, note that the type of hydrogen that causes the classical delayed hydrogen stress cracking is atomic or "nascent" hydrogen, while the formation of molecular (H_2) hydrogen gas in steels or metals is a different type of hydrogen damage.

The Location of the Hydrogen in the Microstructure. In general, the hydrogen is most damaging when it is in the grain boundaries. To compound this effect, the grain boundaries are often the most comfortable and accessible position for the hydrogen atoms, especially those that enter the steel through cleaning or plating processes or heat treating operations. Hydrogen atoms also tend to migrate and reside at many nonmetallic inclusion boundaries (boundaries of internal "impurities" such as sulfides or oxides) or at imperfections (dislocations) in a crystal lattice. It has been well known for decades that proper postplating baking of electroplated parts can minimize the occurrences of cracking due to hydrogen embrittlement. For many years, people thought that the hydrogen was removed from the steel by the baking process, and many older references state that baking removes the hydrogen. In fact, the currently accepted theory explains that the mild heat exposure recommended for postplating baking (generally, approximately 200 °C, or 390 °F) gives the hydrogen atoms that are concentrated in the grain boundaries enough thermal energy to diffuse to nonmetallic inclusion boundaries or dislocations, so called "deep traps." It is much more difficult for the hydrogen to assist in crack propagation when it is in these deep-trap positions.

The Presence of Other Embrittling Elements or Microstructural Phases. The fact is that it is often very difficult, especially within the budget and time allowed for many investigations, to determine with certainty whether other embrittlement phenomena have contributed to the brittle condition under investigation. Antimony, phosphorus, tin, sulfur, lead, and other low-melting-point impurities, which are generally very expensive to analyze, may be contributing to the embrittlement situation. These elements come from the raw steel and, similar to hydrogen, are more damaging when they are present in the grain boundaries.

As with hydrogen, it is very difficult to find these other elements at the grain boundaries that they are supposedly embrittling, because they are generally present in very small amounts. The energy-dispersive x-ray spectroscopy (EDS) unit of a scanning electron microscope (SEM) that is so useful for many failure analysis projects is useless for this analysis, because the layers of antimony, tin, and so on are so thin that they are not detectable with an EDS. They would often be detectable with an Auger spectograph, if one had very clean (pristine) specimens. However, for many projects for small, inexpensive commodity components, it is difficult to convince the parties who have suffered the loss to invest in an Auger analysis, even if a clean-enough part is obtained.

Thermal exposures can also cause intergranular fracture, such as the well-known temper embrittlement of low-alloy steels. Some of the extended baking times that have been recommended for hydrogen embrittlement prevention in electroplated parts may actually increase the concentration of these other low-strength elements in the grain boundaries, even though the temperature ranges suggested are different from those that are said to cause temper embrittlement.

Reference cited in this section

11. P.F. Timmins, Chapter 3, *Solutions to Hydrogen Attack in Steels*, ASM International, 1997

Diagnosing Hydrogen Embrittlement

While the damaging effects of hydrogen on many metals have been well documented for decades, proving that components have been subject to hydrogen embrittlement can still be very challenging. Some of the reasons include the synergistic nature of hydrogen embrittlement with stress and grain-boundary conditions. If the microstructure has a noticeable second phase in the grain boundaries, the experienced and competent analyst

refrains from giving a high-certainty diagnosis of hydrogen embrittlement. The effects of hydrogen embrittlement also can be varied, and many other damage mechanisms can cause fracture appearances that are very similar to those displayed by hydrogen embrittlement.

Another difficulty is the detection of hydrogen as proof, because detecting the miniscule concentrations of hydrogen that could cause the loss of strength is difficult and costly. Analysis of the amount of hydrogen present in a part at the time of the fracture is also virtually impossible, because the hydrogen that facilitated the crack usually evaporates from the fracture surface shortly after the crack opens. Moreover, bulk hydrogen analysis often sheds little light on the concentrations of hydrogen present in the thin layers of material adjacent to what became the crack surfaces.

Other reasons may be more closely related to budgetary constraints than technical limitations. Preservice and early-service failures of commodity steels may strongly suggest hydrogen embrittlement, but budgets may not allow for extensive analysis (depending on the extent of failure). The financial resources available to perform a failure analysis in such cases are often limited, if nobody was injured. In contrast, SCC (where parts are frequently components of costly equipment and often cause expensive, unplanned shutdowns) may require funds for a more comprehensive analysis.

In any case, it is often less important to be certain that the root cause was hydrogen embrittlement than it is to become aware of underlying contributors, such as high hardness or good design to minimize localized tensile stresses in the components. Despite the difficulties in coming to a certain conclusion in a given situation, it is usually not very difficult to determine whether hydrogen embrittlement might have been a major contributing factor in a particular fracture. The following guidelines are given to help plan a failure analysis that is geared to find out whether hydrogen embrittlement was a likely significant factor in an unexpected fracture. Properly performing this analysis may help to diagnose what the problem is, whether or not hydrogen embrittlement is involved.

Streamlining the analysis process does not imply that relevant tests may be left out or that a competent analyst does not have to collect background data and review the data before conclusions are drawn. The purpose of this section is to help make the best use of the failure analysis resources. People involved in failure analysis are often asked if they can do a hydrogen embrittlement test. What this often means is, “Can you do a fracture analysis and determine whether the fracture was due to hydrogen embrittlement?” Of course, it also sometimes means, “Can you test my as-manufactured parts to find out if they are free from the combination of factors that might cause hydrogen embrittlement?” Many other meanings are also possible. It behooves all concerned to clarify the meaning of the question prior to proceeding with the investigation.

General Questions to Consider on the Possibility of Hydrogen Embrittlement. The following items are characteristics that should be considered in evaluating the possibility of HE:

- *Is the subject component broken?* An intact part may be embrittled and just not broken yet. It is difficult to give a certain answer that an intact part is free of embrittling factors, unless a sustained load test is conducted to see if the part breaks. A broken part provides a fracture path to characterize. Intergranular fracture at the initiation site is another potential indicator of HE.
- *Was the part deformed noticeably before or during the cracking?* Embrittlement means that the part broke in a macrobrittle manner and has more or less the same shape after the fracture that it had before, except for the extra surfaces.
- *Has the subject component been in service for more than a week?* Most hydrogen embrittlement happens during assembly or relatively soon after being placed in service. Once something is in service, even if it turns out to be hydrogen embrittlement, there are other questions regarding service conditions that should be included in the investigation, and a more comprehensive and open-structured investigation is called for than the streamlined protocol suggested here.
- *Has the part been electroplated or coated with iron or zinc phosphate?* Electroplating and phosphating can often cause high concentrations of atomic or nascent hydrogen to diffuse into the material. This is a known source of hydrogen.
- *Has the part been acid cleaned?* The acid reacts with the iron and creates atomic hydrogen at the part surface.
- *Does the part have a hardness value over 40 HRC?* Higher-hardness parts are more susceptible to hydrogen embrittlement.

- *Did the part break due to a sudden impact load?* It takes time for the hydrogen to diffuse to the final crack path. Of course, it is possible that the part was already cracked due to fabrication stresses, but nobody noticed until a sudden load was applied. The microfractographic features are different in this case from a part that fractured entirely due to hydrogen embrittlement.
- *Has the part been heavily cold worked?* Heavy cold work, especially in parts that are not subsequently stress relieved, can create residual stress levels that are very near the yield strength of the material. This is an ideal situation for hydrogen-assisted cracks to propagate.

Minimum Test Plan. Even if all the answers to an initial set of questions point to the probability of hydrogen embrittlement being the damage mechanism, other questions must be answered prior to coming to a valid conclusion, and other questions must be answered if an appropriate prevention strategy is desired. These other questions usually require testing. A minimum test plan for a suspected hydrogen-embrittled component follows. The analyst must keep in mind that many other damage mechanisms cause fracture appearances that are very similar to those displayed by hydrogen embrittlement.

Moreover, it is also important to reiterate that hydrogen embrittlement can be manifested in various forms. Less-severe hydrogen embrittlement cracks may even grow by microvoid coalescence, rather than the intergranular cracking typical of high-strength steels. However, most of the hydrogen embrittlement cracks that initiate failure analysis projects do have some intergranular crack growth near the initiation site that grow under a sustained, static stress. In the more subtle mechanical-property degradation cases, the hydrogen causes a significant reduction in ductility, especially during slow strain-rate load test.

Visual examination for fracture surface orientation. The goal is to determine the orientation of the stresses that caused the fracture. Are these stresses in line with what was expected from the loading that was applied? Is there a reason that there would be high residual stresses that could create the crack geometry?

Visual examination, if possible, on larger parts, or optical or electron microscopy on smaller parts to determine crack initiation site. For small clips and threaded fasteners, the fracture surface often looks the same everywhere, so stress analysis principles are required to get an idea of where the origin might be. For example, as previously noted, in an externally threaded fastener, the initiation is usually at or near the thread root. If the crack does not travel through the thread root, it is likely that a manufacturing imperfection, such as a thread rolling lap, is present. This does not mean that the damage mechanism was not hydrogen-assisted cracking. However, if the goal is to determine “why this part cracked,” it is possible that the lap was an important factor that allowed the crack to propagate.

Visual examination should also include focusing in on the general appearance, color changes, nicks, gouges, machining marks, other damage, and so on. Any such features should be studied and thought about in order to form a reasoned explanation of whether they might be indicative of a factor that contributed to the fracture. Even if the reasoning is incorrect, taking the time to formulate a line of reasoning increases the chances of a proper analysis.

Scanning electron microscopy to determine the crack path with respect to the microstructure, especially at the initiation site. Classical hydrogen embrittlement generally creates intergranular cracks, especially in the area near crack initiation. However, as already mentioned, hydrogen can also have other effects that limit ductility on a macroscale but still allow some microductility. The presence of very small ductile dimples on grain-boundary facets can be confusing to the beginning practitioner. It is important not to go to a high magnification before examining the part at lower magnifications. In plated parts, the crack may be intergranular at the initiation site and all around the edge of the fracture surface adjacent to the plated surface layers. The interior of the part may have cleavage or ductile dimples showing some microvoid coalescence.

The presence of plating on the crack surface, easily detectable, in general, with the EDS microchemical analysis feature of a SEM, can indicate that a crack was present prior to the plating operation. If the crack path under the plating is intergranular, it could have been hydrogen-assisted cracking due to hydrogen in the raw material or from a heat treating operation; however, this is rather unusual in most modern steelmaking and proper heat treating operations. Quench cracks are often intergranular and usually happen before the part goes through the plating operation. This might allow plating to be deposited on the crack surface.

Some of these problems are difficult to sort out without spending quite a bit of time doing careful, thoughtful lab work. Sometimes, it is impossible to distinguish hydrogen embrittlement from other forms of embrittlement.

Optical metallography must be performed to determine whether there are any microstructural conditions that might be contributing to intergranular fracture, if the part does show intergranular fracture. For example, if the

part is full of retained austenite or intergranular carbides, that might be a bigger problem than any hydrogen in the part. If the microstructure has a noticeable second phase at the grain boundaries, the experienced and competent analyst refrains from giving a high-certainty diagnosis of hydrogen embrittlement. This emphasizes the importance of a good, general materials background for the failure analyst.

Recommendations on Prevention of Hydrogen Embrittlement. The final stage of an effective investigation is the process of making recommendations to prevent future failures. In fractures suspected of hydrogen embrittlement, recommendations on prevention and remedial action can be most important. Rather than concentrating on trying to prove the exact metallurgical factors that allowed the crack to happen, the project budget may be more effectively spent on educating those involved in the more practical aspects of prevention.

One of the features of hydrogen embrittlement that often tips the balance in favor of additional expenditures for prevention is the fact that, for many applications, the fractures that result from the embrittlement do not happen until several hours or days have passed since the part was assembled. Suppliers of subassemblies are often subject to heavy financial penalties if they supply parts that are obviously defective. There have been many cases where the embrittled parts looked perfectly normal after assembly yet arrived in a cracked condition after a day or two. In industries that subscribe to the just-in-time delivery philosophy, unexpected delays due to broken parts and the associated sorting and part shortage problems can obviously be perceived to be a cause of economic damage.

Following are guidelines for preventing hydrogen embrittlement.

Specify a Proper and Appropriate Hardness Range. To minimize the chance of hydrogen embrittlement, reduce the hardness. Never specify a hardness range on a quench-hardened steel part that cannot be obtained on the alloy in question after effective tempering. While hydrogen embrittlement has been documented in steels as soft as 85 HRB, problems increase in frequency when the specified equivalent hardness range includes values as high as 40 HRC. While much of the literature reports results based on tensile strength values, it is often more practical to speak in terms of hardness levels. Aside from the fact that many of the parts that are subject to hydrogen embrittlement are too small to provide a tensile test coupon, the most susceptible parts are those that have a carburized or carbonitrided case. Even one layer of grains at the surface of the part may allow the part to become affected by hydrogen embrittlement. In this case, even the Rockwell hardness test is invalid. The actual test must be performed on a metallographic cross section, using an appropriate-load Knoop or Vickers microindentation hardness test.

Specify Proper Heat Treating and Tempering. To minimize the chance of hydrogen embrittlement, always temper the parts after hardening heat treatments, or anneal or perform a thermal stress relief after severe cold forming. Residual stresses from hardening and cold forming are key factors that could allow the crack to propagate. Stress relief or tempering minimizes the residual stress level. Lower stress levels equate to higher tolerable levels of residual hydrogen from plating or cleaning operations or even from the raw-steel manufacturing operation.

Ensure Good Cleaning Solution Maintenance. Proper maintenance of acid concentration, temperature, time in the bath, uniformity of any special or proprietary bath additives, as well as proper racking or tumbling to allow draining and rinsing after the cleaning help to minimize the chances of hydrogen embrittlement problems.

Ensure Good Plating Solution Maintenance and/or Proper Selection of Plating Bath Composition. Some plating solutions inherently generate more hydrogen than others. Specifying a low-hydrogen plating process may provide some protection, although an extensive dialogue between the plating chemist and the part specifier may be required to provide significant protection. Proper and frequent maintenance of bath composition and temperature is also very important.

Ensure Proper Postplating Baking Procedure. Most zinc and cadmium electroplating operations have an oven of some sort (batch or continuous) to bake the parts when they are plated. Baking within one hour of plating has been shown to be very effective at minimizing the danger of hydrogen embrittlement. Delaying the baking cycle beyond one hour after the parts exit the plating line, such as at shift-change time or when an employee forgets to bake the parts for a few hours or a few shifts, has been repeatedly shown to reduce the effectiveness of the baking. While the causes of this loss of effectiveness are not well understood, anyone using or manufacturing hardened or cold-formed electroplated parts can avail themselves of the knowledge gained from experience.

A more difficult problem sometimes arises with the use of phosphate coatings. Iron phosphate and zinc phosphate corrosion-resistant coatings work by a combination of sacrificial (galvanic) action and oil barrier layer held between the open lattice of phosphate crystals on the steel part surface. The microcrystalline

structure of the phosphate is very delicate and can be easily damaged by the conveyor systems commonly used in the coating process lines. Most phosphate lines go directly from the phosphating operation to the oil impregnation step, because the platers have found that the parts may not meet the commonly specified salt-spray-resistance requirements if they are detoured to a baking operation first. However, some platers have solved this problem. While zinc and iron phosphate are generally less expensive than zinc plating, the design engineer must consider the possibility that postcorrosion prevention-coating application baking is more readily available for plating than for phosphates. Both plating and phosphate coating can allow damaging amounts of hydrogen to diffuse into the steel.

Following good manufacturing practices, in general, including minimization of surface imperfections, appropriate maintenance of forming, heat treating, and plating line equipment, and so on, goes a long way toward minimizing the chances of a given lot of parts experiencing hydrogen embrittlement. One example that is worth mentioning is laps in threads of externally threaded fasteners. Not only can laps reduce the effective load-bearing cross section, but they can entrap acid cleaning solution and plating solutions. If the parts cannot be completely dried, they can start to corrode in these hidden locations. This is an example where the boundary between hydrogen embrittlement and stress corrosion could become very indistinct.

Perform destructive testing on samples from each production lot to minimize the chances of highly embrittled components going into service. ASTM committees have agreed on protocols for destructive tests of plated steel parts that can detect relatively low levels of embrittlement. However, these tests are not very useful where variability is great within lot hardness. They may give a false sense of security if the hardest parts are not in the sample and may cause unnecessary scrap if the service loads are lower than those required by the standard test method. These destructive tests are also quite time-consuming. They may cause shipments to be delayed if results are required prior to shipping, because the tests may take up to a few days to perform. The rising step load is one of the embrittlement-detection protocols that may be used on parts that are not cracked and that have specific requirements for structural integrity.

Acknowledgments

Portions of this article have been adapted from C.D. Kim, Hydrogen-Damage Failures, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 245–251.

Hydrogen Damage and Embrittlement

References

1. U.K. Chatterjee, S.K. Bose, and S.K. Roy, *Environmental Degradation of Metals,* Marcel Deckker, 2001

2. J.P. Hirth, Theories of Hydrogen Induced Cracking, *Hydrogen Embrittlement and Stress Corrosion Cracking,* R. Gibala and R.F. Hehemann, Ed., American Society for Metals, 1984, p 29–41

3. L. Raymond, Ed., *Hydrogen Embrittlement Testing,* STP 543, American Society for Testing and Materials, 1972

4. L. Raymond, Evaluation of Hydrogen Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 283–290.

5. G.F. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, p 689–736

6. R. Viswanathan and R.I. Jaffie, Clean Steels to Control Hydrogen Embrittlement, *Current Solutions to Hydrogen Problems in Steels,* American Society for Metals, 1982, p 275–278

7. E.E. Stansbury and R.A. Buchanan, *Fundamentals of Electrochemical Corrosion,* ASM International, 2000, p 371

8. A.W. Thompson, The Mechanism of Hydrogen Participation in Ductile Fracture, *Effect of Hydrogen on Behavior of Materials,* American Institute of Mining, Metallurgical, and Petroleum Engineers, 1976, p 467–479

9. "Steel for Hydrogen Service at Elevated Temperatures and Pressures in Petroleum Refineries and Petrochemical Practice," API 941, American Petroleum Institute, July 1970

10. *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., American Society for Metals, 1975, p 235

11. P.F. Timmins, Chapter 3, *Solutions to Hydrogen Attack in Steels,* ASM International, 1997

Stress-Corrosion Cracking

Revised by W. R. Warke

Introduction

STRESS-CORROSION CRACKING (SCC) is a failure process that occurs because of the simultaneous presence of tensile stress, an environment, and a susceptible material. Although manifest mostly in metals, it can also occur in other engineering solids, such as ceramics and polymers. Removal of or changes in any one of these three factors will often eliminate or reduce susceptibility to SCC and therefore are obvious ways of controlling SCC in practice, as is discussed later.

Stress-corrosion cracking is a subcritical crack growth phenomenon involving crack initiation at selected sites, crack propagation, and overload final fracture of the remaining section. Failure by SCC is frequently encountered in seemingly mild chemical environments at tensile stresses well below the yield strength of the metal. The failures often take the form of fine cracks that penetrate deeply into the metal, with little or no evidence of corrosion on the nearby surface or distortion of the surrounding structure. Therefore during casual inspection no macroscopic evidence of impending failure is seen.

Stress-corrosion cracking continues to be a cause of significant service failures. It is very likely that for every alloy there is an environment that will cause SCC, but, fortunately, most of the ones of industrial significance are known and avoidable. Changes in one or more of the three necessary factors—material, stress, and environment—can prevent or mitigate SCC, either in design or after a problem has occurred in the field. Material selection and specification is the first line of defense. Lowering of the applied stresses and elimination of residual stresses can go a long way toward eliminating problems too. Sometimes minor changes or additions to the environment can help. Finally, proper design and operation to avoid such things as splash zones and wet/dry concentration also are important.

Stress-corrosion cracking is a dangerous and severe degradation mechanism, but with proper understanding and care, failures can be avoided. Several processes in addition to SCC, such as corrosion fatigue, hydrogen embrittlement, and fretting corrosion can cause failure of metals exposed to an environment. All of these are

reviewed in other articles in this Volume. This article is intended to include a brief review of stress corrosion and a discussion of the analysis of SCC service failures and their occurrence in commercial alloys in actual service. The reader interested in more detail on the theories and mechanisms of SCC is referred to other Handbook articles (Ref 1, 2), review articles (Ref 3, 4, 5, 6, 7, 8), and the Selected References listed at the end of this article.

References cited in this section

1. R.H. Jones, Stress-Corrosion Cracking, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 145–163

2. G.H. Koch, Stress-Corrosion Cracking and Hydrogen Embrittlement, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 483–506

3. R.N. Parkins, Stress Corrosion Cracking, *Uhlig's Corrosion,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, 2000, p 191–204

4. S.P. Lynch, Failures of Structures and Components by Environmentally Assisted Cracking, *Eng. Fail. Anal.,* Vol 1 (No. 2), June 1994, p 77–90

5. R.N. Parkins, Mechanistic Aspects of Stress Corrosion Cracking, *Parkins Symposium on Fundamental Aspects of Stress Corrosion Cracking,* Minerals, Metals and Materials Society/AIME, 1992, p 3–40

6. R.H Jones and R.E. Ricker, Mechanisms of Stress Corrosion Cracking, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992, p 1–40

7. R.P Gangloff and M.B Ives, *Environment Induced Cracking of Metals,* NACE-10, National Association of Corrosion Engineers, 1988

8. S.W. Stafford and W.H. Mueller, Failure Analysis of Stress Corrosion Cracking, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992, p 417–436

Stress-Corrosion Cracking

Revised by W. R. Warke

General Characteristics of SCC

Based on extensive empirical investigations, a number of special characteristics of SCC have been observed and categorized. As of 1972, the characteristics of SCC were known to include (Ref 9):

- Tensile stress is required. This stress may be supplied by service loads, cold work, mismatch in fitup, heat treatment, and the wedging action of corrosion products.
- Only alloys are susceptible (no pure metals), although there may be a few exceptions to this rule.
- Generally, only a few chemical species in the environment are effective in causing SCC of a given alloy.
- The species responsible for SCC in general need not be present in large quantities or in high concentrations.
- With some alloy/corrodent combinations, temperatures substantially above room temperature may be required to activate some process essential to SCC.
- An alloy is usually almost inert to the environment that causes SCC.

- Stress-corrosion cracks are always macroscopically brittle in appearance, even in alloys that are very tough in purely mechanical tests.
- Microscopically, the fracture mode for SCC is usually different from the fracture mode for plane-strain fractures in the same alloy.
- There appears to be a threshold stress below which SCC does not occur, at least in some systems.

The passage of time has not negated these observations, and they are still considered to be valid. Three additional characteristics have been added:

- Cathodic protection has been successful, in some cases, in preventing crack initiation and in stopping propagation of cracking that has already progressed, to a limited extent.
- Addition of soluble salts containing certain specific anions can inhibit the crack-producing effect of a given environment on a given alloy.
- Certain aspects of the metallurgical structure of an alloy (such as grain size, crystal structure, and number of phases) influence the susceptibility of the alloy to SCC in a given environment.

Stress-corrosion cracks ordinarily undergo extensive branching and proceed in a general direction perpendicular to the stresses contributing to their initiation and propagation. Figure 1(a) shows transgranular, branched chloride SCC in a specimen of type 304 stainless steel in the annealed condition with a weld of type 308 stainless steel. Figure 1(b) shows intergranular caustic SCC of 316L stainless steel.

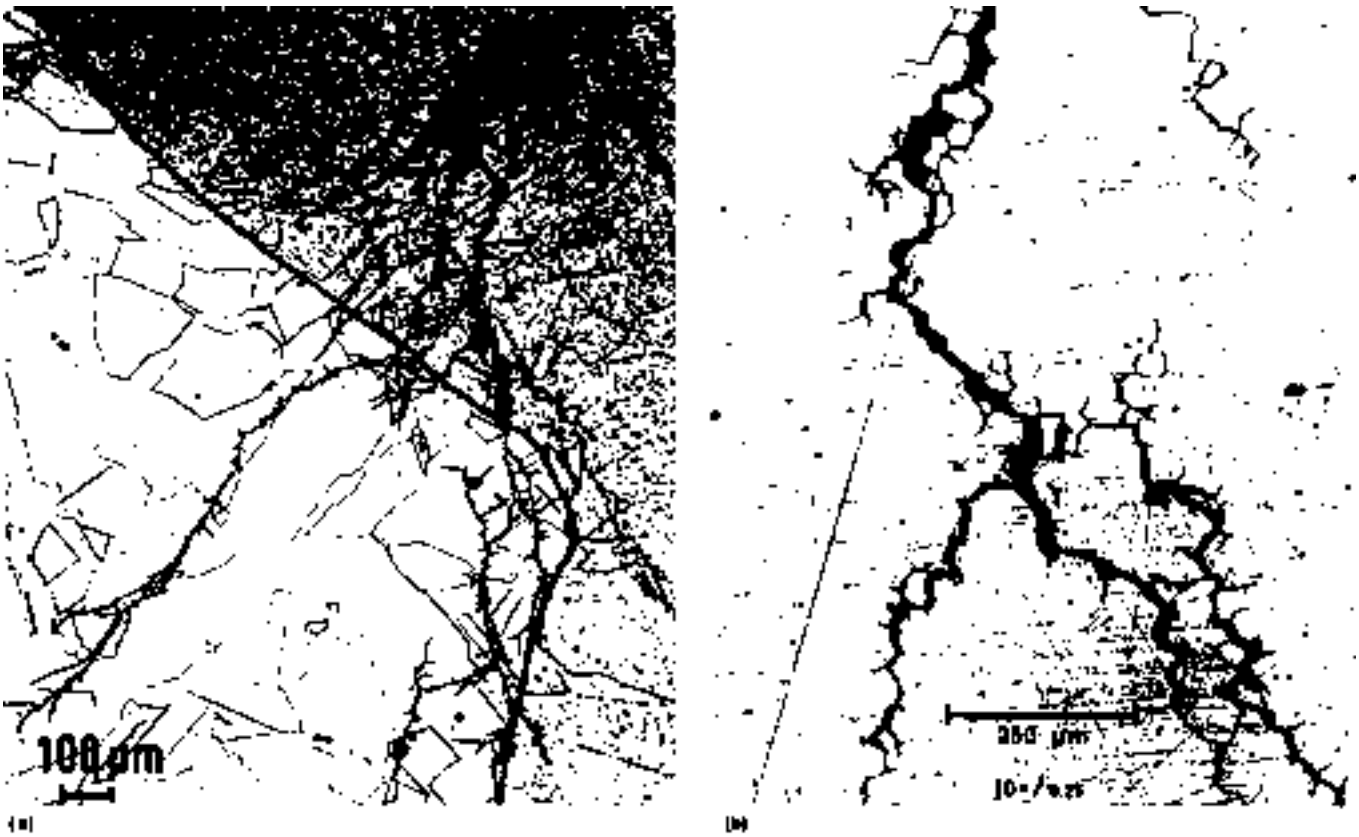


Fig. 1 Branching cracks typical of stress-corrosion cracking (SCC). (a) Chloride SCC of type 304 stainless steel base metal and type 308 weld metal in an aqueous chloride environment at 95 °C (200 °F). Cracks are branching and transgranular. (b) Caustic SCC in the HAZ of a type 316L stainless steel NaOH reactor vessel. Cracks are branching and intergranular.

There are exceptions to the general rule that stress-corrosion cracks are branched. For example, some nonbranched cracks have been observed along with branched cracks in structural steel exposed to contaminated agricultural ammonia (Fig. 2).

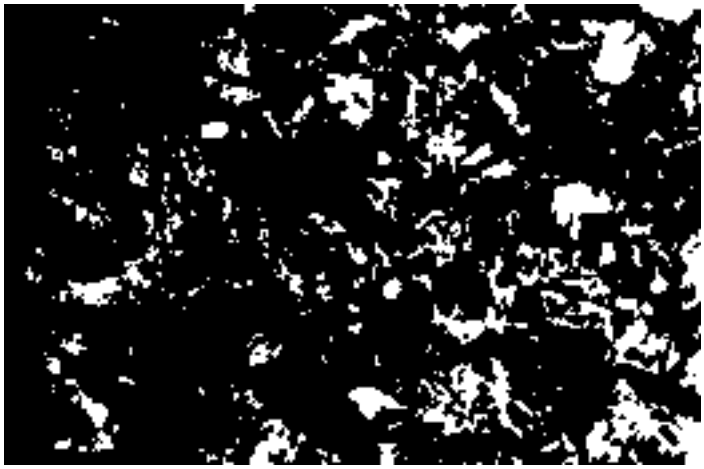


Fig. 2 Picral-etched specimen of structural steel that was exposed to contaminated agricultural ammonia showing nonbranched stress-corrosion cracks. 75×

The surfaces of some stress-corrosion cracks resemble those of brittle mechanical fractures, although they actually are the result of local corrosion in combination with tensile stress. In some metals cracking propagates intergranularly and in others, transgranularly. In certain metals, such as high-nickel alloys, iron-chromium alloys, and brasses, either type of cracking can occur, depending on the metal-environment combination. Features of stress-corrosion cracked surfaces revealed by macroscopic and microscopic examination are discussed in the sections “Macroscopic Examination” and “Microscopic Examination” in this article.

Another characteristic of SCC is the existence of a minimum stress for failure, or *threshold stress,* for smooth specimens and components, and a threshold stress intensity for crack propagation for precracked specimens and components.

Threshold Stress. The stress below which the probability is extremely low that cracking will occur is called the threshold stress, and it depends on temperature, the composition and metallurgical structure of the alloy, and the composition of the environment. In some tests, cracking has occurred at an applied stress as low as approximately 10% of the yield strength; for other metal-environment combinations, threshold stress is approximately 70% of yield strength (Ref 10).

The effect of alloy composition on threshold stress is typified by the graph in Fig. 3, which illustrates the relationship between applied stress and average time-to-fracture in boiling 42% magnesium chloride solution for two 18-8 stainless steels (AISI types 304 and 304L) and two more highly alloyed stainless steels (AISI types 310 and 314). As indicated by the nearly level portions of the curves, the threshold stress in this environment is approximately 241 MPa (35 ksi) for high-alloy stainless steel and approximately 83 MPa (12 ksi) for 18-8 stainless steel.

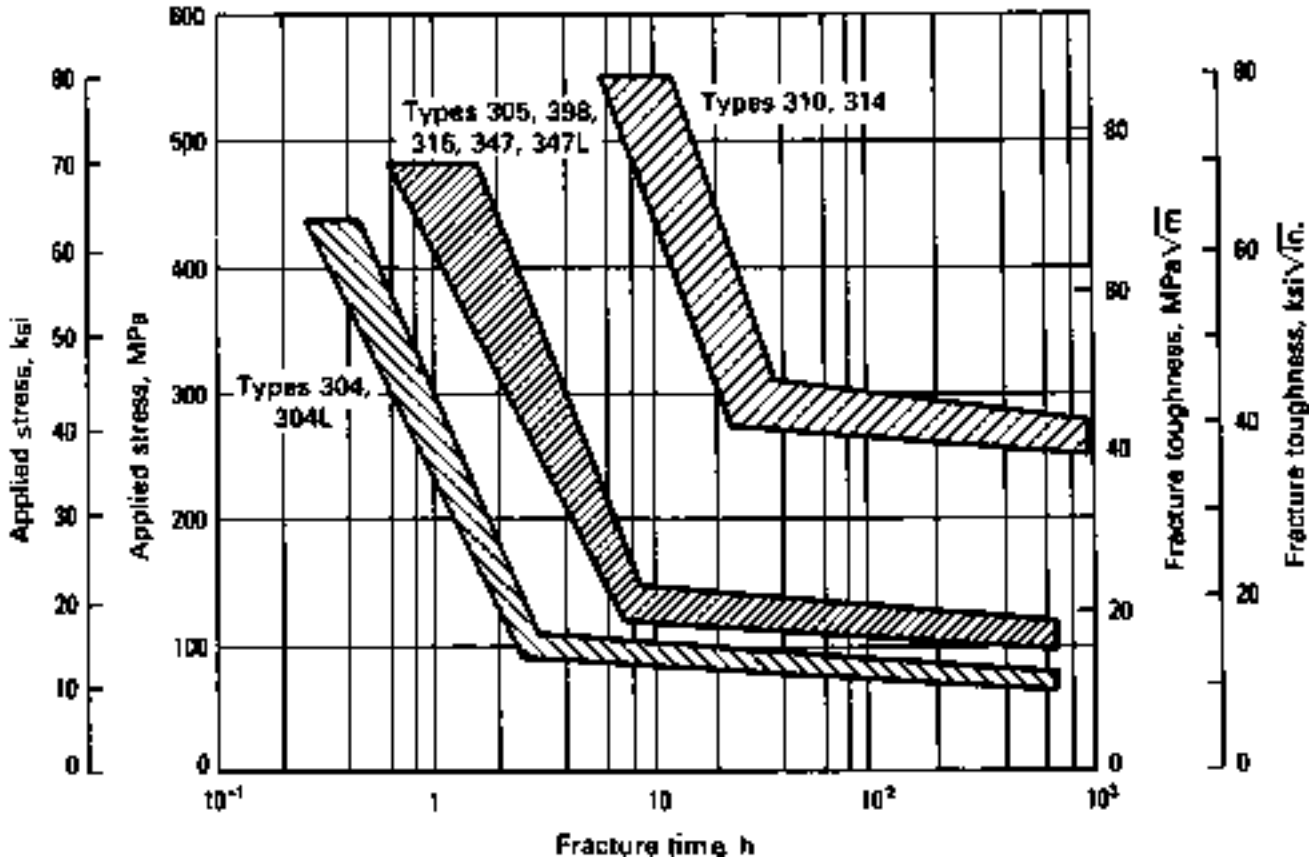


Fig. 3 Relative SCC behavior of austenitic stainless steels in boiling magnesium chloride. Source: Ref 11

Threshold Stress Intensity. For most metal-environment combinations that are susceptible to SCC, there appears to be a threshold stress intensity, K_{ISCC}, below which SCC does not occur. The value obtained for this parameter can vary with the test specimen employed, the sensitivity of crack growth measurement, and the patience of the investigator. Average values of this parameter for maraging steels heat treated to various yield strengths and then exposed to aqueous environments are given subsequently in this article, along with a general description of the relationship of the rate of cracking to stress intensity for maraging steels and for other high-strength steels.

References cited in this section

9. B.F. Brown, A Preface to the Problem of Stress Corrosion Cracking, *Stress-Corrosion Cracking Of Metals—A State Of The Art*, STP 518, H.L. Craig, Ed., ASTM, 1972, p 3

11. D. Sprowls, Evaluation of Stress Corrosion Cracking, *Corrosion*, Vol 13, *ASM Handbook*, ASM International, 1987, p 272

Stress-Corrosion Cracking

Revised by W. R. Warke

Crack Initiation and Propagation

The site of initiation of SCC may be submicroscopic and determined by local differences in metal composition, thickness of protective film, concentration of corrodent, and stress concentration. That is, grain boundary segregation or precipitation can cause the passive layer on the surface of the metal to be locally thinned, allowing pitting or grain boundary corrosion to start. This corrosion pit or trench, produced by chemical attack

on the metal surface, may act as a stress raiser and thus serve as a site for initiation of SCC. A preexisting mechanical crack, surface defect, fabrication flaw, or other surface discontinuity may also trigger SCC. Laboratory investigation has shown that generally there is some localized plastic deformation before cracking occurs, and local barriers to dislocation motion, such as grain boundaries and precipitates, play a role.

The tip of an advancing crack has a small radius, and the attendant stress concentration is great. Using audio-amplification methods, it has been shown in certain materials that a mechanical step or jump can occur during crack propagation. In fact, in one test pings could be heard with the unaided ear (Ref 12). In this test it was demonstrated that the action of both stress and corrosion is required for crack propagation. An advancing crack was stopped when cathodic protection was applied (corrosion was stopped, but the stress condition was left unchanged). When cathodic protection was removed, the crack again started to propagate. This cycle was repeated several times, and the progress of the crack was photographed and projected at the actual speed of propagation.

The role of tensile stressing is important in the rupture of protective films during both initiation and propagation of cracks. These films may be tarnish films, such as those on brasses; thin oxide films; layers richer in the more noble element of a binary alloy, such as a copper-gold alloy; or other passive films. Breaks in the passive film or enriched layer at various points on the surface initiate plastic deformation or incipient cracking. Breaking of a film ahead of an advancing crack permits crack propagation to continue in the metal. Rapid local breaking of the film without particles of film filling the exposed crack is required for rapid crack propagation. Intergranular cracking occurs when grain-boundary regions are anodic to the main body of the metal and therefore less resistant to corrosion because of precipitated phases, depletion, enrichment, or adsorption.

In wrought high-strength heat treatable aluminum alloys, paths of stress-corrosion cracks are always intergranular, because the thermal treatments required to achieve high strength never completely eliminate the electrochemical potential between grain boundaries and grain centers. Stress-corrosion cracking in high-strength aluminum alloys is probably the most widely known example of the effect of grain orientation on the path of cracking. Where the grains are substantially elongated in the rolling direction, stress-corrosion cracks cannot easily propagate perpendicular to this direction, but they propagate very rapidly parallel to the rolling direction. However, newer alloys and tempers have been developed to provide a high resistance to SCC when rolled, extruded, or forged products are stressed in the short-transverse direction.

In wrought austenitic stainless steels, crack paths are usually transgranular if proper heat treatment has been employed. However, if thermal processing has produced sensitization because of carbide precipitation, SCC frequently progresses intergranularly because the chromium-depleted zones along the boundaries of a grain are anodic to the main body of the grain. In wrought martensitic stainless steels, such as AISI types 403, 410, and 431, intergranular cracking is the rule when heat treatment procedures, although not necessarily improper, result in similar carbide precipitation.

The path of SCC in some metals is not governed completely by composition and structure of the metal but is also influenced by the environment. For example, the path of cracking in copper-zinc alloys can be made transgranular or intergranular by adjusting the pH of aqueous solutions in which these alloys are immersed.

The crack propagation behavior of SCC systems is often depicted by the crack growth rate as a function of the applied stress intensity, a *v-K* curve. Figure 4(a) shows a schematic *v-K* curve, and Fig. 4(b) is an actual curve for a stainless steel in two chloride solutions. Three stages are characteristic of these curves:

- *Stage 1:* Crack growth rate increases rapidly with stress intensity above the threshold, K_{ISCC}.
- *Stage 2:* A plateau where velocity is independent of stress intensity
- *Stage 3:* Velocity again increases rapidly with increasing K as K_{Ic} is approached.

The existence of stage 2, the regime where the crack propagation velocity is independent of the driving force, is of particular interest. The plateau velocity is a function of the particular alloy/environment system and may vary by orders of magnitude with the alloy composition and heat treatment, pH, and electrochemical potential. Independence of the driving force indicates that the crack growth rate is controlled by the rate of a chemical reaction, the rate of supply of the embrittling anion or cation to the crack tip, or some other nonmechanical factor.

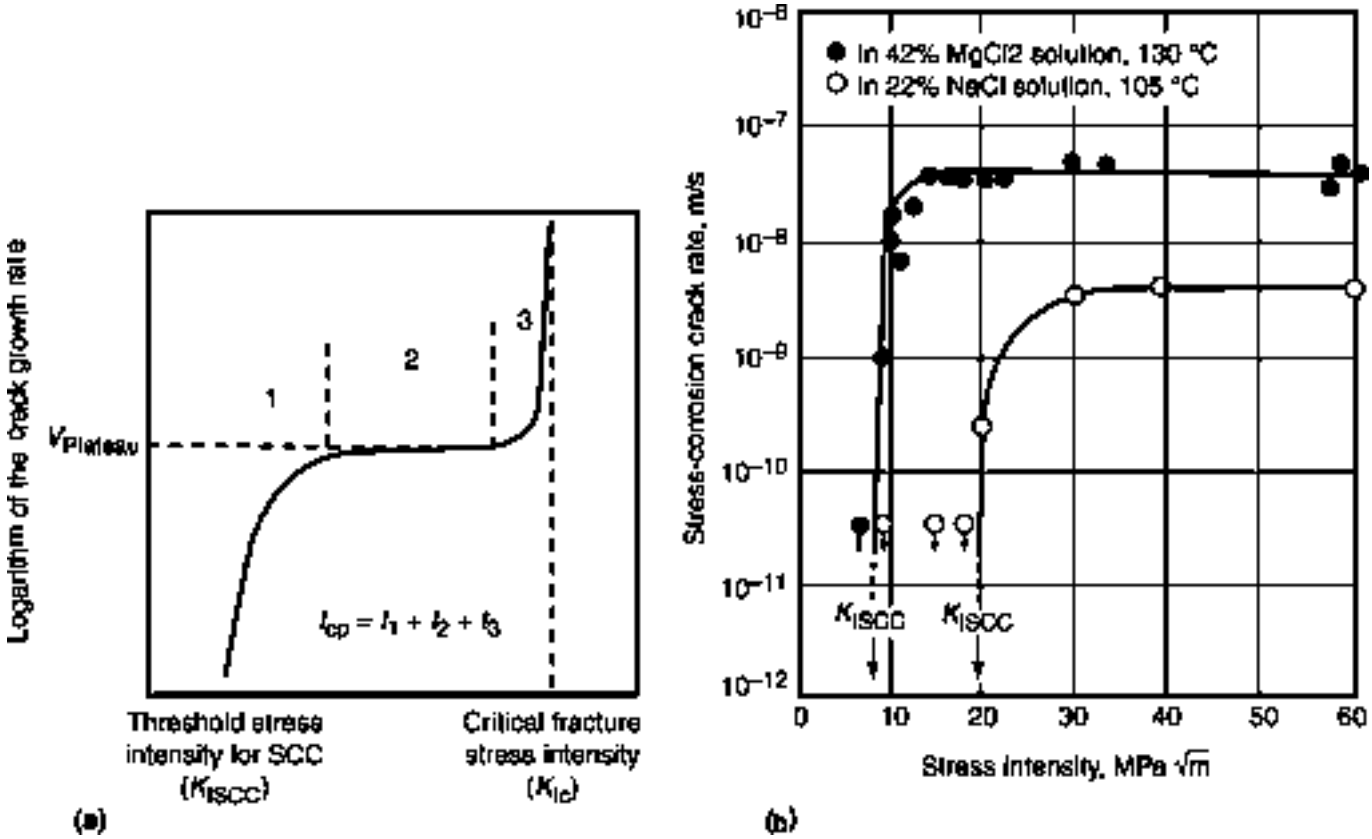


Fig. 4 Stress-corrosion crack growth. (a) Schematic diagram of typical crack-propagation rate as a function of crack-tip stress-intensity behavior illustrating the regions of stage 1, 2, and 3 crack propagation as well as identifying the plateau velocity and the threshold stress intensity. (b) Effect of stress intensity on the growth rate of stress corrosion cracks in type 304L stainless steel exposed to magnesium chloride and sodium chloride solutions

Reference cited in this section

12. W.M. Pardue, F.H. Beck, and M.G. Fontana, Propagation of Stress-Corrosion Cracking in a Magnesium-Base Alloy as Determined by Several Techniques, *Trans. ASM,* Vol 54, 1961, p 539–548

Stress-Corrosion Cracking

Revised by W. R. Warke

Mechanisms of SCC

A number of mechanisms have been proposed to account for the subcritical crack propagation phase of SCC. The following is a brief review of a few of these mechanisms:

- When fresh metal is exposed by plastic deformation at the crack tip, there is a competition between dissolution, which would blunt the crack, and passivation, which would prevent the environment from reaching the metal. Between these two extremes there exists a range where the fresh metal at the crack tip is attacked, but the crack walls are passivated.
- The crack tip repassivates, but the passive layer eventually cracks and the fresh surface is attacked by anodic dissolution, which then repassivates again. This process gives discontinuous growth that could

produce the striations observed in some systems (see the subsequent section on the electrochemical theory).

- The crack progresses by repassivation and cracking of the passive layer with no attack of the base metal. This is called the *tarnish rupture model.*
- There is preferential dissolution of one microstructural constituent (such as a grain boundary) that is anodic with respect to the main body of the metal.
- Adsorption of damaging ions weakens the atomic bonds at the crack tip, allowing them to be pulled apart at a lower-than-normal applied stress. More detail is given in the following paragraphs on the stress-sorption theory.
- Adsorption of damaging ions promotes enhanced, highly localized plasticity that promotes slip at the crack tip, leading to crack growth. This theory is called *adsorption-enhanced plasticity.*
- Hydrogen is generated at local cathodes, enters the metal, diffuses to the crack tip region, and causes propagation by hydrogen stress cracking. This mechanism is no doubt involved in the environmental embrittlement of high-hardness, high-strength carbon, alloy, and stainless steels.

While each of these mechanisms explains some of the observed facts and may be active in one or more alloy/corrodent systems, no single theory has been generally accepted as the universal mechanism completely explaining stress-corrosion crack propagation in all alloy systems and environments. Two major theories are the electrochemical and stress-sorption theories.

Electrochemical Theory

According to the electrochemical theory, galvanic cells are set up between regions of exposed metal surfaces (e.g., metal grains and heterogeneous phases), and anodic paths are thus established. For example, the precipitation of $CuAl_2$ from an Al-4Cu alloy along grain boundaries produces copper-depleted paths in the edges of the grains. When the alloy, stressed in tension, is exposed to a corrosive environment, the ensuing localized electrochemical dissolution of metal, combined with localized plastic deformation, opens up a crack. With sustained tensile stress, protective films that form at the tip of the crack rupture, causing fresh anodic material to be exposed to the corrosive medium, and the SCC is propagated.

Supporting this theory is the existence of a measurable potential in the metal at grain boundaries, which is negative (or active) with respect to the potential of the grains. Furthermore, cathodic polarization will stop the cracking. This theory has been extended to include metals that do not form intermetallic precipitates but for which phase changes or segregation of alloying elements or impurities can occur during plastic deformation of metal at the crack tip. The resulting composition gradient then sets up galvanic cells.

Stress-Sorption Theory

According to the stress-sorption theory, SCC generally proceeds by weakening of the cohesive bonds between surface-metal atoms through adsorption of damaging substances in the environment. Because chemisorption is specific, damaging components are also specific. The surface energy of the metal is said to be reduced, increasing the probability that the metal will form a crack under tensile stress. Adsorption of any kind that reduces surface energy should favor crack formation.

Only a monolayer of adsorbate is needed to significantly decrease the affinity of surface-metal atoms for each other or for atoms of substances in the environment. The only adsorbates presumed to be effective are those that reduce the attractive force of adjoining metal atoms for each other.

Inhibiting anions compete with particles of damaging substances for adsorption sites, thereby making it necessary to apply a more positive potential to the metal to reach a concentration of damaging substances that is adequate for adsorption and resulting cracking of the metal.

Manufacturing Sources of Stress

The preconditions for SCC may involve service stresses, but often the root cause can be attributed to manufacturing. For example, Fig. 5 shows the classic pattern of stress-corrosion crack branching in a type 316L stainless steel screen that was exposed to environment containing toluene (1000 to 2000 ppm), sulfur, chlorine, phosphorus, moisture, and other impurities in the exhaust fumes from large printing presses. The stainless steel screen held activated charcoal to recover the toluene (solvent for printing ink). The screen failed prematurely within one year following installation. The screen was used in an unannealed and unpassivated condition. The stainless steel sheet was perforated by punching the required number of holes. This cold-working operation raised the hardness from approximately 77 HRB to as high as 35 HRC in the necks between adjacent holes. This provided the conditions for premature failure by SCC.

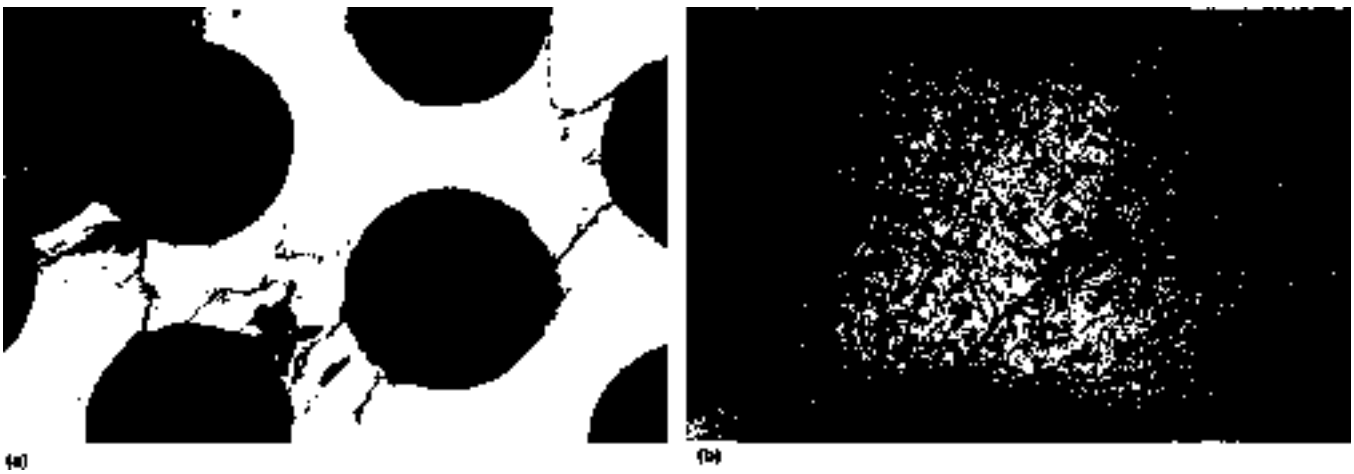


Fig. 5 Branching cracks typical of SCC. Cracking occurred in the work-hardened neck regions of a type 316L stainless steel screen that was exposed to printing press exhaust fumes containing toluene, sulfur, chlorine, phosphorus, moisture, and other impurities. (a) Unetched, picture width ~6 mm. (b) Enlarged view of the cracks; 10% oxalic acid electrolytic etch, picture width ~1.5 mm. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

Preconditions leading to SCC include a variety of materials conditions such as tensile residual stresses, work hardening, surface condition, or stress concentrations. These preconditions also can be introduced in many ways by design and/or in conjunction with various production operations such as welding, thermal processing, surface finishing, fabrication, and assembly. Perhaps of greatest concern is the influence of high levels of residual stress, which can occur from metalworking, casting, machining, and fabrication—in particular, welding. Manufacturing aspects of principal metal forms (wrought, cast, and welded forms) are discussed in separate articles in this Volume, but a few additional examples are given in this article for SCC failures.

Welding

Prevention of SCC failures through proper weld procedures may include appropriate weld preheat, interpass temperature, and postweld heat treatment. For example, Fig. 6 is a micrograph of intergranular SCC in the heat-affected zone (HAZ) of an ER308 stainless steel weld deposit on type 410 martensitic stainless steel that was not tempered following welding. The hardness was 40 to 45 HRC in the HAZ adjacent to this weld, whereas it should have been approximately 20 HRC for good resistance to SCC. Welding of martensitic stainless steels

like type 410 can cause SCC at a later date if it is not feasible to give the HAZ of the weld a satisfactory postweld heat treatment. For type 410 stainless steel, it should be tempered above 500 °C (930 °F), in which case there would be almost no susceptibility to SCC. However, if this steel is tempered at a temperature lower than 500 °C (930 °F), it will be susceptible to SCC when exposed to chlorides and other environments.

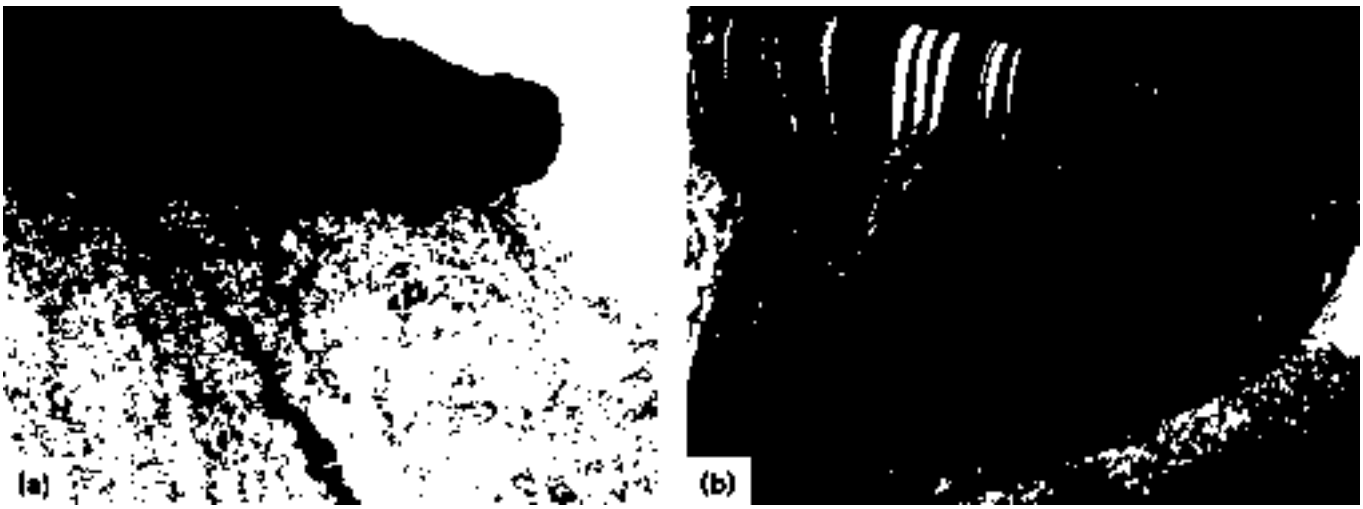


Fig. 6 Cracks in heat-affected zones (HAZs) of type 410 stainless steel beneath weld deposits of ER308 stainless steel. (a) Section through the HAZ that was not tempered after welding showing an intergranular stress-corrosion crack. The weld deposit is at upper right. Electrolytically etched with HCl-methanol. 140×. (b) Fatigue crack initiated by an intergranular underbead crack (arrow). Note beach marks in fatigue region. 3×

Welding is also one of the most frequently encountered sources of residual stress. Shrinkage of weld metal during cooling and the restraint imposed by the adjacent metal and welding fixtures is a source of residual stresses and a possible condition for SCC failures. Example 4 in this article describes a failure from SCC due in part to residual stresses and improper weld procedures. Following is an additional example.

Example 1: SCC of a Type 304 Stainless Steel Pipe Caused by Residual Welding Stresses. A 150 mm (6 in.) schedule 80S type 304 stainless steel pipe (11 mm, or 0.432 in., wall thickness), which had served as an equalizer line in the primary loop of a pressurized-water reactor, was found to contain several circumferential cracks 50 to 100 mm (2 to 4 in.) long. Two of these cracks, which had penetrated the pipe wall, were responsible for leaks detected in a hydrostatic test performed during a general inspection after 7 years of service.

Investigation. The general on-site inspection had included careful scrutiny of all pipe welds by both visual and ultrasonic examination. Following discovery of the two leaks, the entire line of 150 mm (6 in.) pipe, which contained 16 welds, was carefully scanned. Five additional defects were discovered, all circumferential and all in HAZs adjacent to welds. In contrast, scans of larger-diameter pipes in the system (up to 560 mm, or 22 in., in diameter) disclosed no such defects.

Samples of the 150 mm (6 in.) pipe were submitted to three laboratories for independent examination. Inspection in all three laboratories disclosed that all the defects were circumferential intergranular cracks that had originated at the inside surface of the pipe and that were typical of stress-corrosion attack. A majority of the cracks had occurred in HAZs adjacent to circumferential welds. Some cracks had penetrated the entire pipe wall; others had reached a depth of two-thirds of the wall thickness. In general, the cracks were 50 to 100 mm (2 to 4 in.) long. Branches of the cracks that had approached weld deposits had halted without invading the structure (austenite plus δ-ferrite) of the type 308 stainless steel welds. All HAZs examined contained networks of precipitated carbides at the grain boundaries, which revealed that welding had sensitized the 304 stainless steel in those local areas.

Additional cracks, also intergranular and circumferential and originating at the inner surface of the pipe, were discovered at locations remote from any welds. These cracks penetrated to a depth of only one-sixth to one-fourth of the wall thickness through a solution-annealed structure, showing evidence of some cold work at the inner surface.

The water in this heat exchanger was of sufficient purity that corrosion had not been anticipated. The water conditions were:

- Temperature: 285 °C (545 °F)
- Pressure: 7 MPa (1000 psi)
- pH: 6.5 to 7.5
- Chloride content: <0.1 ppm
- Oxygen content: 0.2 to 0.3 ppm
- Electrical conductivity: <0.4 mmho/cm

Analysis of the pipe showed the chemical composition to be entirely normal for type 304 stainless steel. The material was certified by the producer as conforming to ASME SA-376, grade TP304. All available data indicated that the as-supplied pipe had been of acceptable quality.

Several types of stress-corrosion tests were conducted. In one group of tests, samples of 150 mm (6 in.) type 304 stainless steel pipe were welded together using type 308 stainless steel filler metal. Two different welding procedures were used—one consisting of three passes with a high heat input and the other consisting of ten passes with a low heat input. The welded samples were stressed and exposed to water containing 100 ppm dissolved oxygen at 285 °C (545 °F) and 7 MPa (1000 psi). Samples that were welded with the high heat input procedure, which promotes carbide precipitation and residual stress, cracked after 168 h of exposure, whereas samples that were welded with the low heat input procedure remained crack free. These results were considered significant because service reports indicated that the cracked pipe had been welded using a high heat input and a small number of passes and that the larger-diameter piping in the primary loop, which did not crack in service, had been welded with the use of a low heat input, multiple-pass procedure.

Conclusions. The intergranular SCC of the 150 mm (6 in.) schedule 80S type 304 stainless steel pipe exposed to high-temperature, high-pressure, high-purity water in 7 years of heat exchanger service was believed to have been caused by:

- *Stress:* The welding procedure, which employed few passes and high heat input, undoubtedly generated unacceptably high levels of residual stress in the HAZs. The cold working of the internal surface in a sizing operation also set up residual stresses in areas apart from the welds. Both conditions were conducive to SCC.
- *Sensitization:* The high heat input welding caused precipitation of chromium carbides at the grain boundaries in the HAZs. This rendered the steel sensitive to intergranular attack, which, combined with the residual stress, afforded a completely normal setting for SCC.
- *Environment:* Although the level of dissolved oxygen in the water was quite low, it was considered possible that, on the basis of prolonged exposure, oxygen in the range of 0.2 to 1.0 ppm could have provided the necessary ion concentration.

Corrective Measures. All replacement pipe sections were installed using low-heat-input, multiple-pass welding procedures. When stress corrosion is identified as the mechanism of a failure, the proper corrective action can be either a reduction in stress (or stress concentration) or an alteration of the environment. When the cause of failure is inadvertent concentration of a corrodent or inadvertent exposure of the part to a corrosive foreign substance, such as use of a steam line to feed concentrated caustic during chemical cleaning, the preventive measure may be no more complicated than exclusion of the corrodent from that region of the system. In many instances, however, reduction of the level of residual stress is the most effective means of minimizing or preventing SCC.

Metalworking

Work hardening and/or high residual stresses occur from mechanical working or cold-forming operations. Example 22 in this article is one case of damaging tensile residual stress produced by expansion rolling of boiler and heat exchanger tubes into holes in tube sheets. Tensile residual stresses at the surface are generally detrimental in terms of SCC, while compressive residual stresses (from shot peening or other processes) are beneficial. Under some circumstances, severe uniform cold working improves the resistance of a metal to SCC.

For example, cold-drawn steel wire is more resistant to SCC than oil-tempered wire having equal mechanical properties. Also, cold reduction of low-carbon steel to 50% or less of its original thickness makes it relatively immune to cracking in boiling nitrate solutions at 100 to 200 °C (212 to 390 °F) for thousands of hours.

In wrought forms, the relation between the direction of stressing and the grain direction of the metal influences SCC. Transverse stressing is considerably more detrimental than longitudinal stressing, and short-transverse stressing is more detrimental than long-transverse stressing. For example, Fig. 7 shows a fractured lug of a forged 7075-T6 aluminum alloy with cracks that initiated because of stresses acting across the short-transverse grain direction. Intergranular SCC developed when assembly of a pin in the machined hole produced excessive residual hoop stress in the lug.

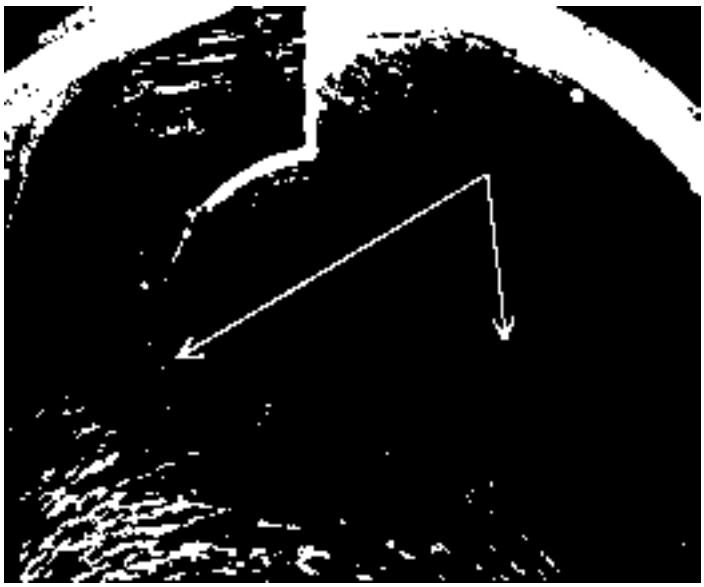


Fig. 7 Fractured lug of an alloy 7075-T6 forging. Arrows show sites at machined hole where stress-corrosion cracks originated because of stress acting across the short transverse grain direction. Keller's reagent. 2×

Severe deformation of the grains in locations of sharp bends or where the metal has become stretched, as in deep-drawn parts, results in a condition that is more vulnerable to failure from mechanisms such as corrosion, fatigue, and SCC. The following is an example of an SCC failure where the effects of sensitization and SCC were intensified as a result of cold forming.

Example 2: Stress-Corrosion Failure of a Strap-Type Clamp Made of 19-9 DL Heat- Resisting Alloy. The clamp shown in Fig. 8(a) was used for securing the hot air ducting system on fighter aircraft. The strap was 0.8 mm (0.032 in.) thick, and the V-section was 1.3 mm (0.050 in.) thick; both were made of 19-9 DL heat-resisting alloy with the following composition:

Element	Composition, %
Carbon	**0.3**
Manganese	**1.1**
Silicon	**0.6**
Chromium	**19**
Nickel	**9**
Molybdenum	**1.25**
Tungsten	**1.2**
Niobium	**0.4**
Titanium	**0.3**
Iron	**Bal**

The operating temperature of the duct surrounded by the clamp was 425 to 540 °C (800 to 1000 °F). The life of the clamp was expected to equal that of the aircraft. After 2 to 3 years of service, the clamp fractured in the area adjacent to the slot near the end of the strap (Fig. 8a) and was returned to the manufacturer to determine the cause of fracture.

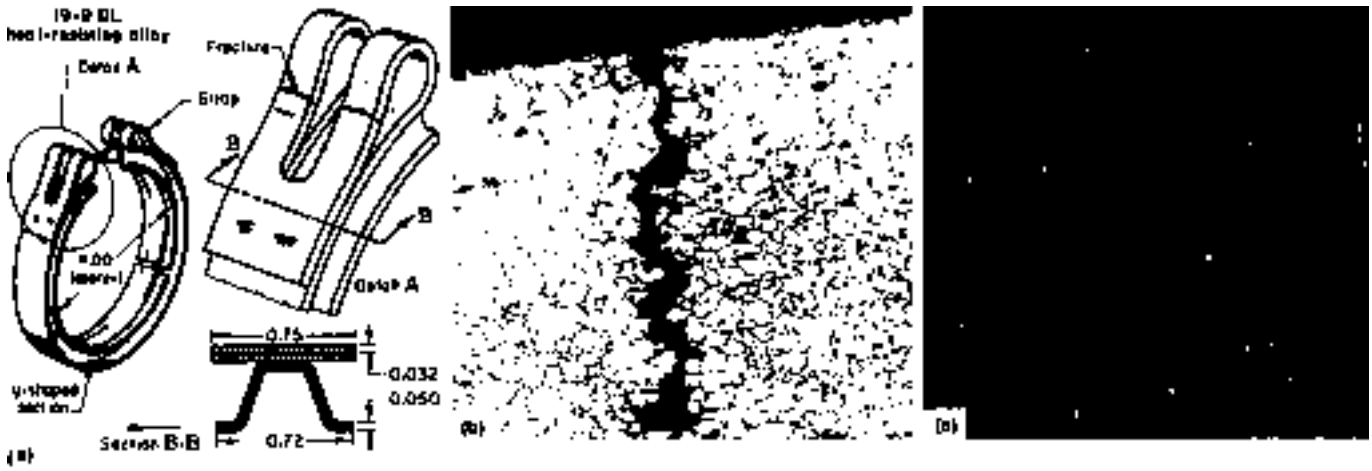


Fig. 8 Heat-resistant alloy clamp for securing the hot air ducting system on fighter aircraft that failed by stress corrosion. (a) Configuration and dimensions (given in inches). (b) Section through the fracture area showing an intergranular crack. Electrolytically etched with oxalic acid. 540×. (c) Specimen of the work metal showing carbide precipitation at grain boundaries and within grains. Electrolytically etched with oxalic acid. 2700×

Investigation. Micrographic examination was made of a section through the strap near the fracture area including an intergranular secondary crack. This section (Fig. 8b) revealed a region of large grains at the surface that was approximately 0.05 mm (0.002 in.) deep. The large grains could have been the result of a sizing pass on the hot-rolled surface. Knoop microhardness values (50 g load) were 345 (25.6 HRC) for the metal in the core and 370 (28.6 HRC) for the metal near the surface.

Electrolytic etching in oxalic acid revealed carbide precipitation at the grain boundaries (Fig. 8c) throughout the material but more prevalent in the area of large grains.

Annealing a portion of the clamp by heating at 980 °C (1800 °F) and water quenching removed the intergranular network of carbides.

Conclusions. The clamp fractured by SCC because the work metal was sensitized. Sensitization occurred during long-term exposure to the service temperature; the effects of sensitization were intensified as a result of cold forming.

Recommendations. Failure of the clamp could have been prevented by using a work metal that was less susceptible to intergranular carbide precipitation.

Stress Concentrations

Stress concentrations that result from various types of deficiencies in manufacture often contribute to mechanical-environmental failures. Some common types of such stress concentrations are:

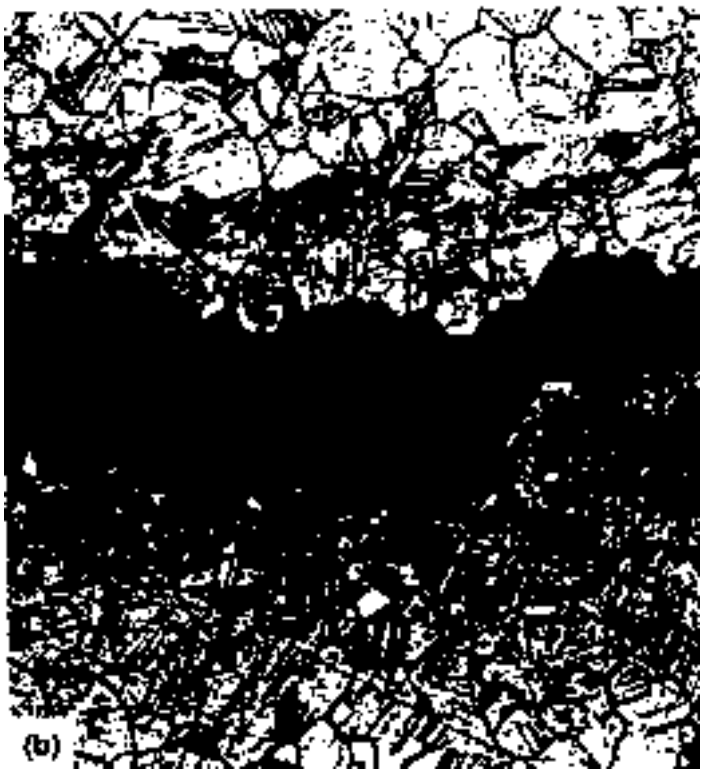
- Those related to design
- Notches or gouges caused by accidental mechanical damage
- Electric-arc strikes
- Cracks produced by incorrect heat treatment, such as quench cracks
- Cracks, undercuts, and poor weld bead geometry from deficiencies in welding
- Inclusions
- Interfaces in layer-type bonded materials (applied by cladding, rolling, electroplating, spraying, brazing, or soldering)

- Case-core interfaces in case-hardened steels
- Severe surface irregularities, such as rolling or forging laps, seams, or slivers

Following is an example of SCC from a stress concentration in which alloy selection and design changes were required.

Example 3: Effect of Alloy Selection and Part Design on SCC of Formed Silicon Bronze in Marine-Air Atmosphere. Electrical contact-finger retainers blanked and formed from annealed copper alloy C65500 (high-silicon bronze A) failed prematurely by cracking while in service in switchgear aboard seagoing vessels. In this service they were sheltered from the weather but subject to indirect exposure to the sea air. About 50% of the contact-finger retainers failed after five to eight months of service aboard ship. Figure 9(a) shows one of the contact retainers that failed and the location of the crack.

(a)



(b)

Fig. 9 Silicon bronze contact-finger retainer that failed from SCC in shipboard service. (a) Overall view of retainer showing cracking in corner (arrow). (b) Specimen taken from failure region showing secondary cracks (arrows). Etched with equal parts NH_4OH and H_2O_2. 250×

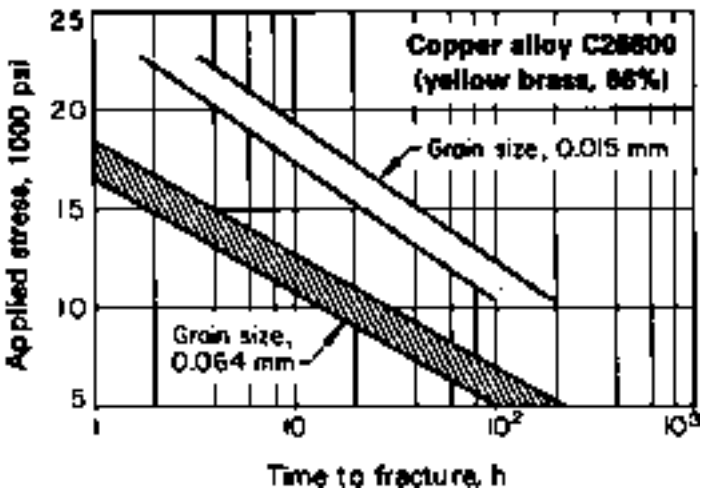


Fig. 10 Effect of grain size on time-to-fracture in ammonia atmosphere. Data are for copper alloy C26800 (yellow brass, 66%) at various values of applied stress.

Nine contact fingers were mounted in each of the two slots of each retainer, and a bank of four to eight retainers held the electrical contacts for a typical circuit breaker system. When a retainer cracked, the usual consequences were open circuits across some contacts, short circuits, and arcing, with resultant overloading and damage to adjacent circuits as well.

Investigation. Emission spectrographic analysis of failed retainers removed from the vessels showed that the composition was within the prescribed range for copper alloy C65500. On each defective retainer examined, a single crack was visible to the unaided eye (Fig. 9a). Each crack was on an end of the retainer, extending between an outer corner of one of the rectangular slotted openings and the periphery of the retainer, and nearly in line with the lengthwise direction of the slot.

Viewing with a low-power stereomicroscope showed that each crack had originated at a slot corner and progressed outward and through the side walls, extending almost or completely to the outermost edge of the retainer. Examination of crack regions at higher magnifications with an optical microscope showed faceted, intergranular fracture patterns and extensive side cracking. No evidence of corrosive attack or corrosion products was visible, and no signs of incipient cracking could be detected at the three other corners of the cracked retainers.

The results of the preceding examination of the cracks were confirmed by microscopic examination of polished-and-etched specimens taken from the cracked regions of several of the retainers. When observed at a magnification of 250×, each of the specimens examined showed an intergranular crack pattern and a microstructure closely resembling that shown in Fig. 9(b).

Conclusions. The cracking was produced by stress corrosion as the combined result of:

- Residual forming and service stresses
- The concentration of tensile stress at outer square corners of the pierced slots
- Preferential corrosive attack along the grain boundaries as a result of high humidity and occasional condensation of moisture containing a fairly high concentration of chlorides (seawater typically contains about 19,000 ppm of dissolved chlorides) and traces of ammonia

Exposure to significant concentrations of ammonia and other corrodents could also have occurred while the vessels were in harbor at industrial locations.

The absence of general surface attack and the formation of faceted, intergranular fracture patterns and extensive side or branch cracking were consistent with the characteristic pattern for failure by SCC of this alloy.

The occurrence of cracking on only one corner of a retainer apparently reflected a characteristically uneven distribution of stress on the four corners of the workpiece during forming and in service, by which one corner was more highly stressed than the three others. Cracking on one corner appeared to relieve the stress on the other corners.

Corrective Measures. Three changes were made to improve the resistance of the contact-finger retainers to SCC. The slots were redesigned to have inside corner radii of 3.2 mm (0.125 in.) to reduce the stress concentration associated with the square corners in the original design, and the formed retainers were shot blasted to remove uneven surface tensile stresses and to put the surface layer in compression. In addition, the retainer metal was changed to a different type of silicon bronze—copper alloy C64700. This alloy has a nominal alloying content of 2% Ni and 0.6% Si instead of the nominal 3% Si content in the alloy C65500 originally used for the contact-finger retainers. Selection of copper alloy C64700 was made on the basis of tests for formability and resistance to SCC (moist ammonia atmosphere) done by the manufacturer after the service failures were encountered.

All of the alloy C65500 retainers in shipboard service were replaced with retainers made to the improved design using alloy C64700. The parts were blanked and formed in the solution, heat treated and 37% cold-reduced tempered, aged after forming, and then shot blasted.

Over a period of several years, only a few isolated SCC failures of the replacement retainers were observed, and these occurred on other types of ships where there was exposure to high concentrations of ammonia and high humidity.

Discussion. Of the various copper alloys considered for the retainers, all except alloy C64700 had to be formed in the annealed temper, with consequent lower strength and excessively unbalanced stress distribution. Alloy C64700, however, could be formed without difficulty in the half-hard (37% cold-reduced) condition. Retainers made from alloy C64700, although not completely immune to SCC in the tests performed by the manufacturer, were much less susceptible to this type of failure than retainers made from alloy C65500.

The excellent formability of 37% cold-reduced alloy C64700 was a key factor in its successful use for the retainers. Cold reduction, either alone or in combination with artificial aging, has been found to minimize the susceptibility of alloy C64700 and other silicon bronzes to SCC. Better as-annealed resistance to SCC and greater improvement by cold reduction were found for alloy C64700 than for alloy C65500.

Plastic deformation produced slip planes, lattice rotation, and curved bands in the grains, thus providing many new sites for attack that were “blind alley” paths for preferential corrosion. These alternate paths allowed only limited penetration of the new sites and decreased corrosion attack at grain boundaries.

Other Examples of Residual Stress

Residual tensile stresses that are damaging or potentially damaging in conjunction with environmental attack can be produced in many different types of manufacturing operations. In heat treating, for example, improper practices that may thermally induce conditions for SCC include:

- Failure to preheat when required
- Overheating during austenitizing or solution treating
- Failure to provide required temperature uniformity in furnaces
- Use of quenching practices too severe for a specific alloy or part shape
- Undue delay in transferring workpieces from the quenchant to the tempering furnace

Many different types of surface-finishing treatments or machining operations also can produce potentially damaging residual stresses. Electroplating, electrical discharge machining, and, under some conditions, conventional grinding and machining can affect surface integrity. Shot peening and surface rolling are often used to produce compressive stresses in metal surfaces, but the magnitude of stress imposed by peening is sometimes not great enough to overcome the effects of extremely high local tensile stresses.

Assembly. Fitup and assembly operations are often sources of tensile stresses. Press fitting, shrink fitting, and assembly by welding are among the major operations in this category. Dimensioning and designing interference fits so as to keep the amount of interference small is desirable in order to avoid creating tensile stresses that could lead to SCC. However, this procedure also makes the assembly more likely to fail by such mechanisms as

fretting, fatigue, and corrosion fatigue; accordingly, a compromise in dimensioning is often the safest procedure.

Bolts, studs, and other threaded fasteners are often pretorqued so that they are under tensile stresses that are significant fractions of the yield strength, making them a frequent victim of SCC. Proper selection of fastener materials for the service environment is essential. The substitution of a straight-threaded connection for a pipe-threaded connection is another corrective measure that has been used to reduce the stress concentrations. In one case, the use of pipe-threaded fittings in making connections to an aluminum alloy 2024-T431 valve body produced excessive stress in the threads and resulted in failure of the valve body by SCC in a service environment of hydraulic fluid that contained small amounts of moisture and chloride. Similarly, stresses on tapered threads on a cast aluminum-zinc alloy connector exposed to a semiindustrial atmosphere caused failure by SCC in less than one month. This connector had been overtorqued during tightening to seal the assembly for a preservice hydrostatic test. In this case, too, substitution of straight-threaded connections reduced the stress concentrations.

Stress-Corrosion Cracking

Revised by W. R. Warke

Sources of Stresses in Service

The normal design stresses that a metal part or assembly was designed to withstand sometimes result in SCC failures, but often, other unexpected stresses introduced in service in various ways are a major contributing factor. Of major concern in connection with failure by mechanical-environmental processes is the combined effect of stress raisers and environment.

Stress concentrations are frequently introduced in service by damage from accidental mechanical impact or local electrical arcing. Stress raisers or notches also frequently result from local wear, fretting, erosion, cavitation, and spalling. Stress raisers can also result from any form of localized corrosion, such as pitting, selective leaching, intergranular attack, and concentration-cell, crevice, or galvanic corrosion.

Thermal expansion effects of various types introduce stresses in service. Exposure of metal parts to high and low temperatures, which is often accompanied by nonuniform heating rates and sharp thermal gradients, is a major source of stress in service. Differential thermal expansion, due to these thermal gradients or dissimilar metals in an assembly, is often an unexpected source of stresses that can lead to SCC.

In heat exchangers, for example, thermal gradients can create strains having equivalent elastic stresses 10 to 20× greater than residual or applied stresses. If the environment on the cold side of the heat exchanger is conducive to stress corrosion, these thermal stresses can produce cracking on this side. Because of the general distribution of the thermal stress, the result can be extensive cracking and sometimes complete fragmentation.

In an AISI type 321 stainless steel ammonia converter in which the temperature produced by the chemical reaction was 910 °C (1670 °F), the temperature of boiler feedwater in an integral cooling jacket surrounding the converter was 205 °C (400 °F). The water jacket had deaerator tubes at the top to remove any steam formed in the jacket, but the tubes did not have sufficient capacity to carry away all of this steam. The resulting accumulation created a steam pocket and a vapor/liquid interface at a level opposite the 910 °C (1670 °F) interior of the converter, and leaks developed at and near the level of the interface.

Removal and examination of portions of the wall in the damaged area revealed the presence of extensive SCC that had completely penetrated the wall from the water side. The tensile stress on the water side of the wall at the vapor/liquid interface was calculated to be approximately 90 MPa (13 ksi); the boiler feedwater in the cooling jacket contained one ppm or less of chlorides.

To correct the problem, it was recommended either that tail gas, which also required heating, be fed through the jacket instead of boiler water or that the deaerator tubes be of a size to ensure flooding of the hot wall at all times. Diffusion of carbon, gaseous carbon compounds, hydrogen, nitrogen, oxygen, and other gases from the environment into the interior of metal parts is also a major source of stress in service.

Corrosion Product Wedging. The formation of corrosion products by local attack in confined spaces produces high stress levels in metals because the corrosion products occupy a larger volume than the metal from which they are formed. The wedging action of corrosion products in cracks and between tight joints in assemblies has been known to generate stresses of 28 to 48 MPa (4 to 7 ksi) and to initiate and propagate failure by mechanical-environmental processes. Where lamellar structures are subject to chemical attack or where weak lamellar structures are subject to stress from entrapped solid or gaseous corrosion products, exfoliation occurs; blisters are produced by gaseous corrosion products entrapped locally just below the surface of a metal part.

Cyclic Stresses. In instances where SCC and corrosion fatigue are competing mechanisms, as might be expected in a component exposed to both cyclic loading and static tensile loading in a corrosive environment, the mechanism of propagation is determined by the stress-intensity factor and the time of exposure. When the maximum stress-intensity factor exceeds K_{ISCC}, both SCC and corrosion fatigue crack propagation may take place. Corrosion fatigue can still take place when the maximum stress intensity is less than K_{ISCC}. The loading conditions responsible for corrosion fatigue cracking are generally known, but the stress situations that typically cause initiation of SCC are not known to the designer because they usually involve sustained tensile stresses that are not anticipated and cannot be measured with precision.

Cyclic stresses of mechanical origin generally arise from the rotary or reciprocating motion of mechanical devices, either from normal operation or from abnormal effects, such as vibration and resonance. Other major sources of cyclic stresses in service are fluid-flow effects, including the von Karman effect and cavitation, and thermal effects. A particularly damaging situation may develope when a small-amplitude cyclic flutter is superimposed on a mean sustained tensile load, such as occurs in the vicinity of a pump station on a pressurized pipeline. Threshold stresses under such situations can markedly decrease. Low-frequency tensile loading, such as occurs during hydrotesting, has been found in some circumstances to enhance stress-corrosion crack growth due to the induced dynamic straining.

Stress-Corrosion Cracking

Revised by W. R. Warke

Metal Susceptibility

Stress-corrosion cracking can be produced in most metals under some conditions. Susceptibility of a given metal to SCC in a specific environment depends on its condition, that is, its overall and local chemical composition and its metallurgical structure, as determined by thermal processing and cold working. Some aspects of the metallurgical condition that are significant include:

- Phase distribution
- Grain size and shape
- Grain boundary precipitation
- Grain boundary segregation
- Cold work
- Inclusion type and distribution

Effects of Composition. One effect of composition on SCC susceptibility, in general, is shown by Fig. 3, where more highly alloyed austenitic stainless steels display longer times to failure and higher threshold stresses than types 304 and 304L. This difference is primarily due to a difference in nickel content. Similar alloy effects could be given for other alloy systems, and some will be presented in the sections to follow.

In considering the effects of metal composition on susceptibility to SCC, the presence of alloying elements and impurities in low concentrations (and even in trace amounts) must also be considered. In general, binary alloys that contain only very small amounts of elements other than the two major constituents are quite susceptible to failure by SCC.

High-purity metals are generally much less susceptible to this type of failure than most commercial grades of metals and alloys. In fact, pure metals were long considered immune to SCC on the basis of experience and several theories of the mechanism of this failure process. However, it was later shown that SCC can be produced in almost any alloy or pure metal in certain environments and at a corrosion potential specific to the metal-environment combination. One example is the SCC of 99.999% pure copper in aqueous ammoniacal solutions containing $Cu(NH_3)_4^{2+}$ complex ions.

Stress-corrosion cracking frequently occurs in commercial alloys such as low-carbon steels, high-strength steels, austenitic stainless steels and other austenitic alloys (especially in the sensitized condition), high-strength aluminum alloys, and brasses and certain other copper alloys (see Tables 1 and 2 and the sections of this article on various classes of metals). A partial list of alloy and environment systems exhibiting SCC (Ref 6) is:

Alloy	Environment
Carbon steel	**Hot nitrate, hydroxide, and carbonate/bicarbonate solutions**
High-strength steels	**Aqueous electrolytes, particularly when containing H_2S**
Austenitic stainless steels	**Hot, concentrated chloride solutions; chloride-contaminated steam**
High-nickel alloys	**High-purity steam**
α-brass	**Ammoniacal solutions**
Aluminum alloys	**Aqueous Cl^-, Br^-, and I^- solutions**
Titanium alloys	**Aqueous Cl^-, Br^-, and I^- solutions; organic liquids; N_2O_4**
Magnesium alloys	**Aqueous Cl^- solutions**
Zirconium alloys	**Aqueous Cl^- solutions; organic liquids; I_2 at 350 °C (660 °F)**

This partial list increases over time. For example, transgranular SCC of copper and SCC of stainless steels and nickel-base alloys in high-purity water can be added to the list (Ref 2). The list can be used as a general guide, but it should be realized that SCC depends on many factors other than the bulk environment. Environments that cause SCC are usually, but not necessarily, aqueous, and specific environmental parameters must be in specific ranges for cracking to occur.

Table 1 Specific ions and substances that have been known to cause stress-corrosion cracking (SCC) in various alloys when present at low concentrations and as impurities

Damaging specific ions and substances	**Alloys susceptible to SCC**	Temperature
Halogen group		
Fluoride ions	Sensitized austenitic stainless steel	**Room**
Gaseous chlorine	High-strength low-alloy steel	**Room**
Fused chloride salt	Zirconium alloys and titanium alloys	**Above melting point of fused salts**
Gaseous iodine	Zirconium alloys	**300 °C (570 °F)**
Gaseous HCl and HBr	High-strength low-alloy steels (rapid crack growth)	**Room**
Halides in aqueous solutions	High-strength aluminum alloys	**Room**
High-strength steels	Room	
Austenitic stainless steels	Hot	
Oxygen group (H_2-O-O_2-H_2 systems)		
O_2dissolved in liquid H_2O	Sensitized stainless steels	**300 °C (570 °F)**
Gaseous hydrogen at ambient pressure	High-strength low-alloy steels	**Room**
Gaseous hydrogen at high temperature and pressure	Low-strength and medium-strength steels	**>200 °C (>390 °F)**
Gaseous H_2O	High-strength aluminum alloys	**Room**
Gaseous H_2O-O_2-H_2	High-strength uranium alloys	**…**
Hydroxides (LiOH, NaOH, KOH)	Carbon steels; Fe-Cr-Ni alloys (caustic	**>100 °C (>210 °F)**

	cracking)	
Oxygen group (S, Se, Te systems)		
Polythionic acids ($H_2S_nO_6$)	Sensitized stainless steels, sensitized Inconel 600	**Room**
H_2S gas	High-strength low-alloy steels	**Room**
Sulfide impurities in aqueous solutions	Medium-strength to high-strength steels (accelerated hydrogen-induced cracking)	**Room**
MnS and MnSe inclusions	High-strength steels (initiation sites for cracking)	**Room**
SO_2 gas with moisture	Copper alloys	**Room**
Nitrogen group		
N_2O_4 liquid	High-strength titanium alloys	**50 °C (120 °F)**
Fuming nitric acid	Pure titanium; high-strength aluminum alloys	**>100 °C (>210 °F)**
Nitrates in aqueous solution	Carbon steels	**>100 °C (>210 °F)**
Nitrogen oxides with moisture	Copper alloys	**Room**
Aerated aqueous NH_3 and ammonium salts in aqueous solution	Copper alloys	**Room**
Nitrogen, phosphorus, arsenic, antimony, and bismuth, as alloying species in metal	Stainless steels (in presence of Cl^-) and copper alloys (in presence of aerated aqueous NH_3); accelerated cracking	**Room**
Arsenic, antimony, and bismuth, as ions in aqueous solutions	High-strength steels, accelerated hydrogen entry and hydrogen-induced cracking	**Room**
Carbon group (C, Si, Ge, Sn, Pb)		
Carbonate ions in aqueous solutions	Carbon steel	**100 °C (210 °F)**
CO-CO_2-H_2O gas	Carbon steel	**…**
Lead ions in aqueous solutions	**High-nickel alloys**	**…**

Table 2 Substances in atmospheric environments that contribute to stress-corrosion cracking (SCC) of various alloys

Damaging substance	**Alloy**	Special aspects
Oxygen	Copper alloys	**Depends on oxygen concentration**
Hydrogen sulfide	Many commercial alloys	**Sources: anaerobic bacteria; breakdown of organic products**
Oxides of nitrogen	Carbon steels; copper alloys	**The oxides produce nitric acid and nitrates**
CO or CO_2, plus moisture	Carbon steels	**…**
Ammonia	Copper alloys	**…**
Arsenic and antimony compounds	Many commercial alloys	**Contained in insecticides and other sprays**
Sulfur dioxide	Copper alloys	**The oxide produces H_2SO_3, H_2SO_4**
Chlorides plus moisture	**Aluminum alloys, austenitic stainless steels, titanium alloys, high-strength steels**	**Marine exposure**

The effect of interstitial elements on the susceptibility of metals to SCC is adequately shown by the behavior of austenitic stainless steels containing trace amounts of elemental nitrogen when exposed to a boiling aqueous solution of magnesium chloride. Below 500 ppm of nitrogen, no cracking occurs; above 500 ppm, cracking occurs with great rapidity. The cracking is transgranular; therefore, the effect of the nitrogen must be attributed either to an influence on the slip process or to the stability of the protective film.

Similar effects are observed in a broad range of alloys when cracking is intergranular. In this circumstance, it is reasonable to assume that impurities are concentrated in grain boundaries and that the chemical reactivity is associated with this segregation. Depending on the chemical interaction between these enriched grain

boundaries and the environment, certain chemical reactions occur preferentially at the boundaries. This reactivity influences both the susceptibility of an alloy to SCC and the paths followed by any cracks that develop.

The concentration of impurities in grain boundaries discussed in the previous two paragraphs does not refer to grain-boundary precipitation effects but rather to concentration of impurities that remain in solid solution at the grain boundaries and make the local chemical composition substantially different from that of the bulk of the material in the grains.

Effects of Metal Structure. In general, any metal with a small grain size is more resistant to SCC than the same metal having a large grain size. This relationship, which applies whether crack propagation is intergranular or transgranular, is shown in Fig. 10 for copper alloy C26800 (yellow brass, 66%) in ammonia.

Elongated grain structures characteristic of high-strength wrought aluminum alloy mill products markedly affect the path of SCC in susceptible alloys and tempers. Alloy 7075-T6, for example, in the form of sheet and plate has a high resistance to SCC when stressed in the direction of rolling or in the long-transverse direction but has a relatively low resistance when stressed in the short-transverse direction (normal to the plane of the plate). In aluminum alloy 7075-T6 extrusions, the threshold stresses in the longitudinal, long-transverse, and short-transverse directions are 414, 221, and 48 MPa (60, 32, and 7 ksi), respectively. For forgings, stressing across a flash line is generally regarded as at least as severe as stressing in the short-transverse direction. A high resistance in the short-transverse direction can be obtained, however, by an extended stabilization treatment to obtain a T7-type (lower strength) temper.

In some metals, SCC follows special crystallographic planes. For example, SCC of titanium proceeds at a slight angle to the basal plane. If the basal plane is perpendicular to the potential plane of crack propagation, resistance to SCC is at a maximum.

Two-phase brasses (those containing more than about 40% Zn) may crack merely in water. Single-phase brasses are resistant to cracking in water, and the presence of a corrodent, such as ammonia or an amine, is necessary to cause cracking in them.

Crystal structure also has an effect on SCC. For example, ferritic and duplex stainless steels (body-centered cubic, bcc) are much more resistant to SCC when exposed to chlorides in aqueous solutions than austenitic stainless steels (face-centered cubic, fcc). The resistance of ferrite-containing cast austenitic stainless steels to cracking when exposed to 205 °C (400 °F) condensate from water containing 800 ppm of chloride is roughly proportional to the volume percentage of ferrite, because pools of ferrite in the austenite matrix interfere with or block the propagation of stress-corrosion cracks.

In addition to the segregation effects discussed, alloys can also be rendered susceptible to intergranular SCC by the precipitation phases in the grain boundaries. The classic example of this is the sensitization of stainless steels and some nickel-base alloys, which causes them to suffer intergranular SCC in some environments. The precipitation of chromium-rich carbides in the grain boundaries results in depletion of the adjacent boundary regions, creating a favorable path for SCC propagation. This phenomenon is described in more detail in a later section of this article.

Welding again plays a significant role in that it causes large changes in the metallurgical structure of both the weld metal and the heat-affected zone. These changes, coupled with the attendant residual stresses mentioned earlier, result in weld zones being the frequent site of SCC in service.

References cited in this section

2. G.H. Koch, Stress-Corrosion Cracking and Hydrogen Embrittlement, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 483–506

6. R.H Jones and R.E. Ricker, Mechanisms of Stress Corrosion Cracking, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992, p 1–40

Revised by W. R. Warke

Environmental Effects

The onset of SCC often depends on the subtleties of environment composition as well as alloy composition and structure. Frequently, minor changes in the environment or the alloy can lead to or prevent cracking. Hence the relationships between environments and alloy types presented in this section provide only general guidelines as to the probability of SCC. The environmental variables involved in SCC are quite numerous and complex. The ones that are commonly considered include:

- Temperature
- Pressure
- Solute species
- Solute concentration and activity
- pH
- Electrochemical potential
- Solution viscosity
- Stirring or mixing

Specific Ions and Substances

Stress-corrosion cracking is often caused by specific ions that are present as components of or impurities in the environment. The concentration of such impurities need not be large. For example, even when in the solution heat-treated or the annealed condition, austenitic stainless steels crack in aqueous solutions that contain as little as 2 ppm of chloride at 200 °C (390 °F). Sensitized austenitic stainless steels crack at room temperature in water containing approximately 10 ppm of chloride or 2 ppm of fluoride. Some copper alloys crack in environments that contain ammonia in similarly low concentrations.

The dependence of the general pattern of SCC on the presence of one or more specific corrosive substances is sometimes called the specific-ion effect. The presence of ions known to induce cracking in a given type of alloy is widely used as a criterion for predicting whether or not cracking will occur. For example, the presence of chlorides in contact with austenitic stainless steels is always regarded as hazardous. Although this is a reasonable criterion for predicting SCC, it is known that specific ions are not always required and that environments as simple as pure water, dry hydrogen, and other pure substances can contribute to SCC. The combination of oxygen and chlorides in trace concentrations in alkaline phosphate-treated boiler water is sufficient to cause SCC of austenitic stainless steels in areas exposed to steam and intermittently wetted by boiler water.

Aircraft parts, both of steel and of aluminum alloys, have failed because of the mildly corrosive action of the atmosphere; water vapor and trace contaminants in air can initiate and propagate SCC and can cause premature failure of such parts. Marine and industrial environments containing small quantities of halide ions are particularly aggressive in causing cracking of high-strength steels, corrosion-resistant steels, and aluminum alloys.

Specific ions and substances that are known to have caused SCC of various alloys when present in low concentrations and as impurities are conveniently classified on the basis of the periodic table. Table 1 lists these damaging specific ions and substances, which are classified as the halogen group, the oxygen group (H_2O-O_2-H_2 systems and S, Se, Te systems), the nitrogen group, and the carbon group. For each damaging substance, the alloys that are susceptible to SCC and the approximate temperature or temperature range at which cracking has been observed are given. Stress-corrosion cracking will not always occur for each listed combination of damaging species, alloy type, and temperature, but these combinations should be considered suspect when the possibility of this type of failure is being assessed.

In environments containing the damaging species of the halogen group, SCC can be caused by these species in the gaseous molecular form, in acids, in fused salts, and as ions in aqueous solutions.

The data presented in Table 1 for the oxygen group demonstrate that simple environments, not necessarily specific ions, are sufficient to cause SCC. The oxygen group is divided into two systems. The first system includes the species based on water, oxygen, hydrogen, and combinations of these and is treated separately because these species are so widely present in the environments to which alloys are exposed. Caustic cracking of low-carbon steels in boilers by hot solutions containing sodium hydroxide, which is listed in this system, is described later in this article. In the second system of the oxygen group, the substances that most frequently cause cracking are those that contain sulfur.

The nitrogen group includes as damaging species some substances that are present not in the environment but as alloying species in stainless steels and copper alloys. These alloying species are the elements nitrogen, phosphorus, arsenic, antimony, and bismuth. The other damaging species in the nitrogen group are present in the environment.

Service Environments

Commercial and consumer equipment is generally operated in two classes of service environments: controlled environments and noncontrolled environments. In controlled environments, the composition of the fluid is controlled; these environments include working (energy transfer) fluids as well as process streams and batches in the chemical and food industries. In operations of this type, the composition of the environment is controlled either to meet and maintain product quality requirements or to prevent corrosion of containers and other types of equipment.

Noncontrolled environments include those to which products of the transportation and manufacturing industries are exposed. Here structures are exposed to a broad range of environments that can be anticipated by the design engineer but can rarely be controlled. As a prerequisite to investigating the possibility of SCC, the properties of a noncontrolled environment should be defined as accurately as possible.

Even in controlled environments, unexpected impurities are sometimes introduced by accidents or improper operations, and SCC can be caused by transient as well as prolonged periods of deviation from normal operating conditions and from normal composition of the environment.

Atmospheric environments contain a variety of damaging or potentially damaging chemical substances. Some of the substances in the atmosphere that contribute to SCC, the alloys they affect, and their special characteristics are given in Table 2.

Working Fluids. Some types of working fluids are the liquids and gases used in power production systems. An important example is pure water used in production of electricity by steam turbine systems. In these systems, water is converted to steam, and the expanded steam is passed through a turbine and subsequently condensed.

These environments are usually treated with corrosion inhibitors to minimize attack of structural materials. A common method is to maintain pH within a range where iron-base alloys are stable by adding such chemicals as phosphates, alkali or ammonium hydroxides, or amines. In addition to control of the pH of such working fluids, additions of reducing substances such as hydrazine and sodium sulfite are made to maintain a low concentration of oxygen, because oxygen accelerates corrosion.

Other heat-transfer fluids include liquid sodium, which is used in liquid-metal-cooled fast-breeder nuclear reactors, and high-temperature gases such as helium, nitrogen, and carbon dioxide, which are used in gas-cooled reactors. Any of these heat transfer fluids, including water, may contain impurities that are possible causes of SCC. In addition to the use of working fluids under essentially steady-state conditions in normal operation, transient changes in the composition of fluids can produce chemical effects that lead to SCC. Such transient changes can be caused by leakage in condensers, pumps, valves, seals, and fittings—for example, leakage of chlorides into a condenser cooled by seawater. Temporary use of a substitute working fluid in a given application can also have unexpected harmful effects.

Process streams and batches differ from working fluids in that they usually are chemicals, foods, or other materials that are being processed into some final product form. The process streams and batches in themselves may not be aggressive with respect to SCC, but may become aggressive if damaging impurities are introduced.

Leached Substances. Stress-corrosion cracking of a metal is frequently caused by damaging substances that have been leached from nonmetallic materials in contact with the metal. Sometimes the substances that cause

cracking are products of thermal decomposition of a material that is innocuous unless heated to its decomposition temperature.

Damaging substances such as chlorides or sulfur compounds can be leached from concrete, gasket materials, insulating materials, and polyvinyl chlorides and similar plastic materials.

Moisture can leach chlorides from the calcium chloride frequently added to concrete to aid in curing. Sulfur compounds and chlorides can leach from gasket materials under conditions of pressure and elevated temperature. Gasket materials include those composed of asbestos, cork, cellulose, or other organic fibers in combination with various binders or impregnants, and the vulcanized elastomeric materials that are used for automotive applications.

The binders used in asbestos and fiberglass insulation frequently contain chloride compounds and are leached in the presence of moisture at room temperature or at elevated temperatures, depending on composition (Example 5 in this article describes SCC of stainless steel piping that resulted from leaching of chlorides from insulation in the presence of moisture). Polyvinyl chloride materials are usually harmless unless they are heated above the decomposition temperature or exposed to ultraviolet radiation.

Concentration of Crack-Inducing Substances. In any environment (liquid, gas, or moist solid), substances capable of inducing SCC are frequently present at harmless levels, either as normal components or as impurities. However, several factors that are related to design or to service conditions can cause these substances to concentrate locally and produce SCC.

Intermittent wetting and drying is a major cause of concentration of damaging and potentially damaging soluble substances in films of moisture on metal surfaces. Where a dilute aqueous solution is transmitted to a metal surface by capillary action through an absorbent fibrous material, the process is called *wicking.* Tests have been done on the effect of temperature on the time required to produce SCC in 1.6 mm (0.062 in.) thick AISI type 304 stainless steel U-bends in a wicking experiment. Cracking was found to occur at much lower temperatures when alternate wetting and drying was used than when the specimens were kept wet continuously.

In service, intermittent wetting and drying is often caused by changes in weather, rise and fall of tides, splashing of liquids as they are pumped into a vessel, and fluctuations in levels of liquids in storage. Also, intermittent condensation and drying occur on metal surfaces in many service environments because of fluctuations in temperature. If soluble substances are available, solutions sufficiently concentrated to cause SCC are readily produced on the metal surfaces.

This phenomenon was found to have caused SCC failure of carbon steel cables on a bridge over a river where the cables were exposed to an industrial atmosphere that contained ammonium nitrate. The cables, which contained 0.7% C, failed after 12 years of service. Tests showed that when specimens from the cables were stressed in tension while immersed in 0.01 *N* ammonium nitrate or sodium nitrate at room temperature, they failed after $3\frac{1}{2}$ to 9 months. No cracking was produced in similar tests in distilled water and in 0.01 *N* solutions of sodium chloride, ammonium sulfate, sodium nitrate, and sodium hydroxide.

Gradual concentration of a damaging substance on a metal surface can result from diffusion effects caused by an imposed electric field, an electric field produced by corrosion processes, or by some other local condition and can lead to SCC in gases, liquids, and moist solid environments such as soil, concrete, or plaster.

Concentration can also occur because of part shape, especially on assemblies in which close fit of parts creates crevices. Crevices provide both sites and proper conditions for concentration of impurities on metal specimens in aqueous media. Such crevices are common in assemblies using gaskets or sealants. Because of the tightness of such crevices, it is impossible for impurities to be removed from them. The impurities are thus concentrated by leaching, decomposition, or corrosion and frequently result in SCC. In heated crevices, such as in tube sheets of steam generators, impurities are concentrated when water flashes to steam, and SCC may occur if hydroxides are present. A similar situation can occur when ground water contacts a warm buried pipeline and the water evaporates, leaving its dissolved impurities behind to concentrate over time.

Dried processing chemicals on metal surfaces, such as those that remain on poorly rinsed surfaces after pickling, cleaning, and similar operations, are available for corrosion attack at a later time when the surfaces are exposed to their normal environments. One example is the effect of residual fluoride on zirconium in corrosion testing. When zirconium specimens exposed to fluoride solutions in corrosion tests are properly cleaned after removal from the solution, they gain weight at a low rate when subsequently exposed to high-temperature water; however, when there is sufficient residual fluoride, the specimen loses weight because of corrosion that can initiate SCC.

Preservice Environments

Many SCC failures are caused by exposure of metal equipment to corrosive environments before the equipment is put into service. These environments include those associated with fabrication, testing, shipment, storage, and installation.

Fabrication Environments. Substances to which metal parts and equipment are exposed during fabrication and processing often leave residues on the metal surfaces or introduce impurities into the metal. Stress-corrosion cracking caused by these residues or impurities frequently begins, and occasionally results in failure, during the production process. Sometimes impending failure is discovered during fabrication; other times it is not discovered until the equipment is in service.

Examples of sources of such deleterious impurities are machining lubricants; protective coatings; fumes from adjacent processing lines; atmospheric impurities; welding and brazing fluxes; cleaning solutions; inspection fluids such as liquid penetrants; containers for processing; lubricants such as copper and lead, used in tube drawing; and plating and chemical-surface treatment solutions.

Testing Environments. In product research and development, metal parts and assemblies are tested to determine load capacity and conformance to other requirements. Failures are sometimes initiated and/or propagated by test environments, such as in hydrostatic testing of pressure vessels, piping, and related equipment. For example, a titanium alloy Ti-6Al-4V vessel that was to contain N_2O_4 in service failed after approximately 13 h of hydrostatic testing at approximately 43 °C (110 °F) at 75% of yield stress. Here the testing environment, which closely simulated the service environment, was the cause of failure. In this case, failure was fortuitous because it established that the material-environment combination was not suitable for reliable performance of the intended application.

The chloride content of water used to hydrotest austenitic stainless steel piping and pressure vessels must be closely controlled. Residual water that fails to drain from the system can concentrate by evaporation and lead to SCC, perhaps at a weld where there are high residual stresses.

Other environmental impurities that may be harmful in hydrostatic testing include microbiological and bacterial contaminants, chlorides, and fluorides. Stress-corrosion cracking of titanium in methanol was first observed when methanol was used as a medium for hydrostatic testing.

Shipping and Storage Environments. During shipping or storage, parts and assemblies may or may not be protected from the atmosphere. There have been instances in which allegedly protective materials have in fact contained species that caused, rather than prevented, SCC. Where suitable protection is not provided, the components can be continually or intermittently in contact with impurities in the surroundings. For example, metal parts shipped by sea are in danger if not fully protected from chlorides. Thus, to prevent initiation and propagation of SCC during shipment or storage, expected exterior environments and protective systems must be thoroughly evaluated and effective protective measures taken.

Example 17 later in this article describes the SCC of aluminum alloy couplings, which occurred in a missile that was on a test stand near the seacoast for several months. The combination of tensile stresses and marine atmosphere caused the failure.

Installation Environments. During installation of equipment, critical materials are often exposed to damaging or potentially damaging environments. Some of these environments are welding flux, marking materials, lubricants (applied to moving parts and to large bolts during installation), machining oils, test environments, and contaminants from fingerprints. Sometimes installation procedures such as bolting or crimping inadvertently introduce residual stresses that can lead to SCC in service.

Stress-Corrosion Cracking

Revised by W. R. Warke

Analysis of SCC Failures

The procedures for analyzing failures suspected to have occurred by SCC are, in general, the same as those for other failures. Ideally, the failed part should be brought to the laboratory immediately after failure, with no

further disturbance and with all of the physical evidence intact. The part should be accompanied by a complete record of the manufacturing history and results of prior testing. Attention must also be given to chemical environments to which the part was exposed during manufacturing, shipment, storage, and service, particularly to such factors as concentration, temperature, pressure, acidity or alkalinity, normal or accidentally introduced impurities, cyclic conditions, and any deliberate or accidental changes in the control of these factors.

On-Site Examination

On-site examination provides an opportunity to observe and test the environmental conditions and mechanical processes involved. The relationship of the failed part to the overall operation as well as possible sources of stress can generally be established. Photographs and sketches are indispensable for recording the location of failure, the extent of fracture or cracking, the presence of corrosion products or foreign deposits, evidence of mechanical abuse, and other significant details. The investigator should look carefully for other stress-corrosion cracks, corrosion pits, mechanical damage, associated welds (fabrication or repair), and design issues.

Sampling

The samples required in addition to the failed parts may include related parts, specimens removed from parts, surface and corrosion deposits, and the environment to which the failed parts were exposed. Selection of samples and removal of specimens should be carefully planned, keeping in mind the main objectives of the investigation. Suitable precautions must be taken to avoid (a) destruction of evidence, for example, corrosion products or other surface deposits that could be of value in investigating the failure; (b) overheating, contaminating, and damaging the part and adjacent and remote related components and structures; and (c) contamination of any of the samples. For example, contrary to other types of failures, components that are suspected of having failed due to SCC should not have preservatives applied to the fracture surfaces.

One technique for collection of associated corrosion products or surface deposits is to tape the flap of a clean envelope to the surface below the material to be sampled and then scrape the deposit with a clean knife or scalpel blade. Glass vials are also useful for collecting and retaining such samples.

In sampling the environment to which the failed part was exposed, the investigator should be guided by the information already obtained about the operating conditions and history of the failure. The general environment should be sampled. Also, appropriate techniques should be used to obtain samples of and to characterize (for example, by pH and the presence of specific contaminants) the environment in immediate contact with the failure region. General information on chemical substances and environments likely to cause SCC in specific metals is provided in Table 1.

Following visual examination, the fracture surface should be stripped with cellulose acetate tape to entrap any fracture surface deposits. The tape can be used for later microanalysis with the aid of an electron microprobe, a scanning electron microscope with an energy-dispersive system, or other suitable analytical equipment. After stripping, the fracture surface should be cleaned (ultrasonically, if possible) and degreased in reagent-grade acetone, which leaves minimal amounts of organic residues on the surface while removing almost all traces of cellulose acetate from stripping.

Observation of Fracture Surface Characteristics

Crack patterns, initiation sites and topography of crack surfaces, and evidence of corrosion all provide clues as to the causes, origins, and mechanisms of fractures.

Crack Patterns. As previously discussed, SCC is almost without exception characterized by multiple branching on both the macroscopic and microscopic levels (for example, Fig. 5). In some cases, a family of cracks occurs, and numerous secondary cracks will be available for examination. The general plane of stress-corrosion crack growth, similar to other subcritical crack growth, is perpendicular to the maximum principal stress, permitting some conclusions to be inferred as to the direction and nature of the stresses causing cracking. Also, the relative size of the subcritical and final fracture areas is an indication of the stress magnitude.

Stress-corrosion cracks may be intergranular, transgranular, or a combination of both. In aluminum alloys and low-carbon steels, intergranular fracture is common, although the fracture path may be immediately adjacent to the grain boundary rather than precisely along it. High-strength steels and α-brasses also usually show grain-

boundary fracture, with some cracking along matrix/twin interfaces in α-brasses. Transgranular fractures showing extensive branching are typical of SCC in austenitic stainless steels of the 18Cr-8Ni type (Fig. 11), and similar transgranular cracks with branches that follow crystallographic planes have been observed in magnesium alloys.

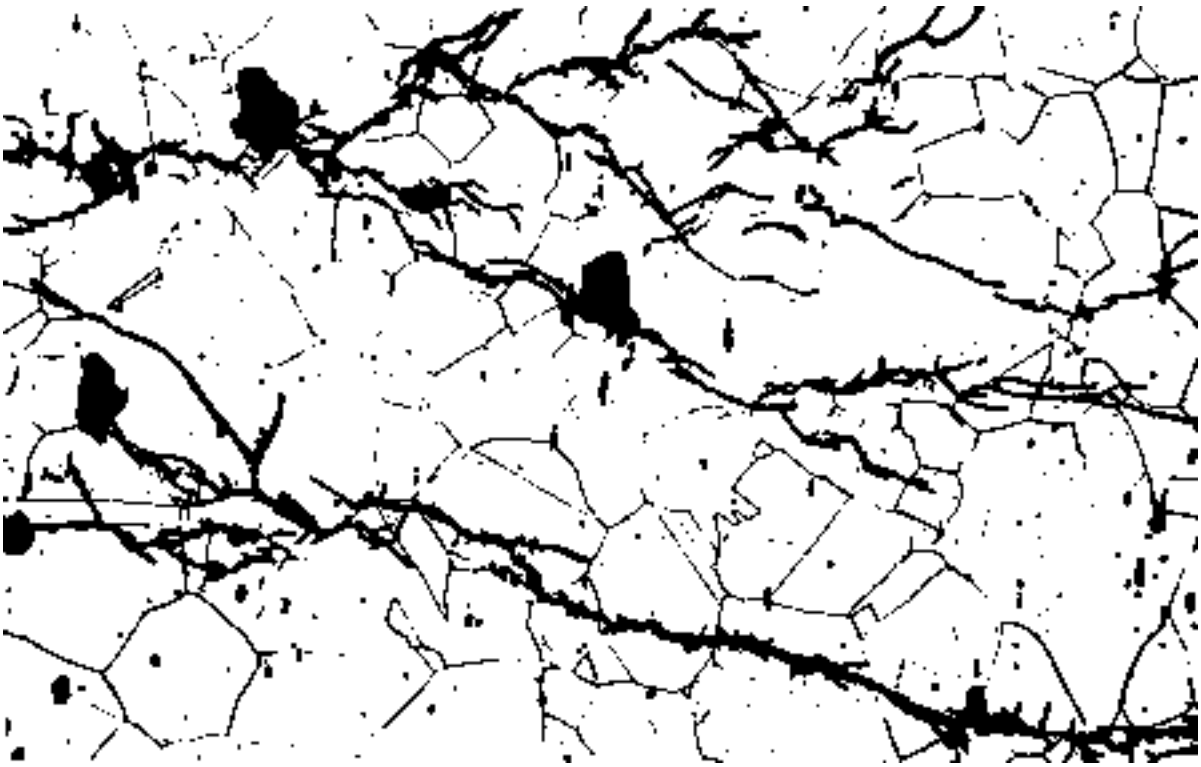


Fig. 11 Annealed type 310 stainless steel after extended exposure to a chloride-containing environment while under load. Structure shows typical mode of transgranular SCC. Electrolytic: 10% chromic acid. 150×

When SCC is transgranular, deviations may occur on a microscopic scale so that the crack may follow microstructural features such as grain and twin boundaries or specific crystal planes. When SCC is intergranular, the presence of flat elongated grains means that there is an easy stress-corrosion path normal to the short-transverse direction, which produces “woody” stress-corrosion surfaces. This behavior is typical of extrusions of high-strength aluminum alloys in which solution treatment does not cause recrystallization.

Some stress-corrosion fractures show progression marks and alternating regions of SCC and overload fracture, with changes of shape of the crack front. The progression marks in the fracture surface shown in Fig. 12 could very easily be confused with fatigue. This fracture occurred in a forging that had not been used in service, and the fracture was due entirely to residual internal stress and atmospheric corrosion. Other features observed on SCC fractures include striations (Fig. 13), cleavage facets, and tongues, which can be easily confused with similar features on cleavage and fatigue fractures.

Fig. 12 Fracture surface of a high-strength aluminum alloy forging that failed from stress corrosion. Progression marks similar to those observed in fatigue fractures are evident. Source: Ref 13

Fig. 13 Stress-corrosion cracking striations on the fracture surface of 316L stainless steel. The arrow indicates the crack-growth direction. 10,000×. Courtesy of I. Le May, Metallurgical Consulting Services Ltd.

The branching pattern of stress-corrosion cracks is a feature that is distinct from corrosion fatigue. For example, Fig. 14 displays a fracture path of intergranular SCC and transgranular corrosion-fatigue cracking of a cartridge brass. The SCC shows some branching, compared to the lack of branching from corrosion fatigue (Fig. 14b).

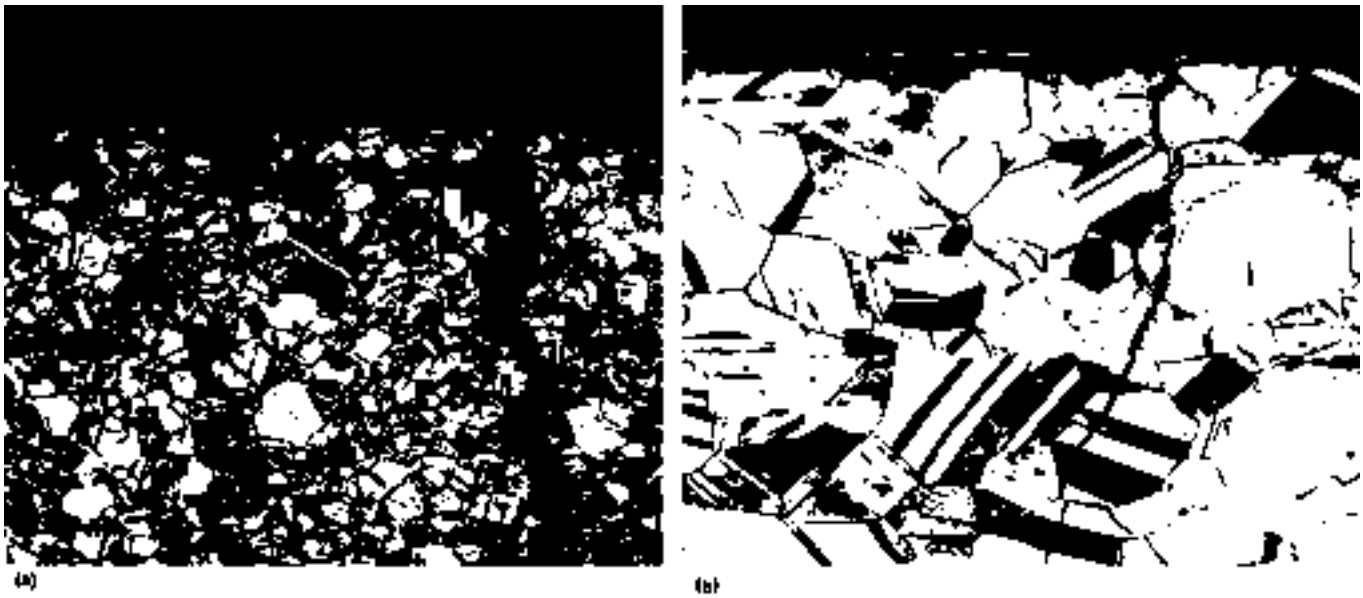


Fig. 14 Comparison of SCC and corrosion fatigue cracks in copper alloy C26000 (cartridge brass, 70%). (a) Typical intergranular stress-corrosion cracks in tube that was drawn, annealed, and cold reduced 5%. The cracks show some branching. H_4OH plus H_2O etch, 150×. (b) Typical transgranular corrosion fatigue crack. Note the lack of branching in the inner, or fatigue, section of the crack. H_4OH plus H_2O etch, 186×

Initiation sites and topography of crack surfaces can usually be ascertained by careful examination. These features are crucial in determining the causes and mechanisms of fractures. When a primary fracture has been damaged or corroded, secondary cracks and incomplete primary cracks may be opened to expose their surfaces for examination.

Corrosion of Fracture Surfaces. Besides establishing the presence and identity of these substances on the fracture surface, an investigation (often extensive) of the environment and service history of the part must be carried out to determine the exact role (if any) of these substances in the failure. Where information on the environment and service history of the part are incomplete or where abnormal transient conditions may have existed, results of the investigation may be inconclusive. In some circumstances, tests that simulate service and investigations of the effects of variations in composition of the environment and in operating conditions are warranted.

Fracture surfaces may also be examined for evidence of local hydrogen embrittlement from hydrogen generated by corrosion. For example, a fracture in a specimen of type 316 stainless steel is shown in two views in Fig. 15. The first view (Fig. 15a) shows the transgranular cracking features typical of austenitic stainless steel SCC in chlorides. The cleavage facets with feather marks were largely characteristic of the entire fracture. However, Fig. 15(b) displays an area of “rock candy” intergranular fracture that may have otherwise been overlooked. This may be evidence of local hydrogen embrittlement from corrosion. In this case, the grain facets show no evidence of corrosion.

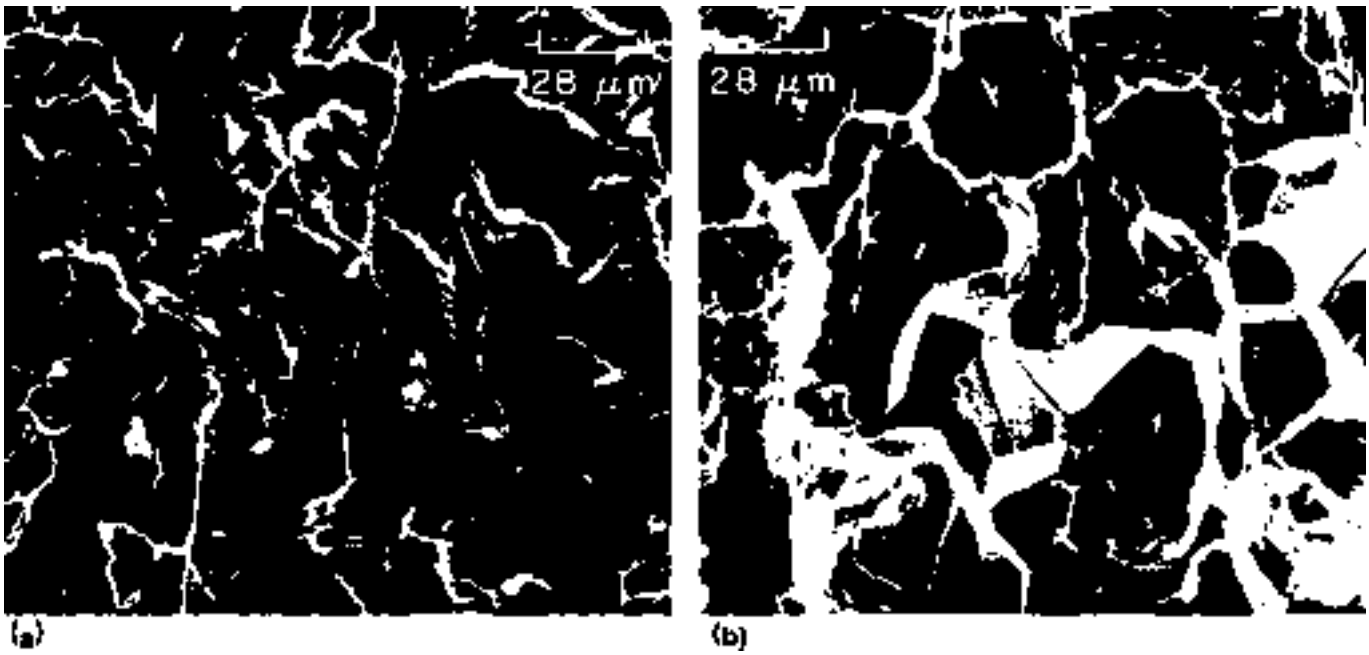


Fig. 15 Surface of a fracture in type 316 stainless steel resulting from SCC by exposure to a boiling solution of 42 wt% $MgCl_2$. The fracture in general exhibited the fan-shaped or feather-shaped transgranular cleavage features shown in (a). In a hasty scrutiny, the presence of local areas of "rock candy" intergranular fracture, such as those in (b), might be missed. This raises the question of whether corrosion-generated hydrogen caused local embrittlement. The separated-grain facets in (b) show no trace of corrosion. Both at 350×

Macroscopic Examination

Macroscopically, fractures produced by SCC always appear brittle, exhibiting little or no ductility even in very tough materials; in this way, they resemble corrosion fatigue and hydrogen stress cracking fractures. Many transgranular stress-corrosion cracks characteristically change fracture planes as they propagate, producing flat facets. Both these flat facets and the grain facets of intergranular stress-corrosion cracks are ordinarily observable at fairly low magnification.

The fracture surfaces usually contain easily identifiable regions of crack initiation, slow, subcritical crack propagation, and final rupture. Final rupture usually occurs by tensile overload. Thus, the area of final fracture often shows some evidence of ductility, such as a shear lip or a herringbone pattern emanating from the zone of slow cracking.

The area of slow crack growth often contains corrosion products or is stained or otherwise discolored with respect to the area of final fast fracture. However, the presence of staining or corrosion products on the fracture surface is by no means positive proof of SCC. Some SCC fractures are not stained or discolored, especially in materials with good corrosion resistance; in addition, many fractures become corroded before inspection can be accomplished.

The zones of slow propagation of stress-corrosion cracks are usually much rougher in appearance than those of corrosion fatigue cracks and often do not contain fatigue beach marks or macroscopic evidence of cold work. However if either the stress component is removed or the environment becomes inactive, crack propagation will discontinue until these factors again become operative. This sometimes results in macroscopic markings similar in appearance to the beach marks characteristic of fatigue fractures.

Both types of fracture can, and often do, initiate at corrosion pits or other stress raisers on the surface of the part. However, if the environment is sufficiently corrosive and if tensile stresses are present, SCC can be initiated even on smooth surfaces, whereas corrosion fatigue is more often initiated at a stress raiser.

Fractures caused by either of these mechanisms always initiate at the surface of the part or at some location where the aggressive species can contact the metal. By corollary, if the crack is found to have a subsurface origin, corrosion fatigue and SCC can usually be eliminated as possible causes of failure.

Cracks produced by stress corrosion and by some hydrogen damage processes generally have macroscopically rough fracture features, which are usually discolored from reaction with the environment. These mechanisms are easily recognized as macroscopically different from corrosion fatigue, but identifying the exact mechanism by macroexamination alone is not always possible. These cracks may not always have distinct origins; there may be much pitting in the region of crack initiation, and the crack may exhibit branching in this region.

Microscopic Examination

Careful correlation of microscopic fracture surface topography with macroscopic fracture surface features is essential. The crack-initiation region and the directions of crack growth must be identified accurately so that information concerning the sequence of events and the micromechanisms of fracture, as observed with electron microscopy, can be correlated with the circumstances of crack initiation and the mechanism of crack propagation. Frequently several different fracture micromechanisms are observed on a single fracture surface. Accordingly, correct identification of the initiating fracture mechanism and of any changes in micromechanism during fracture propagation is of vital importance in arriving at a correct understanding of the failure.

In the scanning electron microscope, intergranular stress-corrosion cracks have the classic rock candy appearance, with little if any evidence of ductility on the grain facets. The transgranular cracks look like cleavage cracks, with flat facets covered by river lines. In some cases the facets have striations perpendicular to the microscopic growth direction, indicating that, on a microscopic scale, the SCC process was discontinuous, even though a macroscopic constant growth rate would be observed. These striations can easily be mistaken for fatigue striations, leading to an eroneous conclusion, and care must be taken to discern the correct mechanism, consistent with all of the evidence.

It is difficult to distinguish between service failures by SCC and by hydrogen damage in high-hardness steel solely from microfractographic evidence. Fractures of both types mainly follow intergranular paths, although SCC is sometimes transgranular, and the surfaces of both types of fracture may be substantially corroded. Also, hydrogen evolution during SCC may be a factor in the cracking process so that the combination of hydrogen embrittlement and corrosion at the crack tip may be instrumental in producing grain-boundary separation.

Chemical Analysis

Chemical analysis is used in investigating cracking failures that may have been caused by SCC or other corrosion-related failures. Impurities and substances present at low concentrations or in trace amounts in the environment are of special interest where SCC is a possibility, because they are the substances that usually induce this type of failure.

Identification of the chemical compounds present in any surface deposits on the metal is important in determining the identity of corrodents and the nature of the damaging corrosion processes. However, great care should be taken to ensure that surface deposits are representative of the service environment and are not residues of extraneous materials introduced during such operations as liquid penetrant or magnetic-particle inspections performed after the failure. X-ray diffraction, coupled with energy-dispersive spectroscopy (EDS) or x-ray fluorescence (XRF) spectroscopy, is particularly useful in identifying surface or corrosion deposits, and can quickly lead to an understanding of the environmental factors involved in the SCC failure. In analyzing SCC failures, the nominal chemical composition of the metal is usually of less concern than the substances present at low concentrations or in trace amounts (usually as impurities) and the metallurgical structure of the metal.

Energy-dispersive spectroscopy in a scanning electron microscope of either the crack faces or crack deposits in metallographic section can detect the presence of the damaging ions or of changes from the metal composition, which can be clues to the cause of SCC. Long counting times may be needed because of the depth of penetration of the electron beam relative to the thickness of the surface layers.

Metallographic Analysis

Aspects of metallographic analysis that require special attention include:

- The crack path relative to the microstucture—intergranular or transgranular, through or around second phases, extent of branching, and so forth
- Preferred grain orientation such that the direction of maximum stress is perpendicular to elongated grain faces
- The possible existence of highly hardened surfaces resulting from improper machining practices, wear, or localized working, which can initiate stress-corrosion cracks in high-strength materials
- The possibility that impurities have been dissolved in the surface layers as a result of high-carbon or high-sulfur activity in the environment
- The possibility that processing has introduced hydrogen into the metal, which can result in hydrogen-induced delayed fracture
- Detailed analysis of grain-boundary regions to determine whether composition, precipitation, or structure had not been properly controlled or had changed such that the grain boundaries became preferential regions for stress-corrosion attack (sensitization of austenitic stainless steel has such an effect)
- The possibility that an impurity may be present that is not specified for the alloy and that would render the alloy particularly susceptible to SCC. This might include, for example, the presence of arsenic or antimony in stainless steel—elements for which chemical analysis is usually not conducted.

It is not sufficient to consider only nominal chemical composition in assessing the potential for SCC. The possibility of surface contamination or changes in surface mechanical properties, as well as other structural changes resulting from heat treatment, must also be carefully considered.

Reference cited in this section

13. P.J. Forsyth and D.A. Ryder, *Metallurgia*, Vol 63, 1961, p 117

Stress-Corrosion Cracking

Revised by W. R. Warke

Simulated-Service Tests

Test Conditions. Once the cause and mechanism of a failure have been established, it is desirable to verify the conclusions experimentally so that appropriate corrective measures can be taken. These tests should simulate as closely as possible the environmental and mechanical conditions to which the failed part was subjected in service. Ideally, this would involve testing a part in service; however, such testing is frequently not feasible because of such factors as time-to-failure or extraneous damage caused by failure of the part. Thus, simulated-service tests may have to be performed in the laboratory.

When conducting simulated-service tests, however, exact duplication of service conditions cannot always be achieved. Also, the size and shape of the failed part may not be conducive to laboratory testing. In this case, carefully selected test specimens should be machined from the original part, if possible, incorporating any design features that may have contributed to the failure, such as sharp bend radii or crevices.

Because many parts fail only after long periods of time, it is often not feasible to perform a simulation test of similar duration. In these cases, the test may be accelerated by increasing the stress, temperature, or concentration of the corrodent. However, changing these conditions may produce a type of failure completely different from that observed in service. Increasing the stress may produce failure by overload rather than SCC, and increasing the temperature or concentration of the corrodent may produce general corrosion or pitting corrosion rather than SCC. Meaningful accelerated testing is accomplished largely by trial and error. Once failure of the test part or specimen has occurred, the failure must be investigated using appropriate techniques to ensure that the failure mechanisms and appearance are representative of that previously encountered.

The following factors should be carefully considered in planning simulated service tests:

- *Environmental factors*
    1. Temperature, which may be steady or fluctuating and may also affect stress
    2. Single-phase or two-phase environment, which may involve alternate wetting and drying
    3. Composition, including major and minor environmental constituents, concentration and changes thereof, dissolved gases, and pH
    4. Electrochemical conditions, which may involve galvanic coupling or applied cathodic protection
- *Mechanical factors*
    1. Loading, which may be static or cyclic. If cyclic, mean stress may be zero, tensile, or compressive; also, if cyclic, stress-wave shape and period must be defined.
    2. Surface damage, which may have occured by denting, gouging, fretting, or wear
- *Combined mechanical and environmental factors*
    1. Surface damage, which may have occured by cavitation or liquid-impingement erosion

The common methods for evaluating the susceptibility of metals to SCC are described in Ref 11 and 14.

References cited in this section

11. D. Sprowls, Evaluation of Stress Corrosion Cracking, *Corrosion*, Vol 13, *ASM Handbook*, ASM International, 1987, p 272

14. W.B. Lisagor, *Environmental Cracking—Stress Corrosion, Corrosion Test and Standards: Applications and Interpretation*, R. Baboian, Ed., ASTM, p 240–252

Stress-Corrosion Cracking

Revised by W. R. Warke

Carbon and Low-Alloy Steels

The susceptibility to SCC of carbon steels generally increases as carbon content decreases. However, decarburized steel and bulk pure iron are resistant to cracking; thus, susceptibility must be at a maximum at some concentration of carbon between 0 and approximately 0.10%. The exact carbon content at which cracking occurs most readily is not known, but the susceptibility to cracking is high in steels containing 0.05% C (Ref 15). The nature and concentration of other alloying elements in low-carbon steels have less effect on general susceptibility to SCC than microstructure does.

Surveys of the stress-corrosion behavior of high-strength low-alloy steels in a variety of environments have shown that the strength of the steel is the most important single indication of sensitivity to SCC. Steels with yield strengths of approximately 1379 MPa (200 ksi) or higher are especially susceptible to this type of failure. On the other hand, SCC failures of mild steel and low-strength alloy steels are also quite common.

SCC Environments

Caustic cracking of steel is a serious stress-corrosion problem and has caused many explosions and other types of failures, particularly in steam boilers. It has also been identified at welds in continuous digesters in the pulp and paper industry. Caustic solutions in the form of 5 to 40% aqueous solutions are also widely used in oil refineries and petrochemical plants to neutralize sulfuric, hydrochloric, and hydrofluoric acid streams. The general conditions that cause caustic cracking are reasonably well understood, but failure by such cracking is still a problem.

Caustic-cracking failures frequently originate in riveted and welded structures near faying surfaces and in splash and wet/dry zones where soluble salts accumulate high local concentrations of caustic soda and silica. Crack propagation is ordinarily intergranular. Failures of this type have been produced at concentrations of sodium hydroxide (NaOH) as low as 5%, but a concentration of 15 to 30% NaOH plus a small amount of oxygen is usually required to induce caustic cracking. Failures typically take place when the operating temperature is in the range of 200 to 250 °C (390 to 480 °F). The concentration of NaOH necessary for producing cracking increases as temperature decreases.

Cracking occurs where corrosion potential is such that only part of the steel surface is covered with an oxide film. Cracking can be prevented by anodic protection, which forms a continuous and more stable oxide film, or by cathodic protection, which completely reduces the oxide film.

Reducing pH and adding strong oxidizing agents to passivate the steel surface are common modifications of boiler water treatment that are used to prevent caustic cracking. Inhibitors such as nitrates, sulfates, phosphates, and tannins have been used. However adjustment of boiler water composition and addition of inhibitors should be done only if the chemical behavior of the materials involved is fully understood and only after suitable laboratory testing, because adjustment of boiler water composition can destroy the effectiveness of many inhibitors. Common practice for most steam boilers is to control pH and to monitor the water supply or maintain an effective concentration of inhibitors.

Reducing grain size has been found to be helpful in decreasing the susceptibility to caustic cracking of the low-carbon steels generally used in boilers. Minor changes in steel composition are of little value because they do not alter sensitivity at grain boundaries—areas that are preferentially attacked in caustic cracking. Stress relieving after welding is advisable in order to avoid excessive residual stress around the welds.

MEA/DEA Cracking. Monoethanolamine (MEA) and diethanolamine (DEA) solutions (typically 20 to 25%) are used in gas fields and refineries to remove hydrogen sulfide and carbon dioxide from produced gas or light hydrocarbon streams. The “rich” solution (loaded with H_2S and CO_2) causes hydrogen stress cracking and hydrogen blistering, while the “lean” solution (stripped of gases) can cause a form of SCC called amine cracking. Cracking has been experienced in both pressure vessels and piping at temperatures down to 40 °C (100 °F). Cracking is intergranular, with the crack surfaces covered by a thin layer of magnetite. The cracking is similar to caustic cracking, but the pH and temperature are lower. Invariably, the cracking was associated with non-stress-relieved welds, so all new equipment is stress relieved after welding and the existing units were heat treated in situ.

Cracking in Nitrate Solutions. Stress-corrosion cracking of carbon and low-alloy steels in nitrate solutions has occurred in tubing and couplings used in high-pressure condensate wells and in storage tanks containing radioactive wastes.

Cracking in nitrate solutions follows an intergranular path. Boiling solutions of several nitrates, including NH_4NO_3, $Ca(NO_3)_2$, $LiNO_3$, KNO_3, and $NaNO_3$, have been found to produce cracking. In general, more acidic solutions have more potent effects. The threshold stress necessary to produce cracking decreases with increasing concentration of the nitrate in the solution. This threshold stress can be quite low. For example, exposure to boiling 4 *N* solutions of the nitrates listed has produced cracking in some carbon and low-alloy steels at tensile stresses lower than 69 MPa (10 ksi). Decreasing temperature increased time-to-failure. By extrapolation of these test results, it can be estimated that failure would occur in about 1000 h at room temperature. Room-temperature tests of bridge cable wire in 0.01 *N* nitrate solutions have produced failures after exposure for several months.

Decreasing pH enhances nitrate cracking, and resistance to cracking can be improved by raising pH. Sodium hydroxide, which causes caustic cracking by itself, can be added to nitrate solutions to retard cracking. The reverse is also true: nitrate additions retard caustic cracking. Cathodic protection can prevent SCC in many nitrate solutions. Anodic polarization is harmful. In addition to sodium hydroxide, several other inhibitors prevent cracking.

In low-carbon steels, carbon content has a strong effect on susceptibility to nitrate-induced cracking. Initiation of SCC is minimized when carbon content is lower than 0.001% (extremely low) or higher than approximately 0.18%. The threshold stress for cracking is low at carbon contents of approximately 0.05% but increases at carbon contents above approximately 0.10%. Accordingly, decarburization of steel surfaces can lead to cracking in nitrate solutions. This effect of carbon content applies primarily to crack initiation. A stress-corrosion crack that starts in a decarburized surface layer will continue to propagate into the higher-carbon interior region of the metal. Tests performed on notched specimens have shown that preexisting cracks will

propagate in steels of high carbon content. Crack initiation and propagation in nitrate solutions appear to be closely related to carbon and, possibly, nitrogen in grain-boundary regions. It has been suggested that cracking proceeds by a mechanism in which carbon-rich areas act as cathodes that sustain dissolution of the adjacent ferrite. As would be expected, changes in composition or processing history of the steel so as to alter carbon distribution can appreciably influence susceptibility to SCC. Strong carbide-forming elements, such as chromium, tantalum, niobium, and titanium, are generally helpful in improving resistance to cracking. Reducing grain size and cooling slowly after annealing are also helpful. Steels with pearlitic or coarse spheroidized microstructures are usually less susceptible to cracking than steels containing fine carbides. Moderate cold working may increase susceptibility to cracking, but severe cold working is beneficial. Just as in caustic cracking, full stress-relief annealing after welding can prevent cracking of welded assemblies, which may contain fairly high residual stresses in the as-welded condition.

Cracking in Ammonia. Carbon steel tanks containing anhydrous ammonia have developed leaks because of SCC. In one case, cracking developed in the interior surface of the head regions of tanks having cold-formed heads, but not in tanks that had been stress relieved after fabrication. Thus, residual tensile stresses, in combination with the applied stresses, were evidently sufficient for cracking to occur. Crack growth in the tanks that failed was slow, and the average service time before detection of leakage was three years.

It has been shown that both plain carbon steels and quenched-and-tempered low-alloy steels are susceptible to SCC in ammonia. In this investigation, both intergranular and transgranular cracking occurred in the quenched-and-tempered steels, but only intergranular cracking in the carbon steels. The higher-strength steels were more susceptible to cracking.

These failures were not produced in pure ammonia, but occurred in ammonia mixed with air or air plus CO_2. It is now known that the critical ingredient for cracking is oxygen and that nitrogen further aggravates cracking. It was discovered that addition of water (0.2 wt%) to liquid ammonia completely eliminated susceptibility to cracking. Water additions have become mandatory by U.S. Department of Transportation ruling for interstate transportation of ammonia in high-strength steel vessels (>620 MPa, or 90 ksi, yield strength). Since the adoption of this ruling, the problem of SCC of steel in ammonia has disappeared in the United States.

Carbonate/Bicarbonate Cracking of Pipelines. This form of SCC is a major problem for gas transmission pipelines. Two forms of cracking have been identified: classical, high-pH SCC and non-classical, low-pH SCC. Cracking generally occurs in older tape wrapped and coal tar or asphalt coated pipe, and disbonding of the coating—so that ground water can get to the metal surface—is a prerequisite for both forms. When the classical form is observed, the pH values of the water under the pipe coating will be in the range of 8.5 to 11 and the carbonate/bicarbonate content of the water will be high. On the other hand, the nonclassical SCC occurs in the pH range of 6 to 8.5 in dilute bicarbonate water. The high-pH SCC will be found within approximately 12 miles of a compressor station, but the low-pH SCC has been seen as far as 75 miles from the nearest compressor.

The conditions responsible are the carbonate-bicarbonate environment, the pipe surface in the critical potential range for cracking (-0.6 to -0.7 V versus the saturated calomel electrode, SCE, for a carbonate-bicarbonate environment), and the applied tensile stress. The cathodic potential also creates the critical potential for SCC. The hoop stress in the circumferential direction of the pipe is created by the internal pressure in the pipe. The heat of compression carried by the gas heats the pipe near the compressor stations, and the classical SCC growth rate increases exponentially with temperature. The high-pH cracks are intergranular and tight, while the low-pH cracks are transgranular with evidence of corrosion of the crack walls. Both forms are promoted by the slow pressure cycling of the pipeline associated with low and high demand periods through the day.

There is no simple control for external SCC, even though one of the three conditions necessary for cracking must be prevented. This means keeping the cathodic potential outside of the critical range for cracking, which is difficult, if not impossible, at all points on a long pipeline. Keeping the temperature as low as possible slows the reaction rate, which helps to prevent cracking. Many of the pipeline failures have been located in the first 16 km (10 miles) downstream of a gas compressor station, which is the high-temperature region on a gas pipeline. Applying a good coating to the pipe surface is also a deterrent, as is shot peening the pipe surface to remove mill scale and provide a good surface for coating bonding. Small pressure fluctuations have been found to promote cracking; thus, pressure control is helpful, but it is difficult to achieve in most situations.

Mitigation involves awareness of the nature of soils along the pipeline, close monitoring of the cathodic protection system on the line, and periodic excavation of sections of the line for magnetic particle examination. Nondestructive inspection for longitudinal SCC cracks using “inspection pigs” passed through the pipeline is a developing technology. Predictive methods are reviewed in Ref 16.

Steam Turbine SCC. Stress-corrosion cracking of carbon and alloy steel steam turbine components has included cast steel casings, rotors, wheels, and bolting. Of particular concern has been the SCC of NiCrMoV alloy steel rotors and wheels. The basic cause is transfer of caustic and other ionic species from the boiler water to the turbine. In addition to caustic (NaOH), these species can include chlorides, sulfates, carbonates, oxygen, carbon dioxide and silica, all of which can cause SCC problems. So the SCC is basically caustic cracking, but the environment is made much worse by the other impurities. Some transfer can occur by carryover, that is, the entrainment of water droplets in the steam. Evaporation of the water in the turbine results in concentrated deposits. Other impurities evaporate from the boiler water and concentrate in any water that may condense in the cooler parts of the turbine. Again alternate wetting and drying will lead to high impurity concentrations. Leakage at bolted flanges or the turbine split line can lead to deposition on and SCC of bolting. Adequate control of steam purity is often difficult in industrial plants because of the variety of sources of make-up water. Selection of more resistant (i.e., lower strength) materials and care in operation are essential for long life.

Steel Case Studies

Example 4: Intergranular SCC of Carbon Steel Pipe Welds in a Kamyr Continuous Digester Equalizer Line (Ref 17). Several cracks in and next to welds leaked within 18 months of normal service. The pipe was Schedule 80 low-carbon steel and was used to transfer kraft liquor at 175 °C (350 °F) in a Kamyr continuous pulp digester. The pipe ranged in diameter from 75 to 125 mm (3 to 5 in.) (NPS 3 to 5) and had been field welded using E6010 root passes and E7018 for the cover passes.

Investigation. Cracks were found to run in both the axial and circumferential directions in the welds and heat-affected zones (HAZs) (Fig 16a) and were localized to these areas. The inside surfaces of the pipes were covered with a thin oxide layer, but there was no evidence of general or pitting corrosion.

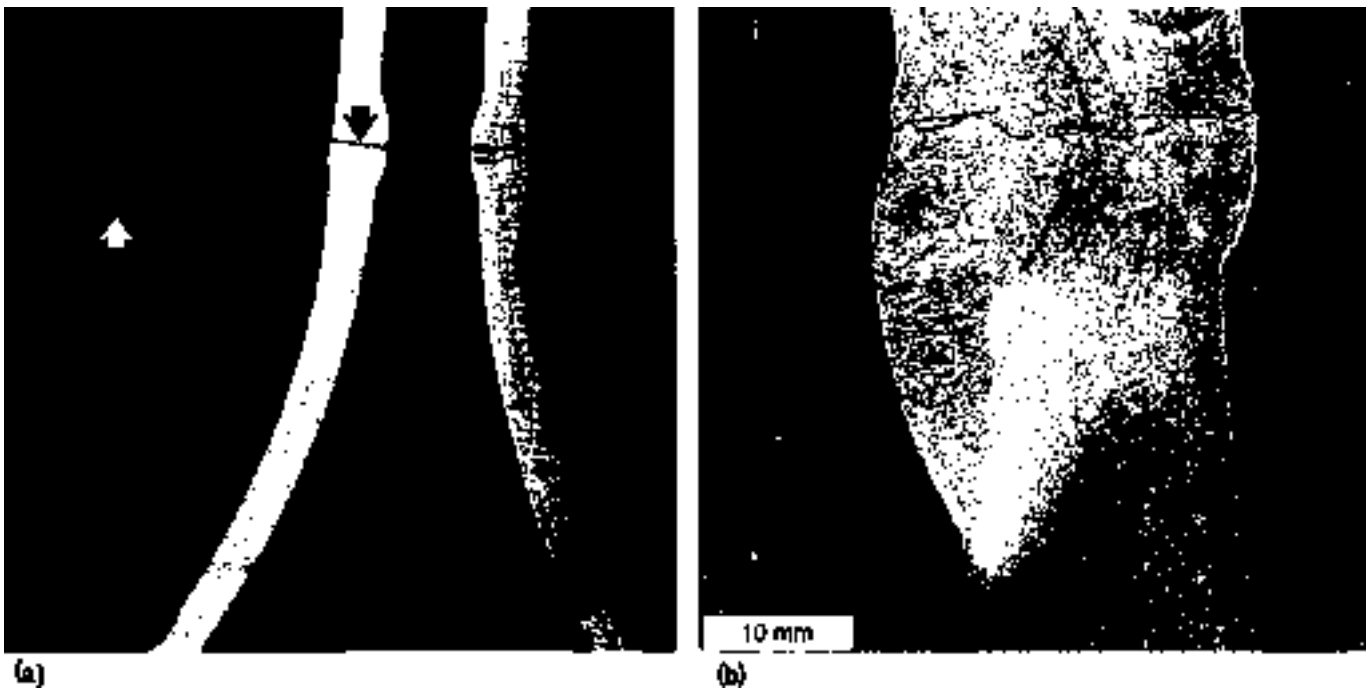


Fig. 16 Stress-corrosion cracks in steel pipe weld. (a) Magnetic particle enhancement of cracks in the weld of an equalizer line elbow section. Cracks were localized in the weld and HAZs. (b) Cracks initiated on the inside surface and propagated through the weld in a multiple branching mode.

Scanning electron microscopy (SEM) examination showed that the subcritical cracks were primarily intergranular, and EDS detected significant levels of sodium in the deposits at the crack tips. Metallographically, the cracks had initiated at the inner surface and propagated through the weld metal to the outside diameter (Fig. 16b). The cracks were intergranular and branched in the ferrite/pearlite microstructure and were oxide filled (Fig. 17). The hardness of the pipe and the weld metal were in the range of 73 to 93 HRB.

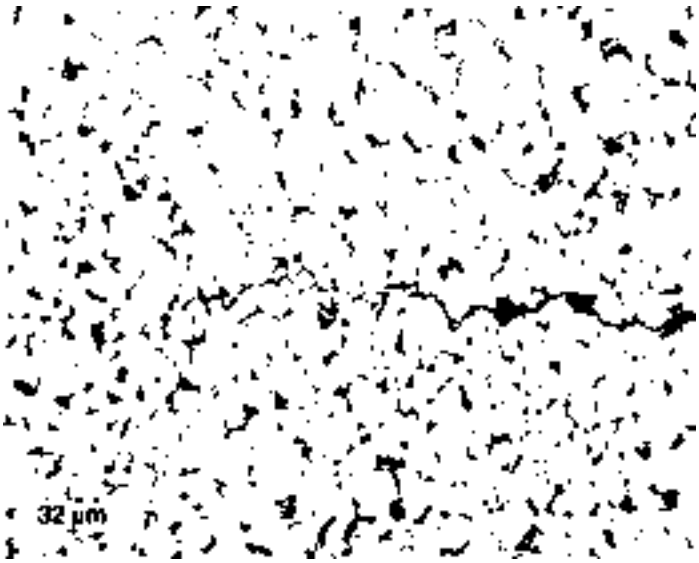


Fig. 17 Tip of an intergranular stress crack that partially penetrated the weld. The crack was covered with oxides.

Discussion. Cracking initiated on the inside diameter surfaces of welds and HAZs. The absence of corrosion damage indicates that carbon steel is suitable for the service, provided the conditions leading to SCC could be eliminated. A primary factor was high residual tensile stresses in the welds and surrounding base metal. Non-stress-relieved welds are known to be susceptible to SCC in caustic environments such as the kraft liquor.

Conclusion. The most probable cause of the intergranular SCC in the pipe welds was high residual stresses that resulted from the welding procedure used and the lack of postweld heat treatment.

Recommendations. The weld procedure should include appropriate controls on weld preheat, interpass temperature, and postweld heat treatment.

Example 5: Leaking Carbon Steel Weldments in a Sulfur Recovery Unit (Ref 18). Two leaks were discovered at a sulfur recovery unit in a refinery. The leaks were at pipe-to-elbow welds in a 152 mm (6 in.) (NPS 6) diameter line, operating in lean amine service at 50 °C (120 °F) and 2.9 MPa (425 psig). Thickness measurements indicated negligible loss of metal, and the leaks were clamped. A year later, 15 additional leaks were discovered, again at pipe-to-elbow welds in lean amine lines. Further nondestructive testing located other cracks, giving a total of 35. These lines had been in service for approximately eight years.

Investigation. Examination revealed that the leaks were stress-corrosion cracks, which had originated at the inside surface. Circumferential cracks were present in the weld metal, the HAZs, and the base metal up to 5 mm (0.2 in.) away from the welds (Fig. 18).

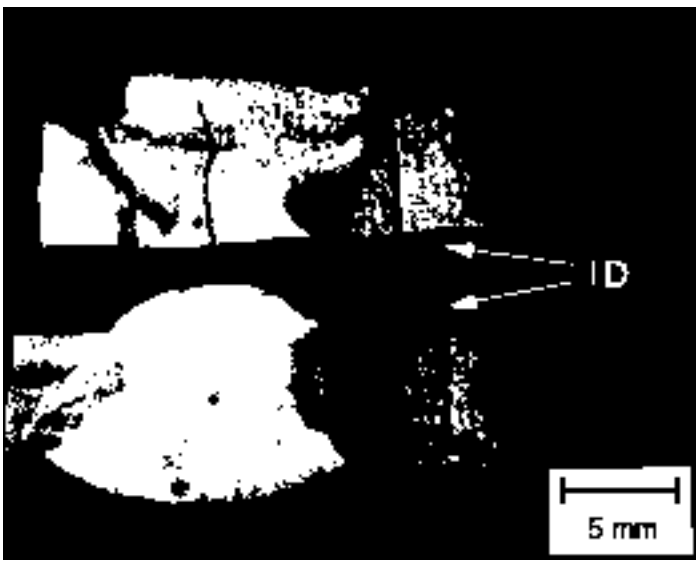


Fig. 18 Cross sections of pipe-to-elbow welds showing stress-corrosion cracks originating from the inside surface of the weld metal and the base metal

Hardnesses were 133 to 230 HB in the weld metal, 182 to 227 HB in the HAZs, and 117 to 198 HB in the base metal, converted from Knoop readings.
The microstructures of the pipes and the elbows were normal, the cracking was intergranular and branched (Fig. 19, 20), and the cracks were oxide filled.

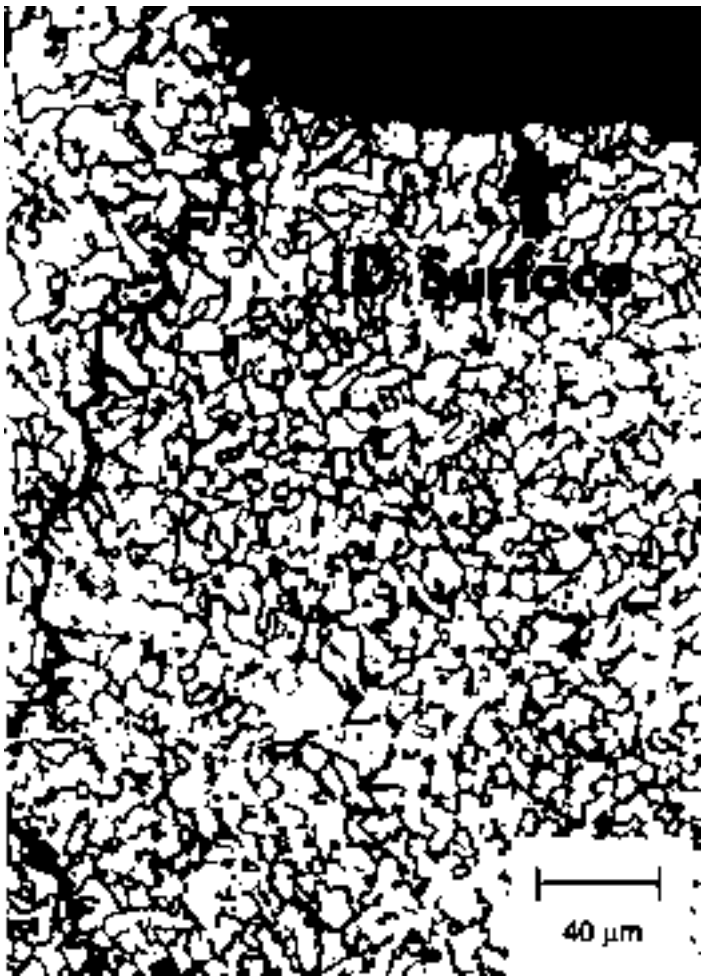


Fig. 19 Micrograph showing tight intergranular SCC originating at the inside surface of a pipe

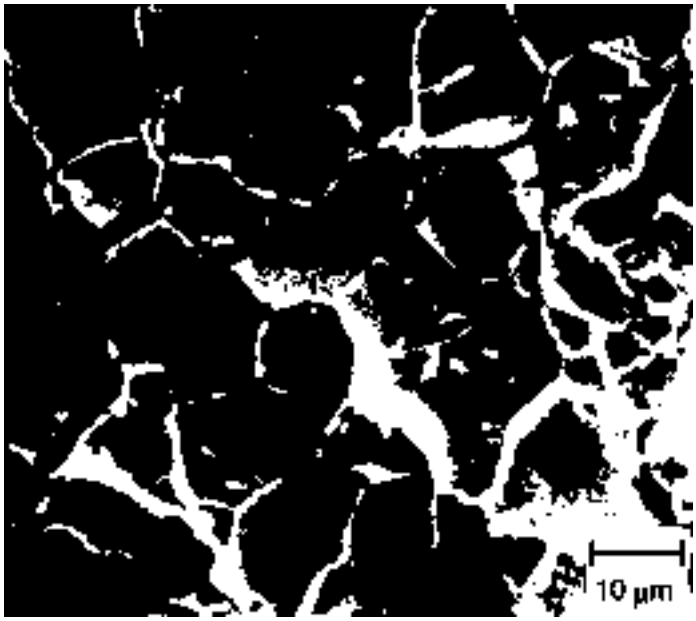


Fig. 20 Scanning electron micrograph showing intergranular SCC (A) and initiation sites for pitting (B) on the inside surface of a pipe

Discussion. It is industry practice to postweld stress relieve piping and pressure vessels in lean amine service if the temperature is expected to be above 95 °C (200 °F). It was therefore considered unlikely that these pipe welds had received such a postweld heat treatment. The observed results were characteristic of lean amine SCC.
Recommendations. It was recommended that all welds in the piping system be inspected using shear wave ultrasonic testing. Most important, all welds in lean amine service were subsequently postweld heat treated.
Example 6: SCC of Carbon Steel Hoppers by Ammonium Nitrate Solution. After 10 to 20 months of service, the carbon steel hoppers on three trucks used to transport bulk ammonium nitrate prills developed extensive cracking in the upper walls. The prills were used in blasting operations and required transportation over rough roads to a mining site.
The prills were discharged from the steel hoppers using air superchargers that generated an unloading pressure of approximately 11 kPa (7 psi). A screw conveyor at the bottom of the hopper assisted in the unloading operation. Each hopper truck held from 9,100 to 11,800 kg (10 to 13 tons) of prills when fully loaded and handled approximately 90,700 kg (100 tons) per month.
The walls of the hoppers were made of 2.7 mm (0.105 in.) thick flat-rolled carbon steel sheet of structural quality, conforming to ASTM A 245 (obsolete specification replaced by A 570 and A 611). The interior surfaces of all three hoppers had been painted before placement in service, after 7 days of service, and after 60 days of service. All exterior surfaces were covered with a gray epoxy coating of the type used at the prill manufacturing plant. There were no specifications for the interior coatings, but one appeared to be a fluorocarbon and another an aluminum pigmented coating.
Investigation. Visual inspection of the internal surfaces of the hoppers revealed extensive rusting under the paint, especially along edges and at corners. This inspection also revealed that most of the cracking was confined to the top areas of the hoppers, around manholes and at circumferential welds in the sheet-steel walls.
The cracking occurred in areas where bare steel had been exposed by chipping and peeling of paint. The cracks were 5 to 20 cm (2 to 8 in.) long and were located in or adjacent to welds. Some cracks were perpendicular to the weld bead, but most followed the weld. In one hopper, cracks were found along the full length of approximately 20 intermittent welds that attached internal bracing to the walls. All cracks in the hopper walls could be seen from the outside, where they appeared as straight lines of rust breaking through the gray external coating.
A rectangular section that included a weld was removed from the upper front edge of the wall of one hopper for metallurgical examination. A transverse section approximately 3.2 mm ($\frac{1}{8}$ in.) from the weld was mounted to permit examination of both the interior and exterior edges of the hopper wall. Examination of metallographic

specimens after polishing and etching with nital revealed an extensive pattern of intergranular cracking (Fig. 21) that had originated at the interior surface.

Fig. 21 Nital-etched specimen of ASTM A 245 carbon steel. Micrograph shows SCC that occurred in a concentrated solution of ammonium nitrate. 100×

Discussion. Ammonium nitrate solutions of certain concentrations are known to crack steel. A concentrated nitrate solution was produced on the upper walls of the carbon steel hoppers by dissolution of ammonium nitrate dust in atmospheric water vapor that had condensed when moist air had entered the hoppers during unloading and during standby periods when the manhole covers were left open.

Conclusions. Failure of the hoppers was the result of intergranular SCC of the sheet-steel walls because of contact with a highly concentrated ammonium nitrate solution. The solution was produced by dissolution of ammonium nitrate dust in condensed atmospheric water vapor that formed intermittently on the upper walls of the tank.

Corrective Measures. A three-coat epoxy-type coating with a total dry thickness of 0.3 mm (0.013 in.) was applied to the interior surfaces of the hoppers. Although high stress levels in the steel resulting from welding, unloading, and traveling on uneven roads contributed to the cracking, it was more economical to coat the interior surfaces of the steel hoppers properly than to design for minimum stress levels or to make a change in material. For example, equivalent aluminum hoppers would cost approximately $2\frac{1}{2}$ times as much as the carbon steel hoppers, and stainless steel hoppers would cost five times as much.

Example 7: Metallurgical Analysis of Steam Turbine Rotor Disc (Ref 19). Numerous cracks observed on the surface of a forged A470 Class 4 alloy steel steam turbine rotor disc from an air compressor in a nitric acid plant were found to be the result of caustic induced SCC. No material defects or anomalies were observed in the sample that could have contributed to crack initiation or propagation or secondary crack propagation. Chlorides detected in the fracture surface deposits were likely the primary cause for the pitting observed on the disc

surfaces and within the turbine blade attachment area. It was recommended that the potential for water carryover or feedwater induction into the turbine be addressed via an engineering evaluation of the water treatment procedures, steam separation equipment, and start-up procedures.

Problem. A steam turbine that drives an air compressor in a nitric acid plant is started using saturated steam but is normally operated using superheated steam. A rotor disc in the turbine was found to contain radial cracks near the rim, when it was magnetic-particle inspected.

The steam turbine disc was fabricated from an A 470 Class 4 alloy steel forging, which is a common Ni-Cr-Mo-V alloy steel steam turbine disc material. Approximately one-third of the disc was available for examination and sectioning.

Investigation. The steam turbine rotor disc segment, which is shown as received for analysis in Fig. 22, was visually examined. Multiple cracks were visible over much of the surface of the damaged disc. The disc surfaces were examined using the dry magnetic-particle techniques. Multiple cracks oriented in the radial direction on the sides of the disc were observed. Figure 23 shows a close-up view of this damage.

Fig. 22 The rotor disc segment as received for analysis

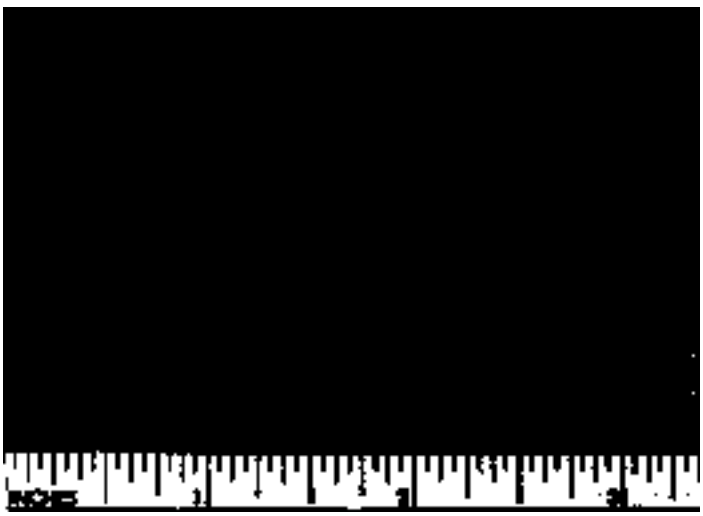


Fig. 23 Macrograph showing a close-up view of the typical magnetic-particle indications observed on the disc side surfaces

The microstructure of the disc was found to consist of tempered martensite and bainite. The micrograph in Fig. 24 shows the typical appearance of the damage; some of the cracking had initiated from the bottoms of corrosion pits. The cracking morphology was mixed mode, that is, both intergranular (around grain) and transgranular (through grain). Figure 25 shows two immediately adjacent cracks exhibiting these two different crack morphologies. The disc surfaces and primary crack paths showed considerable evidence of ongoing corrosion in the form of pitting on the disc surfaces and buildup of corrosion deposits in the larger cracks.

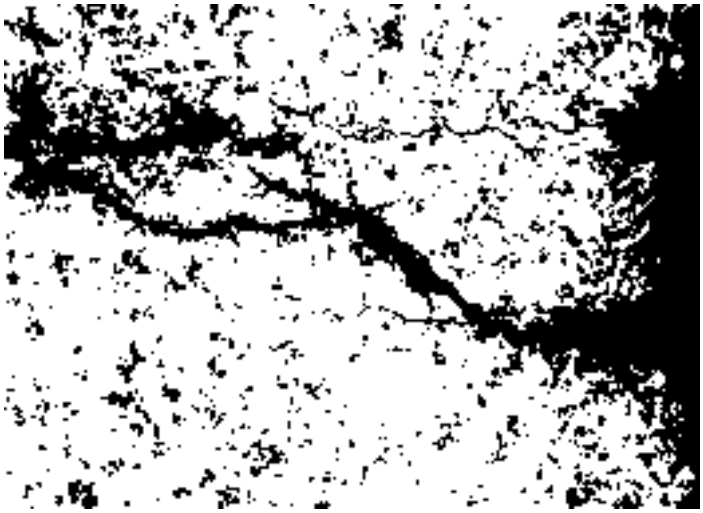


Fig. 24 Micrograph showing a typical example of cracking observed in the blade attachment area. The cracks have propagated from pit bottoms.

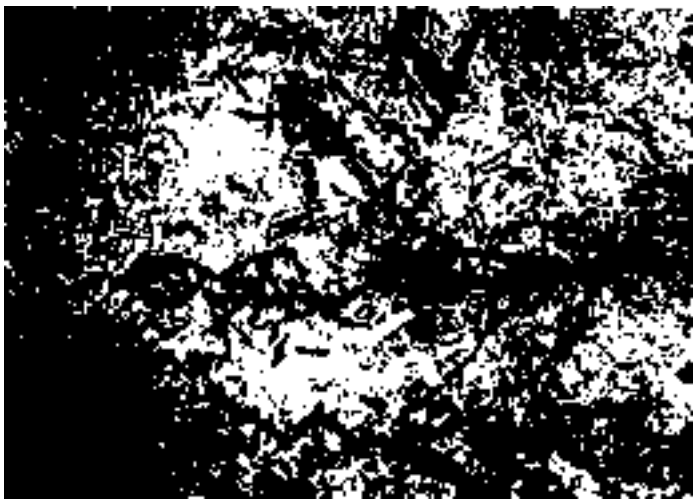


Fig. 25 Micrograph showing two immediately adjacent cracks. The lower crack is primarily intergranular while the upper crack is primarily transgranular.

A sample of the turbine disc segment was analyzed for chemical composition. The results of the compositional analysis of the disc material are presented in Table 3. The compositional requirements for A 470 Class 4 are listed in the table for comparison. The disc material meets the chemical requirements of the indicated specification.

Table 3 Chemical composition of steam turbine rotor disc

Content, wt%		
Element	**Disc**	Chemical requirements per ASTM A 470 Class 4
Carbon	0.27	**0.28 max**
Manganese	0.51	**0.20–0.60**
Sulfur	0.011	**0.018 max**
Phosphorus	0.010	**0.015 max**
Silicon	0.28	**0.15–0.30**
Nickel	3.55	**2.50 min**
Chromium	0.41	**0.75 max**
Molybdenum	0.41	**0.25 min**
Vanadium	**0.09**	**0.03 min**

The deposits present at the termini of laboratory-opened cracks were analyzed in situ using EDS. The typical results of an in situ EDS analysis of the crack surface deposits are presented in Fig. 26. The deposits contained major amounts of iron, silicon, calcium, potassium and chlorine, plus minor to trace amounts of aluminum, sulfur, sodium, magnesium, phosphorus, titanium, chromium, manganese, nickel, and zinc. The iron, nickel, manganese, and chromium are base metal constituents. The other deposit constituents likely came from feedwater carryover.

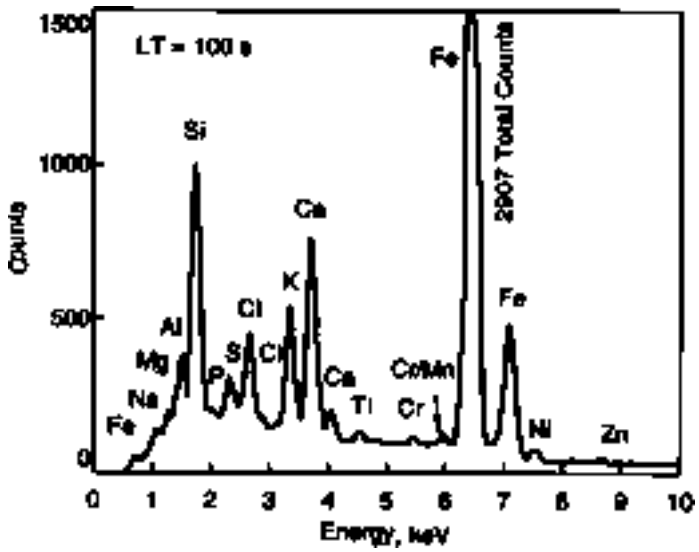


Fig. 26 Results of the in situ energy-dispersive spectroscopy analysis of the deposits present in the fracture terminus areas of the laboratory opened crack sample

The bulk metal hardness of the steam turbine disc was measured using the Rockwell test, and readings averaged 20.2 HRC. This hardness value corresponds to a tensile strength of approximately 758 MPa (110 ksi) per the ASTM A 370 approximate tensile strength conversion. This value meets the specification requirements.

Discussion. The multiple, radially-oriented mixed mode cracking observed in the disc is indicative of caustic-induced SCC. Both sodium and potassium were identified at the fracture terminus areas. The source of the caustic could have been carryover, the introduction of feedwater into the turbine, or severe steam quality problems. Chlorine (as chlorides) was also detected in the fracture surface deposits. The chlorides were likely the primary cause for the pitting observed on the disc surfaces and within the turbine blade attachment area. Both pitting and SCC occur in wet environments.

The disc composition, microstructure, and hardness appeared to be normal. No material defects or anomalies were observed that could have contributed to the cracking.

Conclusions. The most probable cause of the numerous cracks in this steam turbine disc was caustic-induced SCC.

Recommendations. An evaluation of several factors in the operation of the steam system of the plant would be required to identify the source of the caustic. The operation of the steam drum/steam separators and the effectiveness of the plant water treatment program needed to be addressed. Additionally, the procedure of using saturated steam during startup should be assessed with regard to the potential for inducting feedwater into the steam turbine.

References cited in this section

15. K. Bohnenkamp, Caustic Cracking of Mild Steel, *Fundamental Aspects of Stress Corrosion Cracking*, R.W. Staehle, A.J. Forty, and D. van Rooyen, Ed., National Association of Corrosion Engineers, 1969, p 374-383

16. N. Sridhar, D.S. Dunn, and A. Anderko, Prediction of Conditions Leading to Stress Corrosion Cracking of Gas Transmission Lines, *Environmentally Assisted Cracking: Predictive Methods for Risk Assessment and Evaluation of Materials, Equipment and Structures*, STP 1401, R.D. Kane, Ed., American Society for Testing and Materials, 2000

17. D.G. Chakrapani, Intergranular SCC of Carbon Steel Pipe Welds in a Kamyr Continuous Digester Equalizer Line, *Handbook of Case Histories*, Vol 1, ASM International, 1992, p 168

18. J. Gutzeit and J.M. Johnson, “Stress-Corrosion Cracking of Carbon Steel Welds in Amine Service,” Paper 206, presented at Corrosion/86, March 1986, (Houston, TX), National Association of Corrosion Engineers

19. L. Hodas, Metallurgical Analysis of Steam Turbine Rotor Disc, *Handbook of Case Histories in Failure Analysis*, Vol 2, ASM International, 1993, p 307

Maraging Steels

Stress-corrosion cracking of maraging steels has been studied extensively using U-bends, unnotched tensile specimens, machined notched specimens, and fatigue-precracked fracture toughness test specimens. As with high-strength steels, the more severe the notch, the greater is the susceptibility to SCC.

As shown in the following table, the threshold stress intensity needed for SCC of maraging steels in aqueous environments decreases as yield strength is increased.

Yield strength		**Threshold stress intensity**	
MPa	**ksi**	**MPa** $\sqrt{m}$	**Ksi** $\sqrt{in}$
1379	200	110	100
1724	250	44	40
2069	300	11	10

This general trend of decreasing threshold values with increasing yield strength is similar to what is observed for other high-strength steels. At a given yield strength, maraging steels generally have somewhat higher threshold values than other high-strength steels.

Crack-growth rates for maraging steels usually show the three-stage curve characteristic of other high-strength steels. At lower stress intensities, the rate of cracking increases exponentially with stress intensity. At higher stress intensities, a plateau is reached at which crack-growth rate is constant with increasing stress intensity. At still higher stress intensities, near the values for fracture toughness in air, there is a further increase in the rate of SCC with stress intensity. Crack-growth rate at a given stress intensity increases as yield strength increases.

Path of cracking is usually intergranular along prior-austenite grain boundaries. Transgranular cracking has also occurred, and may preferentially follow the martensite-platelet boundaries. Whether or not the crack path depends on stress intensity, as in low-alloy steels, is not clear. Microscopic branching of the main crack into two cracks inclined to the plane of the precrack occurs frequently, but the conditions that cause branching are not known. Crack branching in specimens broken in air has been reported and is believed to be associated with banding of the microstructure. Thus the appearance of crack branching by itself cannot be taken as proof of SCC.

For smooth specimens tested in water, adding chloride ions and increasing pH from 9 to 13 accelerates SCC, probably by accelerating pitting. With precracked specimens these variables generally have much less effect. Cathodic protection has not been found to be a consistently effective way to prevent SCC of maraging steels.

High annealing temperatures that coarsen the microstructure are deleterious. Underaging heat treatments are also generally deleterious. Overaging heat treatments have produced mixed results. Generally, the best practice is probably the use of a low annealing temperature and the standard maraging heat treatment. Cold working either before or after aging improves resistance to SCC. Prestressing so as to leave residual compressive stresses at the notch roots is also helpful.

Variations in major alloying elements have little or no effect on cracking behavior. Of the minor elements, carbon and particularly sulfur are very deleterious. Marked improvement in resistance to cracking can be obtained if the concentration of these elements is held to very low levels.

As with other high-strength steels, there is no completely satisfactory way to prevent SCC in maraging steels other than by reducing the strength level. In some cases, cathodic protection, surface coatings, and prestressing can prevent SCC, but such techniques must be used with care.

Austenitic Stainless Steels

Although the SCC of austenitic stainless steels has been studied extensively, it is not possible to predict the exact conditions under which failure will or will not occur. However, certain behavior patterns have been noted. Austenitic stainless steels should be quenched rapidly from high temperatures to avoid sensitization. Slow cooling from approximately 850 °C (1560 °F) to approximately 550 °C (1020 °F) in heat treatment or welding increases susceptibility to intergranular SCC (or to intergranular corrosion), but rapid cooling through this range prevents such damage. Heating or service exposure in this temperature range will also lead to sensitization of unstabilized grades of austenitic stainless steel. The extent of damage is a function of both time and temperature. In general, the damage is attributable to the precipitation of chromium carbides at grain boundaries, resulting in a depletion of chromium in the adjacent grain boundaries and matrix. The limits of the so-called sensitizing-temperature range cannot be defined exactly, because they are influenced by the composition of the stainless steel, especially the percentages of carbon and of stabilizing carbide-forming elements or combinations of such elements as titanium, niobium, and niobium plus tantalum. Grades containing these elements, such as type 321 or type 347, are termed stabilized grades, while those that do not, such as type 304 or type 316, are called unstabilized. Low carbon grades, such as type 304L, are more resistant to sensitization than the corresponding grade with a higher carbon content.

Also, whether sensitizing is damaging in a given instance depends on the requirements of the specific application, on the environment and stresses to which the steel is exposed, and on the prior thermal and mechanical working history of the steel. Holding at temperatures as low as approximately 400 °C (750 °F) and as high as approximately 900 °C (1650 °F) for prolonged periods has been reported to cause sensitization of unstabilized austenitic stainless steels.

SCC Environments

Chloride SCC. The SCC of austenitic stainless steel in chloride solutions is so widely known and researched as to be almost the classic example of the phenomenon. The cracking is normally transgranular, and in the scanning electron microscope the facets have a very brittle, cleavagelike appearance (e.g., Fig. 15a). Energy-dispersive spectroscopy of crack deposits or of the fracture facets from chloride SCC will generally show the presence of chloride and enrichment of chromium relative to the base metal. This enrichment can be used to differentiate chloride SCC from other types of cracking.

Factors that can influence chloride SCC (Cl-SCC) include steel composition, chloride ion concentration, oxygen concentration, temperature, stress level, and pH. Both stabilized and unstabilized stainless steels are susceptible to Cl-SCC, but the resistance to it increases with increasing nickel content, especially above 20%. It is also possible to use lower-nickel-content duplex stainless steels. These alloys contain 40 to 60% ferrite in their microstructure. Even if stress-corrosion cracks initiate in the austenite phase, they are soon blocked by ferrite and restricted from growing beyond a microscopic size. The same effect occurs in the normal, ferrite-containing cast CF alloys. A further benefit of the duplex stainless steels is their higher strength giving higher design stresses and lighter weight.

In service, Cl-SCC is usually only observed above approximately 65 °C (150 °F), but under certain conditions and in the laboratory, cracking can occur down very close to room temperature. The threshold stresses for Cl-SCC are quite low, so it can be assumed that sufficient stress is always present, whether from the applied loads or pressures or from forming or welding residual stresses.

When chlorides are known to be in the environment, measures to avoid SCC are usually taken, so the chlorides in the case of field failures are often from an unexpected source. Perspiration dripping on warm stainless steel can initiate cracking. Stress relief, coatings, and peening to produce residual compressive stresses are helpful in preventing SCC of austenitic steels. External SCC of stainless steel piping has been traced to rain water entering through damaged insulation jacketing and either leaching chlorides from the insulation or, at seacoast

locations, the rain itself may contain traces of chloride and can reach the pipe surface and concentrate by evaporation. Internal SCC can occur after hydrotesting if the water contains even small amounts of chloride. There will almost always be areas that do not drain, and under the heat of the sun the water will evaporate, the solution will become concentrated, and cracking can initiate and grow.

Caustic SCC. The austenitic stainless steels are also susceptible to SCC in caustic solutions, but the onset is at higher temperatures than for carbon steel—105 to 205 °C (220 to 400 °F), depending on concentration. Again, concentration of the sodium hydroxide by evaporation of water could change a set of conditions from one that is safe to one of a temperature and high pH where SCC would occur.

It was thought for a long time that caustic SCC of stainless steel was always intergranular and therefore could be distinguished from transgranular Cl-SCC by metallography. It has since been found that both types of SCC can have similar crack paths and that this simple rule of thumb is not reliable. When examined by EDS and SEM, caustic SCC cracks will usually contain sodium and be enriched in nickel relative to the base metal.

Polythionic acid stress-corrosion cracking (PASCC) occurs in refinery units processing sulfurous streams. It is a very severe, rapid form of intergranular SCC that attacks sensitized austenitic stainless steels such a weld HAZs in type 304. Polythionic acid is a generic term used to describe acids of the form $H_2S_xO_6$. These acids form during downtime (turnarounds) after air enters the equipment and the temperature drops below the ambient dew point. The iron/chromium sulfide scales on the steel react with oxygen and water to form the corrodent. Testing has shown the damaging specie to be the tetrathionate ion, where *x* in the above formula has the value of four. The thiosulfate ion ($S_2O_3^{-2}$) is just as damaging, producing very high rates of crack propagation. The stresses required are very low, well within the range of welding residual stresses. Polythionic acid SCC can cause severe damage to plant equipment within a day or two following a shutdown. Guidelines are contained in NACE Recommended Practice RP-01-70 for avoiding PASCC during downtime. The practices include blanketing with dry nitrogen or very dry air or neutralizing the acids by spraying with a 2% aqueous sodium carbonate solution. The use of the stabilized grades of stainless steel, types 321 and 347, is helpful, but sensitization can still occur in service in weld HAZs, even in these grades, if they are not thermally stabilized after welding.

High-Temperature Water SCC. In the mid-1970s there was an outbreak of incidents of SCC at the HAZs of pipe welds in boiling water reactor (BWR) service. This cracking also required that the steel in these zones be sensitized, and the residual stresses in the pipe circumferential welds were well able to support cracking, leading to leakage of reactor water. The reactor water contained a small amount of radiolytically produced oxygen, which was sufficient to provide the electrochemical conditions needed for SCC. The presence of chloride, sulfate, and carbonate impurities in the water increased the likelihood of cracking. Mitigation of the pipe cracking attacked each of the three factors involved in SCC. Nuclear grade steels were developed with extra low carbon contents to avoid sensitization and increased nitrogen to restore strength. Welding procedures and in situ thermal treatments were employed to reduce residual stresses. Hydrogen injection into the reactor water was introduced to lower the oxygen content of the water.

Case Studies

Example 8: SCC of Type 316 Stainless Steel Piping. A 680,000 kg (750 ton) per day ammonia unit was shut down following fire near the outlet of the waste heat exchanger. The fire had resulted from leakage of ammonia from the type 316 stainless steel outlet piping.

As Fig. 27(a) shows, the outlet piping immediately downstream from the waste heat exchanger consisted of a 46 cm (18 in.) (NPS 18) diameter flange, a 46 to 36 cm (18 to 14 in.) (NPS 18 to NPS 14) outside-diameter reducing cone, a short length of 36 cm (NPS 14) outside-diameter pipe, and a 36 cm (NPS 14) outside-diameter 90° elbow, all joined by circumferential welds. The flange was made from a casting; the reducing cone, pipe, and elbow were made of 13 mm ($\frac{1}{2}$-in.) thick plate. A 33 cm (13 in.) outside-diameter by 1.6 mm ($\frac{1}{16}$ in.) thick liner wrapped with insulation was welded to the smaller end of the reducing cone. All of the piping up to the flange was wrapped with insulation.

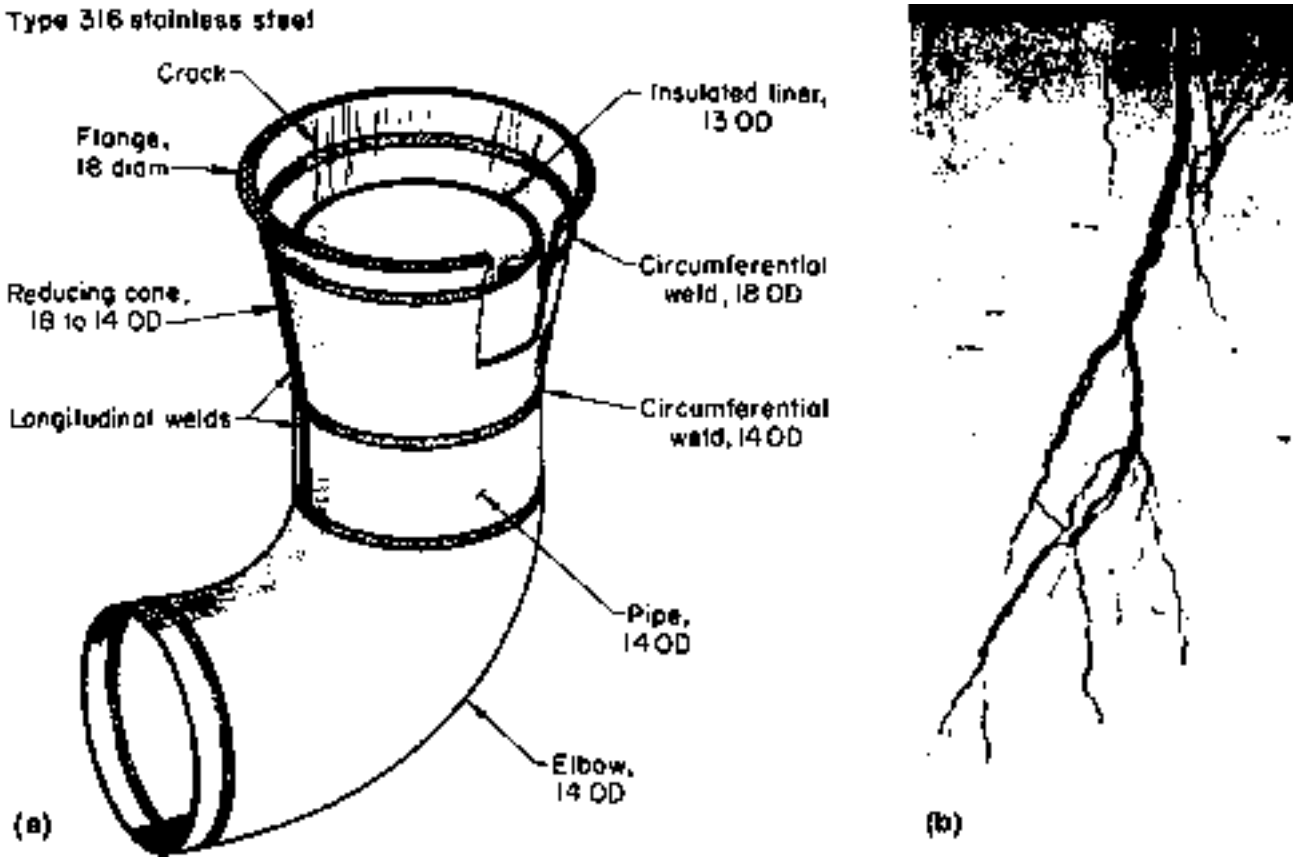


Fig. 27 AISI type 316 stainless steel piping that failed by SCC at welds. Cracking was caused by exposure to condensate containing chlorides leached from insulation. (a) View of piping assembly showing cracks on inner surface of cone. Dimensions given in inches. (b) Macrograph of an unetched section through cone base metal near flange-to-cone weld showing branched cracks. 10×

Investigation. Visual examination disclosed transverse cracks in the weld joining the flange and cone that ran into the base metal on either side. The cracks extended only a short distance into the flange but well into the reducing cone. Cracks also were found in a short section of the longitudinal weld in the 36 cm (NPS 14) outside-diameter pipe between the cone and the elbow.

The piping assembly was cut into the individual components for ease of examination. The components were then liquid penetrant inspected. At the flange end of the cone there were numerous cracks across the inner top surface from the eight o'clock position to the four o'clock position. Only one of these cracks was visible on the outer surface. Figure 27(a) shows several of the cracks on the inner surface of the cone.

Analysis of the insulation wrapped around the liner revealed chlorides in the form of sodium chloride, at a concentration of 190 to 300 ppm, produced by selective leaching.

Metallographic examination of specimens from the cracked area of the cone revealed that both the weld joining the flange and the cone and the base metal of the cone contained several branched, transgranular cracks, all of which had originated at the inner surface (Fig. 27b).

Discussion. The branched, transgranular cracks in the flange-to-cone weld and in the base metal of the cone were characteristic of chloride stress-corrosion cracks in type 316 stainless steel. There were two possible sources of chlorides: boiler feedwater and insulation. The boiler feedwater was an unlikely source because it was monitored regularly to guard against chlorides. However, the insulation material between the cone and its inner liner was known to contain leachable chlorides.

Chlorides will not cause SCC unless an aqueous phase is present; therefore, the temperature of the outlet stream must have dropped below the dewpoint at some time. It is also possible that the absence of insulation at the flange caused a cool spot at which condensation could occur.

Conclusions. The outlet piping failed in the area of the flange-to-cone weld by SCC as the result of aqueous chlorides; the chlorides were leached from the insulation around the liner by condensate.

Recommendations. To prevent future failures, it was recommended that the chlorides be eliminated from the system, that the temperature of the outlet stream be maintained above the dewpoint at all times, or that the type 316 stainless steel be replaced with an alloy more resistant to chloride attack. Use of a more resistant alloy, such as Incoloy 800, would be another reliable approach.

Example 9: SCC in a Neck Liner. A CF-8M (cast type 316) neck liner or manway was removed from the top of a digester vessel. Repeated attempts to repair the part in the field during its life cycle of many years had failed to keep the unit from leaking.

Investigation. Figure 28(a) shows the general appearance of the part. Visual inspection revealed extensive weld repair and evidence of a heavy buildup of material at the abrupt change in diameter to the lower neck of the body. Figure 28(b) is a cross section through the part that shows weld metal, a piece of wrought plate that had been welded in during a repair, and the original casting. Chemistries of the three constituents are presented in Table 4. The casting was a CF-8M modified with the molybdenum level at the top end of the range. The plate was standard 317L material. The filler metal was type 316, although marginal in molybdenum content. The cracking and repair work were concentrated above the change in diameter where the heavy corrosion products and caked-on pulp had collected.

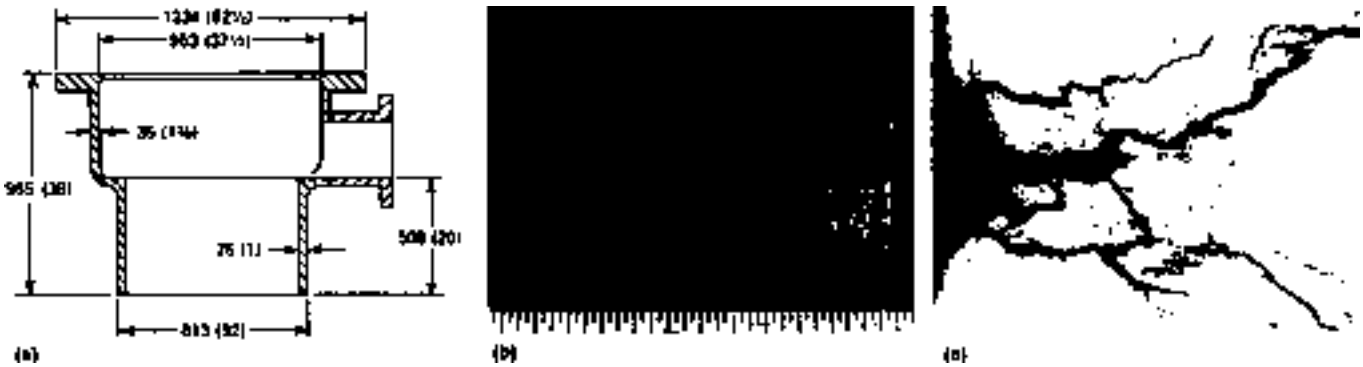


Fig. 28 Stress-corrosion cracking liner of cast neck liner. (a) Illustration of neck liner removed from a pulp digester vessel. Note the abrupt change in cross section that led to a caked-on buildup. Dimensions given in millimeters (inches). (b) Extensive SCC into the 317 plate on the inner surface. (c) Close-up view of the cracks with the classic branched appearance of SCC

Table 4 Chemistries of failed neck liner

Content, %			
Element	**Casting**	**Plate**	Weld
Carbon	0.02	0.02	**0.043**
Manganese	1.26	1.69	**1.50**
Silicon	0.93	0.54	**0.36**
Chromium	19.03	18.47	**17.89**
Nickel	10.32	13.92	**11.54**
Molybdenum	2.81	3.10	**1.91**
Copper	0.09	0.16	**0.13**
Sulfur	0.009	0.018	**0.21**
Phosphorus	**0.030**	**0.026**	**0.031**

Figure 28(c) is a micrograph showing one of the weld-metal cracks in Fig. 28(b). The crack has the branched, transgranular appearance typical of Cl-SCC in stainless steel. Metallography also indicated that prior cracks had not been completely removed before welding.

Discussion. It appears the original casting suffered SCC in a low-oxygen area high in chlorides from repeated wet/dry cycles where there was a buildup of debris. Repair attempts had worsened the situation by increasing the stress levels with the addition of residual stresses and by leaving stress concentrators under the repairs from all of the previously cracked metal not being removed.

Conclusion. The damage to the neck liner was due to Cl-SCC in an area of debris buildup.
Recommendation. The neck liner should be redesigned to eliminate the abrupt change where there was debris buildup. If redesign was impossible, an alloy more resistant to Cl-SCC, such as a duplex stainless steel or a high-molybdenum (4 to 6%) austenitic stainless steel, should be used.
Example 10: Eddy-Current Inspection of Pitting and SCC of Type 316 Stainless Steel Evaporator Tubes in a Chemical Processing Operation. Eddy-current inspection was performed on a vertical evaporator unit used in a chemical processing plant. The evaporator contained 180 tubes, 25 mm (1 in.) in diameter. The tube material was type 316 stainless steel. The shell side fluid was condensate and gaseous methylene chloride, while the tube side fluid was contaminated liquid methylene chloride. Eddy-current inspection revealed 101 tubes that exhibited severe outer surface pitting and cracklike indications near each tube sheet. Several tubes exhibiting strong indications were pulled for visual and metallurgical examination.
Investigation. It was observed that the indications correlated with rust-stained, pitted, and cracked areas on the outer surfaces. The condition was most severe along the portions of the tubes located between the upper tube support and top tube sheet. Figures 29(a) and (b) show a pitted and cracked area before and after dye-penetration application.

Fig. 29 Pitting and stress corrosion in type 316 stainless steel evaporator tubes. (a) Rust-stained and pitted area near the top of the evaporator tube. Not clear in the photograph, but visually discernible, are myriads of fine, irregular cracks. (b) Same area shown in (a) but after dye-penetrant application to delineate the extensive fine cracks associated with the rust-stained, pitted surface. (c) Numerous multibranched, transgranular stress-corrosion cracks initiating from the outer surface pits. 35×. Courtesy of J. P. Crosson, Lucius Pitkin, Inc.

Metallographic examination revealed that the cracking initiated from the outer surface, frequently at pits, and penetrated the tube wall in a transgranular, branching fashion. The crack features were characteristic of Cl-SCC. In many cases the cracking, rather than penetrating straight through the tube wall, veered off in a tangential direction at or about midwall, suggesting the possibility of a change in the residual stress field from tube drawing. Figure 29(c) shows SCC originating from pits on the outer surface of the tube.
Conclusion. The results of the examination indicated that the subject tube failures occurred by way of SCC as a result of exposure to a wet-chloride-containing environment.
Recommendation. A change in tube material was recommended to avoid future failures and loss of service.
Example 11: SCC of Type 316 Stainless Steel Tubing. A steam-condensate line made of type 316 stainless steel tubing that measured 19 mm (0.75 in.) outside diameter by 1.7 mm (0.065 in.) wall thickness had been in service for five to six years when leakage occurred. The line was buried, with no wrapping, in a sandy, caustic soil. The line carried steam condensate at 120 °C (250 °F) with a 2 h heat-up/cool-down cycle. No chemical treatment had been given to either the condensate or the boiler water. Two sections of the tubing, with a combined length of approximately 330 mm (13 in.), were submitted for laboratory analysis of the cause of the leakage.

Investigation. The outside surface of the tubing exhibited no signs of corrosion; however, on one of the sections there was a transverse crack along the edge of a 6.4 mm ($\frac{1}{4}$ in.) diameter pitted area. Two adjacent areas were raised, resembling blisters, and also had transverse cracks along their edges. The cracks were irregular and had penetrated completely through the tube wall. The pitted area was believed to be a spall, and the blisters incipient spalls, resulting from SCC that originated at the inner surface of the tube. The cracks, the spall, and the blisters did not appear to have been related to any effect of the external environment.

The inside surface of the tubing was covered with a brown, powdery scale, was slightly rough, and contained small pits. Some cracks were found that had originated at pits on the inner surface but had not completely penetrated the tubing wall (Fig. 30). The cracks were transgranular and branching, which is typical of chloride-induced SCC of austenitic stainless steel.

Fig. 30 Section through type 316 stainless steel tubing that failed by SCC because of exposure to chloride-contaminated steam condensate. Micrograph shows a small transgranular crack that originated at a corrosion pit on the inside surface of the tubing and only partly penetrated the tubing wall. Etched with aqua regia. 250×

To check for chlorides, the inside of the tubing was rinsed with distilled water, and the rinse water was collected in a clean beaker. A few drops of silver nitrate solution were added to the rinse water, which clouded slightly because of the formation of insoluble silver chloride. This test indicated that there was a substantial concentration of chlorides inside the tubing.

Conclusion. The tubing failed by chloride SCC (Cl-SCC) Chlorides in the steam condensate also caused corrosion and pitting of the inner surface of the tubing. Stress was produced when the tubing was bent during installation.

Recommendations. Water treatment should be provided to remove chlorides from the system. Continuous flow should be maintained throughout the entire tubing system to prevent concentration of chlorides. No chloride-containing water should be permitted to remain in the system during shutdown periods. Bending of tubing during installation should be avoided to reduce residual stress.

Ferritic and Duplex Stainless Steels (Ref 2)

Ferritic stainless steels have been reported to have high resistance to SCC and are highly resistant to SCC in the chloride and caustic environments that crack the common austenitic stainless steels. However laboratory investigations have shown that additions of small amounts of nickel or copper to ferritic steels may make them susceptible to cracking in severe environments. This is especially true for as-welded steels. Nevertheless cracking in chloride or caustic environments in service occurs very rarely.

Materials factors that have been identified as detrimental to the chloride SCC (Cl-SCC) of ferritic stainless steels include the presence of certain alloying elements, sensitization, cold work, high-temperature embrittlement, and the precipitation of α′ (at 475 °C, or 887 °F) (Ref 20). Although it is not clear whether these effects are directly related to Cl-SCC or represent manifestations of other phenomena such as hydrogen embrittlement or stress-aided intergranular corrosion, the effects can be detected in the standard boiling magnesium chloride test. The alloying elements that are detrimental to the resistance of ferritic stainless steel to Cl-SCC include copper, nickel, molybdenum (in the presence of nickel), cobalt (in the presence of molybdenum), ruthenium, sulfur, and carbon (Ref 20).

Any structural features that reduce the ductility of ferritic stainless steels, such as the formation of carbonitrides or α′ and cold work, can reduce the resistance of these steels to SCC. Ferritic stainless steels with chromium contents of 14% or more are subject to embrittlement at temperatures from approximately 370 to 540 °C (700 to 1000 °F). Embrittlement results in loss of room-temperature toughness.

Hydrogen embrittlement also has been identified as the cause of SCC of ferritic stainless steels when they are cathodically polarized in aqueous environments (Ref 21, 22). For example, the super ferritic stainless steels 29-4C and Sea-Cure exhibit hydrogen embrittlement when they are cathodically polarized to potentials in the range of -0.9 to -1.4 V (saturated calomel electrode) (Ref 21). Hydrogen embrittlement of these alloys can also occur when they are exposed to acidic environments or in the presence of hydrogen-ion recombination poisons. The fracture mode of hydrogen embrittlement in these superferritic stainless steels is transgranular cleavage.

Duplex stainless steels with 10 to 28% Cr and 4 to 8% Ni contain both austenite and ferrite. An attractive feature of these alloys is that they have much higher strength than either the austenitic or the ferritic stainless steels. Although duplex stainless steels are susceptible to Cl-SCC, the threshold stress for SCC is generally much higher for duplex stainless steels (Ref 23). The microstructural parameters that have a significant effect on the susceptibility to Cl-SCC are the presence of σ-phase and the ferrite content. The presence of σ-phase has been shown to be detrimental to SCC resistance in boiling 35% $MgCl_2$ at 125 °C (257 °F) (Ref 23). Ferrite has been shown to affect the resistance to Cl-SCC in boiling 42% $MgCl_2$, with the resistance reaching a maximum at 40% and decreasing on either side of this value (Ref 24, 25).

Welding of duplex stainless steels may have a significant effect on the resistance to Cl-SCC. Welding can destroy the duplex structure of an annealed material by forming continuous regions of ferrite along the weld HAZ. In addition, welding can result in the formation of carbonitrides, α′, and precipitate-free zones, which contribute to the susceptibility to SCC.

References cited in this section

2. G.H. Koch, Stress-Corrosion Cracking and Hydrogen Embrittlement, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 483–506

20. A. Sendriks, SCC of Stainless Steels, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992

21. J.F. Grubb and J.R. Maurer, “Use of Cathodic Protection With Superferritic Stainless Steels,” Paper 28, presented at Corrosion/84, National Association of Corrosion Engineers, 1984

22. L.S. Redmerski, J.J. Eckenrod, K.E. Pinnow, and C.W. Kovach, “Cathodic Protection of Seawater Cooled Power Plant Condensors Operating With High Performance Ferritic Stainless Steel Tubing,” Paper 208, presented at Corrosion/85, National Association of Corrosion Engineers, 1985

23. H. Spaehn, *Environment-Induced Cracking of Metals*, R.P. Gangloff and M.B. Ives, Ed., National Association of Corrosion Engineers, 1990, p 449

24. J.E. Truman, *Int. Met. Rev.*, Vol 26 (No. 6), 1981, p 301

25. S. Shimodaira, et al., *Stress Corrosion Cracking and Hydrogen Embrittlement of Iron Base Alloys*, R.W. Staehle, et al., Ed., National Association of Corrosion Engineers, 1977, p 1003

Martensitic and Precipitation- Hardening Stainless Steels

Higher-strength martensitic and precipitation-hardening stainless steels are subject to SCC and, to a greater degree, hydrogen stress cracking. Actually, failure of these alloys under conditions that appear conducive to SCC is so similar to failure caused by hydrogen damage that it is often questionable whether distinguishing between the two mechanisms is meaningful. These steels are susceptible to SCC in aqueous chloride environments with or without sulfides present. The mechanism of SCC in aqueous chloride environments such as marine environments is generally accepted to be hydrogen embrittlement. Martensitic and precipitation-hardening alloys are susceptible to hydrogen embrittlement in marine environments at yield strengths above 1035 MPa (150 ksi).

Susceptibility to failure is determined primarily by metallurgical structure. As expected, increasing the yield strength of martensitic and precipitation-hardening stainless steels increases the probability of cracking. The environments that cause failure are not specific; almost any corrosive environment capable of causing hydrogen evolution can also cause cracking. Even as mild an environment as fresh water at room temperature may cause failure in especially susceptible alloys.

Case Studies

Example 12: SCC Failure of a Sensitized Valve Stem. A 9 cm (3.5 in.) diameter valve stem made of 17-4 PH (AISI type 630) stainless steel, which was used for operating a 61 cm (24 in.) gate valve in a steam power plant, failed after approximately four months of service, during which it had been exposed to high-purity water at approximately 175 °C (350 °F) and 11 MPa (1600 psi). The valve had been subjected to 25 hot cycles to 300 °C (575 °F) and 12.6 MPa (1825 psi) by the manufacturer and to occasional operating tests at 260 °C (500 °F) during service. Valve stems of 17-4 PH stainless steel had previously given satisfactory performance in similar service conditions. The bottom of the stem, from which the valve gate was suspended, was in the shape of a T.

The valve stem was forged from 15 cm (6 in.) square billet stock and was reported to have been solution heat treated at 1040 ± 14 °C (1900 ± 25 °F) for 30 min and either air quenched or oil quenched to room temperature. Following rough machining and magnetic-particle, liquid penetrant, and ultrasonic inspection, the stem was reportedly aged at a temperature from 550 to 595 °C (1025 to 1100 °F) for 4 h and then final machined. The final hardness to be expected from this treatment was less than 35 HRC. Final inspection was done by the liquid penetrant method.

Investigation. Each of the two surfaces of the fracture showed a semicircular stain (Fig. 31a) in a large area where fracture had propagated along prior-austenite grain boundaries without deformation. The stains remained even after ultrasonic cleaning in a solution of detergent in hot water.

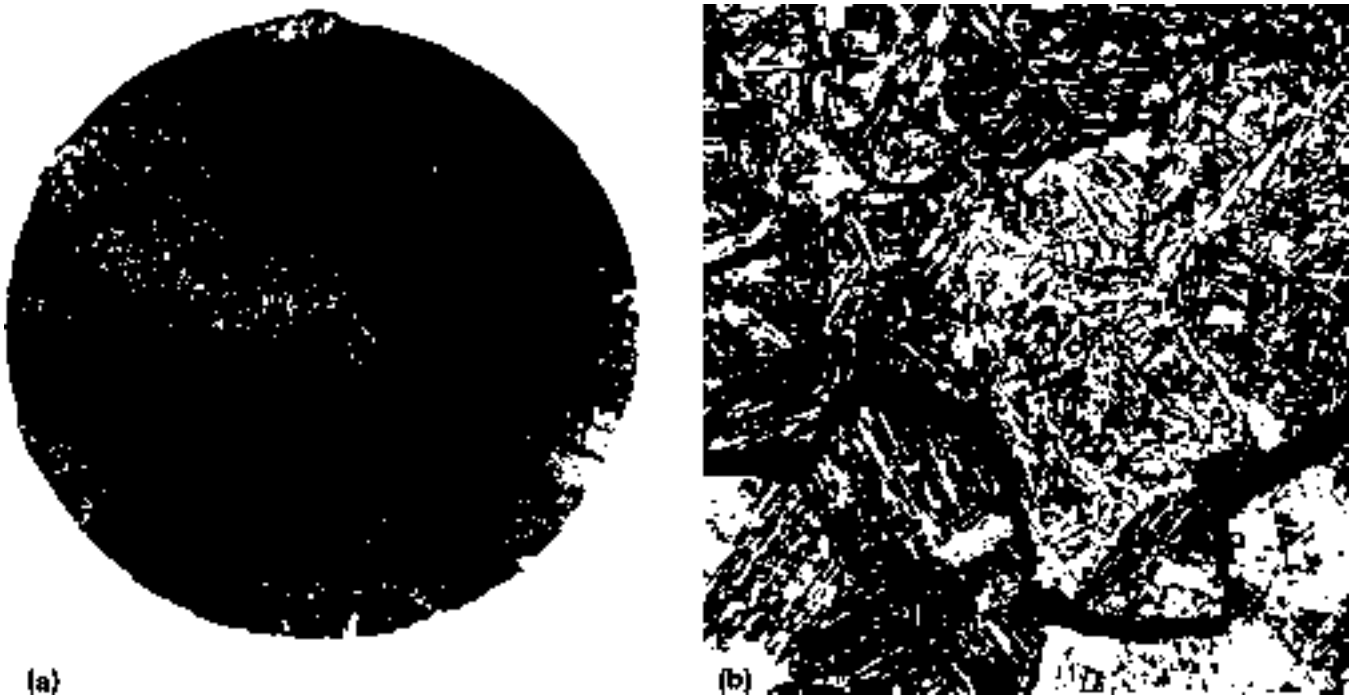


Fig. 31 Power plant gate-valve stem of 17-4 PH stainless that failed by SCC in high-purity water. (a) A fracture surface of the valve stem showing stained area and cup-and-cone shearing at perimeter. 0.7×. (b) Micrograph showing secondary intergranular cracks branching from fracture surface. 50×

Also shown in Fig. 31(a) is a narrow cup-and-cone shear-type area of failure that extends almost around the entire circumference of the fracture surface. This portion of the fracture was transgranular with some deformation, but not enough to affect outside-diameter measurements 6.4 mm ($\frac{1}{4}$ in.) away. In one narrow region of the periphery, intergranular fracture extended to the surface with no cup-and-cone shearing; intergranular SCC began here. This point also coincided with the junction of the stem with a tapered fillet at the T-head, where the stress could have been slightly elevated.

Metallographic examination of the structure adjacent to the fracture revealed a pattern of typical stress-corrosion cracks that radiated from the fracture surface along the grain boundaries (Fig. 31b). A macroetched cross section displayed a region of coarse-grain material (ASTM 2) surrounded by a structure of finer and more uniform grain size that extended to the surface. The coarse grain size was abnormal, suggesting that the solution heat treatment that should have followed forging may have been either omitted or carried out at too high a temperature. The matrix was found to contain small areas of ferrite, which in turn contained fine precipitate. This type of structure has been found in specimens that have been slowly (furnace) cooled through the range of 455 to 400 °C (850 to 750 °F) after aging.

Brinell hardness readings along the length of the stem indicated a uniform hardness throughout. Hardness readings taken from a cross section of the stem 19 mm ($\frac{3}{4}$ in.) from the fracture surface averaged 42 HRC.

Because the hardness seemed too high for the reported aging temperature, a reheat treatment was undertaken as a double check. A portion of the failed stem was re-solution-treated at 1040 °C (1900 °F) for 1 h, oil quenched, then sectioned to permit aging at a variety of times and temperatures. The hardness values resulting from these aging treatments were:

Aging treatment	Hardness, HRC
1 h at 480 °C (900 °F), air cool	**44**
4 h at 495 °C (925 °F), air cool	**41**
4 h at 525 °C (975 °F), air cool	**39**
1 h at 550 °C (1025 °F), air cool	**39**
2 h at 550 °C (1025 °F), air cool	**37**
4 h at 550 °C (1025 °F), air cool	**35**

These results indicated that the stem had been aged at 495 °C (925 °F) or less, rather than at the reported 550 to 595 °C (1025 to 1100 °F).

Conclusions. Failure was by progressive SCC that originated at a stress concentration, followed by ultimate shear of the perimeter when the effective stress exceeded the strength of the structure. Heat treatment had been improper. The solution heat treatment had been either omitted or at too high of a temperature, resulting in a coarse grain size. The aging treatment had been at too low of a temperature, creating a structure having excessive hardness, inadequate ductility, and unfavorable sensitization to SCC.

Recommendations. To prevent failures of this type in other 17-4 PH stainless steel valve stems, the following heat treatments were recommended: after forging, solution heat treat at 1040 °C (1900 °F) for 1 h, then oil quench; to avoid susceptibility to SCC, age at 595 °C (1100 °F) for 4 h, then air cool.

Example 13: Aircraft Attachment Bolt Failure. During a routine inspection on an aircraft assembly line, an airframe attachment bolt was found to be broken. The bolt was one of 12 that attach the lower outboard longeron to the wing carry-through structure. Failure occurred on the right-hand forward bolt in this longeron splice attachment. The bolt was fabricated from PH13-8Mo stainless steel heat treated to have an ultimate tensile strength of 1517 to 1655 MPa (220 to 240 ksi). A water-soluble coolant was used in drilling the bolt hole where this fastener was inserted.

Investigation. Figure 32(a) shows a bolt shank with corrosion evident on surfaces in contact with the splice joint, while Fig. 32(b) shows the fracture and initiation site of the bolt failure. Surface pitting on the bolt shank and subsequent corrosion cracking are shown. Scanning electron microscopy (SEM) examination of the fracture revealed the topography to be intergranular at the initiation site and to a depth of 8.4 mm (0.33 in.) (Fig. 32c). The remainder of the fracture showed a mixed topography of cleavage and ductile dimples (Fig. 32d).

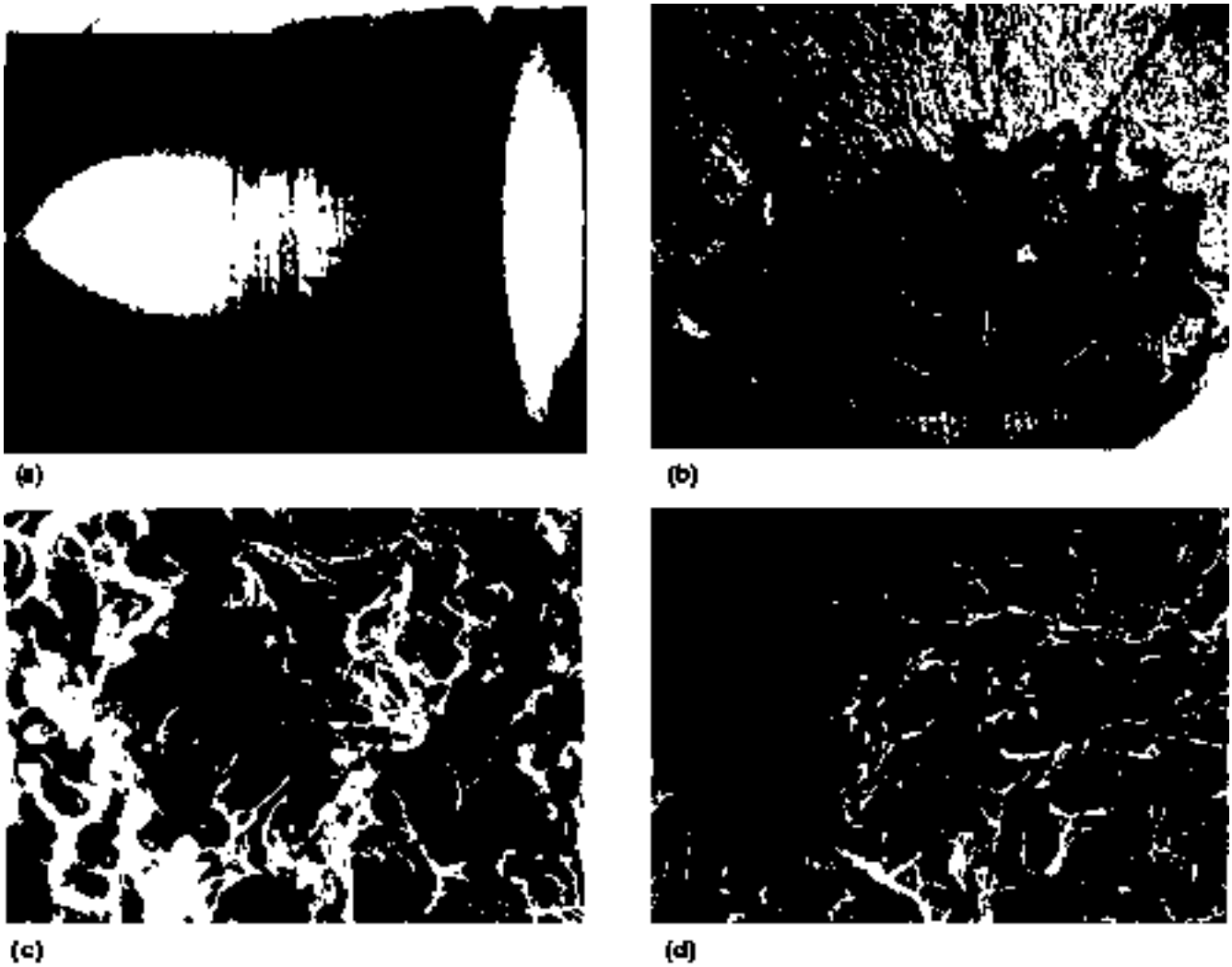


Fig. 32 (a) Parallel lines of corrosion on the shank of a PH13-8Mo stainless steel aircraft attachment bolt. (b) Close-up of fracture surface of bolt showing corroded area. Arrows point to one possible crack arrest line. (c) Scanning electron microscopy fractograph of area B in (b). Note corrosion product (left) and ductile dimples in the center. 265×. (d) Scanning electron microscopy fractograph of area C in (b). Area of fast fracture shows cleavage and dimples. 265×

Examination of corrosion products on the fracture by Auger emission spectroscopy and secondary imaging spectroscopy revealed the presence of elements typically found in tap water. Chemical analysis of the bolt material showed the composition to be within specification limits for PH13-8Mo stainless steel. Rockwell C hardness measurements taken on the bolt produced values ranging from 47 to 48 HRC, which would correspond to the specified ultimate tensile strength of 1517 to 1655 MPa (220 to 240 ksi).

Conclusion. It was concluded from the study that failure of the attachment bolt was caused by stress corrosion. The source of the corrosive media was the water-soluble coolant used in boring the bolt holes.

Recommendations consisted of inspecting for corrosion all the bolts that were installed using the water-soluble coolant at the spliced joint areas, rinsing all machined bolt holes with a noncorrosive agent, and installing new PH13-8Mo stainless steel bolts with a polysulfide wet sealant.

Nickel-Base Alloys

Nickel-base alloys are often used because of their high resistance to corrosion and SCC in various environments. However, SCC can occur under specific combinations of environment, microstructure, and applied stress. The number of environments that have been recognized over the years to cause SCC in nickel-base alloys are listed in Table 5.

Table 5 Evolution of environments that have been determined to cause stress- corrosion cracking (SCC) of nickel and nickel-base alloys

Period	Environment
ca. 1950	**H_2O-HF** **Fluosilicic acid**
ca. 1978	**H_2O-Cl^-** **H_2O-OH^-** **Polythionic acid** **Chromic acid** **Acetic acid** **Molten caustics** **Steam** **Organic liquids** **Liquid Li, Hg, Sn, Zn, Bi, Pb** **H_2O** **H_2**
Late 1970s to present	**H_2O-Cl^--CO_2-H_2S-S_8** **H_2O-Br^-** **H_2O-I^-** **Thiocyanates, tetrathionate, thiosulfate** **Sulfurous acid**

Source: Ref 26

SCC Environments

Cl-SCC of Nickel Alloy. Even though there is data to indicate that Fe-Ni-Cr alloys containing more than approximately 35% Ni will not crack in chloride-containing environments, other studies have shown that cracking can occur under more severe conditions. Figure 33 shows some results for various alloys when tested in autoclaves in a deaerated 20.4% magnesium chloride solution. The temperature for the onset of SCC increases with increasing nickel content, and Alloy 200 and Alloy 600 were immune at the highest temperature tested. The same alloys were tested in sodium chloride and calcium chloride solutions with the same chloride ion concentration; the magnesium chloride solution was the most severe, with sodium and calcium following in that order. The results were summarized as:

- Alloys B, B-2, 600, and 200 showed no SCC.
- Alloys C-276 and C-4 were susceptible in magnesium chloride solution at 232 °C (450 °F).
- Alloy G-3 was susceptible in both sodium and magnesium chloride solution at this temperature.
- Alloys 825, 718, and 800 exhibited SCC only in these two solutions down to 175 °C (350 °F).
- Alloy 28, M-352, L, and 20Cb-3 exhibited the lowest resistance to SCC of the alloys studied.

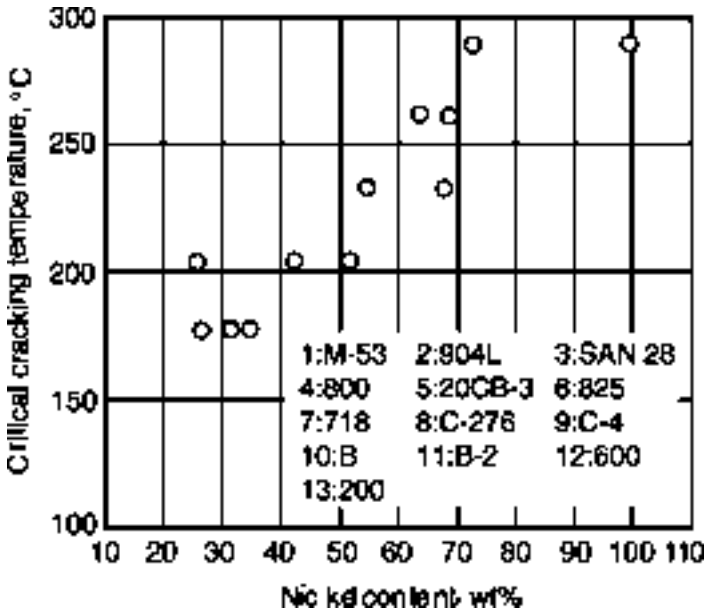


Fig. 33 Effect of nickel on SCC in 20.4% $MgCl_2$ deaerated with nitrogen. Source: Ref 26

Caustic SCC. The nickel-base alloys 600 and 800 are susceptible to caustic SCC at very high temperatures in the sensitized condition. Most of the testing has been done at temperatures above 288 °C (550 °F). A stabilization heat treatment improves resistance, while the presence of sulfate, carbonate, or phosphate ions makes the SCC worse. Alloy 690, with higher chromium and lower carbon contents than 600, is much more resistive, and Alloy 200, commercially pure nickel, is reported to be immune under all conditions encountered in the industrial use of caustic solutions. Alloys 825 and C-276, which also contain Mo, were found to be susceptible in concentrated solutions such as 50% NaOH above 150 °C (300 °F).

Polythionic Acid SCC. Intergranular SCC has been observed in sensitized Alloys 600 and 800 in polythionic acid, sodium tetrathionate, and sodium thiosulfate solutions. Sensitized Alloy 600 samples failed in less than 10 h and at stresses as low as 70 MPa (10 ksi). Research has shown that the grain boundary chromium content in sensitized Alloy 600 can be below 5%, and, when it is, 100% intergranular SCC will occur in tetrathionate solutions. Levels of 8.3% and higher were found to be immune to intergranular fracture. Depending on the degree of sensitization, extremely dilute solutions can cause SCC. Stabilized versions of these alloys and higher-chromium alloys may not be susceptible to this form of SCC.

High-Temperature Water SCC. While boiling water nuclear reactors had their problems with stainless steel piping cracking, the pressurized-water reactors had a different problem of extreme severity. This problem was cracking of the tubes in the steam generators, which are very large, vertical, shell and tube heat exchangers. Again the material was Alloy 600, but in this case it was susceptible to intergranular SCC when in the mill annealed condition. The environment was deaerated high-temperature essentially pure water. Severity increases

with temperature and with trace impurities in the water. From the stress standpoint, most failures have occurred in highly strained areas, such as U-bends, and expanded regions where the tubes were rolled into the tube sheet. *Mill annealed* is a generic term, and different tube manufacturers use different procedures. The most susceptible microstructures were produced by treatments at temperatures less than 950 °C (1740 °F). The resulting structure was very fine grained with relatively heavy intragranular carbide precipitation and almost no grain-boundary carbides. A variety of thermal and mechanical treatments and design modifications have been applied to mitigate the problem, but a number of utilities have had to replace the steam generators at considerable expense.

Another intergranular SCC problem that has occurred in the nuclear power industry is the cracking of the high-strength nickel-base alloys X-750 and 718. These alloys are used for a variety of springs and fasteners in the reactors. Investigations found that crack susceptibility was associated with one particular heat treatment cycle. By changing out the components and replacing them with ones heat treated by a different thermal cycle, the problem has been eliminated.

Case Studies of Nickel Alloys

Example 14: SCC of an Inconel 600 Safe- End on a Reactor Nozzle. Cracking occurred in an ASME SB166 Inconel 600 safe-end forging on a nuclear reactor coolant water recirculation nozzle while it was in service. The safe-end was welded to a stainless-steel-clad carbon steel nozzle and a type 316 stainless steel transition metal pipe segment (Fig. 34). An Inconel 600 thermal sleeve was welded to the safe-end, and a repair weld had obviously been made on the outside surface of the safe-end to correct a machining error. Initial visual examination of the safe-end disclosed that the cracking extended over approximately 85° of the circular circumference of the piece.

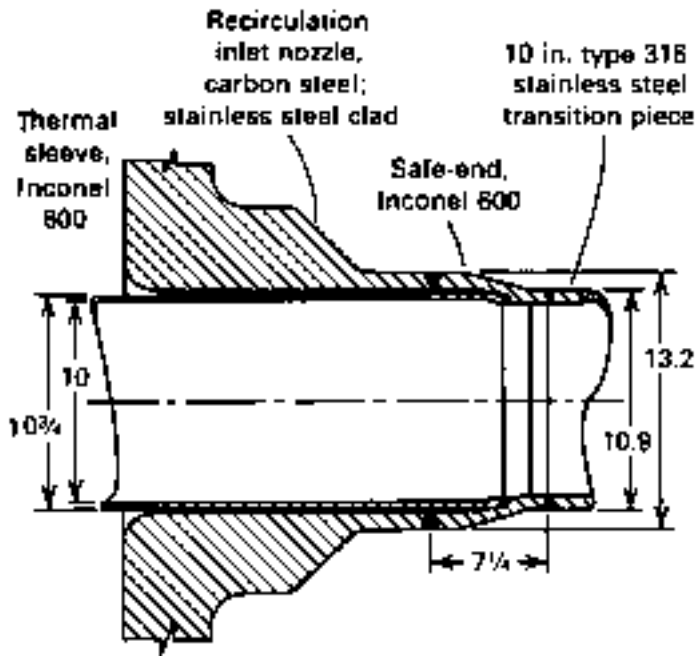


Fig. 34 Cross section through recirculation inlet nozzle of reactor vessel. Shown are the nozzle, the safe-end that failed, and the thermal sleeve that created susceptibility to crevice corrosion. Dimensions given in inches

Investigation. After the safe-end had been cut from the reactor, it was radiographically inspected on-site to confirm the extent of cracking. This inspection indicated major cracking over approximately 280° of the circumference. Limited ultrasonic inspection performed in the laboratory with a 1.5 MHz transducer, directing the incident sound beam from the large end of the safe-end toward the small end, demonstrated that the cracking could be detected nondestructively over approximately the same region that was indicated radiographically.

Eleven longitudinal cross sections were prepared for macroscopic examination, including locations near each end of the through-wall portion of the crack and sections selected on the basis of radiographic crack indications. Major cracking was found in all sections examined. At all locations, the crack originated near the tip of the crevice between the thermal sleeve and the safe-end, adjacent to the attachment weld for the safe-end, and extended outward through the safe-end wall. Over approximately 35° of the circumference, cracking propagated through the repair weld, while over a different 50° of the circumference, the propagation path was entirely within the safe-end base metal, penetrating the outside surface adjacent to the fusion line of the repair weld.

Metallographic examination of selected cross sections showed the effects of repeated thermomechanical cycling during welding on the precipitate structure in the HAZ of the Inconel 600. There were very few matrix precipitates immediately adjacent to the fusion line of the heat shield attachment weld; this condition was true of the entire HAZ near the root pass. The cracking generally initiated in the HAZ some distance away from the tip of the tight crevice and, in its early stages, generally followed a path parallel to the fusion line (Fig. 35). Depending on the location, the crack could be found in regions of heavy or light carbide precipitation. Metallographic cross sections and SEM studies of the fracture surface were in agreement: the cracking was all intergranular in the base metal and intercolumnar where it penetrated the repair weld metal. No secondary cracking was found, nor was any discontinuity found that might have initiated the cracking.

Fig. 35 Polished-and-etched cross section through the failed safe-end of the recirculation inlet nozzle. Note the intergranular nature of the failure. Etched in 8:1 phosphoric acid. 53×

Microanalyses disclosed the presence of sulfur at locations corresponding to the later stages of cracking, but not on the crevice surface between the safe-end and the heat shield. Chlorine was present at one point on the crevice surface. Other deposits on the crevice surface and the crack surface near the crevice were rich in iron.

Discussion. The literature indicated that Inconel 600 is susceptible to intergranular stress corrosion in hot, high-purity water, but it is only likely at stresses above the yield stress of the material. The analysis by the reactor vendor showed the primary bending plus membrane stresses due to service loading to be 78% of the yield

strength. Further butt welds in austenitic piping materials can leave residual stresses of the order of the yield strength; therefore, the mechanism is reasonable. Whether or not the material is in a sensitized condition does not greatly influence susceptibility to cracking, and cracks propagated in the lightly sensitized region adjacent to the fusion line and in the heavily sensitized regions some distance from the fusion line. Therefore the furnace-sensitized condition of the material was not considered to be a significant contributing factor in the cracking problem.

The role, if any, of the observed contaminants could not be clearly identified from this study, although there is a suspicion that they contributed to the failure. Further surface analysis would be required to clarify this point.

The role of the tight crevice between the safe-end and the thermal sleeve thus comes into question. It was considered unlikely that a classical differential aeration cell could develop in the crevice in the uniform high-purity water environment, but such regions could entrap air during outages, causing locally high oxygen levels upon startup. The presence of anionic contaminants could cause acidification of the crevice region, which might also enhance stress-corrosion mechanisms. The crevice is considered to be the principal contributing factor in this case.

Conclusion. These observations led to the conclusion that the cracking mechanism was intergranular SCC.

Example 15: SCC of Alloy X-750 Jet Pump Beams. Jet pumps, which have no moving parts, provide a continuous circulation path for a major portion of the core coolant flow in a boiling water reactor. Part of the pump is held in place by a beam-and-bolt assembly, wherein the beam is preloaded by the bolt. The Alloy X-750 beams had been heat treated by heating at 885 °C (1625 °F) for 24 h and aging at 705 °C (1300 °F) for 20 h. Jet pump beams were found to have failed in two nuclear reactors, and other beams were found to be cracked.

Investigation. Metallurgical examinations of the failed beams revealed that intergranular SCC under sustained bending loading was responsible for the failure. The location of the cracking was consistent with the results of stress analysis of the part.

Analysis also showed that the stress level could be reduced and a significant extension in life obtained by reducing the preload on the beams.

Simulated service testing in oxygenated 288 °C (550 °F) water showed that X-750 was susceptible to SCC in this environment. Extensive research and testing revealed that high-temperature annealing at 1095 °C (2000 °F) for 1 h could reduce the susceptibility of the alloy to SCC in the reactor environment.

Discussion. Alloy X-750, as originally heat treated and at the stress levels imposed in service, was susceptible to intergranular SCC in the high-temperature oxygenated-water environment existing in the boiling water reactor. By changing the heat treatment and lowering the stress, the probability of crack initiation could be extended to more than 40 years.

Recommendations. It was recommended that all utilities operating boiling water reactors address the problem by beam replacement, reheat treatment, or preload reduction. Since this was done, additional cracking of jet pump beams is no longer a concern.

Reference cited in this section

26. N. Sridhar and G. Cragnolino, Stress Corrosion Cracking of Nickel-Base Alloys, *Stress-Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992

Stress-Corrosion Cracking

Revised by W. R. Warke

Aluminum Alloys

Stress-corrosion cracking occurs in high-strength aluminum alloys in ordinary atmospheric and aqueous environments. The relative ratings of resistance to SCC for various alloys and tempers are published by the

Aluminum Association in *Aluminum Standards and Data*, and more detailed guidelines for comparing the high-strength alloys, tempers, and product forms with respect to stressing direction are presented in *MIL-Handbook-5*.

SCC Factors

Although proper selection of both material and structural design and regular inspection in service normally prevent such failures from operating stresses, there are some factors that may be overlooked. These situations frequently result from poor fitup during assembly of structural components or from residual stresses introduced into individual components during such operations as cold forming, welding, and quenching or straightening after heat treatment. Both initiation and propagation of cracking are accelerated by moisture, temperature, chlorides, and other industrial contaminants.

Alloys and Heat Treatment. High-purity and commercially pure aluminum and the relatively low-strength aluminum alloys are not susceptible to SCC. Failures by this mechanism are chiefly associated with heat treated wrought products of the higher-strength alloys used in load-carrying structures, such as aluminum-copper, aluminum-zinc-magnesium, aluminum-zinc-magnesium-copper, and aluminum-magnesium (3% or more Mg) alloys. With these alloys, processing and heat treatment must be controlled to ensure high resistance to SCC. The relative resistance to SCC of various wrought aluminum alloys in relation to strength is given in Table 6.

Table 6 Ratings of resistance to stress-corrosion cracking (SCC) of wrought commercial aluminum alloys

Alloy series	**Type of alloy**	**Strengthening method**	**Tensile strength**		SCC rating[(a)]
			MPa	**ksi**	
1*xxx*	Al	Cold working	69–172	10–25	**A**
2*xxx*	Al-Cu-Mg (1–2.5Cu)	Heat treatment	172–310	25–45	**A**
2*xxx*	Al-Cu-Mg-Si (3–6Cu)	Heat treatment	379–517	55–75	**B**
3*xxx*	Al-Mn-Mg	Cold working	138–276	20–40	**A**
5*xxx*	Al-Mg (1–2.5Mg)	Cold working	138–290	20–42	**A**
5*xxx*	Al-Mg-Mn (3–6Mg)	Cold working	290–379	42–55	**B**
6*xxx*	Al-Mg-Si	Heat treatment	152–379	22–55	**A**
7*xxx*	Al-Zn-Mg	Heat treatment	379–503	55–73	**B**
7*xxx*	**Al-Zn-Mg-Cu**	**Heat treatment**	**517–620**	**75–90**	**B**

(a) A, no instance of SCC in service or in laboratory tests; B, SCC has occurred in service with certain alloys and tempers. Service failures can be avoided by careful design and assembly and proper selection of alloy and temper.

For thick wrought sections, special precipitation heat treatments, such as those that produce T7 tempers in aluminum-zinc-magnesium-copper alloys and T8 tempers in aluminum-copper alloys, have been developed to provide relatively high resistance to SCC in the more critical short-transverse stressing direction. Evaluation of the resistance of improved alloys and tempers in accelerated corrosion tests, however, can be markedly influenced by test procedures, and reliance should therefore be placed mainly on exposure to outdoor atmospheres or other service environments.

Direction of Stressing. Cast products are isotropic with regard to cracking, and directionality is generally not important in the performance of sheet (except in structural applications, such as the stressed skins of aircraft wings), but resistance to cracking of wrought products can vary markedly with direction of stressing. Service failures of extrusions, rolled plate, and forgings have been caused chiefly by tensile stresses acting in the short-transverse direction relative to the grain structure. It is important to consider the direction of stress relative to the grain structure when designing structural components and performing failure analyses. Exceptions may occur in thin forged sections that recrystallize during solution heat treatment and have low directionality.

Effects of Temperature. The resistance to SCC of the non-heat-treatable aluminum-magnesium alloys and the naturally aged (T3 and T4 tempers) heat treated aluminum-copper alloys can be adversely affected in structures subjected to elevated temperatures. The effects depend on the specific alloy and temper, on temperature, and on time at temperature.

Case Studies of Aluminum Alloys

Example 16: SCC of Aluminum Alloy Fittings in a Marine Atmosphere. During a routine inspection, cracks were discovered in several aluminum alloy coupling nuts (Fig. 36a) on the fuel lines of a missile. The fuel lines had been exposed to a marine atmosphere for six months while the missile stood on an outdoor test stand near the seacoast. A complete check was then made, both visually and with the aid of a low-power magnifying glass, of all coupling nuts of this type on the missile.

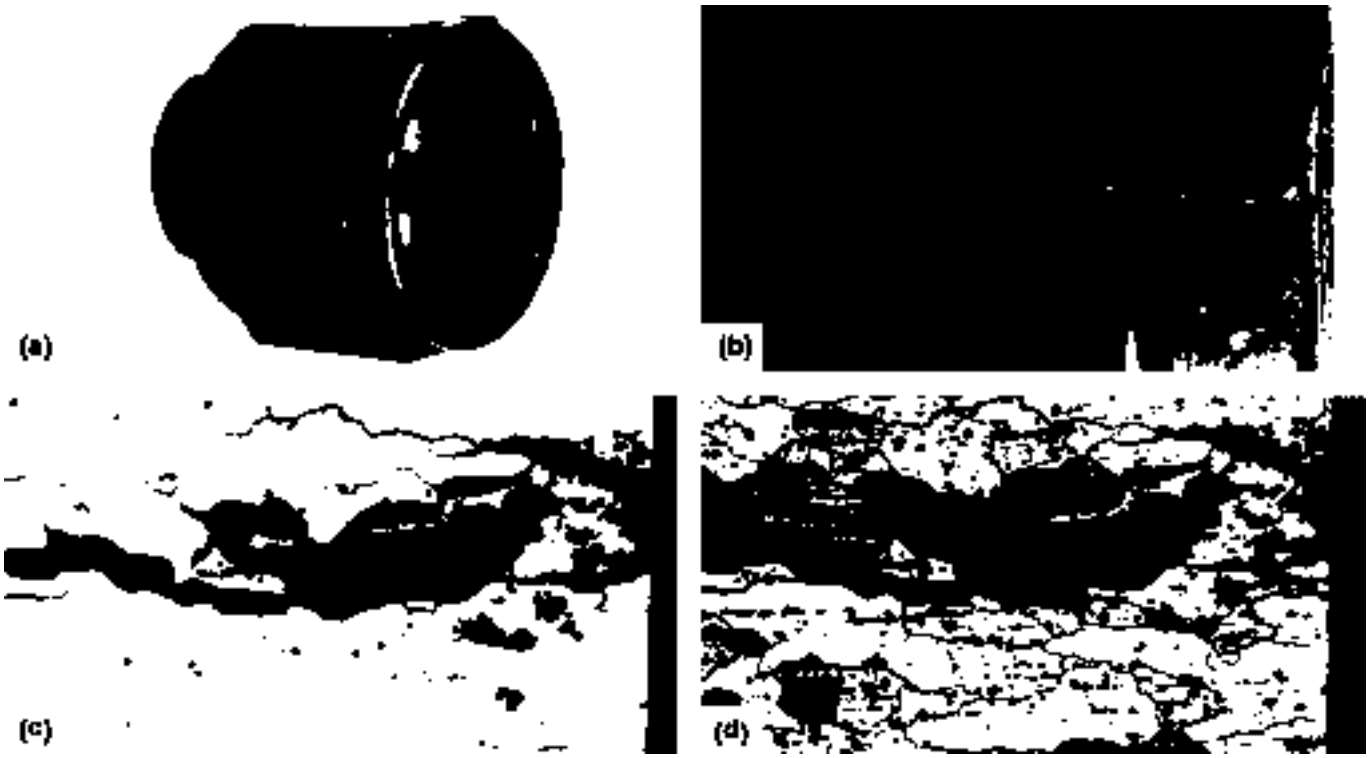


Fig. 36 Aluminum alloy coupling nut that cracked by stress corrosion in a marine atmosphere. (a) Overall view of coupling nut. (b) View of the crack. 6×. (c) and (d) Micrographs of a section through the crack near the origin, showing appearance before and after etching. Both 100×

One to three cracks were found in each of 13 nuts; two of these nuts had been supplied by one vendor, and the other 11 had been supplied by another vendor. There were no leaks at any of the cracked nuts, and further checking using liquid penetrant techniques did not reveal any additional cracks. Subsequent examination of the remaining missiles on test stands at this site showed that some coupling nuts of this type on each of the missiles that had been exposed to the atmosphere for more than about a month had cracked in a similar manner.

Investigation. The 13 cracked coupling nuts, which were on 6.4 to 19 mm ($\frac{1}{4}$ to $\frac{3}{4}$ in.) outside-diameter fuel lines, were removed for further inspection. The nuts had been anodized then dyed for identification.

Visual and low-magnification examination showed that the cracks had originated in the round section at the threaded end and had propagated parallel to the axis of the nut across one of the flat surfaces. An actual-size view of one of the cracked nuts and a view of the crack at 6× are shown in Fig. 36(a) and (b), respectively.

Spectrographic analysis of the 13 cracked nuts identified the material as aluminum alloy 2014 or 2017 (the composition ranges of these two alloys overlap, which prevented exact identification by spectrographic analysis). Chemical analysis of a white deposit removed from one of the cracks showed the presence of a high concentration of chloride.

Metallographic examination of mounted specimens in the polished and polished-and-etched conditions revealed that the cracks, primary and secondary intergranular cracks showing intergranular corrosion, were similar to stress-corrosion cracks observed previously in aluminum alloys of this type. A cross section taken through the failure region of one nut, near the origin of the crack, is shown in Fig. 36(c) as it appeared before etching and in

Fig. 36(d) after etching. The microstructure and longitudinal (axial) grain orientation identified the material of the nuts as bar stock.

Discussion. Tightening of the nuts resulted in sustained tensile stresses in the material. These were in the form of hoop stresses, which are higher than those encountered in most applications of this type of coupling nut because the parts had to be liquid oxygen clean (completely free of lubricants or organic materials); therefore, the nuts had to be tightened to high torque in order to obtain liquid-tight seals.

Conclusions. Cracking of the aluminum alloy coupling nuts was caused by stress corrosion. Contributing factors included use of a material (aluminum alloy 2014 or 2017) that is susceptible to this type of failure, sustained tensile stressing in the presence of a marine (chloride-bearing) atmosphere, and an elongated grain structure transverse to the direction of stress. The elongated grain structure transverse to the direction of stress was a consequence of following the generally used procedure of machining this type of nut from bar stock.

Corrective Measures. The materials specification for new coupling nuts for this application was changed to permit use of only aluminum alloys 6061-T6 and -T651 and 2024-T6, -T62, and -T851 (for identification purposes, the anodic coating on coupling nuts of the new materials was dyed a color different from that used on the aluminum alloy 2014 or 2017 nuts). Alloy 6061 is not susceptible to SCC, and alloy 2024 in the tempers selected has been shown to withstand much higher transverse tensile stresses than alloy 2014-T6 in alternate-immersion tests in 3.5% sodium chloride solution. These tests have shown excellent correlation with exposure to marine atmospheres.

Example 17: SCC of Aircraft Hinge Brackets. Forged aluminum alloy 2014-T6 hinge brackets in naval aircraft rudder and aileron linkages were found cracked in service. The cracks were in the hinge lugs, adjacent to a bushing made of cadmium-plated 4130 steel.

Investigation. Examination of the microstructure in a lug of a typical failed hinge bracket revealed intergranular corrosion and cracking in the region of the cracks discovered in service. The fracture was neither located along the flash line nor oriented in the short-transverse direction, both of which are characteristic of most stress-corrosion failures of forgings.

It was evident from the interference fit of the bushing in the hole of the lug that a hoop stress of sustained tension existed in the corroded area. The overall condition of the paint film on the hinge bracket was good, but the paint film on the lug was chipped. Figure 37(a) shows the crack in the lug (arrow), and Fig. 37(b) shows branched secondary cracking adjacent to and parallel to the fracture surface.

Fig. 37 Aluminum alloy 2014-T6 hinge bracket that failed by SCC in service. (a) Hinge bracket. Actual size. Arrow indicates crack. (b) Micrograph showing secondary cracking adjacent and parallel to the fracture surface. Etched with Keller's reagent. 250×

Conclusions. Failure of the hinge brackets occurred by SCC. The corrosion was caused by exposure to a marine environment in the absence of paint in the stressed area. The stress resulted from the interference fit of the bushing in the lug hole.
Corrective measures taken to prevent further failures consisted of:

- All hinge brackets in service were inspected for cracks and for proper maintenance of paint.
- Aluminum alloy 7075-T6 was substituted for alloy 2014-T6 to provide greater strength and resistance to SCC. Surface treatment for the 7075-T6 brackets was sulfuric acid anodizing and dichromate sealing.
- The interference fit of the bushing in the lug hole was discontinued. The bushings, with a sliding fit, were cemented in place in the lug holes.

Example 18: SCC of a Forged Aircraft Lug. During a routine shear-pin check, the end lug on the barrel of the forward canopy actuator on a naval aircraft was found to have fractured. The lug was forged from aluminum alloy 2014-T6.
Investigation. As shown in Fig. 38(a), the lug had fractured in two places; the original crack occurred at the top, and the final fracture occurred at a crack on the left side of the lug. The surface of the original crack was flaky, with white deposits that appeared to be corrosion products. Apparently, the origin of the failure was a tiny region of pitting corrosion on one flat surface of the lug (back surface, Fig. 38a; arrow there shows location of pitting corrosion region). The opposite flat surface (front, Fig. 38a) had a shear lip at the fracture edge.

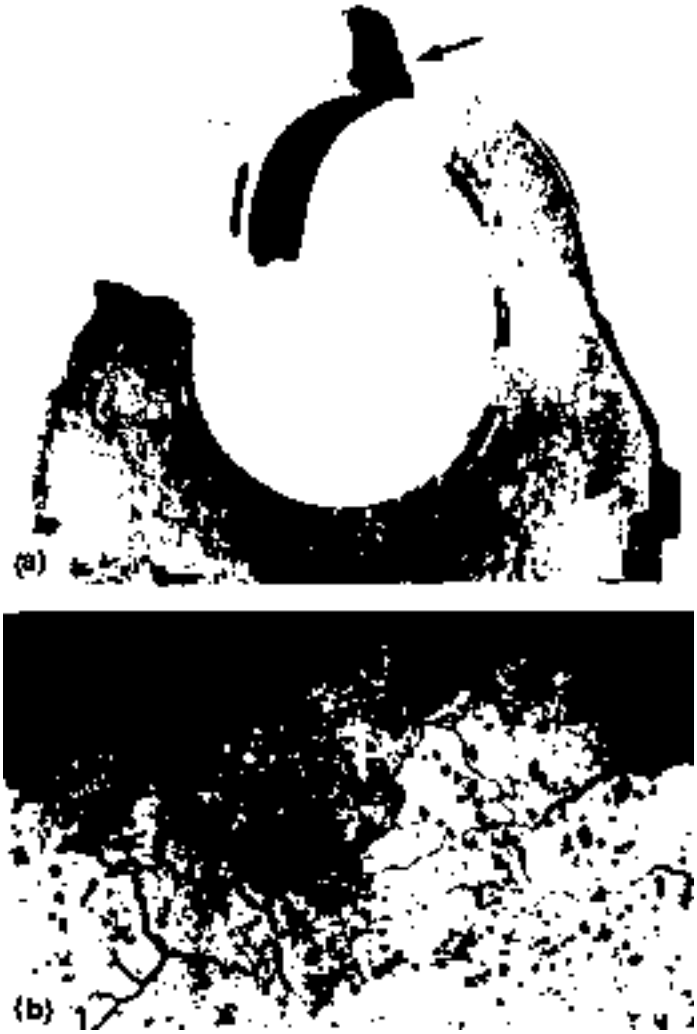


Fig. 38 Forged aluminum alloy 2014-T6 actuator barrel lug that failed by SCC. (a) View of the lug. 2×. Fracture at top was the initial fracture; arrow indicates location of a tiny region of pitting corrosion (on back side of lug) at which failure originated. Final fracture is at left. (b) Micrograph of an etched (Keller's reagent) section through surface of initial fracture showing branched cracking. 140×.

A metallographic section through the initial fracture displayed intergranular branched cracking that had originated at the fracture surface, which was also intergranular (Fig. 38b).

The bolt from the failed lug gave no indication of excessive loading and showed no evidence of deformation.

Conclusions. The cause of failure was SCC resulting from exposure to a marine environment. The fracture occurred in normal operation at a point where damage from pitting and intergranular corrosion acted as a stress raiser, not because of overload. The pitting and intergranular attack on the lug were evidence that the surface protection of the part had been inadequate as manufactured or had been damaged in service and not properly repaired in routine maintenance.

Recommendations. To prevent future failures of this type, the lug and barrel should be anodized in sulfuric acid and given a dichromate sealing treatment, followed by application of a coat of paint primer. During routine maintenance checks, a careful examination should be made for damage to the protective coating, and any necessary repairs should be made by cleaning, priming, and painting. Severely corroded parts should be removed from service.

Example 19: Cracking of Aluminum Alloy Aircraft Undercarriage Forgings. During nondestructive testing by fluorescent dye penetrant and ultrasonic testing, aircraft undercarriage leg forgings were found to be cracked. The forgings were made of 2014 aluminum alloy, and were approximately 1040 mm (41 in.) long and 150 mm (6 in.) in diameter. The cracks were located in the area of the upper attachment point of the undercarriage leg forgings.

Investigation. The major cracks were opened by saw cuts and impact blows. Figure 39 shows a typical opened crack, with one surface having been cleaned. The cracks originated an a region that was unprotected by a steel sleeve. Other forgings contained small cracks in the inside surface at the corresponding location.

Fig. 39 Opened crack in part of an aluminum alloy aircraft undercarriage forging. Note suggestion of beach marks in the fracture, indicating that, in later stages, crack propagation occurred in a series of steps. -0.04×

Metallographic examination showed that the microstructure was normal for the alloy. The small cracks were quite wide near the surface, presumably due to corrosion after crack inititiation (Fig. 40). In all cases examined, the cracks were intergranular and branched—typical of SCC in this alloy (Fig. 41).

Fig. 40 Widening of cracks from the diaphragm region of the part shown in Fig. 39. 10.4×

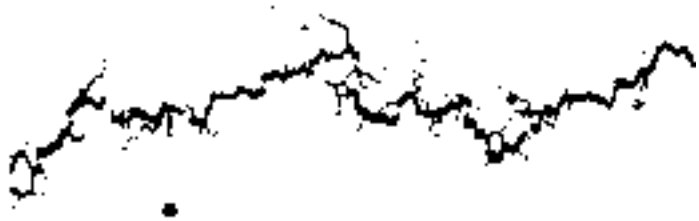


Fig. 41 Typical stress-corrosion crack in an aluminum alloy forging. 277×

A chemical analysis was performed on one of the forgings and it met specification. The hardness was 130 HB, which is indicative of the T6 condition.
Some material was re-heat-treated and, together with as-received material from several forgings, was subjected to SCC testing. There was no significant difference among the as-received and re-heat-treated specimens; all were found to be susceptible to SCC.
Conclusions. The cracks in the forgings were the result of SCC. The alloy was shown to be susceptible and the nature of the cracking was consistent with this failure mode.

Stress-Corrosion Cracking

Revised by W. R. Warke

Copper and Copper Alloys

Copper and copper alloys have excellent corrosion resistance in many industrial environments, in seawater, and in marine atmospheres, but are susceptible to SCC. The first documented case of SCC occurred almost 100 years ago and involved the cracking of brass cartridge cases. These failures took place during the rainy season in India when the British army had stored the cartridges in horse barns, thus leading to the term *season cracking*. The high humidity and ammonia vapors in the horse barns, coupled with the residual stresses from forming the cartridge cases, led to the cracking.

SCC Factors

Effect of Environment. Stress-corrosion cracking in copper alloys occurs most frequently in environments that contain ammonia or amines, in either aqueous solutions or moist atmospheres. Cracking occurs at room temperature and at stress levels as low as 1% of the tensile strength of the alloy. Copper alloys experience SCC in some important industrial environments, including those containing ammonia, citrate, tartrate, moist SO_2, sulfates, nitrates, nitrites, and even pure water.
In aqueous solutions, pH has a strong influence on susceptibility to cracking and on whether crack paths are intergranular or transgranular, as shown for brass in ammoniacal copper sulfate solution in Fig. 42. Cracking occurs most rapidly in nearly neutral solutions, where the crack path is intergranular. The crack path is transgranular in both alkaline and acidic solutions, and the alloy is highly resistant to cracking when pH is less than 4. In ammoniacal solutions, intergranular SCC is normally associated with tarnishing conditions. The intergranular crack facets are generally smooth and featureless. The transgranular facets, on the other hand, have a cleavagelike appearance and may show crack-arrest marks, which are consistent with discontinuous crack propagation. Failure of copper and copper alloys will occur at low stresses in aqueous solutions or moist atmospheres. Figure 43 plots the relation between time-to-fracture at room temperature and initial tensile stress for brass partly immersed in concentrated ammonium hydroxide (curve A), exposed to the vapor of concentrated ammonium hydroxide (curve B), and exposed to a gaseous mixture of ammonia, oxygen, carbon dioxide, and water vapor (curve C).

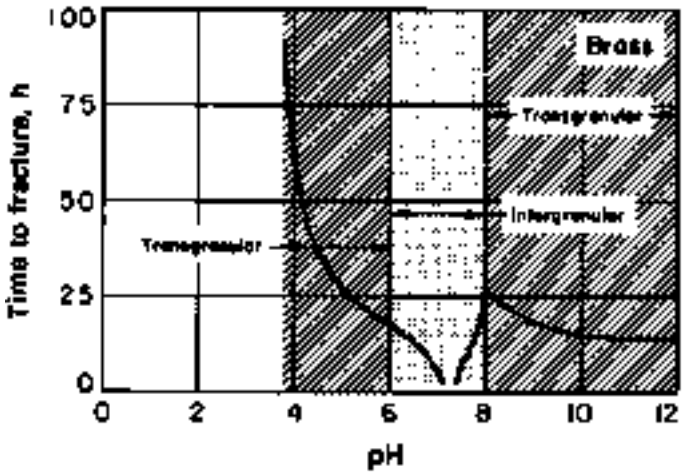


Fig. 42 Effect of pH on time-to-fracture by SCC. Data are for brass in ammoniacal copper sulfate solution at room temperature.

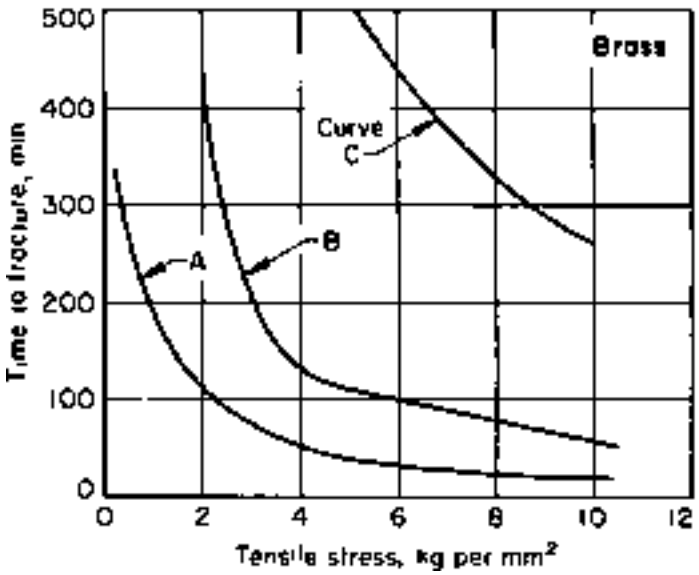


Fig. 43 Effect of initial tensile stress on time-to-fracture by SCC at room temperature of brass in three corrosive environments. Curve A, partly immersed in concentrated ammonium hydroxide; B, exposed to the vapor of concentrated ammonium hydroxide; C, exposed to a gaseous mixture of ammonia, oxygen, carbon dioxide, and water vapor

The presence of such oxidizing substances as dissolved oxygen and cupric, ferric, and nitrate ions accelerates SCC of copper alloys in ammoniacal aqueous solutions. In stress-corrosion tests in which brass was exposed for 20 days to atmospheres containing (by volume) 0.1 to 1% SO_2, the specimens cracked, but similar specimens that had been pretreated with the inhibitor benzotriazole were unaffected. Pretreatment of brass specimens with benzotriazole failed to prevent cracking in tests in which the stressed specimens were exposed for 72 h in an atmosphere containing (by volume) 2% ammonia and in an aged solution of ammonium hydroxide.

Effect of Alloying Elements. High-purity copper is almost immune to SCC in most environments and in the practical range of service stresses. However intergranular cracking of high-purity copper has been observed under some conditions, apparently as a result of segregation of trace impurities at the grain boundaries. Stress-corrosion cracking of pure copper in ammonia and in sodium nitrite has been documented.

The resistance of copper to SCC is greatly reduced by the presence of low concentrations of arsenic, phosphorus, antimony, and silicon as alloying elements. Time-to-fracture for copper containing various concentrations of these alloying elements when stressed at an applied tensile stress of 69 MPa (10 ksi) is plotted in Fig. 44. As the concentration of each alloying element is increased, time-to-failure at first decreases, reaching a minimum between about 0.1 and 1%, then increases.

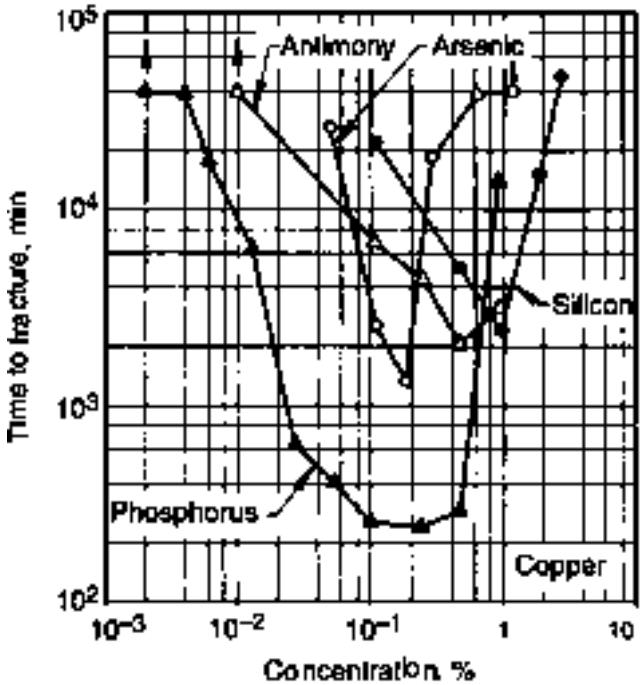


Fig. 44 Effect of presence of low concentration of arsenic, phosphorus, antimony, and silicon on time-to-fracture of copper by SCC under an applied tensile stress of 69 MPa (10 ksi) in a moist ammoniacal atmosphere. Composition of test atmosphere was 80% air, 16% ammonia, and 4% water vapor; temperature was 35 °C (95 °F), which was above the dew point.

Results of similar tests on copper containing higher concentrations of the alloying elements aluminum, nickel, tin, and zinc are plotted in Fig. 45. Except for zinc, a minimum time-to-fracture is reached between approximately 1 and 5%; time-to-failure for copper containing zinc does not show a minimum for concentrations of 0.5 to 40%, but decreases with increasing zinc content for the concentration range shown in Fig. 45. Note that the common 70-30, or cartridge brass, is in the range of very short time-to-failure. In fact, most of the research on SCC of copper alloys has been done on the ammonia cracking of brass.

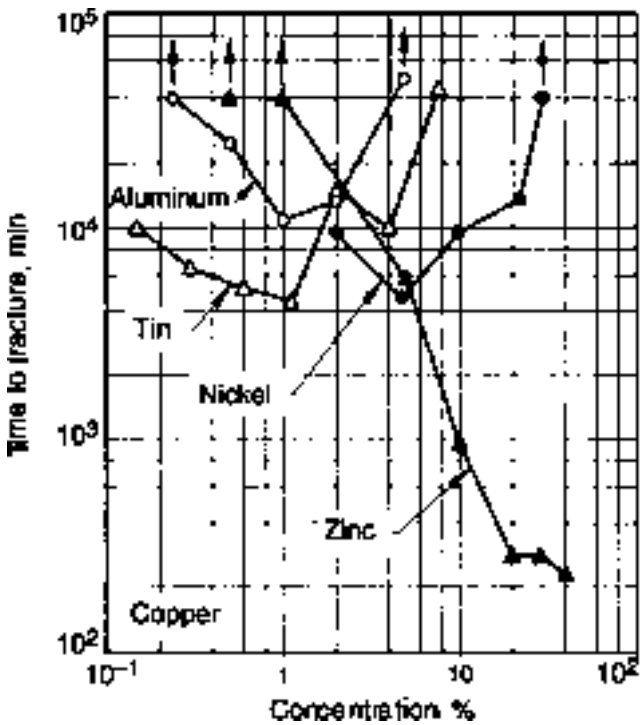


Fig. 45 Effect of presence of low-to-moderate concentrations of aluminum, nickel, tin, and zinc on time-to-fracture of copper by SCC under an applied tensile stress of 69 MPa (10 ksi) in a moist ammoniacal atmosphere. Composition of test atmosphere was 80% air, 16% ammonia, and 4% water vapor; temperature was 35 °C (95 °F), which was above the dew point.

The effect of zinc content on susceptibility of copper-zinc alloys to SCC when exposed to cooling tower water containing amines and dissolved oxygen is discussed in the subsequent case study. The alloys that failed in this application had nominal zinc contents of 25, 26, and 40%. Two replacement cast silicon bronzes that gave satisfactory service contained less than 5 and 1.5% Zn; a satisfactory replacement wrought silicon bronze contained less than 1.5% Zn.

All of the previous discussion has pertained to copper-base alloys having the face-centered cubic or α-structure. It has also been reported that β-brass, containing approximately 50% Zn, will experience SCC in pure water. Both intergranular and transgranular cracking were observed.

Case Studies

Example 20: SCC of Copper Alloy C27000 Ferrules in Storage and in Service in Chemical Plants. A substantial number of copper alloy C27000 (yellow brass, 65Cu-35Zn) ferrules for electrical fuses cracked while in storage and while in service in paper mills and other chemical processing plants. The ferrules, made by three different manufacturers, were of several sizes. One commonly used ferrule was 3.5 cm ($1\frac{3}{8}$ in.) long by 7.5 cm (3 in.) in diameter and was drawn from 0.5 mm (0.020 in.) thick strip.

Investigation. Ferrules from fuses in service and storage in different types of plants, along with ferrules from newly manufactured fuses, were evaluated by visual examination, by determination of microstructure, by examination of any existing cracks, and by the mercurous nitrate test, which is an accelerated test used to detect residual stress in copper and copper alloys.

Ferrules that had been stress relieved after forming but not assembled to fuses passed the mercurous nitrate test; similar unassembled ferrules that had not been stress relieved after forming cracked before the required 30 min immersion period was completed. The mercurous nitrate test also showed that ferrules crimped to fuses were susceptible to cracking whether or not they had been stress relieved after forming. This last observation established that the crimping operation in assembly produced residual stresses, even on stress-relieved ferrules, that were high enough to make the ferrules susceptible to SCC. Metallography showed that the cracks that were found in ferrules on fuses either in service or from storage were of the multiple-branched type characteristic of SCC.

Conclusions. The ferrules failed by SCC resulting from residual stresses induced during forming and the ambient atmospheres in the chemical plants. The atmosphere in the paper mills was the most detrimental, and the higher incidence of cracking of ferrules there was apparently related to a higher concentration of ammonia in conjunction with high humidity.

Recommendations. The fuses were specified to meet the requirements of ASTM B 154. The three manufacturers used different methods to solve the problem. One changed to copper ferrules and fastened them to the fuse tube with epoxy cement (copper has insufficient strength for crimping). Another changed the ferrule material to a copper-iron alloy. The third manufacturer began using copper alloy C23000 (red brass, 85Cu-15Zn), began subjecting the crimped ferrules to a stress-relief anneal using an induction coil, and made long-range plans to change to plated steel ferrules.

Example 21: SCC of Copper Alloy Tube Sheet. Tube sheets of an air compressor aftercooler (Fig. 46a) were found to be cracked and leaking approximately 12 to 14 months after they had been retubed. Most of the tube sheets had been retubed several times previously because of unrelated tube failures. Sanitary (chlorinated) well water was generally used in the system, although filtered process make-up water (river water) containing ammonia was occasionally used.

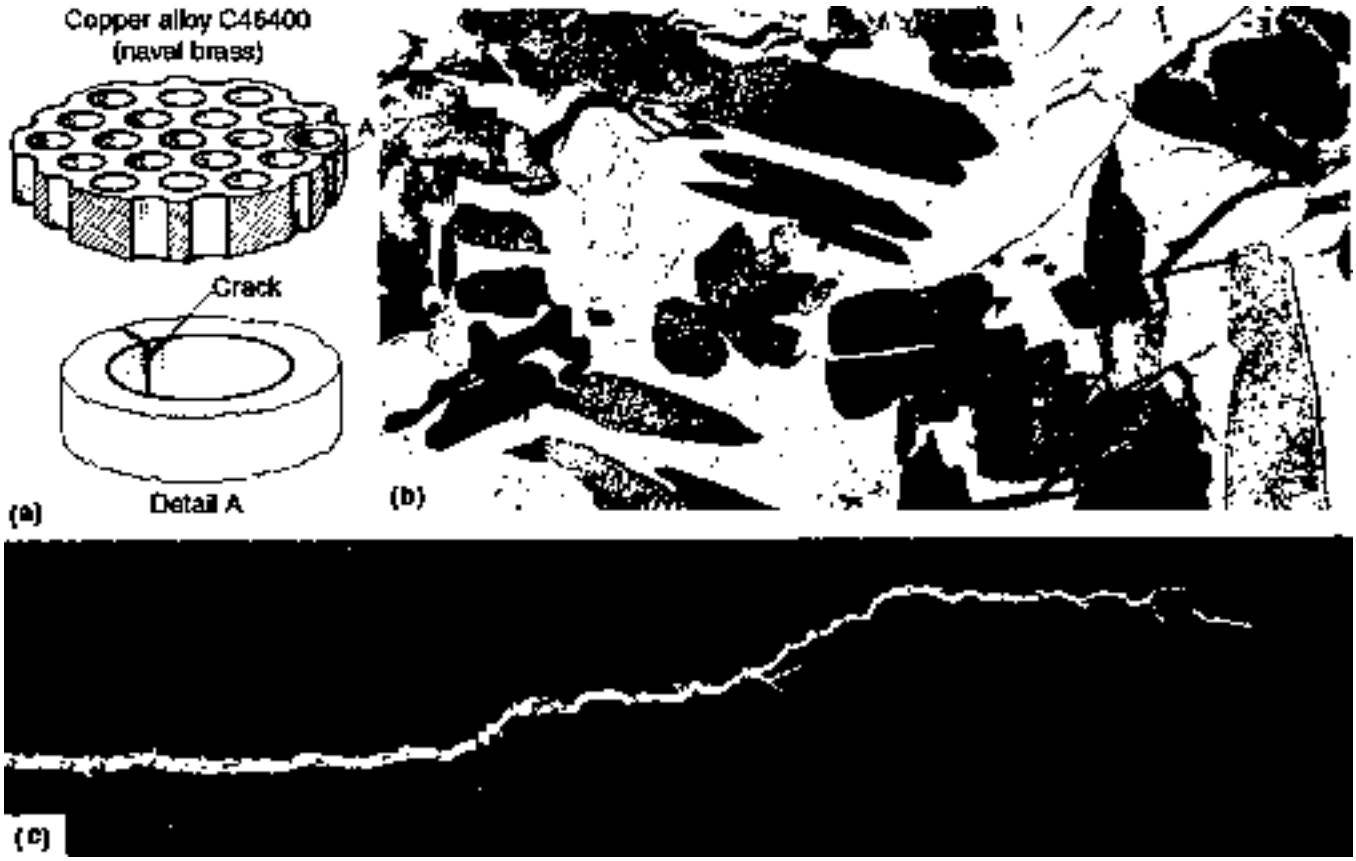


Fig. 46 Tube sheet from an air compressor aftercooler that failed by SCC. (a) Configuration of tube sheet. (b) Micrograph of a specimen etched in 10% ammonium persulfate solution showing intergranular crack propagation. 250×. (c) Macrograph of an unetched specimen showing multiple branching of cracks. 5×

The tube sheets were 5 cm (2 in.) thick. The tubes, which were 19 mm ($\frac{3}{4}$ in.) in outside diameter and 1.3 mm (0.050 in.) in wall thickness, had been expansion rolled into the holes in the tube sheets.

Investigation. One of the cracked tube sheets was removed and submitted for laboratory examination. Chemical analysis showed that the tube sheet was made of copper alloy C46400 (naval brass).

Metallographic examination of sections cut parallel and perpendicular to the tube-sheet surface revealed cracks that had penetrated through more than 90% of the tube-sheet thickness. The cracks had propagated intergranularly through the β-phase as shown in Fig. 46(b), a micrograph of a section taken parallel to the tube-sheet surface. Some dezincification of the β-phase can also be seen in this micrograph. Figure 46(c), a macrograph of a section through a crack that was more than 2.5 cm (1 in.) deep, shows multiple branching.

Sample sections of the tube sheet were subjected to the mercurous nitrate test. Cracks induced in the sample sections during the test indicated the presence of high residual stresses in the tube sheet.

Discussion. The presence of ammonia in the river water used occasionally and the presence of residual internal stresses provided the conditions necessary for SCC of the tube sheets. The multiple branching of the cracks was characteristic of SCC, and the intergranular crack propagation corresponded to the usual course of SCC in brass in the presence of ammonia.

Conclusions. The tube sheets failed by SCC as a result of the combined action of internal stresses and a corrosive environment. The internal stresses had been induced by retubing operations, and the environment had become corrosive when ammonia was introduced into the system by the occasional use of process make-up water. The slight amount of dezincification observed did not appear to have contributed substantially to the failures.

Recommendations. It was made a standard procedure to stress relieve tube sheets before each retubing operation. The stress relieving was done by heating at 275 °C (525 °F) for 30 min and slowly cooling for 3 h to

room temperature. No SCC failures have occurred in this equipment in the several years since the adoption of stress relief before retubing.

Example 22: SCC of Copper Absorber Tubes in an Air-Conditioning Unit. Eddy-current inspection was performed on a leaking absorber bundle in an absorption air-conditioning unit. The inspection revealed cracklike indications in approximately 50% of the tubes. The tube material was phosphorus-deoxidized copper.

Investigation. Two tubes with indications were pulled and examined visually and metallographically to determine the cause of cracking. The outer surfaces of the tubes were irregularly stained with a green-blue-black film, apparently the result of reaction with the shell side lithium bromide solution. The inner surfaces were covered with a thin, crusty green-black deposit that easily flaked off from the surface. No significant corrosive attack was observed on the inner surfaces.

One of the pulled tubes exhibited a large, irregular longitudinal crack, as shown in Fig. 47. The other tube exhibited a very fine longitudinal crack, as shown in Fig. 47(b).

Fig. 47 Longitudinal crack and intergranular stress-corrosion cracks in copper air-conditioning absorber tubes. (a) Longitudinal crack in one of the subject absorber tubes. 0.75×. (b) Macrograph of fine, irregular crack observed on the outer surface of the second absorber tube after light acid cleaning to remove the corrosion product. 2×. (c) Micrograph showing profiles of the primary crack and two fine secondary cracks at the outer surface of the subject absorber tube. The crack profiles are typical of stress-corrosion cracking, that is, intergranular and free of any localized grain deformation. 75×. Courtesy of J.P. Crosson, Lucius Pitkin, Inc.

Metallographic examination revealed that the cracks originated at the outer surface, were intergranular in nature, and were free of any localized grain deformation. Such features are characteristic of SCC in a copper heat exchanger tube. The crack path is shown in Fig. 47(c).

Chemical analysis of the lithium bromide solution revealed significant quantities of nitrates in the solution. Such nitrates are normally added to the lithium bromide solution to act as corrosion inhibitors.

Discussion. The service contractor responsible for maintaining the absorption unit suspected that mercury contamination from a manometer was the cause of the SCC. However, an electron probe microanalysis performed on a microspecimen did not reveal any mercury at or near the cracks.

Cracking was from the shell side or outer surface of the tubes, where the tubes were in contact with nitrate-inhibited lithium bromide. The source of ammonia was apparently the reduction of nitrates by hydrogen evolved during corrosion of the steel shell and/or tubes.

Conclusion. The results of the examination indicated that the absorber tubes failed by SCC initiated by ammonia contamination in the lithium bromide solution.

Example 23: SCC of a Brass Tube in a Generator Air Cooler Unit. An arsenical admiralty brass (UNS C44300) finned tube in a generator air cooler unit at a hydroelectric power station failed. The unit had been in operation for approximately 49,000 h. The cooling medium for the tubes is water from a river. Air flows over the finned exterior of the tubes, while water circulates through the tubes. The cooler unit had been experiencing tube leaks

with increasing frequency for three years. It was extremely difficult to determine the location of the leaks without destroying the cooler. Two previous leaks had been determined to have been caused by poor-quality rolling between the tube and tube sheet. Plant personnel suspected the possibility of fatigue cracks caused by flow-induced vibration.

Investigation. The tube was removed from the cooler by the original equipment manufacturer and returned to the plant for an independent assessment of the failure mode. The tube reportedly contained two leaks, one at a distance of approximately 4 mm (0.16 in.) from the marked end of the tube and the other approximately 50 mm (2 in.) from the marked end. The as-received tube sample was approximately 1.5 m (5 ft) in length, with rectangular external fins to improve heat transfer. No cracks were observed in the tube, but examination of the external surface was limited because of the fins. The internal surface contained a friable deposit.

The end of the tube that reportedly contained the leaks was removed from the remainder of the tube, cleaned of internal deposits, and leak tested; no leaks were found. Subsequently, the small tube section was cut longitudinally along the minor axis of the tube. The internal surface was examined using a stereomicroscope at magnifications of 7 to 45×. Several small cracks were observed. One small longitudinal crack, propagating from the inside diameter to the outside diameter, was observed at the very end of the tube. Two small transverse cracks were observed within a few millimeters of the end, with one being a through-wall crack. This through-wall crack had been covered by the plug during leak testing.

Sections of the cracked region of the tube were prepared for metallographic examination to evaluate the microstructure, crack morphology, and inside and outside diameter surface conditions. The typical microstructure of the cooler tube consisted of equiaxed alpha grains with annealing twins, as is normal for an admiralty brass in the annealed condition.

The transverse through-wall crack contained branching secondary cracks. The mode of cracking was transgranular. Branched transgranular crack paths are characteristic of SCC. The cracks appeared to initiate in pits. Figure 48 shows a micrographic montage of the longitudinal crack. This crack, like the other, was transgranular.

Fig. 48 Micrographic montage of a longitudinal crack in a brass tube from a generator air cooler. The inside diameter of the tube is at the bottom. Etched in potassium dichromate. 100×

An additional segment was removed from the tube and chemically analyzed. The composition of the tube is presented in Table 7, along with the chemical requirements for UNS C44300, an arsenical admiralty brass. The cooler tube met the chemical requirements of the specification.

Table 7 Results of chemical analysis

Element	Composition, wt%	
	Cooler tube	Chemical requirements for UNS C44300 brass[(a)]
Copper	70.86	**70.0–73.0**
Tin	0.95	**0.9–1.2**
Arsenic	0.03	**0.02–0.10**
Phosphorus	<0.01	**NR**
Antimony	<0.05	**NR**
Iron	<0.05	**0.06 max**
Lead	<0.05	**0.07 max**
Zinc	**bal**	**bal**

(a) NR, no requirement

A sample of the internal tube deposits that had been collected prior to hydrotesting was analyzed using energy-dispersive spectroscopy (EDS). Crystalline compounds in the deposit were identified using powder x-ray diffraction (XRD) techniques. The internal deposits consisted of silicon oxide (SiO_2), potassium aluminum silicate ($KAlSiO_4$), calcium carbonate ($CaCO_3$), and ammonium copper sulfite hydrate [$(NH_4)_7Cu(SO_2)_4 \cdot 5H_2O$]. In addition to the elements contained in these compounds, EDS detected minor to trace amounts of titanium, manganese, iron, and zinc. Most of these deposits are common minerals found in water. Ammonium copper sulfite hydrate indicated the possible presence of ammonia. Therefore, approximately 0.5 g of internal deposit were analyzed for ammonia by the distillation/nesslerization method. The analysis confirmed the presence of ammonia in a concentration of 368 mg/g.

Discussion. The branched transgranular cracking in the brass cooler tube was a result of SCC. The stresses in the tube were probably a combination of hoop stresses from the internal pressure, residual stresses from manufacturing, and possibly flow-induced stresses. The corrosive agent was ammonia, which was found in the internal deposits in the form of ammonium copper sulfite hydrate, a corrosion product of ammonia and copper in the brass.

Conclusion. The cause of the tube leaks was ammonia-induced SCC. Because the cracks initiated on the inside surfaces of the tubes and because the river water is not treated before it enters the coolers, the ammonia was likely present in the river water and probably concentrated under the internal deposits.

Recommendations. Two possible options are available to alleviate the cracking: either the ammonia can be eliminated or an alternate material can be used that is resistant to ammonia corrosion as well as to chlorides and sulfur species. Unless the ammonia is biologically produced, the first option is not practical; the ammonia could originate from any company upstream of the plant; locating the source would be very difficult, and continuous monitoring of the river water would be required. The second option is more viable. Alternate tube materials include a 70Cu-30Ni alloy or a more expensive titanium alloy.

Revised by W. R. Warke

Magnesium Alloys

Commercial alloys of magnesium are generally resistant to failure by SCC in their major applications, in which they are exposed to the atmosphere.

Effects of Environment. Exposure to the atmosphere has been shown to have little effect on SCC of some magnesium alloys, whether in marine, industrial, urban, or rural environments, while others are quite susceptible in these environments. Rainfall, dew, and high humidity accelerate SCC of magnesium alloys. Distilled water and other aqueous environments also produce SCC in susceptible alloys. The standard environment for accelerated SCC testing of magnesium alloys is a solution of sodium chloride and potassium chromate.

Effects of Alloying. High-purity magnesium as well as magnesium alloys containing manganese, rare earths, thorium, or zirconium, but no aluminum or zinc, can be made to fail by SCC only by exposure to such corrodents as dilute aqueous fluoride solutions at stresses higher than those encountered in normal service. Experience and testing have shown that magnesium-zinc alloys that contain zirconium or rare earths, such as wrought ZK60 and ZE10, have intermediate SCC resistance. Aluminum-containing magnesium alloys, including those containing both aluminum and zinc, are the most susceptible to SCC, with failure times decreasing as the aluminum content increases, and the peak susceptibility occurring at approximately 6% Al. Thus the most commonly used commercial alloys, the AZ alloys, have the greatest susceptibility to SCC.

Effect of Product Form. When judged on the basis of safe stress as a fraction of tensile yield strength, castings are generally somewhat less susceptible to SCC than wrought products of the same composition. Although wrought magnesium alloy products usually show some anisotropy in their mechanical properties, their resistance to SCC is not appreciably influenced by the direction of applied stress in relation to the direction of working.

Path of cracking. Stress-corrosion cracking in magnesium alloys is generally transgranular and branched, but in magnesium-aluminum alloys, intergranular cracking sometimes occurs. The crack path may change with changes in environment for a given alloy and processing and with changes in alloy processing for a given alloy composition and environment. For example, grain boundary precipitation of a magnesium-aluminum intermetallic compound, which is cathodic to the matrix, promotes the intergranular crack path. This precipitation leading to intergranular SCC can occur during slow cooling of the alloy from elevated temperatures.

The transgranular stress-corrosion cracks are typically crystallographic in appearance in the scanning electron microscope and are believed to occur by discontinuous cleavage. Very fine, parallel striations perpendicular to the direction of cracking may be discerned on the facets.

Failures and Preventive Measures. Stress-corrosion failures of magnesium alloy structures in service, which are infrequent, are usually caused by residual tensile stresses introduced during fabrication. Sources of such stresses are restrained weldments, interference fits, and casting inserts. It has been reported that approximately 10 to 60 magnesium aerospace components per year failed by SCC in the period from 1960 to 1970. Seventy percent of these failures involved either the high strength magnesium-aluminum casting alloy AZ91C-T6 or the wrought alloy AZ80-F.

Effective preventive measures include stress-relief annealing of structures that contain residual stresses produced by welding or other methods of fabrication, cladding with a metal or alloy that is anodic to the base metal, and applying protective inorganic coatings, paint, or both to the surface. Magnesium-zinc alloys with no aluminum have intermediate susceptibility, and alloys that contain neither aluminum nor zinc are the most resistant to SCC.

Titanium and Titanium Alloys

Titanium is an inherently active metal that forms a thin protective oxide film. The apparent stability and high integrity of the film in environments that cause SCC of most structural alloys make titanium and its alloys resistant to SCC in boiling 42% magnesium chloride and boiling 10% sodium hydroxide solutions, which are commonly used to induce SCC in stainless steels.

A number of environments in which some titanium alloys are susceptible to SCC and the temperatures at which cracking has been observed are listed in Table 8. Some of these environments are discussed in the following paragraphs.

Table 8 Environments and temperatures that may be conducive to stress-corrosion cracking (SCC) of titanium alloys

Environment	Temperature
Nitric acid, red fuming	**Ambient**
Hot dry chloride salts	**260–480 °C (500–900 °F)**
Cadmium, solid and liquid	**Ambient to 400 °C (750 °F)**
Chlorine	**Elevated**
Hydrogen chloride	**Elevated**
Hydrochloric acid, 10%	**Ambient to 40 °C (100 °F)**
Nitrogen tetroxide	**Ambient to 75 °C (165 °F)**
Methyl and ethyl alcohols	**Ambient**
Seawater	**Ambient**
Trichloroethylene	**Elevated**
Trichlorofluorethane	**Elevated**
Chlorinated diphenyl	**Elevated**

Aqueous Environments

Titanium alloys exhibit a wide range of behaviors in sodium chloride solutions and in seawater, and most data on SCC of titanium alloys have been generated for these environments.

Metallurgical Factors. Commercial purity grades are immune to aqueous SCC. Other grades, such as mill-annealed Ti-6Al-4V, require the presence of a preexisting crack, which then can grow by SCC in seawater. Still other grades, such as mill-annealed Ti-8Al-1V-1Mo, will experience SCC in the presence of either a notch or a precrack. A few alloys, such as step-cooled Ti-8Al-1V-1Mo, will exhibit SCC in the absence of any stress concentration. Susceptibility increases with increasing aluminum content, due to the presence of the ordered Ti_3Al phase. Oxygen and tin increase susceptibility to SCC. These elements and step-cooling also promote the formation of the ordered titanium-aluminum phase.

Crack propagation in α/β alloys is generally transgranular and cleavagelike in appearance in the alpha phase and ductile tearing in the beta phase. In β-alloys, cracking may be either intergranular or transgranular, depending on the alloy and microstructure.

Environmental Factors. A number of environmental factors influence the SCC of titanium alloys in aqueous media:

- *Type and concentration of ionic species:* Chloride, bromide, and iodide ions increase susceptibility, decreasing K_{ISCC} and increasing the crack growth rate. Some other species can have a neutral or inhibitive effect.
- *pH:* Low pH also decreases K_{ISCC} and increases the crack growth rate.

- *Temperature:* The plateau crack growth rate increases with increasing temperature.
- *Potential:* Either anodic or cathodic polarization of susceptible alloys inhibits SCC by increasing K_{ISCC}. However, increasing potential causes the crack velocity to increase.

Hot Dry Chloride Salts

Hot-salt SCC of titanium alloys was first observed in the mid-1950s.
Metallurgical Factors. All commercial alloys, but not unalloyed titanium, have some degree of susceptibility to hot-salt cracking. The most susceptible alloys are Ti-8Al-1Mo-1V and Ti-5Al-2.5Sn, and the common Ti-6Al-4V has intermediate susceptibility.
Cracking is branched, stepwise, and discontinuous, and is predominantly intergranular.
Environmental Factors. Hot-salt SCC of titanium alloys is a function of temperature, stress, and time of exposure. In general, hot-salt cracking has not been encountered at temperatures below approximately 260 °C (500 °F); greatest susceptibility occurs at approximately 290 to 425 °C (550 to 800 °F), based on laboratory tests. Time-to-failure decreases as either temperature or stress level is increased. Cracking requires the presence of both water and oxygen and will not occur with dry salts or in vacuum.
Stress-corrosion cracking has been observed in chlorides, bromides, and iodides, but not in fluorides or hydroxides. With chlorides, the severity increases with the cation in the following order: magnesium, strontium, cesium, calcium, potassium, barium, sodium, and lithium.

Other SCC Environments

Chlorine, Hydrogen Chloride, and Hydrochloric Acid. In these environments, the mechanism of cracking is not completely understood, although it appears that both oxygen and water must also be present for cracking to occur.
Nitrogen Tetroxide. In 1964, a Ti-6Al-4V pressure vessel containing high-purity liquid N_2O_4 failed during hydrotest. It has been found that N_2O_4 containing small amounts of dissolved oxygen causes cracking of titanium and some titanium alloys. No cracking occurs if the nitrogen tetroxide contains a small percentage of nitric oxide (NO) or water. The cracking may be transgranular, intergranular, or both, depending on alloy composition. It is intergranular in α-alloys and becomes mixed with increasing β-phase volume fraction.
Methanol and Other Organic Solvents. The space program also led to an awareness of SCC of titanium alloys in anhydrous methanol in the early 1960s, as a result of catastrophic failures of pressurized missile tanks. Alcohols containing small amounts of water, chloride, bromide, and iodide promote cracking at ambient temperatures. Greater concentrations of water inhibit cracking. Higher alcohols may induce cracking, but to a lesser extent; the longer the chain, the less reactive the alcohol becomes. Unalloyed titanium and Ti-6Al-4V are considered to be immune to cracking in alcohols other than methanol.
Alloys that are relatively immune to SCC in aqueous environments, such as commercially pure titanium (grades 1 and 2) and β-alloys, will exhibit intergranular SCC in methanol. Stage I cracking, at low stress levels, in other alloys can be intergranular as well. Alloys that are susceptible to transgranular SCC in aqueous environments will crack in the same way in alcohol.
Stress-corrosion cracking of highly susceptible titanium alloys has also been reported in chlorinated solvents and fluorinated hydrocarbons. Cracking in the chlorinated solvents is transgranular and similar in appearance to that in aqueous environments.

Case Studies

Example 24: Reaction Control System Oxidizer Pressure Vessels (Ref 27). Nitrogen tetroxide, a storable hypergolic oxidizer, was used in the service propulsion system (SPS) and in the reaction control system (RCS) on the Apollo program. The SPS provided the propulsion for orbit insertion, lunar flight, return from the moon, and deorbit for entry. The RCS provided for vehicle attitude control through roll, pitch, and yaw engines. Titanium alloy Ti-6Al-4V was chosen for pressure vessels in both systems as a result of laboratory tests on corrosion, stress corrosion, and impact ignition with N_2O_4.
Qualification testing of the SPS pressure vessel for 46 days of exposure under a membrane stress of 690 MPa (100 ksi) was completed without problems in mid-1964. In January 1965, an RCS pressure vessel of the same

alloy, protected from N_2O_4 by a teflon positive expulsion bladder, cracked in six adjacent locations. The cracks were parallel to each other and perpendicular to the maximum stress. The fracture surface had a red stain and was flat and brittle in appearance. Because no N_2O_4 was supposedly in contact with the vessel, it was believed that misprocessing may have caused the fractures. Subsequent testing of ten pressure vessels without bladders in June of that year was conducted to ascertain if misprocessing could be bracketed to certain manufactured pressure vessel lots over a period of time. Within 34 h after testing, one of the pressure vessels blew up (Fig. 49), and within a few days most of the others had failed.

Fig. 49 Stress-corrosion failure of an Apollo Ti-6Al-4V reaction control system (RCS) pressure vessel due to nitrogen tetroxide. (a) Failed vessel after exposure to pressurized N_2O_4 for 34 h. (b) Cross section through typical stress-corrosion cracks. 250×

The cause of the failure was immediately suspected to be stress corrosion, but the aggressive fluid that contacted these surfaces was not immediately determined. Cracks occurred very close together on the inside of the pressure vessel, and the number of cracks per unit area was proportional to the local stress. Pressure vessels of the same design, when tested with N_2O_4 by Rockwell on the West Coast, did not fail. The RCS tank contractor, Bell Aerosystems Company, was soon able to demonstrate coupon failures in N_2O_4, but Rockwell could not, although over 300 specimens were cleaned and tested under more than 40 variables, including various contaminants.

Chemical testing of the propellants used at Bell and Rockwell revealed no differences or out-of-specification conditions. Existing chemical techniques were not capable of accurately quantifying all species present in the N_2O_4, especially compounds of nitrogen; however, cooling of the propellants revealed a color difference in N_2O_4 between the supplies at Bell and Rockwell. The N_2O_4 at Rockwell, when cooled to -18 °C (0 °F), turned green because of the presence of nitric oxide (NO), but the N_2O_4 at Bell was yellow. The green color resulted from a mixture of N_2O_4 (yellow) and dissolved NO as N_2O_3 (blue). Rockwell supplies of N_2O_4 had been purchased earlier than those of Bell Aerospace.

Investigation revealed that a change in the military specification MIL-P-26539 had been made during this time period to improve the specific impulse of N_2O_4 by oxygenating the trace quantities of residual NO. This simple

change had a devastating effect on its stress-corrosion behavior with titanium. Testing indicated as little as 0.2% NO was probably sufficient to inhibit the SCC of titanium.

Specifications were changed thereafter to require a minimum NO content of 0.6%; present grades contain 1.5 to 3% NO. Addition of NO to existing supplies solved the problem.

Early in the investigation, it was believed that no stress corrosion could occur in solutions as nonconductive as N_2O_4 (specific conductivity at 25 °C, or 80 °F, is $3.1 \times 10^{-13}\ \Omega^{-1}\ cm^{-1}$). However, because of the low conductivity of N_2O_4, only closely spaced local cathodes and anodes could carry corrosion currents. This resulted in a large number of cracks (up to 70 cracks/in.) rather than a single crack. This investigation also illustrated how past stress-corrosion test results or pressure vessel qualification can be voided by minor chemical changes in the corroding medium. This is further illustrated in the following discussion.

Example 25: Service Propulsion System Fuel Tanks (Ref 27). The storable hypergolic fuel used for the SPS on the Apollo Service Module was a blend of 50% hydrazine and 50% unsymmetrical dimethyl hydrazine. It was contained in two titanium Ti-6Al-4V pressure vessels approximately 1.2 m (4 ft) in diameter and 3 m (10 ft) long. Because hydrazine compounds are toxic and dangerous to handle, a “referee” fluid with similar density and flow characteristics was used in system checkout testing. Methanol was chosen as a safe fluid based on subcontracted studies conducted for the program. Methanol had been successfully used as a fluid in a tri-flush cleaning process for propellant systems at that point in time. Among the other advantages, it had a low explosive potential, was miscible with both fuel and water, and would leave surfaces residue free.

During an acceptance test of the Apollo spacecraft 101 service module prior to delivery, an SPS fuel pressure vessel (SN054) containing methanol developed cracks adjacent to the welds (Fig. 50). The test was stopped. This acceptance test had been run 38 times on similar pressure vessels without problems. Failure analysis could not reveal the cause of the cracking. The fractures had branching cracks characteristic of stress corrosion, yet fracture faces exhibited a somewhat featureless quasi-cleavage appearance without typical stress-corrosion features. The adjacent material was ductile and within chemistry.

Fig. 50 Stress-corrosion cracking of a solution-treated and aged Ti-6Al-4V Apollo service propulsion system (SPS) fuel pressure vessel during a system checkout test. Fluid test medium was methanol. (a) Cross section adjacent to weld in cracked vessel. 65×. (b) Another crack near the same weld. 65×

Misprocessing of pressure vessels during manufacturing was suspected, because the supplier had previously demonstrated that pickup of cleaning agents and contaminants on a Ti-6Al-4V pressure vessel prior to a heat treat aging cycle would result in delayed stress-corrosion failure. These included such contaminants as fingerprints, chlorinated kitchen cleansers, and liquid hand soap. A similar crack adjacent to the weld that occurred on the first development pressure vessel for this program was thought to be caused by such contamination.

The spacecraft 017 service module was then put into test. An additional test was initiated to ensure that no marginal SPS pressure vessels would pass through system checkout. This additional test consisted of 25 pressure cycles, followed by a 24 h pressure hold. After only a few hours into the hold cycle, the replaced SPS pressure vessel failed catastrophically.

The investigation conducted after this failure disclosed that titanium will undergo SCC in methanol. Attack is promoted by crazing of the protective oxide film. It was learned that minor changes in the testing procedures could inhibit or accelerate the reaction. For example, the addition of 1% H_2O inhibited the reaction completely. It could be restarted by a 5 ppm addition of chloride. Initial stress-corrosion testing in the laboratory was performed with available aluminum text fixtures, with 300-series stainless steel, and then with titanium. No failures occurred in aluminum fixtures (apparently it provided cathodic protection). Failures in stainless steel fixtures occurred at the specimen pin areas, not at the sharp notches used to initiate stress cracking. Failures occurred in only a few hours with stainless steel because it apparently accelerated the reaction by galvanic coupling. Specimens tested to the same stress levels in titanium fixtures often took several times as long to fail.

The obvious solution to the problem was to replace the methanol with a suitable alternate fluid. Isopropyl alcohol was chosen after considerable testing. This incident further resulted in the imposition of a control specification (MF0004-018) for all fluids that contact titanium for existing and future space designs.

This failure further illustrates the importance of trace chemicals in accelerating or inhibiting a failure. It also points out that test fixture materials can affect test results. Although it is preferable to use fixtures of the same alloy, it may not always be critical in the screening of possible stress-corroding fluids, because it is often desirable to accelerate stress-corrosion testing during the screening phase. This acceleration can be carried out by significantly increasing the stress (using a notch), by raising the temperature (usually), or by making the specimen the anode in a fluid. It is notable that failures involving stainless steel fixtures did not occur when 1% H_2O was added to methanol, nor did failures occur in isopropyl alcohol, benzene, Freon TF, Freon MF (E.I. DuPont de Nemours & Co.), and distilled water and MMH, even if hydrogen chloride was added or bubbled through the solutions.

Reference cited in this section

27. L.S. Korb, Corrosion in the Aerospace Industry, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 1074–1075

Stress-Corrosion Cracking

Revised by W. R. Warke

References

1. R.H. Jones, Stress-Corrosion Cracking, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 145–163

2. G.H. Koch, Stress-Corrosion Cracking and Hydrogen Embrittlement, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* ASM International, 1996, p 483–506

3. R.N. Parkins, Stress Corrosion Cracking, *Uhlig's Corrosion,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, 2000, p 191–204

4. S.P. Lynch, Failures of Structures and Components by Environmentally Assisted Cracking, *Eng. Fail. Anal.,* Vol 1 (No. 2), June 1994, p 77–90

5. R.N. Parkins, Mechanistic Aspects of Stress Corrosion Cracking, *Parkins Symposium on Fundamental Aspects of Stress Corrosion Cracking,* Minerals, Metals and Materials Society/AIME, 1992, p 3–40

6. R.H Jones and R.E. Ricker, Mechanisms of Stress Corrosion Cracking, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992, p 1–40

7. R.P Gangloff and M.B Ives, *Environment Induced Cracking of Metals,* NACE-10, National Association of Corrosion Engineers, 1988

8. S.W. Stafford and W.H. Mueller, Failure Analysis of Stress Corrosion Cracking, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992, p 417–436

9. B.F. Brown, A Preface to the Problem of Stress Corrosion Cracking, *Stress-Corrosion Cracking Of Metals—A State Of The Art,* STP 518, H.L. Craig, Ed., ASTM, 1972, p 3

10. M.G. Fontana and N.D. Greene, *Corrosion Engineering,* McGraw-Hill, 1967

11. D. Sprowls, Evaluation of Stress Corrosion Cracking, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 272

12. W.M. Pardue, F.H. Beck, and M.G. Fontana, Propagation of Stress-Corrosion Cracking in a Magnesium-Base Alloy as Determined by Several Techniques, *Trans. ASM,* Vol 54, 1961, p 539–548

13. P.J. Forsyth and D.A. Ryder, *Metallurgia,* Vol 63, 1961, p 117

14. W.B. Lisagor, *Environmental Cracking—Stress Corrosion, Corrosion Test and Standards: Applications and Interpretation,* R. Baboian, Ed., ASTM, p 240–252

15. K. Bohnenkamp, Caustic Cracking of Mild Steel, *Fundamental Aspects of Stress Corrosion Cracking,* R.W. Staehle, A.J. Forty, and D. van Rooyen, Ed., National Association of Corrosion Engineers, 1969, p 374-383

16. N. Sridhar, D.S. Dunn, and A. Anderko, Prediction of Conditions Leading to Stress Corrosion Cracking of Gas Transmission Lines, *Environmentally Assisted Cracking: Predictive Methods for Risk Assessment and Evaluation of Materials, Equipment and Structures,* STP 1401, R.D. Kane, Ed., American Society for Testing and Materials, 2000

17. D.G. Chakrapani, Intergranular SCC of Carbon Steel Pipe Welds in a Kamyr Continuous Digester Equalizer Line, *Handbook of Case Histories,* Vol 1, ASM International, 1992, p 168

18. J. Gutzeit and J.M. Johnson, "Stress-Corrosion Cracking of Carbon Steel Welds in Amine Service," Paper 206, presented at Corrosion/86, March 1986, (Houston, TX), National Association of Corrosion Engineers

19. L. Hodas, Metallurgical Analysis of Steam Turbine Rotor Disc, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 307

20. A. Sendriks, SCC of Stainless Steels, *Stress Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992

21. J.F. Grubb and J.R. Maurer, "Use of Cathodic Protection With Superferritic Stainless Steels," Paper 28, presented at Corrosion/84, National Association of Corrosion Engineers, 1984

22. L.S. Redmerski, J.J. Eckenrod, K.E. Pinnow, and C.W. Kovach, "Cathodic Protection of Seawater Cooled Power Plant Condensors Operating With High Performance Ferritic Stainless Steel Tubing," Paper 208, presented at Corrosion/85, National Association of Corrosion Engineers, 1985

23. H. Spaehn, *Environment-Induced Cracking of Metals,* R.P. Gangloff and M.B. Ives, Ed., National Association of Corrosion Engineers, 1990, p 449

24. J.E. Truman, *Int. Met. Rev.,* Vol 26 (No. 6), 1981, p 301

25. S. Shimodaira, et al., *Stress Corrosion Cracking and Hydrogen Embrittlement of Iron Base Alloys,* R.W. Staehle, et al., Ed., National Association of Corrosion Engineers, 1977, p 1003

26. N. Sridhar and G. Cragnolino, Stress Corrosion Cracking of Nickel-Base Alloys, *Stress-Corrosion Cracking: Materials Performance and Evaluation,* R.H. Jones, Ed., ASM International, 1992

27. L.S. Korb, Corrosion in the Aerospace Industry, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 1074–1075

Stress-Corrosion Cracking

Revised by W. R. Warke

Selected References

- B.F. Brown, Ed., *Stress-Corrosion Cracking in High-Strength Steels and in Titanium and Aluminum Alloys,* Naval Research Laboratory, Washington, 1972
- J. Hochmann, J. Slater, and R.W. Staehle, Ed., *Stress-Corrosion Cracking and Hydrogen Embrittlement of Iron Base Alloys,* National Association of Corrosion Engineers, 1975
- R.H. Jones, Ed., *New Techniques for Characterizing Corrosion and Stress Corrosion,* TMS, Warrendale, PA, 1995
- *Chemistry and Electrochemistry of Corrosion and Stress Corrosion Cracking: A Symposium Honoring the Contributions of R.W. Staehle,* R.H. Jones, Ed., TMS, Warrendale, PA, 2001
- R.H. Jones, Ed., *Environmental Effects on Engineered Materials,* Marcel-Dekker, Inc., New York, 2001
- Jones and Baer, Ed., *New Techniques for Characterizing Corrosion and Stress Corrosion,* TMS, Warrendale, PA, 1996
- Jones and Gerberich, Ed., *Modeling Environmental Effects on Crack Growth Processes,* TMS, Warrendale, PA, 1986
- R.H. Jones, H. Li, and J.P. Hirth, Effects of Mixed Mode I/III Loading on Environmental-Induced Cracking, *Corrosion,* Vol 57 (No.1), 2001, p. 52
- *Environmentally Assisted Cracking: Predictive Methods for Risk Assessment and Evaluation of Materials, Equipment and Structures,* STP 1401, R.D. Kane, Ed., American Society for Testing and Materials, West Consohocken, PA, 2000
- A.I. Kovaleo, V.P. Mishing, D.L. Wainstein, and V.V. Zabilsky, Effect of Residual Stress on Hydrogen Embrittlement and Stress Corrosion Cracking, *Handbook of Residual Stress and Deformation in Steel,* G. Totten, M. Howe, and T. Inoue, Ed., ASM International, 2002
- T. Magnin, D. Tanguy and D. Delafosse, Physical and Numerical Modeling of the Stress Corrosion Cracking Behavior of Austenitic Stainless Steel, Al-Mg and Nickel Base Alloys in PWR and Chloride Solutions *Environment-Induced Cracking of Metals,* R.P. Gangloff and M.B. Ives, Ed., NACE-10, National Association of Corrosion Engineers, 1988, p 45

- *Parkins Symposium on Fundamental Aspects of Stress Corrosion Cracking,* Minerals, Metals and Materials Society/AIME, 1992
- J.C. Scully, *The Theory of Stress Corrosion Cracking in Alloys,* North Atlantic Treaty Organization, Brussels, 1971
- R.W. Staehle, Bases for Predicting the Earliest Failures Due to Stress Corrosion Cracking, *Proceedings of Chemistry and Electrochemistry of Corrosion and Stress Corrosion Cracking,* R.H. Jones, Ed., TMS, Warrendale, PA, 2001, p K-1
- J.A. Whittaker, *A Survey on the Stress Corrosion of Copper Based Alloys,* International Copper Research Association, New York, 1965

Liquid Metal and Solid Metal Induced Embrittlement

William R. Warke

Introduction

METAL-INDUCED EMBRITTLEMENT is a phenomenon in which the ductility or fracture stress of a solid metal is reduced by surface contact with another metal in either liquid or solid form. This phenomenon from the presence of another metal in the molten state has been known for some time as liquid metal embrittlement. More recently, the term *liquid metal induced embrittlement* (LMIE) has been accepted as more descriptive of the situation. Furthermore, those cases where such embrittlement extends below the melting temperature of the embrittling metal, formerly known as stress alloying, are now termed *solid metal induced embrittlement* (SMIE). In either case, the metal-induced embrittlement is the result of subcritical crack growth in the structural metal, when it is stressed in tension in the presence of specific lower-melting-point metal and alloys.
Metal-induced embrittlement has been known for many years and has been identified as the cause of numerous failures. Reference 1 cites some of the reviews published over the years and notes that there are at least four distinct forms of LMIE (Ref 2):

- *Type 1:* Instantaneous fracture of a metal under an applied or residual tensile stress when in contact with certain liquid metals. This is the most common type of LMIE.
- *Type 2:* Delayed failure of a metal when in contact with a specific liquid metal after a certain time interval at a static load below the ultimate tensile stress of the metal. This form involves grain-boundary penetration by the liquid metal and is less common than type 1.
- *Type 3:* Grain-boundary penetration of a solid metal by a specific liquid metal, which causes the solid metal to eventually disintegrate. Stress does not appear to be a prerequisite for this type of LMIE in all observed cases.
- *Type 4:* High-temperature corrosion of a solid metal by a liquid metal causes embrittlement, which is an entirely different problem from types 1 to 3.

For the purposes of this Volume, this article summarizes some of the characteristics of metal-induced embrittlement. As noted, it shares many of these characteristics with other forms of environmentally induced cracking, such as hydrogen stress cracking and SCC. This article also briefly reviews some commercial alloy systems where LMIE or SMIE has been documented and describes some examples of cracking due to these phenomena, either in manufacturing or in service. The selected references at the end of the article provide a sample of literature citation from the 1990s on failures associated with metal-induced embrittlement. For more details on the occurrence, characteristics, and proposed mechanisms of these embrittlement phenomena, readers

should consult other pertinent references, such as other articles on LMIE (Ref 3) and SMIE (Ref 4) in *Corrosion,* Volume 13 of *ASM Handbook,* and Ref 1.

References cited in this section

1. G.V. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook,* ASM International, 1990, 689–736

2. N.S. Stoloff, *Liquid Metal Embrittlement in Surfaces and Interfaces,* Part II, Syracuse University Press, 1968, p 157–182

3. M.H. Kamdar, Liquid Metal Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 171

4. M.H. Kamdar, Solid Metal Induced Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 184

Liquid Metal and Solid Metal Induced Embrittlement

William R. Warke

Characteristics of SMIE and LMIE

Solid metal induced embrittlement and liquid metal induced embrittlement both require the simultaneous presence of:

- A susceptible structural metallic material
- Surface wetting by an embrittling metal or alloy
- Tensile stress, either applied or residual

The following are some general characteristics of SMIE and LMIE:

- Embrittlement is manifested as a loss in ductility and a decrease in fracture stress with no observable change in the flow characteristics of the structural metal, as subcritical crack growth or as delayed failure under static load.
- For any given combination of structural metal and embrittling metal, there exist temperatures below and above which there is no embrittlement. The low-temperature behavior is generally believed to be controlled by the thermally activated transport of the embrittling species to growing crack tips, and the high-temperature limit is generally believed to be controlled by the decrease in strength of the structural metal. The temperature range of embrittlement is sometimes called the LMIE trough.
- Both the lower-onset temperature and the high-temperature limit are raised by increasing strain rate.
- The embrittling species must have access to potential initiation sites and to growing crack tips. For initiation to occur, the embrittler must wet (or adhere in solid form) to the structural metal surface; the passive layer can be removed locally, either chemically or by strain.
- There is no correlation between aggressiveness of the embrittling metal toward the structural metal in the absence of stress and the tendency for cracking under stress. Low mutual solubility, little tendency for compound formation, and strong binding energies are considered favorable for embrittlement.
- Embrittlement is favored by barriers to dislocation motion (grain boundaries or precipitates), high strength, and grain-boundary segregation. Thus, pure metals and single crystals are less susceptible to embrittlement but are not immune.

- The brittle fracture stress varies with the inverse square root of the average grain diameter.
- The fracture path is usually intergranular in polycrystalline metals having cubic crystal structures and is frequently transgranular in hexagonal close-packed metals.
- Changes in the alloy composition of either the structural or the embrittling metal have a strong influence on the embrittlement severity and extent.
- For LMIE, subcritical crack growth rate is extremely rapid; plateau velocities of inches per second (0.1 m/s) have been measured in fracture-mechanics-type tests. These tests also gave very low threshold stress intensities for crack growth. Because of the rapid crack growth, failure of small specimens or components is virtually instantaneous, once a sufficient stress is present.
- In SMIE, smooth or notched specimens or components, under static loads, exhibit increasing time to failure with decreasing stress and a minimum stress for failure (i.e., a threshold).
- It has been proposed that crack growth rate in SMIE is controlled by thermally activated surface self diffusion of embrittler atoms over adsorbed layers on the crack faces. Therefore, rate and time to failure are strongly temperature dependent.

Occurrence of SMIE and LMIE

Metal-induced embrittlement exhibits another characteristic, which has been termed *specificity*. This term is meant to indicate that any given structural metal or alloy is only affected by certain specific embrittling metals or alloys. Other potential embrittlers do not cause either SMIE or LMIE or, at least, have not been tested, if at all, under conditions where embrittlement has manifested itself. For the purposes of service failure analysis, it is only necessary to list those embrittling metals that are active for commercially important metals and alloys and that are liable to be encountered in service (Table 1). This list is not exhaustive, and it should not be assumed that the absence of a potential embrittler from the table means that it is not capable of causing a service failure. Specificity does not have a strong theoretical basis, and, under some conditions, a supposed nonembrittler might cause a failure.

Table 1 Metals that have been shown to cause liquid metal induced embrittlement, solid metal induced embrittlement, or both, of commercial alloys having the indicated base metals

Structural alloy base metal	Embrittling metal
Aluminum	**Hg, Ga, Na, In, Sn, Bi, Cd, Pb, Zn**
Copper	**Hg, Ga, In, Li, Bi, Cd, Pb**
Nickel	**Hg, In, Li, Sn, Pb, Zn, Ag**
Steels	
Ferritic/martensitic	**Hg, Ga, In, Li, Sn, Cd, Pb, Zn, Te, Cu**
Austenitic	**Hg, Li, Zn, Cu**
Titanium	**Hg, Cd, Ag, Au**

It should be pointed out that alloys of these embrittler metals can also be embrittling. In fact, for service failures, the embrittling metal is probably not pure but is impure or an alloy. For example, tin- and lead-base bearing alloys and solders also cause metal-induced embrittlement. Furthermore, some alloy additions to embrittling metals have been observed to increase the severity of embrittlement, while other additions have had the opposite effect.

Failure Analysis of SMIE and LMIE

Field failures that are the result of SMIE or LMIE can be analyzed using the standard techniques described elsewhere in this Volume. This article simply describes some of the unique features that assist in arriving at a clear conclusion that SMIE or LMIE is the most probable cause of the problem.

Visual Examination. In general, the subcritical crack follows a path such that its fracture surface is perpendicular to the maximum principal stress.

Solid metal induced embrittlement fractures exhibit a subcritical crack growth area and a final fracture area of different texture. The subcritical area may or may not be discolored and exhibit a characteristic color or sheen.

In the case of LMIE, the fracture is, of course, covered by a visible layer of the embrittler metal. At room temperature, mercury or its alloys are liquid, and other metals are present as a resolidified coating. Mercury-induced failures of aluminum alloys are characterized by the generation of significant quantities of white aluminum oxide. A small amount of aluminum dissolves in the mercury and oxidizes at the mercury-air interface.

Scanning Electron Microscope (SEM) Fractography. Solid metal induced embrittlement fractures of most commercial metals and alloys are classically intergranular. Near the final fracture area, the grain facets are usually smooth and the grain edges sharply defined. As the source of the embrittler is approached, the fracture takes on a rounded appearance, because there is a layer of embrittler metal on the facets. The presence of this layer can be detected by energy-dispersive spectroscopy (EDS), with decreasing intensity nearer to the final fracture area. Long counting times and/or low electron accelerating voltages may be needed, because the embrittler layer can be much thinner than the penetration depth of the electron beam. This layer can also be detected by other surface analytical techniques, such as Auger electron spectroscopy or electron spectroscopy for chemical analysis.

Liquid metal induced embrittlement fractures are not very easy to examine in the SEM. Removal of the coating of resolidified metal on the fracture surface is required, and this process may damage the underlying surface. The best that can be expected is to ascertain the fracture path.

Metallography. Metallographic examination of SMIE and LMIE failures parallels what was described previously for SEM examination. In SMIE, a thin layer of embrittler metal is often detectable on the crack faces near the origin. The crack path in both SMIE and LMIE is characteristic of the particular structural metal and embrittler but is usually intergranular, especially for body-centered cubic and face-centered cubic alloys. Secondary cracks and crack branching are often observed, and, in LMIE, these cracks and branches are filled to the tip with resolidified embrittler metal. Liquid metal induced embrittlement is often confirmed by EDS in a SEM of the material in the primary and secondary cracks in metallographic sections.

Liquid Metal and Solid Metal Induced Embrittlement

William R. Warke

LMIE and SMIE Service Failures

The following examples of LMIE and SMIE cracking and fracture in manufacture and in service alert the reader to the potential for problems resulting from this phenomenon. Such failures are much less common than those due to other fracture mechanisms, but, when they do occur, they can lead to serious consequences. Because they are less common, they are more unexpected than others and can more easily be misinterpreted. Additional examples also are available in the Selected References list at the end of this article.

Aluminum-Mercury. The embrittlement of aluminum alloys by mercury is an easily studied model system and has been the subject of much research. However, there have been relatively few examples of such failures having occurred in service. One case has been described of mercury-induced cracking of aluminum alloy piping in a natural gas processing plant (Ref 5). As-produced natural gas from the wells was cryogenically treated to remove higher-value heavier hydrocarbons, called natural gas liquids (NGL), from the methane. Mercury at the parts per billion (ppb) level was produced with the natural gas and normally was removed by adsorption beds upstream of the cryogenic plant. After some years of operation, a leak was detected in the area of the heat exchangers in the cryogenic unit, and a leak was found in a vertical nominal pipe size (NPS) 10, Schedule 80, 6061 aluminum alloy pipe. This pipe was the first one that the gas passed through, which operated at a temperature below the melting point of mercury. The pipe had been girth welded with 5356 filler metal using backing strips, and the leak occurred through one of these welds behind the backing strip. It was concluded that solid mercury had condensed out on the pipe walls during cryogenic operation. Then, when the plant was occasionally brought up to ambient temperatures for various reasons, the mercury melted and ran down, to be trapped behind the backing strips. These temperature excursions also caused stress and strain in the piping, which cracked the passive layer, allowed the mercury to wet the weld metal at the root of the weld, and led to LMIE crack propagation. Other mercury-contaminated welds were detected by x-ray radiography and removed, and the mercury absorption material was replaced.

A similar failure has been reported in an aluminum alloy piping elbow in a cryogenic unit at an ethylene plant (Fig. 1) (Ref 6, 7). In this case, the mercury was received in shipments of feedstock that contain up to 40 ppb of mercury. The failure occurred in the girth welds at an elbow; these welds had been made using 5183 filler metal. Examination clearly showed that the failure was due to mercury condensation, and testing confirmed that the base metal and the weld metal were susceptible to embrittlement by mercury.

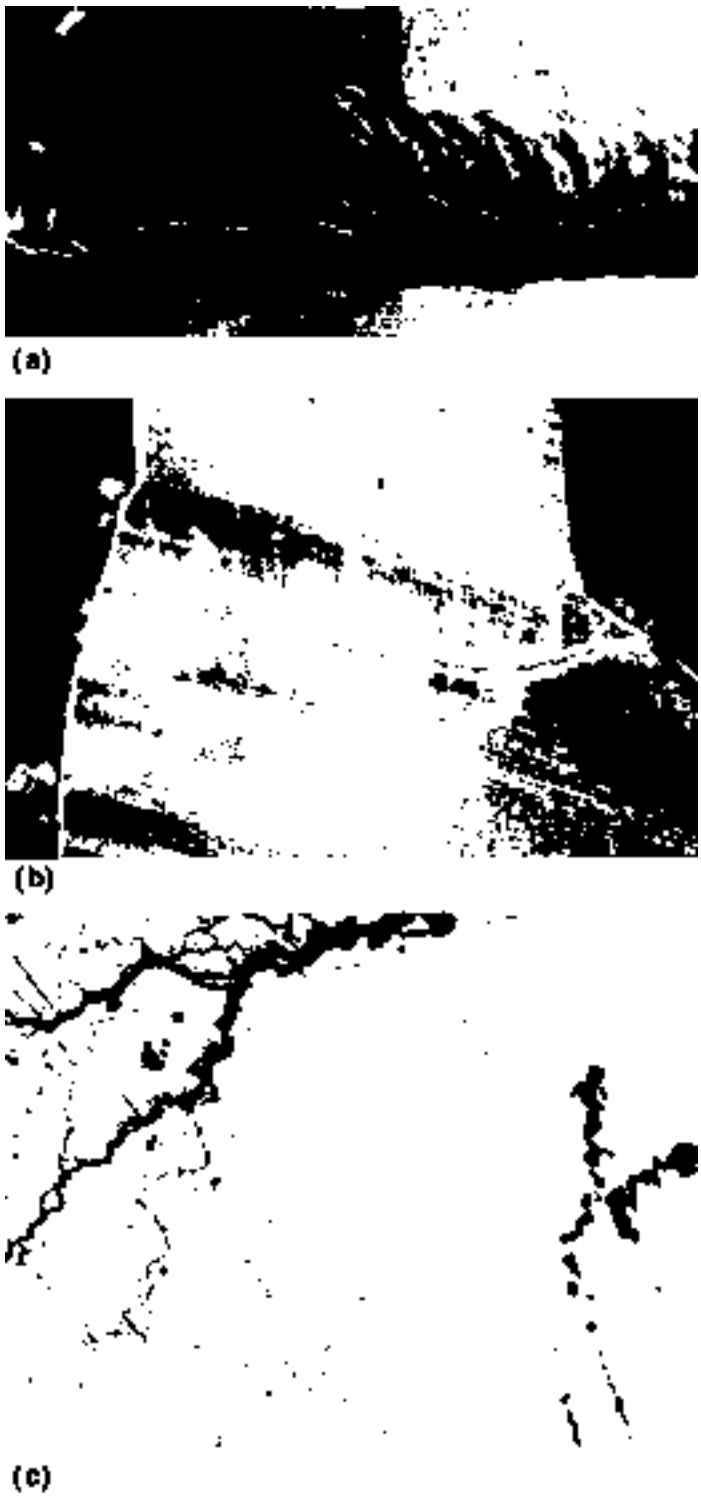


Fig. 1 Mercury-induced cracking of aluminum alloy piping. (a) Weld cap with through-wall branched cracking. (b) Cross section at the through-wall crack location. (c) Branched, intergranular cracking at a crack tip. 27×. Source: Ref 7

Aluminum-Lead. Sustained load cracking (SLC) and delayed failures have been observed at ambient temperatures in 6351-T6 aluminum alloy, which contains small amounts of lead (>10 ppm). It is perhaps unclear whether these failures should be classified as SMIE or grain-boundary segregation. The lead is contained in the alloy and apparently diffuses to grain boundaries near stress concentrations, where it exists as metallic lead and leads to subcritical crack growth by intergranular separation. Failures have been observed in self-contained breathing apparatus, in self-contained underwater breathing apparatus cylinders (Ref 8, 9), and in piping at a NGL plant (Ref 10).

The latter case involved two half-pipe headers on a finned plate heat exchanger in a natural gas liquids plant after 13 years of service. A 28 cm (11 in.) long crack was found in one 6351-T6 aluminum alloy, NPS 4, Schedule 40 header, and another partial crack was subsequently found by ultrasonic inspection. Chemical analysis confirmed the alloy and indicated a lead content of 12 ppm. Metallographic examination showed the cracking to be intergranular through very coarse-grained regions of the pipe walls (about 10 grains through the 6.0 mm, or 0.237 in., wall thickness). Other coarse-grained regions were located 90° circumferentially from the cracked areas. It was concluded that the pipes had been manufactured by extrusion through a bridged die, and that the extreme grain growth had resulted from the heat and strain of welding in the die.

Coarse grains mean less grain-boundary area per unit volume, so a small amount of lead could build up to a higher level in the boundaries. Lead-induced SLC was identified as the most probable cause of failure.

Copper-Mercury. The embrittlement of copper and copper alloys by mercury is another classic case of LMIE. Brass is very severely embrittled by mercury, and a number of studies have been carried out on this system. Again, however, reports in the literature of service failures are relatively rare. One report is provided in Ref 11, which describes the failure of three C22000 commercial bronze rupture discs by mercury embrittlement. The discs were part of a flammable gas-cylinder safety device designed to fail in a ductile mode when cylinders experience higher-than-design pressures. The subject discs failed prematurely (Fig. 2) below design pressure in a brittle manner. Fractographic examination using SEM indicated that failure occurred intergranularly from the cylinder side. Energy-dispersive spectroscopy analysis indicated the presence of mercury on the fracture surface, and mercury was also detected using scanning Auger microprobe analysis. The mercury was accidentally introduced into the cylinders during a gas-blending operation through a contaminated blending manifold. Replacement of the contaminated manifold was recommended, along with discontinued use of mercury manometers, the original source of mercury contamination.

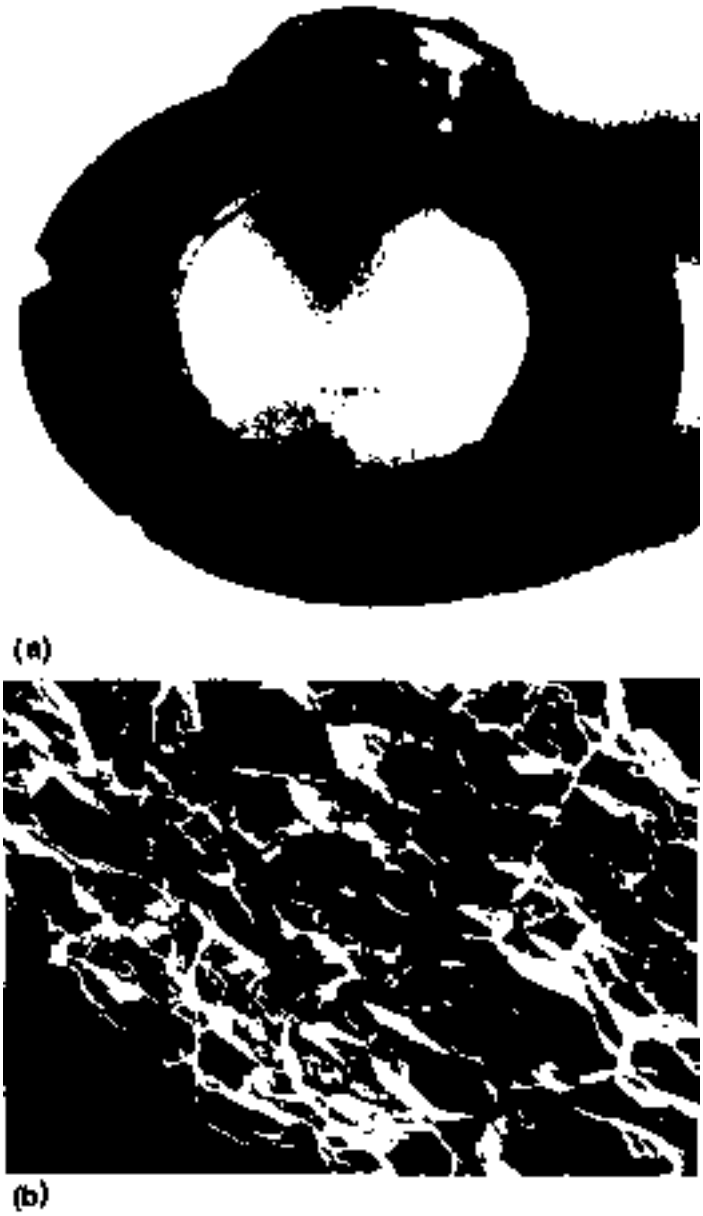


Fig. 2 Mercury-induced embrittlement of bronze rupture discs. (a) Premature, atypical rupture of a rupture disc. (b) SEM fractograph of a failed rupture disc, showing intergranular crack propagation. 554×. Source: Ref 11

Carbon or Alloy Steel and Tin. Spacecraft separation springs were cold coiled from 13 mm (0.5 in.) 9254 alloy steel high-strength rod (Ref 12). The springs were electroplated with tin and subsequently given a hydrogen embrittlement relief heat treatment at 200 °C (400 °F). Dye-penetrant inspection revealed numerous cracks near one end of the springs. Metallographic examination and EDS revealed that the cracks were filled with tin. It was concluded that the ends of the springs that were resting on the hearth had been heated to or near to the melting point of the tin (230 °C, or 450 °F) and the springs had cracked due to the tin and the residual stresses from cold forming.

It has also been reported (Ref 13) that tin-plated steel bolts, loaded to 90% of the yield strength, failed in a short time at 290 °C (550 °F).

Carbon or Alloy Steel and Cadmium. Cadmium SMIE and LMIE of high-strength steel has been recognized for many years, particularly in the fastener (Ref 14) and aircraft industries (Ref 15). The term *stress alloying* was used to describe the phenomenon. A rough guideline suggested for the maximum recommended use temperature of cadmium-plated fasteners is 230 °C (450 °F). However, other tests have shown cadmium SMIE cracking down to 160 °C (325 °F) or even 120 °C (250 °F), depending on the strength level and test conditions. Examples of service failures due to cadmium-induced embrittlement are given for a fastener problem in Fig. 3 and an aircraft component problem in Fig. 4.

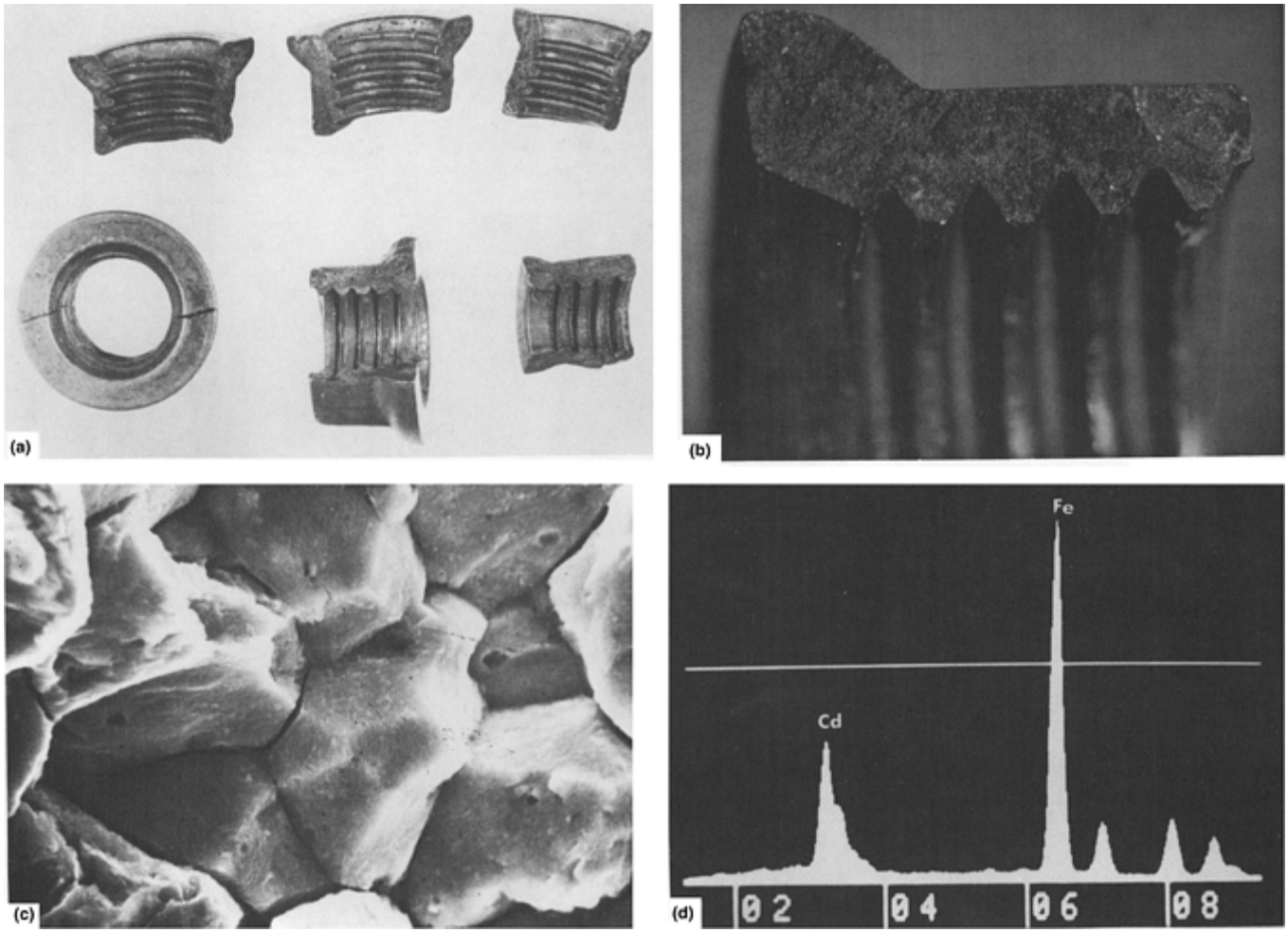


Fig. 3 Service failure of a low-alloy steel nut by LMIE. Cadmium-plated, 4140 low-alloy steel (44 HRC) nuts were inadvertently used on bolts for clamps used to join ducts that carried hot (500 °C, or 930 °F) air from the compressor of a military jet engine. (a) The nuts were fragmented or severely cracked. (b) The fracture surfaces of the nuts were oxidized and tarnished; purple, blue, and golden-yellow colors were sometimes evident. (c) SEM revealed brittle intercrystalline fracture surfaces. (d) Energy-dispersive x-ray analysis showed that the fracture surfaces were covered with a thin layer of cadmium. Failure was therefore attributed to LMIE by cadmium.

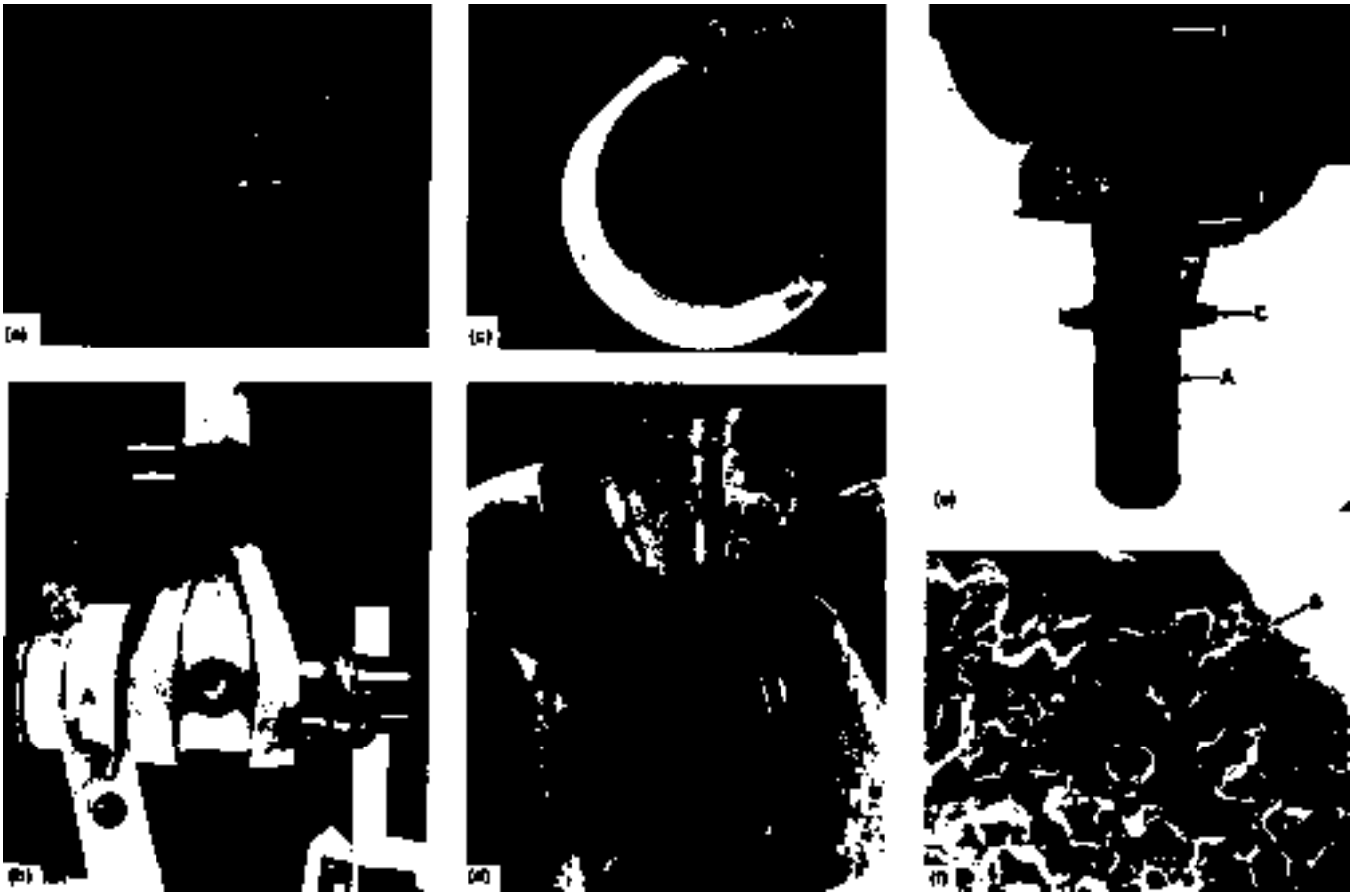


Fig. 4 Failed nose landing gear socket assembly due to LMIE. (a) Overall view of the air-melted 4330 steel landing gear axle socket. Arrow A indicates the fractured lug; arrow B, the bent but unfailed lug. Arrow C indicates the annealed A-286 steel interference-fit plug containing the grease fitting that was removed from the fractured lug. (b) The segment of the fractured lug (A) that remained attached to the launch bar. (c) The forward section of the lug shown in (b) after removal from the launch bar. Arrow A indicates the primary fracture side; arrow B, the secondary overload side. (d) The fracture face looking at the forward end of the lug [arrow A in (c)]. Region A is the grease hole that accepts the interference-fit plug; the hole was inadvertently cadmium plated approximately three-fourths of its length during original part processing and before the required brush tin plate. Arrows B indicate approximate origins; arrows C, the extent of intergranular crack growth predominantly from liquid metal embrittlement; Regions D, final catastrophic overload. (e) Shown are the interference-fit plug (arrow A) with a brush tin plate and a dry film (MoS_2) finish, the threaded grease fitting (arrow B), and the sealing washer (arrow C). Arrow D indicates an upset area that occurred during shrink-fit insertion of the plug. The upset area provided high contact pressure between the plug and the hole, resulting in a high hoop tension stress condition. (f) A typical section adjacent to the interference-fit hole (arrow B) showing intergranular crack growth and the liquid metal diffusion of almost pure tin with a trace of cadmium at the grain boundaries (arrows C). SEM. 320×

After a fire in an electrical switchgear building, some cadmium-plated, high-hardness (47 HRC) steel fuse-holder clips were found to have broken. The fractures were intergranular, and cadmium was identified on the fracture face and in secondary cracks by EDS. The cadmium on the surface appeared to have melted. It was concluded that the clips had failed by cadmium SMIE or LMIE during the fire and not previously.

Carbon or Alloy Steel and Lead. Breyer (Ref 16) referred to several service failures that had occurred as a result of lead-induced embrittlement of leaded or lead-contaminated steels. These failures included:

- Radial cracking of a gear during an induction heating operation to harden the teeth
- Failure of leaded steel dies
- Failure of shafts during warm punch straightening after heat treatment
- Failure in the seventh stage of the compressor of jet airplane engines
- Failure of fourth-stage compressor wheels of helicopter jet engines

In an investigation of a series of failures of diesel locomotive axles, it was found that the cause was LMIE by the lead-base Babbitt metal in the traction motor suspension bearings (Ref 17). The Babbitt was a Pb-Sb-Sn alloy. Due to loss of lubrication, the bearings heated to a temperature above the Babbitt melting temperature, and LMIE cracks initiated and propagated into the axle surface. Fatigue cracking then proceeded from the LMIE cracks, leading to ultimate final fracture of the axle when a critical size was reached.

Carbon or Alloy Steel and Zinc. There are occasional anecdotal reports of steel components fracturing when immersed in a hot dip galvanizing bath. These failures are believed to be associated with high hardness and high residual stresses often found in welding.

Penetration of molten zinc can also occur when previously galvanized parts are welded.

Carbon or Alloy Steel and Copper. “Copper checking” is a form of LMIE that occurs during hot working of some copper-containing steels. Any hot working operation in which the steel has been reheated in an oxidizing atmosphere can lead to this form of hot shortness and surface cracking. A layer of metallic copper is left at the interface between the iron oxide scale and the steel, and when the metal is strained, the copper penetrates the grain boundaries.

A study of some other locomotive axles was reported in Ref 18. In these axles, the temperature of the bearings became sufficiently high to melt the bronze backing of the Babbitt metal bearing. It was concluded that the Babbitt metal melted at approximately 260 °C (500 °F) and was displaced or volatilized. Frictional heating continued until the bronze bearing shells melted at 900 to 925 °C (1650 to 1700 °F), and the molten copper alloy penetrated the grain boundaries (Fig. 5) of the stressed axles. Because of the high temperature, the axles were weak, and the final fractures occurred by torsion.

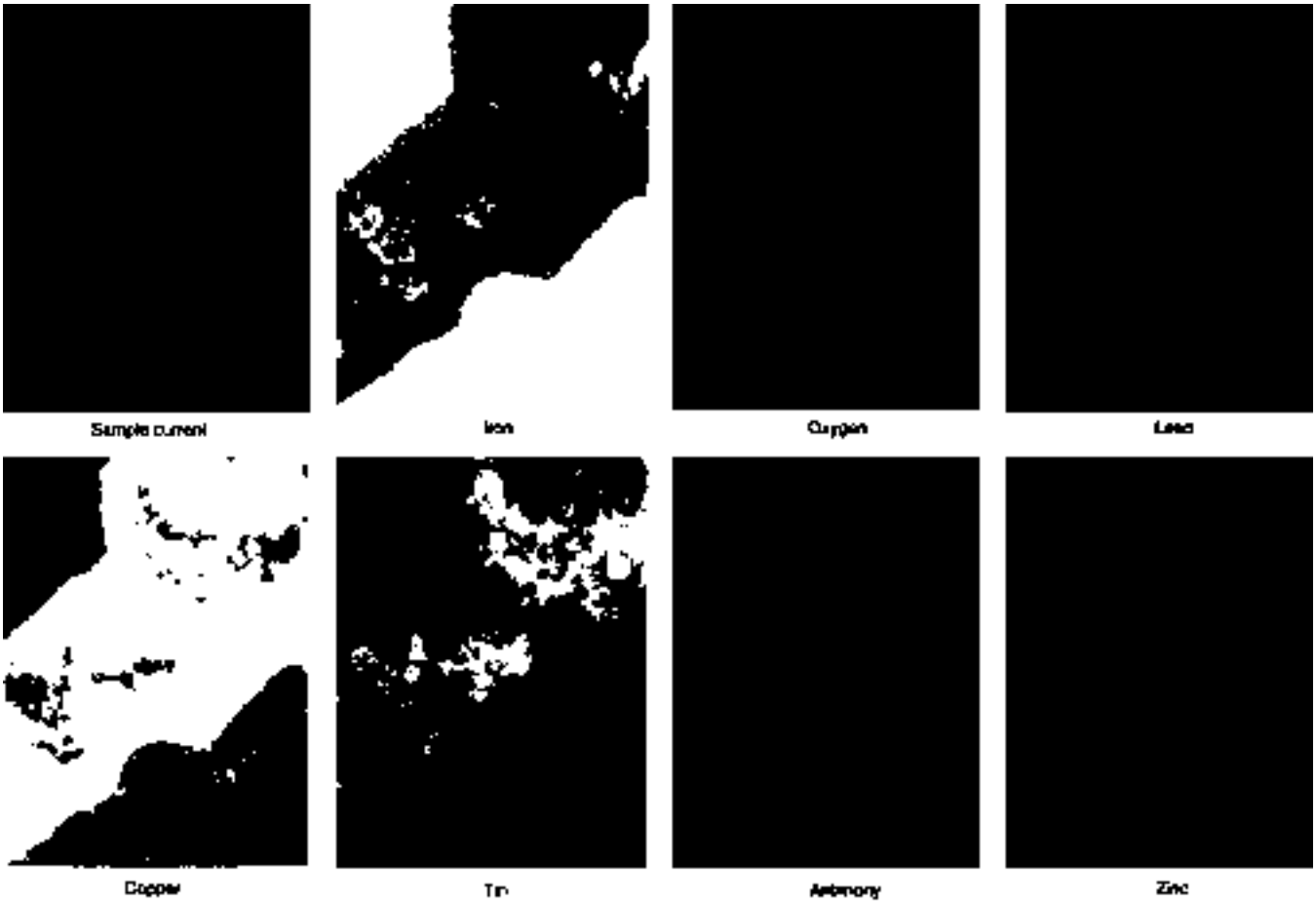


Fig. 5 X-ray elemental composition maps made with the electron microprobe of a copper penetration crack in a failed locomotive axle. 270×

Stainless Steel and Zinc. In the spring of 1975, a major disaster occurred in a chemical plant in Flixborough in the United Kingdom. The primary failure was determined to be the result of the collapse of a temporary pipe linking two reactors (Ref 19). Subsequently, secondary failures were found in another section of NPS 8, Type 316L stainless steel pipe. The results of various metallurgical investigations of these secondary failures are summarized in Ref 20. There was a 125 cm (50 in.) long rupture in an elbow, an 8 cm (3 in.) crack in a straight section of pipe, and three areas containing numerous fine cracks. The 125 cm rupture was due to creep rupture. The other cracking, however, was determined to be due to zinc-induced embrittlement. Zinc was detected by EDS on the faces of the 8 cm crack, and zinc was found in crack branches off of the main crack. The three areas of fine cracking were also concluded to be due to zinc LMIE. The zinc was thought to have come from the melting during the fire of the coating on galvanized steel components in the structures above the NPS 8 line.

A similar failure was detected in some Type 321 stainless steel piping on restarting the unit following a fire (Ref 21). Metallographic examination and electron probe microanalysis showed that zinc was present in the cracks and on the surface of the pipe. During the fire, considerable galvanizing was lost from platforms and structures above the bare pipe that cracked.

Another incident of zinc LMIE occurred in a chemical plant under construction (Ref 22, 23). The piping was Type 304 and Type 316L stainless steel. After the pipe was put in place but not welded, the structural steel was painted with zinc-rich primer by both spraying and brushing. The piping was then welded. Several leaks were identified during hydrotest, and subsequent dye-penetrant examination revealed numerous cracks in the heat-affected zones of the welds. Metallography and EDS again revealed the presence of zinc in the intergranular cracks. The pipe surfaces were found to contain spatters of the zinc-rich primer, and when welding was performed through spattered regions, LMIE cracks formed adjacent to the weld passes.

Stainless Steel and Copper. In Ref 24, heat-affected zone cracking at butt welds in thin sheets of Type 304L stainless steel when the weld area had been contaminated with copper particles is reported. This reference also discusses cracking at tube-to-sheet welds due to copper contamination from adjacent tooling. The cracks were demonstrated to be from copper penetration along grain boundaries. The copper was visible in metallographic

sections and was analyzed by EDS in the SEM. Copper LMIE of austenitic and duplex stainless steels has been documented in the literature; no other reports of cracking in manufacture or service were available.

Titanium-Cadmium. Liquid metal induced embrittlement and SMIE of titanium alloys by cadmium has been studied primarily with regard to fasteners (Ref 25). An example of LMIE is shown in Fig. 6. Another study found cracking of titanium bolts in the head-to-shank fillet after torquing and exposure to 150 °C (300 °F) for cadmium, nickel-cadmium, and vacuum cadmium plating. Cadmium SMIE of titanium alloys has been observed at temperatures as low as 90 °C (200 °F).

Fig. 6 Failed Ti-6AI-4V shear fastener. The fasteners were cadmium plated for galvanic compatibility with the aluminum structure. (a) Photograph showing failure at the head-to-shank fillet. (b) Intergranular fracture morphology. Failure was attributed to LMIE caused by excessive temperature exposure while under stress. SEM split screen. Left: 105×. Right: 1050×

References cited in this section

5. E.J. Sperling and W.R. Warke, Metal Induced Cracking of Aluminum Alloy Piping, *PVP,* Vol 288, *Service Experience and Reliability Improvement: Nuclear Fossil and Petrochemical Plants,* American Society of Mechanical Engineers, 1994, p 329

6. D.R. McIntyre, J.J. English, and G. Kobrin, “Mercury Attack of Ethylene Plant Alloys,” Paper 106, presented at Corrosion/89, National Association of Corrosion Engineers, 1989

7. J.J. English and D.J. Duquette, Mercury Liquid Embrittlement Failure of 5083-O Aluminum Alloy Piping, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993 p 207

8. J. Notaro, M.D. Rana, and R.O. Tribolet, “Sustained-Load-Cracking Behavior of DOT 3 AA6061 and 6351 Aluminum Cylinder Neck and Shoulder Materials,” report submitted to U.S. Department of Transportation, 1988

9. A.I. Roberts, Safety Advisory; High Pressure Aluminum Composite Hoop-Wrapped Cylinders, Research and Special Programs Administration, Department of Transportation, *Fed. Reg.,* Vol 59, (No. 142), 1994, p 38028

10. E.J. Sperling and W.R. Warke, Metal Induced Cracking of Aluminum Alloy Piping, *PVP,* Vol 288, *Service Experience and Reliability Improvement: Nuclear Fossil and Petrochemical Plants,* American Society of Mechanical Engineers, 1994, p 332

11. W.R. Watkins, Jr., Liquid Metal Embrittlement of Bronze Rupture Discs, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 130

12. A.J. Babecki, Some Failure Problems in Spacecraft and Space Boosters, *Proc. William Hunt Eisenman Conf. on Failure Analysis,* American Society for Metals, 1968

13. J. Glackin, "Corrosion Resistance Properties and Susceptibility to Stress Alloying of Tin Plated EWB 22 Bolts and Tin-Nickel Plated NAS 624 Bolts," SPS Laboratory Report 974, SPS Laboratories, 1963

14. T.C. Baumgartner, "Fastener Performance at Elevated Temperatures," presented at the Design Engineering Conf., American Society of Mechanical Engineers, 1967

15. E.M. Kennedy, Jr., "The Effect of Cadmium Plating on SAE 4340 Steel in the Presence of Stress Concentrations at Elevated Temperatures," WADC TR 58-108, ASTIA 151075, 1958

16. N.N. Breyer, Some Effects of Certain Trace Elements on the Properties of High Strength Steels, *Proc. 31st Electric Furnace Conf.,* 1974

17. I. Le May, A.K. Koul, and R.V. Dainty, Fracture Mechanisms in a Series of Locomotive Axle Failures, *Materials Characterization,* Vol 26, 1991, p 235

18. G.F. Vander Voort, Failures of Locomotive Axles, *Failure Analysis and Prevention,* Vol 11 *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 715–728

19. D.R.H. Jones, *Engineering Materials 3, Materials Failure Analysis, Case Studies and Design Implications,* Pergamon Press, 1993, p 283

20. A.H. Cottrell and P.R. Swann, A Metallurgical Examination of the Eight-Inch Line, *The Chemical Engineer,* 1976, p 266

21. J.E. Cantwell and R.E. Bryant, Failure of Refinery Piping by Liquid Metal Attack and Materials Experience with Smokeless Flares, *Proc. 38th Midyear Meeting of the API,* American Petroleum Institute, 1973

22. T. Shinohara and K. Matsumoto, Welding Cracks of Zn-Contaminated Stainless Steel Pipe, *Corros. Sci.,* Vol 22, 1982, p 723

23. J.M. Johnson, M.R. Berry, and J. Gutzeit, Zinc Embrittlement of Stainless Steel Welds, *Embrittlement by Liquid and Solid Metal,* Kamdar, Ed., TMS/AIME, 1984

24. R.K. Holbert, Jr., A.G. Dobbins, and R.K. Bennett, Jr., Copper Contamination Cracking in Thin Stainless Steel Sheet, *Weld. J.,* Aug 1987, p 38

25. R.L. Sproat, "The SPS Study of the Cracking of Titanium Bolts in the Presence of Cadmium Plating," SPS Laboratory Report 1880, SPS Laboratories, 1969

References

1. G.V. Vander Voort, Embrittlement of Steels, *Properties and Selection: Irons, Steels, and High-Performance Alloys,* Vol 1, *ASM Handbook*, ASM International, 1990, 689–736

2. N.S. Stoloff, *Liquid Metal Embrittlement in Surfaces and Interfaces,* Part II, Syracuse University Press, 1968, p 157–182

3. M.H. Kamdar, Liquid Metal Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 171

4. M.H. Kamdar, Solid Metal Induced Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 184

5. E.J. Sperling and W.R. Warke, Metal Induced Cracking of Aluminum Alloy Piping, *PVP,* Vol 288, *Service Experience and Reliability Improvement: Nuclear Fossil and Petrochemical Plants,* American Society of Mechanical Engineers, 1994, p 329

6. D.R. McIntyre, J.J. English, and G. Kobrin, "Mercury Attack of Ethylene Plant Alloys," Paper 106, presented at Corrosion/89, National Association of Corrosion Engineers, 1989

7. J.J. English and D.J. Duquette, Mercury Liquid Embrittlement Failure of 5083-O Aluminum Alloy Piping, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993 p 207

8. J. Notaro, M.D. Rana, and R.O. Tribolet, "Sustained-Load-Cracking Behavior of DOT 3 AA6061 and 6351 Aluminum Cylinder Neck and Shoulder Materials," report submitted to U.S. Department of Transportation, 1988

9. A.I. Roberts, Safety Advisory; High Pressure Aluminum Composite Hoop-Wrapped Cylinders, Research and Special Programs Administration, Department of Transportation, *Fed. Reg.,* Vol 59, (No. 142), 1994, p 38028

10. E.J. Sperling and W.R. Warke, Metal Induced Cracking of Aluminum Alloy Piping, *PVP,* Vol 288, *Service Experience and Reliability Improvement: Nuclear Fossil and Petrochemical Plants,* American Society of Mechanical Engineers, 1994, p 332

11. W.R. Watkins, Jr., Liquid Metal Embrittlement of Bronze Rupture Discs, *Handbook of Case Histories in Failure Analysis,* Vol 2, ASM International, 1993, p 130

12. A.J. Babecki, Some Failure Problems in Spacecraft and Space Boosters, *Proc. William Hunt Eisenman Conf. on Failure Analysis,* American Society for Metals, 1968

13. J. Glackin, "Corrosion Resistance Properties and Susceptibility to Stress Alloying of Tin Plated EWB 22 Bolts and Tin-Nickel Plated NAS 624 Bolts," SPS Laboratory Report 974, SPS Laboratories, 1963

14. T.C. Baumgartner, "Fastener Performance at Elevated Temperatures," presented at the Design Engineering Conf., American Society of Mechanical Engineers, 1967

15. E.M. Kennedy, Jr., “The Effect of Cadmium Plating on SAE 4340 Steel in the Presence of Stress Concentrations at Elevated Temperatures,” WADC TR 58-108, ASTIA 151075, 1958

16. N.N. Breyer, Some Effects of Certain Trace Elements on the Properties of High Strength Steels, *Proc. 31st Electric Furnace Conf.,* 1974

17. I. Le May, A.K. Koul, and R.V. Dainty, Fracture Mechanisms in a Series of Locomotive Axle Failures, *Materials Characterization,* Vol 26, 1991, p 235

18. G.F. Vander Voort, Failures of Locomotive Axles, *Failure Analysis and Prevention,* Vol 11 *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 715–728

19. D.R.H. Jones, *Engineering Materials 3, Materials Failure Analysis, Case Studies and Design Implications,* Pergamon Press, 1993, p 283

20. A.H. Cottrell and P.R. Swann, A Metallurgical Examination of the Eight-Inch Line, *The Chemical Engineer,* 1976, p 266

21. J.E. Cantwell and R.E. Bryant, Failure of Refinery Piping by Liquid Metal Attack and Materials Experience with Smokeless Flares, *Proc. 38th Midyear Meeting of the API,* American Petroleum Institute, 1973

22. T. Shinohara and K. Matsumoto, Welding Cracks of Zn-Contaminated Stainless Steel Pipe, *Corros. Sci.,* Vol 22, 1982, p 723

23. J.M. Johnson, M.R. Berry, and J. Gutzeit, Zinc Embrittlement of Stainless Steel Welds, *Embrittlement by Liquid and Solid Metal,* Kamdar, Ed., TMS/AIME, 1984

24. R.K. Holbert, Jr., A.G. Dobbins, and R.K. Bennett, Jr., Copper Contamination Cracking in Thin Stainless Steel Sheet, *Weld. J.,* Aug 1987, p 38

25. R.L. Sproat, “The SPS Study of the Cracking of Titanium Bolts in the Presence of Cadmium Plating,” SPS Laboratory Report 1880, SPS Laboratories, 1969

Liquid Metal and Solid Metal Induced Embrittlement

William R. Warke

Selected References

- W.G. Ashbaugh, Liquid Metal Embrittlement II, *Mater. Perform.,* Vol 32 (No. 2), Feb 1993, p 88–89
- P.H. Au-Yeung, J.T. Lukowski, L.A. Heldt, and C.L. White, An Auger Spectrometric Study of the Crack Tip Surface Chemistry for Liquid Metal Embrittlement: Beta Brass Embrittled by Gallium, *Scr. Metall. Mater.,* Vol 24 (No. 1), Jan 1990, p 95–100
- J. Barralis and V. Traskine, “Liquid Metal Induced Degradation of Alloy 7010,” ICAA5, Part 3, Conf. Aluminum Alloys—Their Physical and Mechanical Properties, 1–5 July 1996 (Grenoble, France), and published in *Mater. Sci. Forum,* Vol 217–222 (No. 3), 1996, p 1611–1616
- D.W. Cameron, Failure in Large Gas Turbines Due to Liquid-Metal Embrittlement, *Mater. Charact.,* Vol 33 (No. 1), July 1994, p 37–43

- R.E. Clegg, P.J.L. Fernandes, and D.R.H. Jones, Failure by Liquid Metal Induced Embrittlement, *Eng. Fail. Anal.,* Vol 1 (No. 1), March 1995, p 51–63
- R.E. Clegg and D.R.H. Jones, Liquid Metal Embrittlement in Failure Analysis, *Sov. Mater. Sci.,* Vol 27 (No. 5), Sept-Oct 1991, p 453–459
- R.E. Clegg and D.R.H. Jones, "The Effect of Cold Work on the Liquid Metal Induced Embrittlement of Brass by Gallium," ICF9, Conf. Advances in Fracture Research, 1–5 April 1997 (Sydney, Australia), and published in *Failure Analysis, Remaining Life Assessment, Life Extension and Repair,* Vol 1, Elsevier Science Ltd., Oxford, U.K., 1997, p 439–445
- C.P. Dillon, Liquid Metal Cracking of Stainless Steels in Chemical Plants, *Mater. Perform.,* Vol 29 (No. 11), Nov 1990, p 54–55
- H.E. Ebert, Liquid Metal Embrittlement of Flange Connector Studs in Contact with Cadmium, *Handbook of Case Histories in Failure Analysis,* Vol 1, ASM International, 1992, p 335–337
- P.J.L. Fernandes and D.R.H. Jones, Specificity in Liquid Metal Induced Embrittlement, *Eng. Fail. Anal.,* Vol 3 (No. 4), Dec 1996, p 299–302
- P.J.L. Fernandes and D.R.H. Jones, The Effect of Crack Blunting in Liquid Metal Environments on K_{IEAC} Determined by the Rising Load Test, *Eng. Fail. Anal.,* Vol 3 (No. 3), Sept 1996, p 227–230
- R.E. Hargrave, Unusual Failures Involving Copper Deposition in Boiler Tubing, *Corrosion,* Vol 47 (No. 7), July 1991, p 555–567
- J.J. Krupowicz, "Effect of Heat Treatment on Liquid Metal-Induced Cracking of Austenitic Alloys," Conf. Slow Strain Rate Testing for the Evaluation of Environmentally Induced Cracking: Research and Engineering Applications, 18–19 May 1992 (Pittsburgh, PA), and Report STP 1210, ASTM, 1993, p 193–201
- J.J. Lewandowski, Y.S. Kim, and N.J.H. Holroyd, Lead-Induced Solid Metal Embrittlement of an Excess Silicon Al-Mg-Si Alloy at Temperatures of -4 °C to 80 °C, *Metall. Trans. A,* Vol 23 (No. 6), June 1992, p 1679–1689
- J.T. Lukowski, D.B. Kasul, L.A. Heldt, and C.L. White, Discontinuous Crack Propagation in Germanium Induced Liquid Metal Embrittlement of Beta Brass, *Scr. Metall. Mater.,* Vol 24 (No. 10), Oct 1990, p 1959–1964
- C.D. Lundin, C.Y.P. Qiao, and Y. Kikuchi, Intergranular Liquation Cracking in an Advanced Austenitic Stainless Steel, *Conf. Heat-Resistant Materials,* 23–26 Sept 1991 (Fontana, WI), ASM International, 1991, p 71–79
- S.P. Lynch, A Commentary on Mechanisms of Environmentally Assisted Cracking, *Second Int. Conf. on Corrosion-Deformation Interactions,* CDI '96, 24–26 Sept 1996 (Nice, France), Institute of Materials, 1997, p 206–219
- C.J. McMahon, Jr., Diffusion-Controlled Brittle Fracture in Overheated Steels, *Proc. R.I. Jaffee Memorial Symp.,* Conf. Clean Steels Technology, 2–5 Nov 1992 (Chicago, IL), ASM International, 1992, p 3–8
- A.P. Reynolds and G.E. Stoner, Cleavage Crystallography of Liquid Metal Embrittled Aluminum Alloys, *Metall. Trans. A,* Vol 22 (No. 8), Aug 1991, p 1849–1855
- Y.-J. Su, Y.-B. Wang, and W.-Y. Chu, "Mechanism of Liquid Metal Embrittlement for Aluminum in Hg+3 at.% Ga," Corrosion 98, 22–27 March 1998 (San Diego, CA), National Association of Corrosion Engineers, 1998
- R.L. Turissini, T.V. Bruno, E.P. Dahlberg, and R.B. Setterlund, Mercury Liquid Metal Embrittlement Causes Aluminum Plate Heat Exchanger Failure, *Mater. Perform.,* Vol 37 (No.6), June 1998, p 59–60
- E. Vergeldt, A.A. Junai, and B.H. Kolster, Liquation Cracking of High Strength Aluminum Alloy 7075, *Joining/Welding 2000 (Assemblage/Soudage 2000),* 1–2 July 1991 (The Hague, the Netherlands), Pergamon Press, Oxford, U.K., 1991, p 229–241
- M.D. Wright, G. Goszczynski, and F. Peca, Embrittlement of Alloy 400 by Lead in Secondary Side Steam Generator Environments, *Seventh Int. Symp. Environmental Degradation of Materials in Nuclear Power Systems—Water Reactors,* Vol 1, 7–10 Aug 1995 (Breckenridge, CO), National Association of Corrosion Engineers, 1995, p 209–217
- P. Zhong and H. Yan, A Case Study of Bi-Sn-Induced Embrittlement, *Eng. Fail. Anal.,* Vol 3 (No. 4), Dec 1996, p 241–247

High-Temperature Corrosion-Related Failures

Robert B. Pond, Jr., M-Structures, Inc.; David A. Shifler, Naval Surface Warfare Center

Introduction

WHEN FAILURES OCCUR, the unscheduled outages result in loss of reliability and increased economic costs. A failure investigation can determine the primary cause of material, component, or system degradation. Based on this analysis, corrective action to mitigate or eliminate future failures can be initiated. Analysis of the failure generally requires the identification of the mode of degradation, whether it is mechanical, chemical, electrochemical, or a combination of several factors. This failure analysis can be complex, because more than one failure mechanism may be operative. The educational background and experience of the failure analyst can influence the failure investigation. The analyst must assess the importance of contributory causes to the rupture and the techniques required for the examination. A complex investigation may require several experts in different disciplines of engineering, physical science, metallurgy, or an understanding of corrosion science and engineering.

High-temperature corrosion may occur in numerous environments and is affected by factors such as temperature, alloy or protective coating composition, time, and gas composition. High-temperature corrosive environments may include power-generation plants (coal, oil, natural gas, and nuclear), pulp and paper mills, waste incineration sites, numerous industrial chemical processes, diesel engines (power, land vehicle, and maritime), and gas turbine (land-based, marine shipboard, and aircraft).

Predicting corrosion of metals and alloys or coated alloys is often difficult because of the range of composition of the corrosive gaseous or molten environments. Moreover, corrosion prediction is further complicated, because materials often degrade in a high-temperature environment by more than one corrosion mechanism.

High-Temperature Corrosion-Related Failures

Robert B. Pond, Jr., M-Structures, Inc.; David A. Shifler, Naval Surface Warfare Center

High-Temperature Corrosion Mechanisms

High-temperature corrosion of materials may occur through a number of potential degradation processes:

- Oxidation
- Carburization and metal dusting
- Sulfidation
- Hot corrosion
- Chloridation
- Hydrogen interactions
- Molten metals
- Molten salts
- Aging reactions, such as sensitization
- Environmental cracking (stress-corrosion cracking and corrosion fatigue)

The presence of molten salts or metals may induce other mechanisms, such as galvanic corrosion, crevice corrosion, and pitting corrosion. Impingement by solid particles may contribute to erosion-corrosion or accelerate corrosion in the various gaseous or molten environments.

Oxidation. Most metals and alloys are often oxidized on exposure to temperatures above 300 °C (570 °F) in environments containing more than 1 vol% O_2. Alloys are protected, to varying degrees, by the oxide layer. The thickness of the oxide is proportional to the parabolic rate kinetics. Oxidation of carbon steels from air or steam forms an oxide scale along the metal surface that grows and thickens with time. Gas composition influences the rate of oxidation. The effects of oxygen concentration and temperature are specific to each alloy. Scale growth rate increases with a rise in temperature. Table 1 provides suggested maximum temperatures for various alloys.

Table 1 Suggested maximum temperatures for various alloys in continuous, isothermal oxidizing atmospheres without excessive scaling

Alloy designation	Temperature maximum		Ref
	°C	°F	
Carbon and alloy steels			
1010 steel	480	900	1
SA-178, grade A	510[a]	950[a]	2
SA-192; SA-210, grade A	510[b]	950[b]	2
SA-209, grade T-1A	525[b]	975[b]	2
$1\frac{1}{4}$Cr-$\frac{1}{2}$Mo-Si; SA-213, grade T-11	565[b]	1050[b]	2
$2\frac{1}{4}$Cr-1Mo-Si; SA-213, grade T-22	600[b]	1115[b]	2
7Cr-Mo	650	1200	3
9Cr-1Mo-V; SA-213, grade T-91	650[a]	1200[a]	2
9Cr-Mo	675	1250	3
Stainless steels			
201, 202	845	1555	4
301	900	1650	4
302, 304	925	1695	4
309	1095	2000	4
310	1150	2100	4
316, 317, 321, 347	925	1695	4
330	1150	2100	4
405, 410	705	1300	4
416	675	1245	4
420	620	1150	4
430	815	1500	4
440	760	1400	4
442	980	1795	4
446	1095	2005	4
Superalloys			
N-155 iron-base superalloy	1040	1900	1
S-816 cobalt-base superalloy	980	1800	1
Hastelloy X nickel-base superalloy	1205	2200	1
HX (17Cr-66Ni-bal Fe)	1150	2100	1

(a) Seamless tube.
(b) Electric resistance welded tube

Iron oxides alone are not protective above 550 °C (1020 °F) (Ref 5). Chromium, aluminum, and/or silicon assist in forming scales, which are more protective at higher temperatures (Ref 6). Chromium oxide is stable up to 983 °C (1800 °F), above which the oxide volatizes to gaseous CrO_3.

The temperatures listed in Table 1 may require modification, depending on the severity of the oxidation rate as influenced by the actual composition of environment and other operational factors. Preferential grain-boundary oxidation, thermal cycling, the presence of moisture, and chloride- or sulfur-containing gases cause service lives much shorter than those predicted from isothermal oxidation. Loss of oxide adhesion and integrity may be

caused by mechanical damage via spallation or cracking of the oxide film from cyclic oxidation of chromia and alumina oxides formed on nickel-base alloys due to the mismatch of thermal expansion coefficients between the oxide and the base alloy (Ref 7). Small additions of lanthanum, yttrium, tantalum, ceria, zirconia, and/or niobia improve the scale adhesion and scale resilience to cyclic oxidation (Ref 5).

Carburization is the formation of carbide corrosion products on exposure to temperatures above approximately 760 °C (1400 °F) in gases containing methane, carbon monoxide, hydrocarbons, carbon, or other carbonaceous compounds. Carburization alone usually does not result in corrosion or alloy wastage, but adsorption and diffusion of carbon into the base alloy can lead to significant changes in the alloy mechanical properties and thus possible alloy embrittlement. Variables that affect the carburization rate are the temperature, exposure time, alloy composition, and partial pressures of H_2, CH_4, and H_2S. Hydrogen sulfide tends to slow alloy decarburization rates.

Carbon dioxide produced in an oxidizing environment is less corrosive than carbon monoxide, CO, which is the prevalent carbonaceous gaseous species in a reducing environment. Under simultaneous oxidizing and carburizing conditions, where hydrocarbons are present in the gas stream and/or with the introduction of oily components, stainless steels and nickel-base alloys are susceptible to severe carburization called “green rot” (Ref 8). Green rot occurs when internal chromium carbides are formed during carburization; subsequent oxidizing of the carbides at the grain boundaries may result in embrittlement.

Metal dusting can be described as a catastrophic form of carburization occurring under conditions where the carbon activity, a_c, of the gaseous atmosphere is greater than that of the metal. The degradation takes place in strongly carburizing atmospheres with carbon activities greater than 1 (a_c » 1) and at intermediate temperatures between 400 and 900 °C (750 and 1650 °F). Although metal dusting has taken place at temperatures as high as 1095 °C (2000 °F) in strongly reducing gaseous environments, metal dusting generally occurs at temperatures from 480 to 815 °C (900 to 1500 °F).

Deterioration and material wastage by metal dusting in carburizing gases result in pitting and overall wastage of stainless steels, with or without an oxidizing environment (Ref 9, 10). Corrosion products in the pits consist of metal carbides, metal oxides, and graphite. The attack is localized, and carburization is confined to the pit.

Slight modifications in the process conditions may significantly affect the occurrence and extent of metal dusting attack. The phenomenon can produce rapid metal wastage, producing pits and grooves as the affected metal disintegrates into a mixture of powdery carbon and metal particles (Fig. 1). Metal dusting corrosion has negatively impacted the efficiency and productivity of processes within numerous industries. These include petrochemical processing (processes involving steam reforming of methane or natural gas, such as ammonia and methanol production), direct iron-ore reduction, coal degasification, and heat treating.

Fig. 1 Metal dusting attack on the inner wall of a stainless steel tube

There is no iron- or nickel-base alloy that is absolutely resistant to metal dusting attack. Research is focused on the reaction chain from the formation of alloy substrate oxide layers and deteriorating conditions. Accordingly, carbon and coke deposition and carbide formation and decomposition are investigated. For nickel-base alloys, the following mechanism was proposed (Ref 11, 12, 13, 14):

1. Carbon transfer from the gas phase and dissolution of carbon into the metal phase at oxide defect sites
2. Formation of a supersaturated solution of carbon in the nickel-iron matrix
3. Deposition of graphite on the alloy surface in different orientations
4. Growth of graphite into the metal phase by carbon atoms from the solid solution, attaching to graphite planes growing vertical to the metal surface
5. Destruction of the metal phase by the inward-growing graphite under transfer of metal particles into the "coke" layer
6. Further graphite deposition from the gas phase on these catalytically active metal particles (iron, nickel).

Chemical reactions producing such high carbon activities, for example, in reformer applications, are:

$$\text{Hydrogen reformer: } CH_4 + H_2O \rightleftharpoons CO + 3H_2 \qquad \text{(Eq 1)}$$

In a temperature regime between 550 and 850 °C (1020 and 1560 °F), the prevailing gas composition may be due to the following reactions:

$$\text{Redox reaction: } CO + H_2 \rightleftharpoons 3H_2O + C \qquad \text{(Eq 2)}$$

$$\text{Boudouard reaction: } 2CO \rightleftharpoons CO_2 + C \qquad \text{(Eq 3)}$$

In cases where carbon activities are calculated to be a_c » 1 for the reactions given in Eq 2 and 3, metal dusting may occur as a catastrophic reaction of carbon monoxide and carbon with metallic materials. Reaction kinetics of metal dusting have been reported to be controlled by the temperature and the carbon-monoxide-to-carbon-dioxide ratio in the gas (Ref 15).

Nickel-base alloys are more resistant to carbon diffusion than most other alloys, and alloys with higher nickel content are superior to low-content nickel alloys. Carburization appears to be appreciable, with a penetration of 0.2 to 0.8 mm (0.008 to 0.031 in.) in 50 to 55 days, in nickel alloys at 650 to 700 °C (1200 to 1300 °F). Nickel-chromium alloys containing titanium, niobium, and aluminum are better than basic nickel-chromium alloys in carbon dioxide atmospheres at 700 to 800 °C (1300 to 1470 °F). The alumina scale-forming alloys appear to be much more resistant to carburization than the chromia scale-formers (Ref 16).

Nitridation of alloys in ammonia environments is well known in ammonia and heat treating industries (Ref 17). Nitridation attack by N_2 is known in the powder metal industry. It has been shown that the kinetics of nitridation in N_2 are much higher than found in contact with NH_3 at high temperatures (980 to 1090 °C, or 1800 to 2000 °F) (Ref 18). Heat treating furnaces and accessories for heat treating powder metal products are typically exposed to N_2 or N_2-H_2 atmospheres at elevated temperatures (~1090 °C, or 2000 °F). Both stainless steel and nickel-base alloys may suffer from severe nitridation attack, resulting in loss of creep strength and/or ductility. Alloys are susceptible to nitridation attack in combustion environments even when such environments are oxidizing. Increasing iron in the alloy increases its susceptibility to nitridation attack. Titanium and aluminum are also detrimental to nitridation attack (Ref 18).

Carbon-Nitrogen Interaction. This type of high-temperature interaction can be found in centrifugally cast furnace tubes of alloys HK-40 (26Cr-20Ni) and HL-40 (30Cr-20Ni) used extensively in ethylene-pyrolysis furnaces for many years. In service, thick deposits of carbon or coke form on the intertube walls, and it is necessary to burn these deposits away periodically by means of steam-air decoking. Examination of failed furnace tubes has established an interaction between carbon diffused from the surrounding furnace atmosphere and nitrogen already present in the tubes. Metallographic examinations and chemical analyses have shown that a high nitrogen content can cause microscopic voids along grain boundaries, which in turn lead to premature failure by stress-rupture cracking. Void formation occurs only when carbon and nitrogen exceed a certain critical level.

It has been determined that, in general, the inward diffusion of carbon from the carburizing atmosphere forces the migration of nitrogen present in the tube toward the outside surface. However, there may be exceptions. In

one hydrogen-reformer furnace, for example, there was a joint inward diffusion of carbon and nitrogen, because the methane-gas feed contained 18% N. Analyses showed that the tubes contained 0.83% C and 0.156% N at the inside surface and only 0.42% C and 0.094% N at the outside surface.

Example 1: High-Temperature Degradation of a Gas Turbine Transition Duct. The presence of carbides and carbonitrides in a gas turbine transition duct is seen in Fig. 2, from an investigation reported in Ref 19. The transition duct was part of a 100 MW power-generation gas turbine. The duct was fabricated from several panels of a modified nickel alloy, IN-617. After six years of operation, two such ducts failed during the next two years, causing outages. Failure was in the form of a total collapse of the duct. Carbides and carbonitrides were found in all of the transitions examined. Oxidation, oxide penetration, and oxide spallation also caused thinning of the duct wall. It was felt that the high oxygen and nitrogen partial pressures of the gases within the duct, combined with the high temperatures, facilitated nitrogen pickup. In some cases, continuous grain-boundary precipitation was observed.

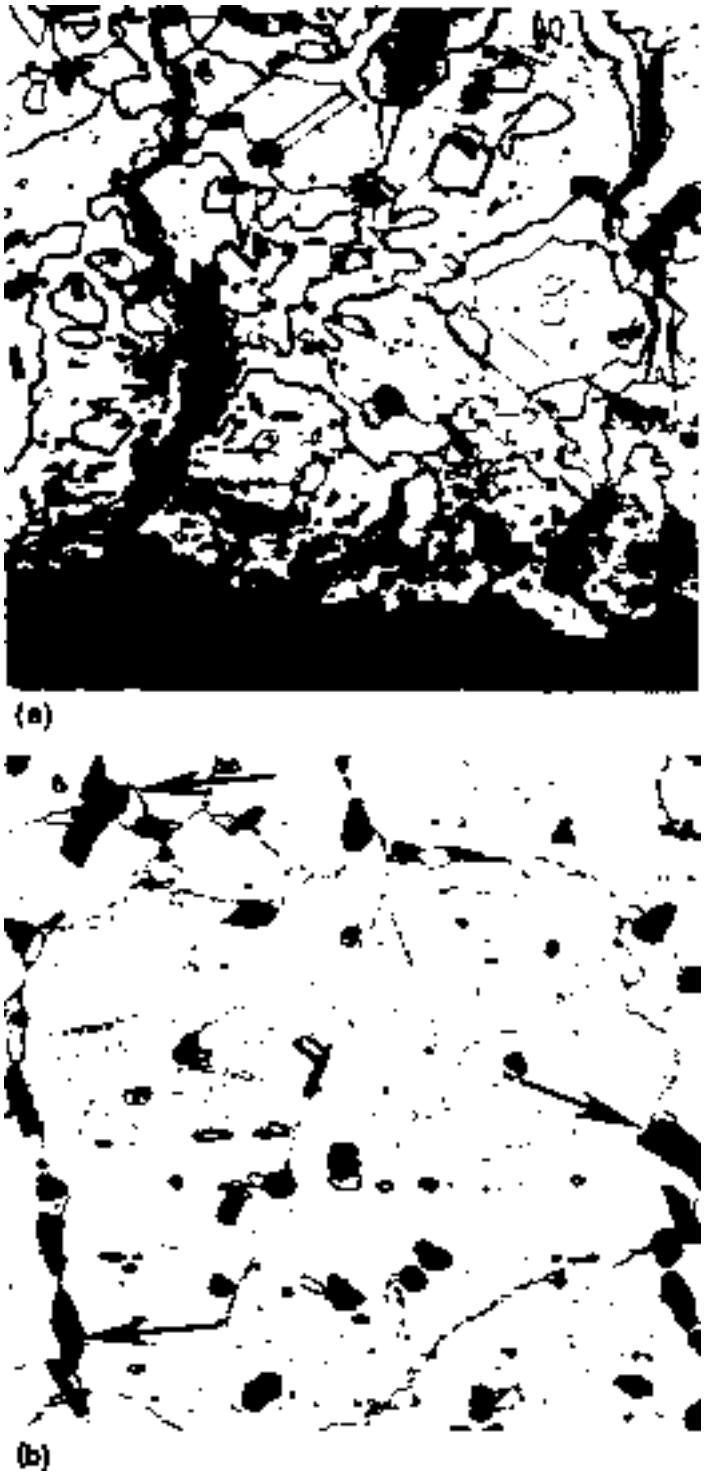


Fig. 2 High-temperature degradation of a gas turbine transition duct. (a) Carbide, carbonitride precipitates, and oxide pentration along grain boundary. (b) Creep cracking along grain-boundary

precipitates (arrows) on IN-617 panel. Creep cavities along grain boundaries link up and lead to preferential cracking. Both micrographs show an area approximately 0.38 mm (0.015 in.) wide. Source Ref 19

Sulfidation is a reaction of a metal or alloy with a sulfur-containing species to produce a sulfur compound that forms on or beneath the surface of the material. Sulfidation may involve the formation of oxides plus sulfides in reducing gases such as hydrogen-hydrogen sulfide mixtures. In most environments, Al_2O_3 or Cr_2O_3 form preferentially to the sulfides, but destructive sulfidation attack occurs at oxide crack sites (Ref 20). The reduced species, H_2S, is more corrosive than the oxidized species, SO_2. Gaseous reactions or molten phases, such as low-melting-point metal-sulfides eutectics, may cause sulfidation. Sulfidation of nickel results in the formation of eutectic Ni_3S_2 that has a melting point of 635 °C (1175 °F), with sulfur solubility of 0.005% (Ref 21). The eutectic formation causes the initiation and propagation of intergranular corrosion. The sulfidation rate of metals and alloys is usually many times higher than the corresponding oxidation rate. These metal sulfides generally are more complex and have lower melting temperatures than the corresponding oxides (Ref 6).

Adding chromium improves resistance to sulfide attack and increases the temperature limit in a sulfidizing environment. Chromium sulfide formed below the surface of the protective oxide layer may lead to depletion of chromium and breakaway corrosion. Once sulfur has reacted with the alloy, sulfur tends to preferably react with chromium or aluminum to form sulfides; this may cause redistribution of the scale-forming elements or interfere with the formation of protective oxide scales. Nickel-chromium alloys have been successfully used at service temperatures of 750 to 1200 °C (1400 to 2200 °F). Manganese, silicon, and aluminum additions improve the resistance of nickel alloys to sulfur attack (Ref 21).

Scaling rates of high-temperature stainless steels are higher in dry sulfur dioxide environments than in air, due to the formation of chromium sulfides below the protective chromia scale. The introduction of water vapor into air/sulfur dioxide environments accelerates sulfidation attack (Ref 22). Sulfidation of Ni-18Cr and Ni-28Cr alloys was observed to initiate by the formation of NiS, which prevented the continuous formation of Cr_2O_3 at sites with sufficient sulfur concentrations to form NiS (Ref 23). The higher chromium content of the Ni-28Cr alloy reduced the substrate area that was depleted of chromium (Ref 23).

Sulfur can be transported through the Al_2O_3 and Cr_2O_3 protective scales under certain conditions, resulting in discrete sulfide precipitates observed at the scale/alloy interface (Ref 24). Sulfide concentrations above approximately 10 to 20 ppm have been shown to modify the adhesion strength and cyclic oxidation resistance of the protective oxide scales. Generally, CoCr binary alloys have a resistance to sulfidation that is superior to that of NiCr binary alloys because of the slower outward diffusion of cobalt ions through the scale during sulfidation and the relatively high melting point of the Co-Co_4S_3 eutectic (Ref 25).

Example 2: Failure of a Nickel-Base Alloy Incinerator Liner. An alloy IN-690 (N06690) incinerator liner failed after only 250 h of service burning solid waste. The root cause of the failure was overfiring during startup and sulfidation of the nickel-base alloy. Figure 3 shows the perforation of the liner near a patch made of alloy 160. The original wall was 6.35 mm (0.250 in.) and was reduced to 1.27 mm (0.050 in.) thickness. The accelerated oxidation of the liner is shown in Fig. 4. Sulfidation is seen in Fig. 5. This sulfidation weakens the sheet, allowing the cracks to propagate. The thickness is approximately 0.8 mm (0.031 in.). A more detailed view of the sulfidation (Fig. 6) reveals the formation of chromium sulfides (gray area indicated by arrow) by the sulfur, which diffused in the alloy along the grain boundaries.

Fig. 3 Perforation near alloy 160 patch showing that the original 6.35 mm (0.250 in.) wall thickness had been reduced to 1.27 mm (0.050 in.) or less in the general area of failure

Fig. 4 Incinerator environment has led to accelerated oxidation of the IN-690 liner approximately 100 to 150 µm deep. Oxidation first initiates along intergranular paths. Width represents approximately 0.572 mm (0.0225 in.)

Fig. 5 Sulfidation penetration into IN-690 liner approximately 50 to 250 µm deep. The sulfidized weakened structure of the alloy has led to cracking.

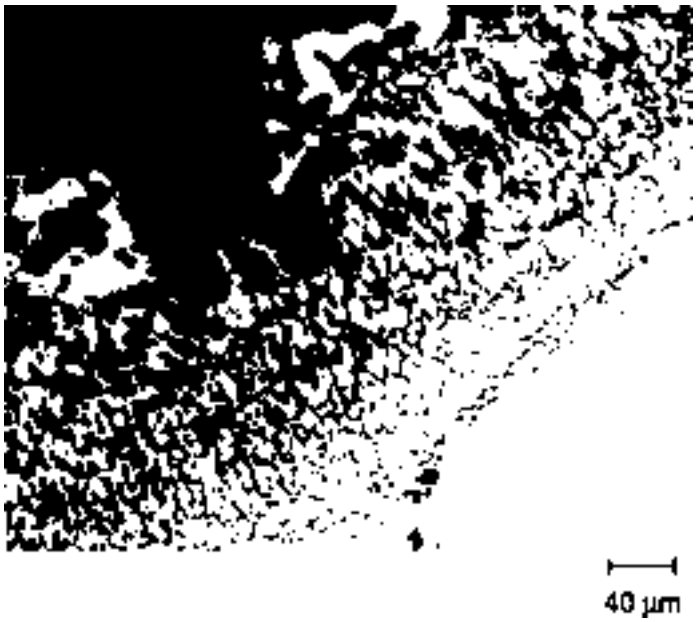


Fig. 6 Formation of chromium sulfides (gray areas, such as marked by arrow) along the surface, caused by diffusion of sulfur species along the grain boundaries of IN-690 liner. As expected, sulfide concentration decreases with depth, due to diffusion limitations. Precipitates formed along the diffusion front.

Example 3: Perforation of a Nickel-Base Alloy Kiln (Ref 26). A kiln, 7.6 m (25 ft) long with a 1 m (3 ft) internal diameter and a 6.3 mm (0.25 in.) wall thickness, is used to regenerate spent charcoal returned by water utilities. This charcoal contains up to 0.57% S and 2.04% Cl. The kiln is made of Inconel 601 (N06601) welded using Inconel 617 (N06617) as a filler alloy. Wet charcoal is fed in at one end of the kiln and travels while being tumbled within the inclined rotating vessel. Temperatures range from 480 °C (900 °F) (Zone 1) to 900 °C (1650 °F) (Zones 2 and 3). Steam is introduced at the discharge end at 95 g/s (750 lb/h), 34 to 69 kPa (5 to 10 psi), and 125 °C (260 °F). The kiln developed perforations within eight months of operation.

Discussion. The sulfur and chlorine in the charcoal attacked the Inconel 601, forming various sulfides and chlorides (Fig. 7). The tendency is for the sulfur to attack the nickel first and then the chromium to form their respective sulfides. Figure 8 shows the light yellow (light gray) nickel sulfide grains and the dark chromium sulfide grains in a bright chromium-depleted matrix. The accompanying energy-dispersive spectroscopy spectra (Fig. 9) show the chemical makeup of these phases.

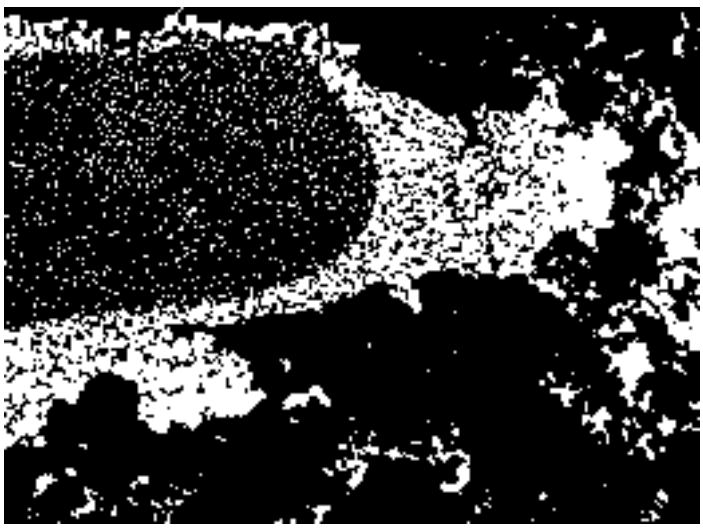


Fig. 7 Sulfidation and chloridation attack on nickel alloy of charcoal-regeneration kiln. See also Fig. 8. Region 1 is an area of chromium sulfide islands (dark phase) interspersed in chromium-depleted region (bright phase). Region 2 has angular phase (consisting mostly of nickel sulfide) in the chromium-depleted region (bright phase). Dark, outer region contains cavities, corroded metal, and various debris, including mounting compounds. Far left region (not shown) is base metal. Micrograph shows an area approximately 6 mm (0.24 in.) wide. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

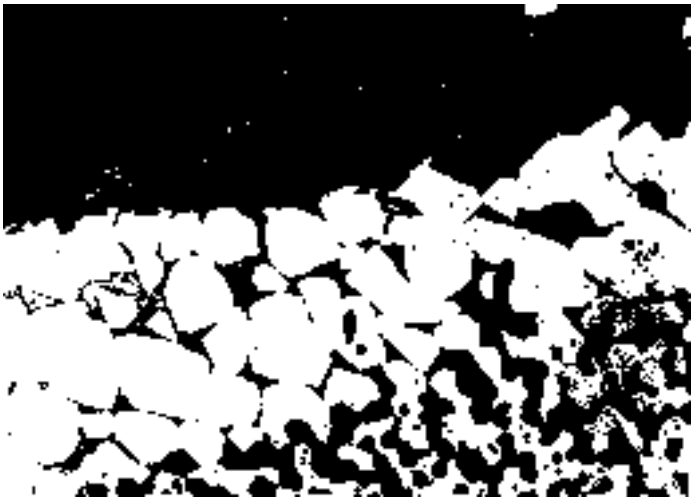


Fig. 8 Sulfidation and chloridation attack on nickel alloy of charcoal-regeneration kiln, with greater magnification (at ~44×). Lower right is region of chromium sulfide islands (dark phase) interspersed in chromium-depleted region (bright phase). Middle region has angular phase (consisting mostly of nickel sulfide) in the chromium-depleted region (see energy-dispersive spectroscopy output in Fig. 9). Dark regions at top of the photo are cavities, corroded metal, and various debris. Micrograph shows an area approximately 1.5 mm (0.06 in.) wide. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

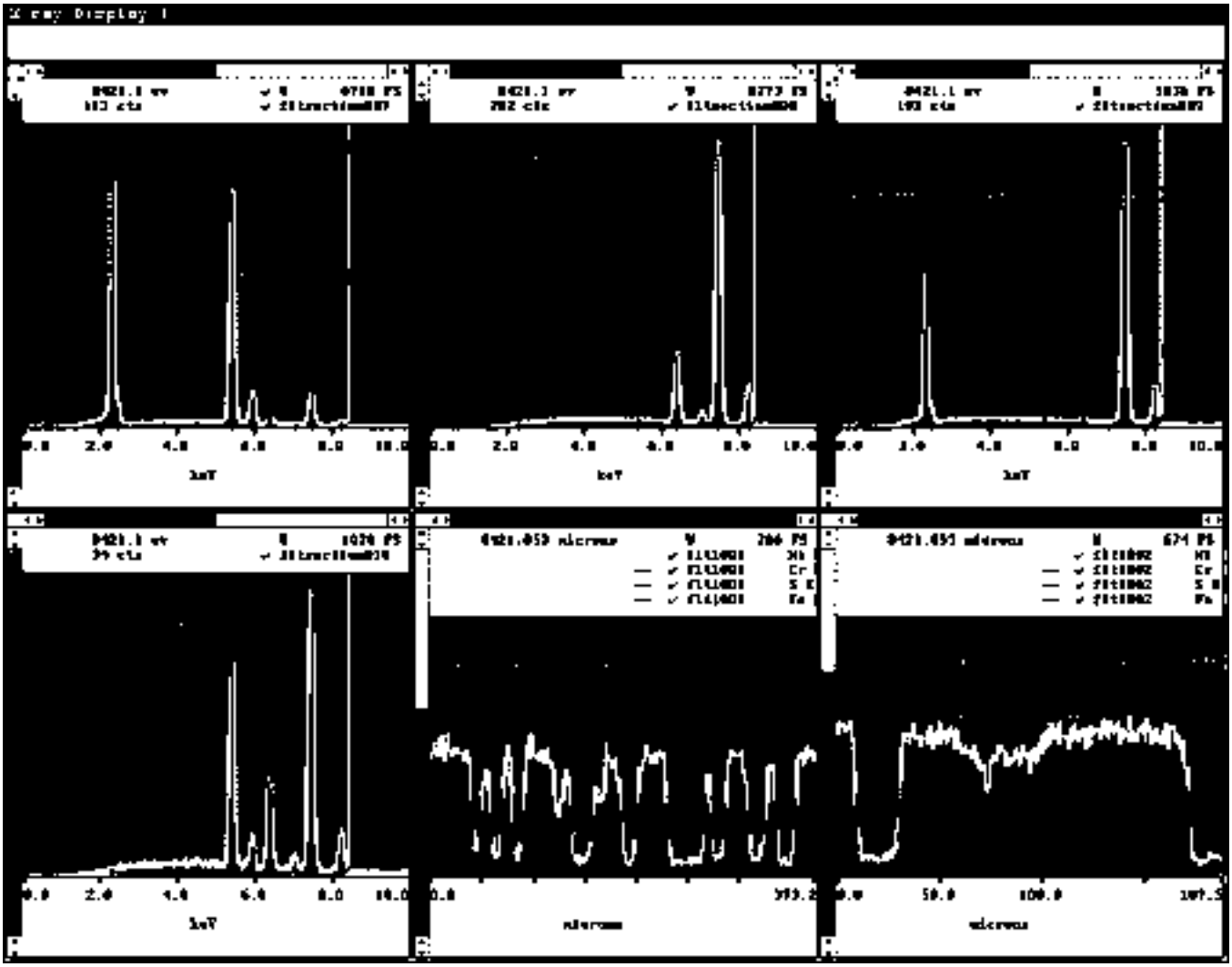


Fig. 9 Energy-dispersive spectroscopy corresponding to areas of Fig. 8. Courtesy of Mohan Chaudhari, Columbus Metallurgical Services

Little or no chlorides were detected in the corrosion debris, because these compounds have low melting and sublimation temperatures. At the kiln operating temperatures, they are believed to have been lost to off-gassing.

It was concluded that Inconel 601 of the kiln was affected by severe sulfidation attack. The welds, where Inconel 617 was used, were not attacked because of the higher resistance to sulfidation of the alloy. Inconel 617 has lower nickel content and higher cobalt and molybdenum content.

Recommendations. On-site testing was suggested, and test coupons of various alloys were installed before fabricating another kiln. The suggested alloys were RA85H, 800HT, HR-120, Haynes 556, and HR-160 (listed in ascending order of estimated life).

Hot corrosion is a complex process involving both sulfidation and oxidation (Ref 27). Hot corrosion is a form of accelerated oxidation that affects alloys and coatings exposed to high-temperature gases contaminated with sulfur and alkali metal salts (Ref 28). These contaminants combine in the gas phase to form alkali metal sulfates; if the temperature of the alloy is below the dewpoint of the alkali sulfate vapors and above the sulfate melting points, molten sulfate deposits are formed (Ref 28). Molten sodium sulfate is the principal agent in causing hot corrosion (Ref 29, 30).

Chloride salts can act as a fluxing agent and dissolve protective oxide films. Air in a marine environment ingested into the combustion zone of any marine gas turbine engine is laden with chlorides, unless properly filtered. Sodium chloride has been viewed as an aggressive constituent in the hot corrosion of gas turbine components in the marine environment (Ref 31, 32, 33).

Two forms of hot corrosion exist. Type 1, high-temperature hot corrosion, occurs through basic fluxing and subsequent dissolution of the normally protective oxide scales by molten sulfate deposits that accumulate on the surfaces of high-temperature components such as hot section turbine blades and vanes. High-temperature hot corrosion usually occurs at metal temperatures ranging from 850 to 950 °C (1560 to 1740 °F). Type 1 hot

corrosion involves general broad attack caused by internal sulfidation above 800 °C (1470 °F); alloy depletion is generally associated with the corrosion front. This attack involves molten ash and deposits containing sodium sulfate and/or potassium sulfate salts. Very small amounts of sulfur and sodium or potassium can produce sufficient Na_2SO_4. In gas turbine environments, a sodium threshold level below 0.008 ppm by weight precluded type 1 hot corrosion. Other impurities, such as vanadium, phosphorus, lead, chlorides, and unburned carbon, can be involved in lowering salt melting temperatures and accelerating hot corrosion. High chromium content (>25 to 30% Cr) is required for good resistance to hot corrosion. Nickel alloys with both chromium and aluminum show improved resistance to hot corrosion.

Type 2, low-temperature hot corrosion, occurs in the temperature range of 650 to 750 °C (1200 to 1380 °F), where SO_3 is relatively high or melts are deficient in the oxide ion concentration, leading to acidic fluxing that results in pitting attack. Sulfides are found in the pitted area (Ref 25). Low-temperature hot corrosion may involve a gaseous reaction of SO_3 or SO_2 with CoO and NiO, which results in pitting from the formation of low-melting mixtures of Na_2SO_4 in Ni-Cr, Co-Cr, Co-Cr-Al, and Ni-Cr-Al alloys (Ref 34). If other reactants are added, melting temperatures can be further lowered. The relative hot corrosion resistance of several alloys was evaluated in incinerator environments (Ref 20, 35, 36).

Example 4: Failure of Gas Turbine Last- Stage Bucket. Turbine buckets in a 37.5 MW gas turbine made of Udimet 500 superalloy failed in service (Ref 37). The power plant was located 1 km (0.6 miles) from the Pacific Ocean and operates on No. 2 diesel fuel, which was supplied by tanker ship. Turbine bucket failures occurred on three units after 2500 to 6400 h of operation.

Investigation. The buckets fractured at the tip shroud. These integral tip shrouds interlock with adjacent buckets to form a continuous band. A cross section of a single foil, with its internal grain structure and loading, is shown in Fig. 10. Metallographic examination revealed that two different microstructures existed in the shroud area: columnar grains and equiaxial grains, as sketched in the figure. This is due to the geometry of the casting and its heat flow consequences during solidification. The entire fracture surface was covered with dark scale; the airfoil surface and the underside of the shroud had thick scale ranging from dark-rust color to black, with a thinner layer on top of the dark layer. This thin layer was white in some areas. Pitting was evident, especially at the shroud fillet. There was erosion-corrosion that resulted in a maximum reduction in thickness of 0.91 mm (0.036 in.) at the airfoil tip.

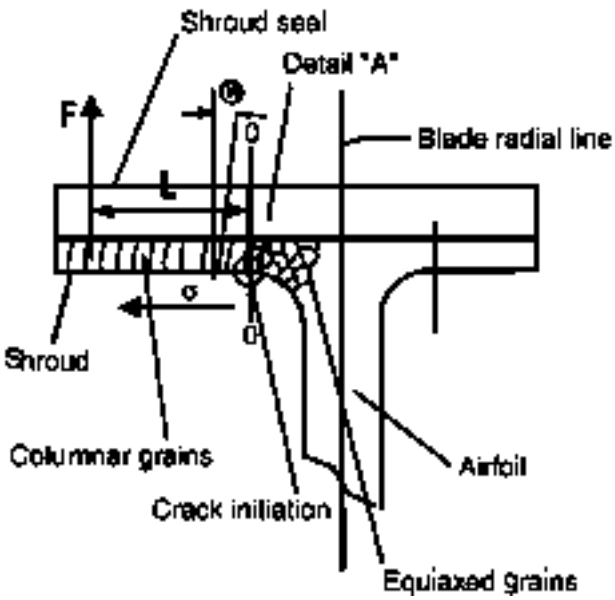


Fig. 10 Relation between microstructural elements in a bucket tip shroud and its principal stresses

Discussion. The differing microstructure of the airfoil results in changes in mechanical properties.

Type 2 hot corrosion was present at these third-stage buckets, which operate at approximately 650 °C (1200 °F). Other contributing factors were:

- Ingestion of sodium salts, which formed sodium sulfate
- Sulfur in the fuel
- Presence of potassium sulfate, other sulfates, and chlorides

- Presence of salt in air, due to coastal location

Stress analysis (finite-element modeling) on the shroud indicated that, due to the thinning by erosion-corrosion and scale formation, the stress levels increased from 430 MPa (62 ksi) at the central portion (where the crack occurred) to 550 MPa (80 ksi). From Fig. 11, it is seen that this change in stress would reduce the time to rupture by a factor of 1000. The columnar grains on the shroud tips are perpendicular to the applied force and intersected the entire thickness of the shroud. Fine-grain surface structure was lost to the type 2 corrosion. The long grain boundaries provided an easy path for crack initiation and propagation. The transition from columnar to equiaxial grains caused a step change in stress. Because normal operation includes cycling of loads and temperatures, the shroud tip fractured due to thermomechanical fatigue in its degraded state.

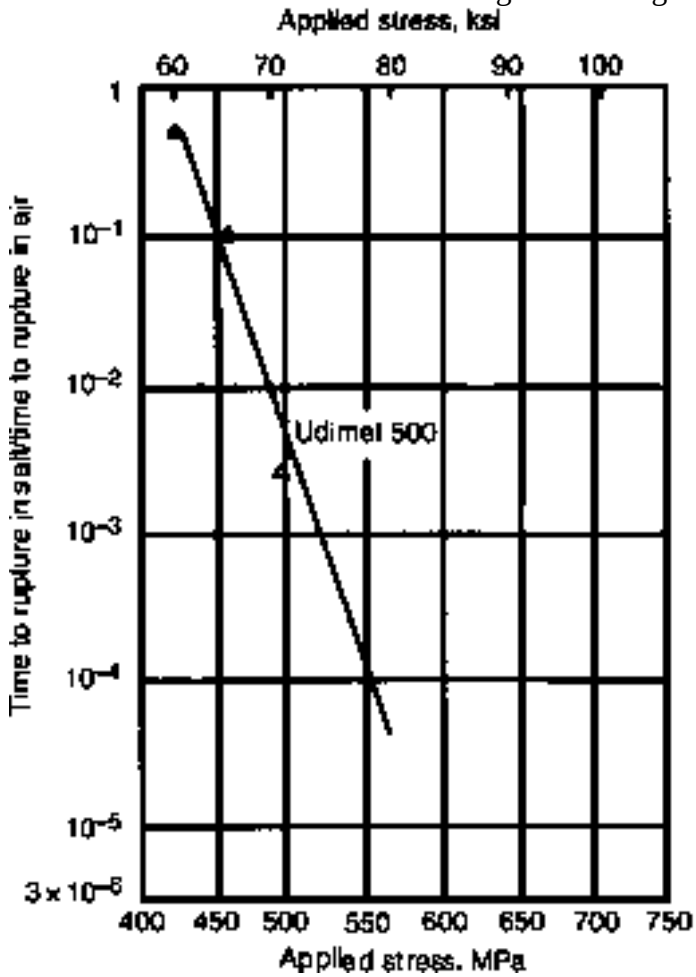


Fig. 11 Degradation of rupture for Udimet 500 due to hot corrosion at 705 °C (1300 °F)

Recommendations. Gas turbines operating in a marine environment with operating temperatures in the range of type 2 corrosion (650 to 750 °C, or 1200 to 1380 °F) should have special chromium- or silicon-rich coating to minimize corrosion. Fuel delivery, handling, and treatment should be high quality, to maintain fuel contamination within design limits. Inlet air filtration must be designed for the coastal site. If possible, changing the bucket tip by increasing its thickness and changing the casting technique to eliminate the columnar grain structure would reduce the stress and make the design more tolerant of corrosion.

Chloridation. Chlorides accumulate rapidly on metallic surfaces of test samples. Typical deposits contained 21 to 27% Cl when the flue gas contained 40 to 140 ppm HCl (Ref 38). Municipal wastes were characterized as having a 0.5% halide dry content, of which 60% was derived from organic, polymer sources (Ref 39). Chloride salts have melting temperatures as low as 175 °C (350 °F), which can act as a fluxing agent that dissolves protective oxide films. High-temperature components exposed to a marine environment are laden with chlorides. Molten $SnCl_2$ (melting point, 246 °C, or 475 °F), $SnCl_2$ + NaCl (melting point, 199 °C, or 390 °F), $ZnCl_2$ (melting point, 283 °C, or 541 °F), eutectic $PbCl_2/FeCl_3$ (melting point, 175 °C, or 347 °F), and eutectic $ZnCl_2$/NaCl (melting point, 262 °C, or 504 °F) (Ref 40) may cause rapid corrosion of carbon steel, if present. Attack by halogens at elevated temperatures occurs through the volatility of the reaction products. Oxides that

form in combustion gases are porous and prone to fracture. Stainless steels are generally passive, but surface pitting may occur in chloride-containing environments. Nickel-base alloys can be expected to have superior corrosion resistance, as compared to stainless steel alloys (Ref 41). Clay containing aluminum silicate may inhibit chloride-related corrosion by raising the melting points of chloride salts through the formation of sodium aluminum silicates, which expel HCl and SO_3 (Ref 42). Problems with chlorides can be mitigated if plastics and other sources of halogens are removed or minimized from the waste stream. Increasing the oxygen content and adding water vapor have also reduced the corrosion rate of various alloys by chlorides in a simulated waste incinerator environment (Ref 38).

Hydrogen Interactions. Steam may decompose on metal surfaces at elevated temperatures to form hydrogen and oxygen. Selected applications may be able to produce atomic hydrogen and atomic oxygen. Loss in tensile ductility of steels and nickel-base alloys has been observed in gaseous environments with a total hydrogen content of 0.1 to 10 ppm at -100 to 700 °C (-150 to 1300 °F). Hydrogen attack or hydrogen damage occurs when gaseous hydrogen diffuses into the steel and reacts with the carbides to form methane, $4H + Fe_3C \bullet CH_4 + 3Fe$, and internal microcracks that lead to brittle rupture. The reaction rate is dependent on hydrogen concentration, diffusion, and total gaseous pressure and occurs at 200 to 600 °C (390 to 1110 °F). Hydrogen damage has been observed in utility boilers as low as 316 °C (600 °F) (Ref 43). Atomic hydrogen diffuses readily in steel, which may cause cracking and decarburization of steel. Hydrogen damage can occur in other high-strength alloys, resulting in loss of tensile ductility. Nickel alloys are much less susceptible to hydrogen damage than ferrous-base alloys.

Hydrogen damage also results from fouled heat-transfer surfaces. There is some disagreement as to whether hydrogen damage can occur only under acidic conditions or whether it can happen under alkaline and acidic conditions as well. Hydrogen damage may occur from the generation of atomic hydrogen during rapid corrosion of the waterside tube surface, although it may occur with little or no apparent wall thinning. The atomic hydrogen diffuses into the tube steel, where it reacts with tube carbides (Fe_3C) to form gaseous methane (CH_4) at the grain boundaries. The larger methane gas molecules tend to concentrate at the grain boundaries. When methane gas pressures exceed the cohesive strength of the grains, a network of discontinuous, intergranular microcracks is produced. Often, a decarburized tube microstructure is associated with hydrogen damage as observed by metallographic examination.

Hydrogen damage has been incorrectly referred to in the literature and in practice as hydrogen embrittlement; actually, the affected ferrite grains have not lost their ductility. However, because of the microcrack network, a bend test indicates brittle-like conditions. Hydrogen damage is one problem that began to occur when one failure mechanism (caustic embrittlement) was remedied, and the thrust of the utility industry moved toward higher-pressure boilers (Ref 43). Hydrogen damage and caustic gouging are experienced at similar boiler locations and, usually, under heavy waterside deposits. Hydrogen damage from low-pH conditions may be distinguished from high-pH conditions by considering the boiler-water chemistry. A low-pH condition can be created when the boiler is operated outside of normal recommended water chemistry parameter limits. This is caused by contamination such as condenser in-leakage (e.g., seawater or recirculating cooling water systems incorporating cooling towers), residual contamination from chemical cleaning, and the inadvertent release of acidic chemicals into the feedwater system. A mechanism for concentrating acid-producing salts (departure from nucleate boiling, deposits, waterline evaporation) must be present to provide the low-pH condition. Low pH dissolves the magnetite scale and may attack the underlying base metal through gouging:

$$M^+Cl^- + H_2O = MOH(s) + H^+Cl^-$$

Hydrogen damage may be eliminated by proper control of the water chemistry and by removal of waterside deposits where concentrating boiler solids could occur.

In high pH conditions, concentrated sodium hydroxide dissolves the magnetite according to the following reaction (Ref 43).

$$4NaOH + Fe_3O_4 \rightarrow 2NaFeO_2 + Na_2FeO_2 + 2H_2O$$

After the protective oxide is destroyed, water or sodium hydroxide can react with iron to form atomic hydrogen.

$$1)\ 3Fe + 4H_2O \rightarrow Fe_3O_4 + 8H$$

$$2)\ Fe + 2NaOH \rightarrow Na_2FeO_2 + 2H$$

Nickel and nickel-base alloys are susceptible in gaseous hydrogen environments. The same factors that affect hydrogen interactions in ferrous alloys are also operative for nickel-base alloys, although to a slightly lesser degree, because face-centered cubic (fcc) metals and alloys have a greater number of slip planes and have reduced solubilities for hydrogen compared to body-centered cubic alloys (Ref 44). Hydrogen in nickel-base alloys may lead to intergranular, transgranular, or quasi-cleavage cracking. The Fe-Ni-Cr (Incoloy) and Inconel alloys show ductility reductions when exposed to hydrogen, particularly age-hardenable alloys (Ref 44).

Molten metal corrosion may cause dissolution of an alloy surface directly, by intergranular attack, or by leaching. Liquid metal attack may also initiate alloying, compound reduction, or interstitial or impurity reactions. Carbon and low-alloy steels are susceptible to various molten metals or alloys, such as brass, aluminum, bronze, copper, zinc, lead-tin solders, indium, and lithium, at temperatures from 260 to 815 °C (500 to 1500 °F). Plain carbon steels are not satisfactory for long-term use with molten aluminum. Stainless steels are generally attacked by molten aluminum, zinc, antimony, bismuth, cadmium, and tin (Ref 45). Nickel, nickel-chromium, and nickel-copper alloys generally have poor resistance to molten metals such as lead, mercury, and cadmium. In general, nickel-chromium alloys also are not suitable for use in molten aluminum (Ref 46). Liquid metal embrittlement (LME) is a special case of brittle fracture that occurs in the absence of an inert environment and at low temperatures (Ref 47). Decreased stresses can reduce the possibility of failure in certain embrittling molten alloys. Stainless steels suffer from LME by molten zinc. Small amounts of lead embrittle nickel alloys, but molybdenum additions appear to improve lead LME resistance. There are no known metals or alloys totally immune to attack by liquid aluminum (Ref 48). The selection of fabricating processes must be chosen carefully for nickel-base superalloys. Liquid metal embrittlement can occur when brazing precipitation-strengthened alloys such as Unified Numbering System (UNS) N07041 (Ref 49). Many nickel superalloys crack when subjected to tensile stresses in the presence of molten (boron-silver) brazing filler alloys.

Molten salts are often involved in sulfidation, chloridation, and hot corrosion, as discussed previously. The type of environment and the component metal temperatures are important factors in the promotion of fireside or external corrosion in tube steels in boilers. The corrosiveness of the environment depends on the surface temperature and the condition of and/or the corrosive ingredients in the medium.

In the presence of molten ash products, the oxide, even in oxidizing environments, becomes unstable and dissolves. Alkali sulfates deposited on the fireside surfaces of boilers may react with SO_3 or SO_2 to form mixtures of alkali pyrosulfates (m.p. 400–480 °C) or alkali-iron trisulfates (550 °C, or 1030 °F) that cause fireside corrosion of reheater and superheater tubes (Ref 35). Molten sodium pyrosulfates ($Na_2S_2O_7$: m.p. 400 °C, or 750 °F) or potassium pyrosulfates ($K_2S_2O_7$: m.p. <300 °C, or <570 °F) react with carbon steels by dissolving the protective oxide film and then reacting with iron to cause accelerated wastage.

Waterwall fireside corrosion may develop when incomplete fuel combustion causes a nonoxidizing, reducing condition. Incomplete combustion causes the release of volatile sulfur and chloride compounds, which causes sulfidation and accelerated metal loss. Poor combustion conditions and steady or intermittent flame impingement on the furnace walls may favor an environment that forms sodium and potassium pyrosulfates and has a melting point below 425 °C (800 °F). Chlorides increase the corrosion rate.

Metal attack occurs along the crown of the tube and may extend uniformly across several tubes. Corroded areas are characterized by abnormally thick iron oxide and iron sulfide scales. Corrosion in supercritical boilers produces circumferential grooving or cracking in the waterwall external surface, with fingerlike penetrations into the tube wall. Verification of waterwall fireside corrosion involves analyzing the fuel, the completeness of combustion, and the evenness of heat transfer. Carbon content in the ash that is greater than 3% is indicative of corrosive combustion conditions.

Reducing conditions tend to lower the melting temperature of any deposited slag, which also increases the solvation of the normal oxide scales. A stable gaseous sulfur compound under reducing conditions is H_2S. Under the reducing environment, iron sulfide is the expected corrosion product of iron reacting with pyrosulfates.

Coal ash corrosion results when a molten ash of complex alkali-iron trisulfates forms on the external surfaces of tubes (reheater and superheater tubes) in the temperature range of 540 to 700 °C (1000 to 1300 °F). Liquid trisulfates solubilize the protective iron oxide scale and expose the base metal to oxygen, which produces more oxide and subsequent metal loss (according to a mechanism proposed by Reid, Ref 50):

$$3K_2SO_4 + Fe_2O_3 + 2SO_3 = 2K_3Fe(SO_4)_3$$

$$9Fe + 2K_3Fe(SO_4)_3 \rightleftarrows 3K_2SO_4 + 4Fe_2O_3 + 3FeS$$

$$4FeS + 7O_2 \rightleftarrows 2Fe_2O_3 + 4SO_2$$

$$2SO_2 + O_2 \rightleftarrows 2SO_3$$

$$3K_2SO_4 + Fe_2O_3 + 2SO_3 \rightleftarrows 2K_3Fe(SO_4)_3$$

The greatest metal loss creates flat sites at the interface between the fly-ash-covered half and the uncovered half (2 and 10 o'clock positions to the gas flow). Corrosive coal typically contains a significant amount of sulfur and sodium and/or potassium compounds.

Visually, ferritic steel exhibits shallow grooves (referred to as alligatoring or elephant hide). A sulfur print photographic paper dipped in a 2% sulfuric acid solution, the excess solution wiped off, and a tube cross section then placed on the paper reveals the presence of sulfides at the metal/scale interface. Coal ash corrosion reduces the effective wall thickness, thereby increasing tube stress. This, combined with temperatures within the creep range, may cause premature creep stress ruptures because of coal ash corrosion.

Several corrective options to reduce failure through coal ash corrosion in boilers are employing thicker tubes of the same material; shielding tubes with clamp-on protectors; coating with thermal sprayed, corrosion-resistant materials; blending coals to reduce corrosive ash constituents; replacing tubes with higher-grade alloys or coextruded tubing; lowering final steam outlet temperatures; redesigning affected sections to modify heat transfer; and adding $CaSO_4$ or $MgSO_4$ to bind SO_3 to a less-corrosive form.

Fireside corrosion also can occur in oil-fired boilers. Oil ash corrosion is believed to be a catalytic oxidation of the material by reaction with vanadium pentoxide. Only a few ppm of vanadium accelerates this reaction. Vanadium pentoxide and sodium oxide form a liquid with a low melting point of approximately 540 °C (1000 °F). Sodium oxide also reacts with sodium trioxide to form sodium sulfate, which also forms a low-melting-point liquid (Ref 51). Magnesium additions mitigate oil ash corrosion.

Aging Reactions. Long-term aging effects can cause metallurgical changes that alter the mechanical and corrosion properties of materials. Graphitization is a microstructural change that can occur in carbon and low-alloy steels. Pearlite (alternating ferrite and Fe_3C platelet structure) normally found in such steels transforms in time to graphite and ferrite at temperatures above 425 °C (800 °F), which causes loss of ductility and promotes embrittlement. Alloying additions of at least 0.7% Cr and 0.5% Mo to steels markedly improve resistance to graphitization. Weld sites are particularly susceptible to graphitization. Heat treatments or operating conditions that heat ferritic and austenitic stainless steels to 500 to 900 °C (930 to 1650 °F) cause segregation at the grain boundaries by the precipitation of complex chromium carbides, the formation of sigma phase, or the precipitation of other deleterious phases. The properties of precipitation-strengthened superalloys (processed at 535 to 815 °C, or 1000 to 1500 °F) such as UNS N07041 may be altered by brazing (Ref 49). Normal welding can induce susceptibility of stainless steels without postweld heat treatment. Careful control of heat treatment can avoid intergranular corrosion of unstabilized stainless steels containing more than 0.03% C. Nickel-base superalloys are also susceptible to grain-boundary segregation. Nickel-base superalloys can be embrittled by less than 20 ppm sulfur segregation to alloy grain boundaries.

Grain-boundary segregation may result in localized corrosion if the grain boundaries are not as resistant as the alloy matrix or if segregation causes depletion of corrosion-resistant elements. Pitting is another form of localized corrosion. Both dissolved oxygen and carbon dioxide can promote pitting of carbon steels. An alkaline pH can diminish corrosion, with the evolved gases likely to be neutral to acidic. Ash deposition, thermally induced stress, and particulate erosion may cause conditions that accelerate localized corrosion by cracking the protective oxide layer.

Iron, nickel, and cobalt superalloys owe their unique strength, up to a fairly high threshold temperature, to the presence of strengthening particles such as inert oxide dispersoids, carbides, coherent ordered precipitates, and solid-solution strengthening agents in a fcc (with appropriate nickel additions) structure. Introduction of precipitates into the alloy through solid-solution reactions requires specific alloy chemistries and heat treatments to achieve the desired mechanical properties (Ref 46, 52). However, at prolonged exposures to high temperatures (760 to 1000 °C, or 1400 to 1830 °F), the phase rule alone suggests that a large number of new phases could nucleate within multicomponent superalloys to form a number of microstructural instabilities (Ref 53). These instabilities can include transformation of a metastable phase to a more stable but incoherent phase,

as in carbide transformation; stress-induced differential diffusion of solute atoms (coarsening of $M_{23}C_6$ at grain boundaries); or dissolution of strengthening precipitates in critical areas of the microstructure. Structural instabilities of superalloys at elevated temperatures can also include the formation of new phases, such as sigma, mu, and Laves phases, which are brittle and possess cracklike morphologies that can adversely affect alloy tensile and stress-rupture ductility and impair resistance of the alloy to crack growth (Ref 53).

Intergranular corrosion occurs at or adjacent to grain boundaries, with little corresponding corrosion of the grains. Intergranular corrosion can be caused by impurities at the grain boundaries or enrichment or depletion of one of the alloying elements in the grain-boundary area. Austenitic stainless steels, such as type 304, become sensitized or susceptible to intergranular corrosion when heated in the range of 500 to 800 °C (930 to 1470 °F). In this temperature range, $Cr_{32}C_6$ precipitates deplete chromium from the area along the grain boundaries (below the level required to maintain stainless properties). The depleted area is a region of relatively poor corrosion resistance. Controlling intergranular corrosion of austenitic stainless steels occurs through employing high-temperature solution heat treatment; adding elements (stabilizers) such as titanium, columbium, or tantalum that are stronger carbide formers than chromium; and lowering the carbon content below 0.03%.

Protective coatings discussed later in this chapter and superalloy substrates generally have widely diverse compositions due to the intended balancing between mechanical properties and corrosion resistance. Depending on the actual temperature, chemical gradients between the coating and superalloy substrate may cause interdiffusion phenomena to occur. Interdiffusion between the coating and the substrate (Ref 54) can modify the oxidation and corrosion resistance of the coating and the mechanical properties of the coating/substrate system. Interdiffusion slowly changes the chemical composition of both the coating and the alloy and subsequently, may alter the microstructure of the coating and/or the substrate. The formation of metal sulfides in the protective coating may eventually lead to diffusion into the substrate via grain boundaries or other relatively easy diffusion pathways in the coating that may alter the long-term performance of a coating/alloy system in a hot corrosion environment compared to that observed in an oxidation environment.

Interdiffusion is a complex process that requires use of multivariable diffusion couples in a phase-field approach (Ref 55) or knowledge of diffusion pathways in multicomponent phase diagrams (Ref 56) to model the process and provide an understanding of the contributions to interdiffusion. These methods are based on Onsager's formalism of Fick's law, which includes diffusion interactions among the chemical species (Ref 57).

Environmental cracking can occur with a wide variety of metals and alloys in specific environments. Environmental cracking is defined as the spontaneous, brittle fracture of a susceptible material (usually quite ductile itself) under tensile stress in a specific environment over a period of time. Stress-corrosion cracking (SCC) and corrosion fatigue are some of the forms of environmental cracking that can lead to failure.

Stress-corrosion cracking results from the conjoint, synergistic interaction of tensile stress and a specific corrodent for the given metal. The tensile stress may be either applied (such as caused by internal pressure) or residual (such as induced during forming processes, assembly, or welding). Stress-corrosion cracking involves the concentration of stress and/or the concentration of the specific corrodent at the fracture site. Stress-corrosion cracking fractures may be oriented either longitudinally or circumferentially, but the fractures are always perpendicular to the stresses. In boiler systems, carbon steel is specifically sensitive to concentrated sodium hydroxide; stainless steel is sensitive to concentrated sodium hydroxide, chlorides, nitrates, sulfates, and polythionic acids; and some copper alloys are sensitive to ammonia and nitrites. Stress-corrosion cracking failures produce thick-walled fractures that may be intergranular, transgranular, or both. Branching is often associated with SCC. Normally, gross attack of the metal by the corrodent is not observed in SCC failures. Failures caused by SCC are difficult to see with the naked eye. Metallographic and chemical analyses are performed to identify the constituents in the alloy and the bulk corrosion products. Stress-corrosion cracking failures frequently have been experienced immediately after chemical cleanings and initial startups.

Copper alloys, specifically brasses, are susceptible to SCC. Most failures are steamside failures that have occurred where high concentrations of ammonia and oxygen exist, such as in the air removal section of the condenser. Stress-corrosion cracking failures are found often at inlet or outlet ends, where the tubes have been expanded into the tubesheet.

Stress-corrosion cracking may be reduced by either removing applied or residual stresses or by avoiding concentrated corrodents. The reduction of corrodents by avoiding boiler upsets and inleakage is generally the most effective means of diminishing or eliminating SCC. A change in tube metallurgy also may reduce SCC.

Stress-corrosion cracking is a brittle fracture of an otherwise ductile material that occurs under the combined action of a residual or applied tensile stress and an alloy-specific corrodent. Alloy composition, heat treatment,

the exposure time, temperature, and solution media affect SCC. Fluctuating temperatures can accelerate SCC of austenitic stainless steels in the presence of chlorides, particularly at high stress points from design, operation, or fabrication. Nickel alloys have improved resistance to environmental cracking over steels and stainless steels, although nickel alloys are not immune to SCC. Stress-corrosion cracking of certain nickel alloys has occurred in high-temperature halogen ion environments, high-temperature waters, and high-temperature alkaline systems and environments containing acids and H_2S. Stress-corrosion cracking is dependent on the stress level, chloride concentration, alloy composition, and temperature. Susceptible temperature ranges vary from 175 to 330 °C (350 to 625 °F). Higher temperatures tend to lower the chloride threshold concentration and stress levels required for cracking. Ferritic stainless steels are usually not considered susceptible to SCC. Stressed martensitic stainless steel grades can crack in chloride environments.

Fatigue. Boiler tube cyclic stresses are stresses periodically applied to boiler tubes that can reduce the expected life of the tube through the initiation and propagation of fatigue cracks. The environment within the boiler would suggest that corrodents would also interact with the tube to assist in this cracking process. The number of cycles required to produce cracking is dependent on the level of strain and the environment. Vibration and thermal fatigue failures describe the type of cyclic stresses involved in the initiation and propagation of these fractures.

Vibration fatigue cracks originate and propagate as a result of flow-induced vibration. This occurs where tubes are attached to drums, headers, walls, seals, or supports. Circumferential orientations are common to vibration fatigue cracks. Crack initiation sites generally occur on the fireside (external) tube portions.

Thermal fatigue cracks develop from excessive strains induced by rapid cycling and sudden fluid temperature changes in contact with the tube metal across the tube wall thickness. This can be caused by rapid boiler startups above proper operation parameter limits. Water quenching by spraying from condensate in the sootblowing medium and from the bottom ash hoppers also can induce thermal fatigue cracks. Sudden cooling of tube surfaces causes high tensile stresses, because cooled surface metal tends to contract; however, this metal becomes restricted by the hotter metal below the surface. Procedures that can reduce these sudden cyclic temperature gradients diminish or eliminate thermal fatigue cracking.

Corrosion fatigue failures are caused by the combined effects of a corrosive environment and cyclic stresses of sufficient magnitude. Corrosion fatigue usually lowers the minimum stress level and/or the number of cycles that will cause failure than in an inert environment. Corrosion fatigue cracks may develop at stress concentration sites (stress raisers) such as pits, notches, or other surface irregularities. Corrosion fatigue is commonly associated with rigid restraints or attachments. Corrosion fatigue cracking most frequently occurs in boilers that operate cyclically. The fracture surface of a corrosion fatigue crack is thick-edged and perpendicular to the maximum tensile stress. Multiple parallel cracks are usually present at the metal surface near the failure. Microscopic examination detects straight, unbranched cracks. The cracks often are wedge-shaped and filled with oxide. The oxide serves to prevent the crack from closing and intensifies the stresses at the crack tip during tensile cycles, thereby assisting in the crack growth.

Corrosion fatigue cracking can be reduced or eliminated by controlling cyclic tensile stresses (reducing or avoiding cyclic boiler operation or extending startup and shutdown times), redesigning tube restraints and attachments where differential expansion could occur, controlling water chemistry to reduce the formation of stress raisers such as pits, and removing residual stresses by heat treatment, if possible, by altering the alloy either by chemistry or microstructural changes by directional solidification or single crystal production, or applying a compressive strain on the outer surface of a alloy component by peening or similar process.

Frequent cycling may reduce the design life of high-temperature components from that predicted for operation at steady state. Cyclic temperature conditions accelerate attack mainly through the loss of the protective scale (Ref 4). Depending on the severity of the thermal cycles on a high-temperature component, spallation of thermal barrier, chromized, and aluminized coatings is probable. In certain instances, such as in nickel-base superalloys in high-temperature oxidizing environments, the aggressive conditions can slow the fatigue fracture propagation. Stainless steels have good corrosion fatigue limits when compared to other steels.

Erosion-corrosion is the acceleration or increase of deterioration of a metal because of the relative movement between a corrosive fluid and the metal surface. Erosion may accelerate corrosion of high-temperature components discussed previously. Generally, mechanical wear, abrasion, or abrupt changes in flow direction are involved. Metal or alloy loss results when the protective, passive surface films are damaged or worn, and rapid attack comes about. Metals or alloys that are soft and readily damaged or mechanically worn, such as copper or lead, are susceptible to erosion. Other factors controlling the rate of metal loss are related to the

quantity, impact angle, speed, and density of the eroding medium. Areas subject to erosion are pipe or tube bends, elbows and tees, valves, pumps, blowers, propellers, impellers, condenser or feedwater heater tubes, orifices, turbine blades, nozzles, wear plates, ducts and vapor lines, and equipment subject to sprays. Fireside tube surfaces within the boiler are affected by fly ash erosion, sootblower erosion, falling slag erosion, coal particle erosion, and steam erosion from another tube failure.

The appearance of erosion-corrosion is characterized by grooves, gullies, waves, and rounded holes, which usually exhibit a directional pattern. Increases in velocity may cause no increase in attack until a critical velocity is reached. Erosion-corrosion can occur on metals and alloys that are completely resistant to a particular environment at low velocities. Greater velocity may either increase or decrease attack, depending on the nature of the corrosion or passivating mechanisms. Increasing the oxygen, carbon dioxide, or hydrogen sulfide concentration in contact with the metal surface may accelerate attack of steel. Reduced attack may occur from increased velocity either by raising the diffusion or transfer of ions that diminish the stagnant surface film, or by preventing deposition of salt or dirt that could cause crevice corrosion.

Stress-Rupture Failures. Although not strictly a corrosion-related failure, stress ruptures are often involved when the temperature of a component is subjected to high temperatures where changes in the microstructure, and relatively rapid oxidation, are in play. In boiler environments, stress and temperature influence the useful life of the tube steel. The strength of the boiler tube within the creep range decreases rapidly when the tube metal temperature increases. Creep entails a time-dependent deformation involving grain sliding and atom movement. When sufficient strain has developed at the grain boundaries, voids and microcracks develop. With continued operation at high temperatures, these voids and microcracks grow and coalesce to form larger and larger cracks, until failure occurs. The creep rate increases, and the projected time to rupture decreases when the stress and/or the tube metal temperature is increased. High-temperature creep generally results in a longitudinal, fishmouthed, thick-lipped rupture that has progressed from overheating over a long period of time. High-temperature creep can develop from insufficient boiler coolant circulation, long-term elevated boiler gas temperatures, inadequate material properties, or as the result of long-term deposition. A thick, brittle magnetite layer near the failure indicates long-term overheating.

Failure from high-temperature creep can be controlled by restoring the boiler components to boiler design conditions or by upgrading the tube material (either with ferritic alloys containing more chromium or with austenitic stainless steels). Failure from overheating caused by internal flow restrictions or loss of heat-transfer capability can be eliminated by removal of internal scale, debris, or deposits through flushing or chemical cleaning.

Boiler tubes exposed to extremely high temperatures—metal temperatures of 450 °C (850 °F) and often exceeding 730 °C (1350 °F) for a brief period—also fail from overheating. The overheating may occur from a single event or a series of brief events. Short-term overheating, generally, is related to tube pluggage, insufficient coolant flow due to upset conditions, and/or overfiring or uneven firing patterns. Loss of coolant circulation may be caused by low drum water levels or by another failure located downstream in the same tube. Inadequate coolant circulation can result in departure from normal nucleate boiling in horizontal or sloped tubes when steam bubbles forming on the hot tube surface interfere with the flow of water coolant to the surface. This restricts the flow of heat away from the tube. Normally, the short-term overheating failure experiences considerable tube deformation from bulging, metal elongation, and reduction of wall thickness. Normally, the rupture is longitudinal, fish-mouthed, and thin-lipped. Often, the suddenness of the rupture bends the tube. Very rapid overheating may produce a thick-lipped failure. A metallurgical analysis can determine tube metal temperature at the moment of rupture. Heavy internal deposits or scale often are absent from a short-term overheating failure.

Because short-term overheating failure is caused most often by boiler upsets, rectifying any abnormal conditions can eliminate these ruptures. If restricted coolant flow or plugged sections are responsible, the boiler should be inspected and cleaned. Boiler regions with high heat flux may be redesigned with ribbed or rifled tubing to alleviate film boiling. Boiler operation should be monitored to avoid rapid startups, excessive firing rates, low drum levels, and improper burner operation.

References cited in this section

1. M.G. Fontana, *Corrosion Engineering*, 3rd ed., McGraw-Hill, 1986, p 526

2. S.C. Stultz and J.B. Kitto, Ed., *Steam: Its Generation and Use,* 40th ed., Babcock & Wilcox, 1992, p 6–21

3. Elevated Temperature Failures, *Failure Analysis and Prevention,* Vol 11, *ASM Handbook,* American Society for Metals, 1986, p 271

4. A.J. Sedriks, *Corrosion of Stainless Steels,* 2nd ed., Wiley-Interscience, 1996, p 399

5. P. Elliot, Practical Guide to High-Temperature Alloys, *Mater. Perform.,* Vol 28, 1989, p 57

6. F.H. Stott, G.C. Wood, and J. Stringer, Influence of Alloying Elements on the Development and Maintenance of Protective Scales, *Oxid. Met.,* Vol 44, 1995, p 113

7. M. Shütze, Mechanical Properties of Oxide Scales, *Oxid. Met.,* Vol 44, 1995, p 29

8. R.A. Smith, *Corrosion,* 2nd ed., L.L. Sheir, Ed., Newes-Butterworths, London, 1976, p 7:101

9. W.B. Hoyt and R.H. Caughey, High Temperature Metal Deterioration in Atmospheres Containing Carbon-Monoxide and Hydrogen, *Corrosion,* Vol 15, 1959, p 627t

10. P.A. Lefrancois and W.B. Hoyt, Chemical Thermodynamics of High Temperature Reactions in Metal Dusting Corrosion, *Corrosion,* Vol 19, 1963, p 360t

11. H.J. Grabke, R. Krajak, and E.M. Müller-Lorenz, Metal Dusting of High Temperature Alloys, *Mater. Corros.,* Vol 44, 1993, p 89

12. H.J. Grabke, Metal Dusting of Low and High Alloyed Steels, *Corrosion,* Vol 51 (No. 9), 1995, p 711

13. H.J. Grabke, R. Krajak, E.M. Müller-Lorenz, and S. Strauβ, Metal Dusting of Nickel Base Alloys, *Mater. Corros.,* Vol 47, 1996, p 495

14. H.J. Grabke, C.B. Brancho-Troconis, and E.M. Múller-Lorenz, Metal Dusting of Low Alloy Steels, *Mater. Corros.,* Vol 45, 1995, p 215

15. C.M. Schillmoller, Solving High-Temperature Problems in Oil Refineries and Petrochemical Plants, *Chem. Eng.,* Vol 93 (No. 1), (6 Jan 1986)

16. G.Y. Lai, *Proceedings in High Temperature Corrosion of Energy Systems,* M.F. Rutman, Ed., The Metallurgical Society, 1980, p 45

17. K.M. Verma, H. Ghosh, and J.S. Rai, *Br. Corros. J.,* Vol 13 (No. 4), 1978, p 173

18. G.Y. Lai, Nitridation of Several Combustor Alloys in a Simulated Gas Turbine Combustion Environment, *Advanced Materials and Coatings for Combustion Turbines,* V.P. Swaminathan and N.S. Cheruvu, Ed., ASM International, 1994, p 113

19. V.P Swaminathan and S.J. Lukezich, Degradation of Transition Duct Alloys in Gas Turbines, *Advanced Materials and Coatings for Combustion Turbines,* V.P. Swaminathan and N.S. Cheruvu, Ed., ASM International, 1994, p 106

20. S. Mrowec, The Problem of Sulfur in High-Temperature Corrosion, *Oxid. Met.,* Vol 44, 1995, p 177

21. R.A. Smith, *Corrosion,* 2nd ed., L.L. Sheir, Ed., Newes-Butterworths, London, 1976, p 7:90

22. *Materials of Construction for Shipboard Waste Incinerators,* National Research Council, National Academy of Sciences, 1977, p 91–92

23. W.H. Hatfield, *J. Iron Steel Inst.,* Vol 115, 1927, p 483

24. I.G. Wright, High Temperature Corrosion, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 99

25. K.N. Strafford and P.K. Datta, Design of Sulphidation Resistant Alloys, *Mater. Sci. Technol.,* Vol 5, Aug 1989, p 765

26. S. Lampman, ASM International, and M.D. Chaudhari, Columbus Metallurgical Service, private communication May 2002

27. A.U. Seybolt, P.A. Bergman, and M. Kaufman, "Hot Corrosion Mechanism Study, Phase II," final report, Contract N600 (61533) 65595, MEL 2591, Assignment 87-122, prepared for NSRDC, Annapolis, MD (30 June 1966 to 31 Oct 1967)

28. S.R.J. Saunders, in *Guidelines for Methods of Testing and Research in High Temperature Corrosion: Working Party Report,* H.J. Grabke and D.B. Meadowcroft, Ed., European Federation of Corrosion, No. 14, The Institute of Materials, London, 1995, p 85

29. E.L. Simons, G.V. Browning, and H.A. Liebhafsky, *Corrosion,* Vol 11 (No. 12), 1955, p 505t

30. R.K. Ahluwalia and K.H. Im, *J. Inst. Energy,* Vol 64, (No. 461), 1991, p 186

31. R.L. Jones, Naval Research Laboratory Report 6170-859, 1975

32. J.F. Condé and B.A. Wareham, *Proceedings: Gas Turbine Materials in the Marine Environment Conference,* MCIC-75-27, The Technical Working Group for Metals, Materials Advisory Group/U.S. Army Material Command in Cooperation with the National Association of Corrosion Engineers, 1974, p 73

33. D.A. Shores, D.W. McKee, and H.S. Spacil, *Proceedings: Properties of High Temperature Alloys with Emphasis on Environmental Effects,* PV 77-1, Z.A. Foroulis and F.S. Pettit, Ed., The Electrochemical Society, 1976, p 649

34. K.T. Chiang, F.S. Pettit, and G.H. Meier, Low Temperature Corrosion, *High Temperature Corrosion,* NACE-6, R.A. Rapp, Ed., National Association of Corrosion Engineers 1983, p 519

35. G. Wacker and W.L. Wheatfall, Corrosion Problems Associated with Shipboard Waste Incineration Systems, *1976 Triservice Corrosion Conference,* 26–28 Oct 1976 (Philadelphia, PA), DoD Triservice Committee on Corrosion, 1976, p 185

36. P. Elliott, Materials Performance in High-Temperature Waste Combustion Systems, *Mater. Perform.,* Vol 32, 1993, p 82

37. Z. Mazur, J. Kubial, and C. Marino-Lopes, Failure Analysis of Gas Turbine Last Stage Bucket Made of Udimet 500 Superalloy, *Pract. Fail. Anal.,* Vol 2 (No. 2), ASM International, April 2002, p 51

38. P. Ganesan, G.D. Smith, and L.E. Shoemaker, "The Effects of Excursions of Oxygen and Water Vapor Contents on Nickel-Containing Alloy Performance in a Waste Incineration Environment," Paper 248, Corrosion/91, National Association of Corrosion Engineers International, 1991

39. H.H. Krause, *J. Mater. Energy Syst.,* Vol 7, 1986, p 322

40. H.H. Krause, Hot Corrosion in Boilers Burning Municipal Solid Waste, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 997–998

41. G.D. Smith, P. Ganesan, and L.E. Shoemaker, "The Effect of Environmental Excursions within the Waste Incinerator on Nickel-Containing Alloy Performance," Paper 280, Corrosion/90, National Association of Corrosion Engineers International, 1990

42. C.T. Wall, J.T. Groves, and E.J. Roberts, How to Burn Salty Sludges, *Chem. Eng.,* Vol 7 (14 April 1975)

43. F. Wilkes and D.A. Shifler, "Historical Perspectives of Intergranular Corrosion of Boiler Metal by Water," Paper 44, Corrosion/93, National Association of Corrosion Engineers International, 1993

44. B. Craig, Hydrogen Damage, *Corrosion,* Vol 13, *Metals Handbook,* 9th ed., ASM International, 1987, p 163–171

45. R.M. Davison, T. Debold, and M.J. Johnson, Corrosion of Stainless Steels, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 547–565

46. R.A. Smith and S.R.I. Saunders, *Corrosion,* 3rd ed., L.L. Sheir, R.A. Jarman, and G.T. Burstein, Ed., Butterworth-Heinemann, London, 1994, p 7:136

47. M.H. Kamdar, Liquid Metal Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 171–184

48. M.G. Fontana and N.D. Greene, *Corrosion Engineering,* 2nd ed., McGraw-Hill, 1978, p 294

49. E.F. Bradley, Ed., *Superalloys: A Technical Guide,* ASM International, 1988

50. W.T. Reid, *External Corrosion and Deposits: Boiler and Gas Turbines,* American Elsevier, 1971

51. D.N. French, *Metallurgical Failures in Fossil Fired Boilers,* Wiley-Interscience, 1983, p 203

52. J.K. Tien and G.S. Ansell, Ed., *Alloy and Microstructural Design,* Academic Press, 1976

53. J.K. Tien and S. Purushothaman, The Metallurgy of High Temperature Alloys, *Proceedings: Properties of High Temperature Alloys—With Emphasis on Environmental Effects,* Vol 77-1, Z.A. Foroulis and F.S. Pettit, Ed., The Electrochemical Society, 1976, p 3

54. A. Sanz, L. Llanes, J.P. Bernadou, M. Anglada, and M.B. Rapaccini, in *Elevated Temperature Coatings: Science and Technology II,* N.B. Dahotre and J.M. Haampikian, Ed., The Minerals, Metals, and Materials Society, 1996, p 373

55. E. Basuki, A. Crosky, and B. Gleeson, *Mater. Sci. Eng. A,* Vol 224, 1997, p 27

56. K. Wu, Y. Wang, and J.E. Morral, in *Elevated Temperature Coatings: Science and Technology IV,* N.B. Dahotre, J.M. Hampikian, and J.E. Morral, Ed., The Minerals, Metals, and Materials Society, 2001, p 133

57. L. Onsager, *Phys. Rev.,* Vol 37, 1931, p 405; *Phys. Rev.,* Vol 38, 1932, p 2265

Robert B. Pond, Jr., M-Structures, Inc.; David A. Shifler, Naval Surface Warfare Center

Protective Coatings

The failure analyst may encounter protective coatings and thermal barriers when investigating components designed for high-temperature applications. Protective coatings may be higher alloys applied to a lower alloy by cladding or plasma spraying; diffusion coatings, such as nickel, cobalt, or iron aluminides, or overlay coatings, such as iron, cobalt, nickel, or cobalt-nickel Cr-Al-Y or Cr-Al-Hf coatings; or thermal barrier coatings (TBCs), such as yttria-stabilized zirconia. Some of the more common types are described as follows.

Aluminides. High-temperature coatings provide a barrier from corrosive or highly oxidative environments that may degrade the substrate alloy, thus supplementing the inherent corrosion resistance of most high-temperature alloys. To provide optimal performance, coatings need to have good adherence to the substrate as well as ductility to resist thermal cycling; in addition, they must avoid defects that may provide easy pathways to deleterious gases and maintain corrosion resistance over the design life of the high-temperature component.

Aluminide coatings on transition metals and alloys, such as iron, nickel, and cobalt, are β-FeAl, β-NiAl, and β-CoAl, respectively. Diffusion aluminide coatings are typically processed via pack cementation, slurry-diffusion, or chemical vapor deposition (Ref 58). Other processes may include plasma spraying, hot dipping, or electrophoresis. Subsequent heat treatments help develop the proper mechanical properties and trigger further diffusion within the coating. The principal protective oxide generated by the high-temperature oxidation of these aluminide intermetallics is Al_2O_3, which causes subsequent improved oxidation and corrosion resistance of aluminide coatings, although less-protective oxides can form if other alloying elements are present in the substrate, either in solution or as precipitated phases (Ref 59).

Two diffusional growth mechanisms characterize the pack aluminizing process on nickel-base superalloys, depending on the relative aluminum activity: low-activity aluminide and high-activity aluminide (Ref 60).

Low-activity coatings grow predominately by outward diffusion of nickel from the substrate to form a two-zone structure. The outer zone is a single-phase β-NiAl layer saturated with other substrate alloying elements, such as chromium, cobalt, titanium, and molybdenum, that diffuse outwardly with the nickel from the substrate (Ref 60). The content of any foreign element in the outer zone is limited by its solubility in NiAl. β-NiAl stoichiometry varies with aluminum content, decreasing from the outer surface toward the inner coating zone. The inner zone is β-NiAl containing a variety of second phases, such as carbides and sigma phase (Ref 60). The effects of nickel outward diffusion can lead to the enrichment of elements, such as chromium, cobalt, and molybdenum, in certain cases, in the substrate underlying the coating proper (Ref 60).

High-activity coatings grow by inward diffusion of aluminum, resulting in the formation of a Ni_2Al_3 layer. A single diffusion layer typifies this coating type, prior to heat treatment. Because the coating is formed by predominantly inward aluminum diffusion, the substrate structure is reproduced in the coating, with all alloy elements being diluted in proportion to the amount of aluminum introduced. High-activity coatings subjected to heat treatment convert the initial Ni_2Al_3 layer into more ductile and oxidation-resistant β-NiAl. During this transformation, aluminum diffuses from the Ni_2Al_3 to meet and react with the nickel diffusing from the substrate, forming a three-zone coating structure. This gives rise to an outer zone after heat treatment that consists of a β-NiAl matrix saturated with foreign elements and containing numerous precipitates, due to their lower solubility in the β-NiAl phase: a single-phase β-NiAl in the middle zone, along with varying amounts of substrate alloying elements (Ref 60); and an inner zone with β-NiAl with a variety of dispersed phases created by outward diffusion of the nickel from the substrate, similar to the diffusion zone in a low-activity coating.

Modified aluminide coatings are altered by the addition of a secondary element, such as chromium, platinum, a reactive element, or a combination of these elements (Ref 58). These coatings are generated either by incorporating the secondary element into the coating by pack codeposition, or by deposition of the modifying element prior to an aluminization process by electroplating, electrophoresis sputtering, chemical vapor deposition, or pack cementation (Ref 58).

Chromium-modified aluminide coatings result in a microstructure that contains layers of chromium and NiAl (Ref 61). One commercial chrome-modified aluminide coating displayed a three-zone structure after thermal

treatment: an outer zone, consisting of a fine dispersion of α-Cr in a hyperstoichiometric β-NiAl matrix; a middle zone, containing stoichiometric β-NiAl; and an inner zone, with α-Cr in a hypostoichiometric β-NiAl (Ref 62).

Platinum, a relatively inert metal, is added to improve the hot corrosion resistance of an aluminide coating. This effect is strongly influenced by processing procedures for the coatings (Ref 63). Platinum has shown significant benefits on the oxidation resistance of both low-activity and high-activity platinum aluminide coatings (Ref 64). Platinum promotes the formation of thinner and purer Al_2O_3 scales during the early stages, lowers the rate of β-NiAl degradation, and delays the onset of scale spallation. Platinum-modified aluminide coatings develop a spectrum of two-zone and three-zone structures of standard aluminides, depending on processing and the heat treatments involved (Ref 65).

A chromium-silicon aluminide coating is formed by inward diffusion of the aluminum and silicon constituents generated during heat treatment from a molten metallic slurry on the surface of the substrate (Ref 65). Aluminum diffuses to nickel, and silicon diffuses to chromium to form a composite monolayered coating consisting of β-NiAl and chromium silicide ($CrSi_2$) (Ref 66). A multistage process, developed to redistribute these key elements and structures, generates a multilayered coating that is more resistant to cracking and more stable at higher temperatures (Ref 67). The silicon content is limited on the outer surface of the coating, and layers of fine chromium silicides are distributed throughout the β-NiAl in discrete bands (Ref 67).

Overlay Coatings. Coatings of MCrAlY are applied by either electron beam physical vapor deposition (EB-PVD) or by thermal or plasma spraying, where M is iron, cobalt, nickel, or a combination of these elements. Typically, these coatings initially contained cobalt or nickel, with 15 wt% Cr, 12 wt% Al, and 0.1 to 0.9 wt% Y. To improve the coating resistance to type 2 hot corrosion, the cobalt was partially replaced with nickel, the aluminum content was reduced somewhat, and the chromium content was increased from 15 to 20 wt% to almost 32 wt%. Hafnium was used as a replacement for yttrium in some MCrAlX coatings, where X is the element (yttrium or hafnium) used to advance coating adhesion. In recent years, MCrAlY coatings have been air plasma sprayed as a cost-savings procedure and for reportedly better control of composition. There are various procedures and methods by which plasma spraying may be done, many of which are proprietary or patent-protected. The principal oxide in the outer portion of a MCrAlY overlay coating is Al_2O_3. Yttrium is more prevalent in the outer scale and promotes oxidation resistance under thermal cycling conditions by improving scale adhesion. Hafnium also promotes oxidation resistance and prevents sulfur from segregating to the coating/substrate interface, thus providing one mechanism of improving the overlay coating adhesion.

Nickel, cobalt, or both, and chromium are found in the inner half of the coating nearer to the coating/substrate interface. Increases in chromium content improve both the hot corrosion resistance and the oxidation resistance of the overlay coating but tend to make the coating less ductile.

Thermal barrier coatings usually consist of a porous layer of zirconium oxide (ZrO_2) that is stabilized in the tetragonal crystalline structure by approximately 8 wt% yttria (Y_2O_3). The yttria-stabilized zirconia is applied by using either plasma spraying or EB-PVD over a bond coat that is usually MCrAlY (where M is cobalt and/or nickel) or diffusion aluminide. The bond layer is typically 75 to 125 μm, and the TBC is 125 to 375 μm thick. Application of the TBC system to the hot section component in a gas turbine has been able to decrease metal surface temperature to approximately 150 °C (300 °F). Thermal barrier coatings also smooth out hot spots, thus reducing thermal fatigue stresses. A thermally grown oxide (TGO) layer is formed during fabrication and grows thicker on exposure of the TBC.

Thermal barrier coatings that have been applied by EB-PVD tend to produce a columnar structure with fine, ribbonlike voids, porosity within the columns aligned normal to the plane of the coating, and superior compliance to residual stresses. Plasma-sprayed TBCs have disklike voids aligned parallel to the coating plane and promote superior insulating efficiency.

There are a number of degradation modes that can limit the life of a TBC (Ref 68) as a result of oxidation due to the possible porosity network in the TBC. Electron beam physical vapor deposition TBCs tend to spall at the TGO/bond coat interface or in the alumina layer (Ref 69). It was observed that the intermittent water vapor degraded the adherence of alumina to the bond coat under a TBC layer more severely than alumina formed on the same bond coat in the absence of a TBC (Ref 68).

Hot corrosion exposure to vanadium salts and sulfur trioxide promote corrosion of stabilized zirconia by acid leaching of yttria. Loss of yttria destabilizes the zirconia, and thermal cycling causes rapid failure of the coating. This is potentially serious when burning low-grade fuels in a marine environment (Ref 70). For engine use, ZrO_2 must be stabilized in its high-temperature tetragonal structure to avoid a destructive monoclinic-to-

tetragonal phase transformation at 1100 °C (2010 °F). Scandia (Sc_2O_3) improves the vanadate hot corrosion resistance of stabilized zirconia.

Ceramics. In some high-temperature applications, refractories are being used to replace superalloys. Refractories are mainly high-melting-point metallic oxides but also include substances such as carbides, borides, nitrides, silicides, and graphite (Table 2). Maximum service temperatures are always less than the melting points of pure ceramics, because refractories usually contain minor constituents, and actual incinerator and pyrolytic environments are vastly different from simple oxygen atmospheres. Refractories are used to act as a corrosion barrier or as a thermal barrier between the high-temperature environment. Refractory suitability depends on its resistance to abrasion; maximum service temperature; corrosion resistance to liquids, gases, and slag; erosion resistance; spalling resistance; resistance to thermal cycling, shock, and fatigue from high thermal gradients; oxidation or reduction reactions; its mechanical strength; and its inspection and maintenance requirements.

Table 2 Melting points of common refractories

Name(s)	**Formula**	Melting point	
		°C	°F
Corundum, alpha alumina	Al_2O_3	2054	**3729**
Baddeleyite, zirconia	ZrO_2	2700	**4892**
Periclase, magnesia	MgO	2852	**5166**
Chromic oxide, chromia	Cr_2O_3	2330	**4226**
Lime, calcia	CaO	2927	**5301**
Cristobalite, silica	SiO_2	1723	**3133**
Carborundum, silicon carbide	SiC	2700[(a)]	**4892[(a)]**
Silicon nitride	Si_3N_4	1800[(b)]	**3270[(b)]**
Boron nitride	**BN**	**3000[(a)]**	**5430[(a)]**

(a) Sublimes.

(b) Dissociates

There are often property trade-offs required to attain the desired optimal, serviceable refractory. Insulating refractories generally have lower density, strength, corrosion resistance, and erosion resistance and higher porosity and void fraction than stronger, less insulating refractories. Refractories are generally susceptible to thermal shock and spalling. Monolithic refractories are often more susceptible to thermal shock than shaped refractories. Dense refractories tend to have high thermal conductivities, low porosity, relatively high strength, and improved corrosion resistance but tend to be susceptible to thermal shock. Insulating refractories have higher porosity, which improves thermal shock resistance but provides numerous pathways for molten and gaseous materials to promote corrosion and spalling.

The corrosion resistance of refractories is based on the thermodynamics of the potential corrosion reactions, the reaction rates and kinetics of these possible reactions, the composition and form of the refractory material, and the surface chemistry of the refractory with corrosive gaseous, liquid, or solid environments. The spontaneous direction of the possible corrosion chemical reactions can be calculated, which can indicate which reactions are theoretically possible. The reaction kinetics are dependent on the specific environment (gaseous, liquid, or solid), surface chemistry, the material microstructure and composition, refractory phase(s), and the temperature. All ceramics or refractories used in an oxygen-rich environment are oxides or develop a protective oxide on their surface (such as SiO_2 on SiC or Si_3N_4). Although oxide ceramics are inert and resist oxidation and reduction, they are not chemically inert. Ceramic corrosion between ceramic oxides and molten salt deposits generally can be explained as oxide/acid-base reactions (Ref 71). To determine if a particular slag is acidic or basic, the ratio of lime to silica in the slag is commonly used. As a general rule, the slag is basic if the CaO/SiO_2 ratio (or $CaO + MgO/SiO_2 + Al_2O_3$ ratio) is greater than one. If the ratio is less than one, the slag is considered acidic (Ref 72). Gases and liquids can penetrate deeply into highly porous refractories, which can exacerbate spalling and accelerate corrosion. Oxidizing gases, such as NO_x, Cl_2, O_2, CO_2, H_2O, and SO_2, reducing gases (NH_3, H_2, C_xH_y, CO, and H_2S), and vapors of volatile elements and compounds may react with different refractories. Hydrodynamic mass transport by either convection or diffusion can markedly affect the corrosion rate of a refractory in an environment. Nonwetting refractories are relatively resistant to corrosion with liquids or gases. Molten aluminum, iron, silica, and other constituents derived from the waste stream can

react with the crucible refractory to form new compounds or phases that can have different melting temperatures, dimensional stabilities, and corrosion resistances. The new phases or compounds can cause dimensional changes, which can cause problems in the chamber. High-purity refractories are more corrosion resistant, because certain minor impurity components have microstructures, phases, and defect structures that can introduce significant corrosion. High-purity refractories improve corrosion resistance by avoiding low-melting eutectics. Alkali impurities should be minimized as much as possible. Solids, either as fine particulates or dust, can cause degradation of refractories through abrasion or deposition.

Alumina-base refractories are usually available with silica and generally offer good corrosion resistance. However, high-alumina refractories degrade rapidly in hydrofluoric acid (HF), chlorides, molten or strong hot aqueous alkalis, and some molten metals (Ref 73). Alkalis flux the alumina-silica system. Chlorides cause increased porosity and removal of alumina or silica. Chromia (Cr_2O_3) improves the corrosion resistance of high-alumina refractories by reducing slag or liquid penetration (Ref 74). Favorable performance of alumina-chromia occurs even when the slag has a high iron content, but the use of chromia may introduce environmental questions. Silica contained in high-alumina refractories reacts with iron oxide and alkalis from slag to form a glassy surface layer over the refractory, which eventually exposes new surfaces to further reaction. Silicon carbide is not attacked by acids but reacts readily with fused caustic, halogens, and certain metal oxides. Silicon carbide is also abrasion resistant and less prone to thermal shock than other refractories (Ref 75). The operating temperature of silicon carbide should be limited to 1300 °C (2370 °F) in reducing N_2-H_2-CO environments when no steam is present (Ref 76).

Hafnium-base ceramics (HfC, HfN, and HfB_2) are being developed for temperature applications exceeding 2000°C (3630 °F) (Ref 77, 78). Both pure and alloyed hafnium compounds (tantalum, tungsten, and vanadium) were proposed for simple, oxidizing environments.

Intermetallic Alloys. In recent years, alloying and processing have been used to create advanced materials that control the ordered crystal and grain-boundary structure, microstructural features, and composition to overcome the brittleness inherent in ordered intermetallics (Ref 79, 80, 81). Results have led to a number of intermetallic alloys (usually aluminides of iron, nickel, and titanium) with low densities, relatively high melting points, a useful degree of ductility, and high-temperature strength. These materials are useful in high-temperature applications, particularly in oxidizing environments, because sufficient aluminum is present to form thin, compact, and protective alumina (Al_2O_3) scales.

Aluminides can exist over a range of compositions, but the degree of order decreases as the composition increasingly deviates from stoichiometric intermetallic compounds. Additional elements may be added without losing ordered structure. Titanium aluminides are usually based on Ti_3Al (α-2) and TiAl (γ). The oxidation of γ-TiAl-base alloys is complex and strongly affected by the oxidizing environment, alloy microstructure, and presence of trace alloying additions. At temperatures exceeding 800 °C (1470 °F), γ-based alloys typically develop a fast-growing, mixed oxide of TiO_2 + Al_2O_3. Excessive alumina scale growth of γ-TiAl-base alloys cannot be sustained, because a metastable Z-phase is formed at 1000 °C (1830 °F) in pure oxygen (Ref 82). Further information on titanium aluminide alloys may be found in Ref 83.

Excellent resistance to oxidation and carburization is exhibited by Ni_3Al alloys, while Fe_3Al alloys exhibit significantly better oxidation and sulfidation resistance than nickel-base alloys (Ref 84). The poor ductility of Ni_3Al at room temperature is due to low cohesive strength at the grain boundaries. Intergranular fracture can be stopped by small additions of boron. Chromium added to boron-doped Ni_3Al can avoid a drop in ductility observed at 700 °C (1290 °F), which is thought to be caused by hydrogen embrittlement (Ref 85). Hydrogen embrittlement also causes a loss in ductility of Fe_3Al (Ref 86). Both mechanical and oxidation properties can be further improved by the addition of alloying elements.

References cited in this section

58. R. Streiff, in *Elevated Temperature Coatings: Science and Technology II,* N.B. Dahotre and J.M. Hampikian, Ed., The Minerals, Metals, and Materials Society, 1996, p 407

59. G.W. Goward, Protective Coatings for High Temperature Alloys: State of Technology, *Proceedings of the Symposium on Properties of High Temperature Alloys,* PV 77-1, Z.A. Foroulis and F.S. Pettit, Ed., The Electrochemical Society, 1976, p 806

60. G.W. Goward and D.H. Boone, *Oxid. Met.,* Vol 3 (No. 5), 1971, p 475

61. G. Marijnissen, in *High-Temperature Protective Coatings,* M. Khobaib and R.C. Krutenat, Ed., The Metallurgical Society, 1984, p 27

62. G. McAdam and S.G. Russo, The Performance of Two Chrome-Modified Nickel Aluminides Subjected to Burner-Rig Testing, Paper 314, *Proceedings of the 13th International Corrosion Congress* (Clayton Vic, Australia), Australasian Corrosion Association, 1996

63. R. Streiff, D.H. Boone, and L.J. Purvis, in *Surface Engineering,* NATO ASI Series, R. Kossowski and S.C. Singhal, Ed., Martinus Nijhoff Publishers, 1984, p 469

64. Y. Niu, W.T. Wu, D.H. Boone, J.S. Smith, J.Q. Zhang, and C.L. Zhen, in *High Temperature Corrosion III,* R. Streiff, J. Stringer, R.C. Krutenat, and M. Caillet, Ed., Editions de Physique, Paris, 1993, p 511; and *J. Phys. (France) IV,* Vol 3, 1993, p 511

65. R. Streiff, O. Cerelier, and D.H. Boone, *Surf. Coat. Technol.,* Vol 32, 1987, p 111

66. B.G McMordie and A.K. Isaacs, "Slurry Aluminide Coating Processes for Gas Turbine Components at Overhaul," presented at The Annual Gas Turbine Conference and Exhibition, 1 Nov 1991 (Harrogate, U.K.), The Institution of Diesel and Gas Turbine Engineers

67. D. Barry, M.C. Meelu, B.G. McMordie and T.A. Kircher, Enhancing Performance of Silicon-Modified Slurry Aluminides on Turbine Components Operating in Marine Environments, Paper 95-GT-359, presented at International Gas Turbine & Aeroengine Congress, American Society of Mechanical Engineers, 1997

68. R.L. Jones, *High Temp. Technol.,* Vol 6, 1988, p 187

69. M.J. Stiger, N.M. Yanar, F.S. Pettit, and G.H. Meier, Mechanism of Failure of EB-PVD Thermal Barrier Coating Induced by High Temperature Oxidation, *Elevated Temperature Coatings: Science and Technology III,* J.M. Hampikian and N.B. Dahotre, Ed., TMS, 1999, p 51

70. S. Bose and J. DeMasi-Marcin, Thermal Barrier Coating Experience in Gas Turbine Engines at Pratt & Whitney, NASA CP 3312, *Thermal Barrier Coating Workshop,* NASA, 1995, p 63

71. R.L. Jones and R.F. Reidy, Development of Hot Corrosion Resistant Scandia-Stabilized Thermal Barrier Coatings, *Elevated Temperature Coatings: Science and Technology I,* N.B. Dahotre, J.M. Hampikian, and J.J. Stiglich, Ed., TMS, 1995, p 23

72. *Handbook of Refractory Practice,* 1st ed., Harbison-Walker/Indresco Inc., 1992, PR-12

73. L.M. Sheppard, Corrosion Resistant Ceramics for Severe Environments, *Ceram. Bull.,* Vol 70, 1991, p 1146

74. H.E. Wolfe and P.A. Beaulieu, "Update on Refractories for Hazardous Waste Incineration," Paper 551, Corrosion/89, National Association of Corrosion Engineers International, 1989

75. *Ceram. Ind.,* Vol 154, 1996, p 114

76. D.P. Butt, R.E. Tressler, and K.E. Spear, Corrosion of Silicon Carbide Materials in Metallurgical Heat-Treatment Environments, *Am. Ceram. Soc. Bull.,* Vol 71, 1992, p 1683

77. E.J. Wuchina and M.M. Opeka, Oxidation of Hf-Base Ceramics, *High Temperature Corrosion and Materials Chemistry: Proceedings of the Per Kofstad Memorial Symposium,* PV 99-38, M.J. McNallan, E.J. Opila, T. Maruyama, and T. Narita, Ed., The Electrochemical Society, 2000, p 477

78. E.J. Wuchina and M.M. Opeka, The Oxidation Behavior of HfC, HfN, and HfB_2, *High Temperature Corrosion and Materials Chemistry III: Proceedings of the International Symposium,* PV 99-38, M.J. McNallan and E.J. Opila, Ed., The Electrochemical Society, 2001, p 136

79. High-Temperature Ordered Intermetallic Alloys, *Materials Research Society Symposia Proceedings,* Vol 39, C.C. Koch, C.T. Liu, and N.S. Stoloff, Ed., Materials Research Society, 1985

80. High-Temperature Ordered Intermetallic Alloys II, *Materials Research Society Symposia Proceedings,* Vol 81, N.S. Stoloff, C.C. Koch, C.T. Liu, and O. Izumi, Ed., Materials Research Society, 1987

81. High-Temperature Ordered Intermetallic Alloys III, *Materials Research Society Symposia Proceedings,* Vol 133, C.T. Liu, C.C. Koch, A.I. Taub, and N.S. Stoloff, Ed., Materials Research Society, 1989

82. E. Copland, B. Gleeson, and D.J. Young, Oxidation of γ-TiAl Base Alloys and the Subsurface Formation of Z-$Ti_{50}Al_{30}O_{20}$, *Proceedings on High Temperature Corrosion and Materials Chemistry,* PV 98–9, P.Y. Hou, M.J. McNallan, R. Oltra, E.J. Opila, and D.A. Shores, Ed., The Electrochemical Society, 1998, p 446

83. Titanium Aluminides, *Materials Properties Handbook: Titanium Alloys,* R. Boyer, G. Weisch, and E.W. Collings, Ed., ASM International, 1994, p 1011

84. C. Deevi and V.K. Sikka, *Intermetallics,* Vol 4, 1996, p 357

85. R.N. Wright and V.K. Sikka, *J. Mater. Sci.,* Vol 23, 1988, p 4315

86. N.S. Stoloff and D.J. Duquette, *JOM,* Vol 45, 1993, p 30

High-Temperature Corrosion-Related Failures

Robert B. Pond, Jr., M-Structures, Inc.; David A. Shifler, Naval Surface Warfare Center

---

References

1. M.G. Fontana, *Corrosion Engineering,* 3rd ed., McGraw-Hill, 1986, p 526

2. S.C. Stultz and J.B. Kitto, Ed., *Steam: Its Generation and Use,* 40th ed., Babcock & Wilcox, 1992, p 6–21

3. Elevated Temperature Failures, *Failure Analysis and Prevention,* Vol 11, *ASM Handbook,* American Society for Metals, 1986, p 271

4. A.J. Sedriks, *Corrosion of Stainless Steels,* 2nd ed., Wiley-Interscience, 1996, p 399

5. P. Elliot, Practical Guide to High-Temperature Alloys, *Mater. Perform.,* Vol 28, 1989, p 57

6. F.H. Stott, G.C. Wood, and J. Stringer, Influence of Alloying Elements on the Development and Maintenance of Protective Scales, *Oxid. Met.,* Vol 44, 1995, p 113

7. M. Shütze, Mechanical Properties of Oxide Scales, *Oxid. Met.,* Vol 44, 1995, p 29

8. R.A. Smith, *Corrosion,* 2nd ed., L.L. Sheir, Ed., Newes-Butterworths, London, 1976, p 7:101

9. W.B. Hoyt and R.H. Caughey, High Temperature Metal Deterioration in Atmospheres Containing Carbon-Monoxide and Hydrogen, *Corrosion,* Vol 15, 1959, p 627t

10. P.A. Lefrancois and W.B. Hoyt, Chemical Thermodynamics of High Temperature Reactions in Metal Dusting Corrosion, *Corrosion,* Vol 19, 1963, p 360t

11. H.J. Grabke, R. Krajak, and E.M. Müller-Lorenz, Metal Dusting of High Temperature Alloys, *Mater. Corros.,* Vol 44, 1993, p 89

12. H.J. Grabke, Metal Dusting of Low and High Alloyed Steels, *Corrosion,* Vol 51 (No. 9), 1995, p 711

13. H.J. Grabke, R. Krajak, E.M. Müller-Lorenz, and S. Strauβ, Metal Dusting of Nickel Base Alloys, *Mater. Corros.,* Vol 47, 1996, p 495

14. H.J. Grabke, C.B. Brancho-Troconis, and E.M. Múller-Lorenz, Metal Dusting of Low Alloy Steels, *Mater. Corros.,* Vol 45, 1995, p 215

15. C.M. Schillmoller, Solving High-Temperature Problems in Oil Refineries and Petrochemical Plants, *Chem. Eng.,* Vol 93 (No. 1), (6 Jan 1986)

16. G.Y. Lai, *Proceedings in High Temperature Corrosion of Energy Systems,* M.F. Rutman, Ed., The Metallurgical Society, 1980, p 45

17. K.M. Verma, H. Ghosh, and J.S. Rai, *Br. Corros. J.,* Vol 13 (No. 4), 1978, p 173

18. G.Y. Lai, Nitridation of Several Combustor Alloys in a Simulated Gas Turbine Combustion Environment, *Advanced Materials and Coatings for Combustion Turbines,* V.P. Swaminathan and N.S. Cheruvu, Ed., ASM International, 1994, p 113

19. V.P Swaminathan and S.J. Lukezich, Degradation of Transition Duct Alloys in Gas Turbines, *Advanced Materials and Coatings for Combustion Turbines,* V.P. Swaminathan and N.S. Cheruvu, Ed., ASM International, 1994, p 106

20. S. Mrowec, The Problem of Sulfur in High-Temperature Corrosion, *Oxid. Met.,* Vol 44, 1995, p 177

21. R.A. Smith, *Corrosion,* 2nd ed., L.L. Sheir, Ed., Newes-Butterworths, London, 1976, p 7:90

22. *Materials of Construction for Shipboard Waste Incinerators,* National Research Council, National Academy of Sciences, 1977, p 91–92

23. W.H. Hatfield, *J. Iron Steel Inst.,* Vol 115, 1927, p 483

24. I.G. Wright, High Temperature Corrosion, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 99

25. K.N. Strafford and P.K. Datta, Design of Sulphidation Resistant Alloys, *Mater. Sci. Technol.,* Vol 5, Aug 1989, p 765

26. S. Lampman, ASM International, and M.D. Chaudhari, Columbus Metallurgical Service, private communication May 2002

27. A.U. Seybolt, P.A. Bergman, and M. Kaufman, “Hot Corrosion Mechanism Study, Phase II,” final report, Contract N600 (61533) 65595, MEL 2591, Assignment 87-122, prepared for NSRDC, Annapolis, MD (30 June 1966 to 31 Oct 1967)

28. S.R.J. Saunders, in *Guidelines for Methods of Testing and Research in High Temperature Corrosion: Working Party Report,* H.J. Grabke and D.B. Meadowcroft, Ed., European Federation of Corrosion, No. 14, The Institute of Materials, London, 1995, p 85

29. E.L. Simons, G.V. Browning, and H.A. Liebhafsky, *Corrosion,* Vol 11 (No. 12), 1955, p 505t

30. R.K. Ahluwalia and K.H. Im, *J. Inst. Energy,* Vol 64, (No. 461), 1991, p 186

31. R.L. Jones, Naval Research Laboratory Report 6170-859, 1975

32. J.F. Condé and B.A. Wareham, *Proceedings: Gas Turbine Materials in the Marine Environment Conference,* MCIC-75-27, The Technical Working Group for Metals, Materials Advisory Group/U.S. Army Material Command in Cooperation with the National Association of Corrosion Engineers, 1974, p 73

33. D.A. Shores, D.W. McKee, and H.S. Spacil, *Proceedings: Properties of High Temperature Alloys with Emphasis on Environmental Effects,* PV 77-1, Z.A. Foroulis and F.S. Pettit, Ed., The Electrochemical Society, 1976, p 649

34. K.T. Chiang, F.S. Pettit, and G.H. Meier, Low Temperature Corrosion, *High Temperature Corrosion,* NACE-6, R.A. Rapp, Ed., National Association of Corrosion Engineers 1983, p 519

35. G. Wacker and W.L. Wheatfall, Corrosion Problems Associated with Shipboard Waste Incineration Systems, *1976 Triservice Corrosion Conference,* 26–28 Oct 1976 (Philadelphia, PA), DoD Triservice Committee on Corrosion, 1976, p 185

36. P. Elliott, Materials Performance in High-Temperature Waste Combustion Systems, *Mater. Perform.,* Vol 32, 1993, p 82

37. Z. Mazur, J. Kubial, and C. Marino-Lopes, Failure Analysis of Gas Turbine Last Stage Bucket Made of Udimet 500 Superalloy, *Pract. Fail. Anal.,* Vol 2 (No. 2), ASM International, April 2002, p 51

38. P. Ganesan, G.D. Smith, and L.E. Shoemaker, “The Effects of Excursions of Oxygen and Water Vapor Contents on Nickel-Containing Alloy Performance in a Waste Incineration Environment,” Paper 248, Corrosion/91, National Association of Corrosion Engineers International, 1991

39. H.H. Krause, *J. Mater. Energy Syst.,* Vol 7, 1986, p 322

40. H.H. Krause, Hot Corrosion in Boilers Burning Municipal Solid Waste, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 997–998

41. G.D. Smith, P. Ganesan, and L.E. Shoemaker, “The Effect of Environmental Excursions within the Waste Incinerator on Nickel-Containing Alloy Performance,” Paper 280, Corrosion/90, National Association of Corrosion Engineers International, 1990

42. C.T. Wall, J.T. Groves, and E.J. Roberts, How to Burn Salty Sludges, *Chem. Eng.,* Vol 7 (14 April 1975)

43. F. Wilkes and D.A. Shifler, “Historical Perspectives of Intergranular Corrosion of Boiler Metal by Water,” Paper 44, Corrosion/93, National Association of Corrosion Engineers International, 1993

44. B. Craig, Hydrogen Damage, *Corrosion,* Vol 13, *Metals Handbook,* 9th ed., ASM International, 1987, p 163–171

45. R.M. Davison, T. Debold, and M.J. Johnson, Corrosion of Stainless Steels, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 547–565

46. R.A. Smith and S.R.I. Saunders, *Corrosion,* 3rd ed., L.L. Sheir, R.A. Jarman, and G.T. Burstein, Ed., Butterworth-Heinemann, London, 1994, p 7:136

47. M.H. Kamdar, Liquid Metal Embrittlement, *Corrosion,* Vol 13, *ASM Handbook,* ASM International, 1987, p 171–184

48. M.G. Fontana and N.D. Greene, *Corrosion Engineering,* 2nd ed., McGraw-Hill, 1978, p 294

49. E.F. Bradley, Ed., *Superalloys: A Technical Guide,* ASM International, 1988

50. W.T. Reid, *External Corrosion and Deposits: Boiler and Gas Turbines,* American Elsevier, 1971

51. D.N. French, *Metallurgical Failures in Fossil Fired Boilers,* Wiley-Interscience, 1983, p 203

52. J.K. Tien and G.S. Ansell, Ed., *Alloy and Microstructural Design,* Academic Press, 1976

53. J.K. Tien and S. Purushothaman, The Metallurgy of High Temperature Alloys, *Proceedings: Properties of High Temperature Alloys—With Emphasis on Environmental Effects,* Vol 77-1, Z.A. Foroulis and F.S. Pettit, Ed., The Electrochemical Society, 1976, p 3

54. A. Sanz, L. Llanes, J.P. Bernadou, M. Anglada, and M.B. Rapaccini, in *Elevated Temperature Coatings: Science and Technology II,* N.B. Dahotre and J.M. Haampikian, Ed., The Minerals, Metals, and Materials Society, 1996, p 373

55. E. Basuki, A. Crosky, and B. Gleeson, *Mater. Sci. Eng. A,* Vol 224, 1997, p 27

56. K. Wu, Y. Wang, and J.E. Morral, in *Elevated Temperature Coatings: Science and Technology IV,* N.B. Dahotre, J.M. Hampikian, and J.E. Morral, Ed., The Minerals, Metals, and Materials Society, 2001, p 133

57. L. Onsager, *Phys. Rev.,* Vol 37, 1931, p 405; *Phys. Rev.,* Vol 38, 1932, p 2265

58. R. Streiff, in *Elevated Temperature Coatings: Science and Technology II,* N.B. Dahotre and J.M. Hampikian, Ed., The Minerals, Metals, and Materials Society, 1996, p 407

59. G.W. Goward, Protective Coatings for High Temperature Alloys: State of Technology, *Proceedings of the Symposium on Properties of High Temperature Alloys,* PV 77-1, Z.A. Foroulis and F.S. Pettit, Ed., The Electrochemical Society, 1976, p 806

60. G.W. Goward and D.H. Boone, *Oxid. Met.,* Vol 3 (No. 5), 1971, p 475

61. G. Marijnissen, in *High-Temperature Protective Coatings,* M. Khobaib and R.C. Krutenat, Ed., The Metallurgical Society, 1984, p 27

62. G. McAdam and S.G. Russo, The Performance of Two Chrome-Modified Nickel Aluminides Subjected to Burner-Rig Testing, Paper 314, *Proceedings of the 13th International Corrosion Congress* (Clayton Vic, Australia), Australasian Corrosion Association, 1996

63. R. Streiff, D.H. Boone, and L.J. Purvis, in *Surface Engineering,* NATO ASI Series, R. Kossowski and S.C. Singhal, Ed., Martinus Nijhoff Publishers, 1984, p 469

64. Y. Niu, W.T. Wu, D.H. Boone, J.S. Smith, J.Q. Zhang, and C.L. Zhen, in *High Temperature Corrosion III,* R. Streiff, J. Stringer, R.C. Krutenat, and M. Caillet, Ed., Editions de Physique, Paris, 1993, p 511; and *J. Phys. (France) IV,* Vol 3, 1993, p 511

65. R. Streiff, O. Cerelier, and D.H. Boone, *Surf. Coat. Technol.,* Vol 32, 1987, p 111

66. B.G McMordie and A.K. Isaacs, "Slurry Aluminide Coating Processes for Gas Turbine Components at Overhaul," presented at The Annual Gas Turbine Conference and Exhibition, 1 Nov 1991 (Harrogate, U.K.), The Institution of Diesel and Gas Turbine Engineers

67. D. Barry, M.C. Meelu, B.G. McMordie and T.A. Kircher, Enhancing Performance of Silicon-Modified Slurry Aluminides on Turbine Components Operating in Marine Environments, Paper 95-GT-359, presented at International Gas Turbine & Aeroengine Congress, American Society of Mechanical Engineers, 1997

68. R.L. Jones, *High Temp. Technol.,* Vol 6, 1988, p 187

69. M.J. Stiger, N.M. Yanar, F.S. Pettit, and G.H. Meier, Mechanism of Failure of EB-PVD Thermal Barrier Coating Induced by High Temperature Oxidation, *Elevated Temperature Coatings: Science and Technology III,* J.M. Hampikian and N.B. Dahotre, Ed., TMS, 1999, p 51

70. S. Bose and J. DeMasi-Marcin, Thermal Barrier Coating Experience in Gas Turbine Engines at Pratt & Whitney, NASA CP 3312, *Thermal Barrier Coating Workshop,* NASA, 1995, p 63

71. R.L. Jones and R.F. Reidy, Development of Hot Corrosion Resistant Scandia-Stabilized Thermal Barrier Coatings, *Elevated Temperature Coatings: Science and Technology I,* N.B. Dahotre, J.M. Hampikian, and J.J. Stiglich, Ed., TMS, 1995, p 23

72. *Handbook of Refractory Practice,* 1st ed., Harbison-Walker/Indresco Inc., 1992, PR-12

73. L.M. Sheppard, Corrosion Resistant Ceramics for Severe Environments, *Ceram. Bull.,* Vol 70, 1991, p 1146

74. H.E. Wolfe and P.A. Beaulieu, "Update on Refractories for Hazardous Waste Incineration," Paper 551, Corrosion/89, National Association of Corrosion Engineers International, 1989

75. *Ceram. Ind.,* Vol 154, 1996, p 114

76. D.P. Butt, R.E. Tressler, and K.E. Spear, Corrosion of Silicon Carbide Materials in Metallurgical Heat-Treatment Environments, *Am. Ceram. Soc. Bull.,* Vol 71, 1992, p 1683

77. E.J. Wuchina and M.M. Opeka, Oxidation of Hf-Base Ceramics, *High Temperature Corrosion and Materials Chemistry: Proceedings of the Per Kofstad Memorial Symposium,* PV 99-38, M.J. McNallan, E.J. Opila, T. Maruyama, and T. Narita, Ed., The Electrochemical Society, 2000, p 477

78. E.J. Wuchina and M.M. Opeka, The Oxidation Behavior of HfC, HfN, and HfB_2, *High Temperature Corrosion and Materials Chemistry III: Proceedings of the International Symposium,* PV 99-38, M.J. McNallan and E.J. Opila, Ed., The Electrochemical Society, 2001, p 136

79. High-Temperature Ordered Intermetallic Alloys, *Materials Research Society Symposia Proceedings,* Vol 39, C.C. Koch, C.T. Liu, and N.S. Stoloff, Ed., Materials Research Society, 1985

80. High-Temperature Ordered Intermetallic Alloys II, *Materials Research Society Symposia Proceedings,* Vol 81, N.S. Stoloff, C.C. Koch, C.T. Liu, and O. Izumi, Ed., Materials Research Society, 1987

81. High-Temperature Ordered Intermetallic Alloys III, *Materials Research Society Symposia Proceedings,* Vol 133, C.T. Liu, C.C. Koch, A.I. Taub, and N.S. Stoloff, Ed., Materials Research Society, 1989

82. E. Copland, B. Gleeson, and D.J. Young, Oxidation of γ-TiAl Base Alloys and the Subsurface Formation of Z-$Ti_{50}Al_{30}O_{20}$, *Proceedings on High Temperature Corrosion and Materials Chemistry,* PV 98–9, P.Y. Hou, M.J. McNallan, R. Oltra, E.J. Opila, and D.A. Shores, Ed., The Electrochemical Society, 1998, p 446

83. Titanium Aluminides, *Materials Properties Handbook: Titanium Alloys,* R. Boyer, G. Weisch, and E.W. Collings, Ed., ASM International, 1994, p 1011

84. C. Deevi and V.K. Sikka, *Intermetallics,* Vol 4, 1996, p 357

85. R.N. Wright and V.K. Sikka, *J. Mater. Sci.,* Vol 23, 1988, p 4315

86. N.S. Stoloff and D.J. Duquette, *JOM,* Vol 45, 1993, p 30

Biological Corrosion Failures

Thomas R. Jack, NOVA Chemicals Ltd.

Introduction

MICROORGANISMS can directly or indirectly affect the integrity of many materials used in industrial systems. Most metals, including iron, copper, nickel, aluminum, and their alloys, are more or less susceptible to damage (Ref 1, 2, 3). Only titanium and its alloys appear to be generally resistant (Ref 4). This review focuses initially on the mechanisms of microbially induced or influenced corrosion (MIC) of metallic materials as an introduction to the recognition, management, and prevention of microbiological corrosion failures in piping, tanks, heat exchangers, cooling towers, and so on. Numerous reviews of MIC have appeared over the last decade (Ref 2, 5, 6, 7, 8, 9, 10, 11, 12). Two recent publications (Ref 13, 14) present broader discussions of MIC, including a useful introduction to microbial problems seen with nonmetallic materials such as polymers, composites, concrete, glass, wood, and stone.

Viable microorganisms can be found over a surprisingly wide range of temperature, pressure, salinity, and pH (Ref 1). In the 1950s, pioneering work by Zobell isolated sulfate-reducing bacteria (SRB) that grew at 104 °C (219 °F) and pressures of 1000 bar from oil-bearing geological formations deep underground (Ref 15). Microbial communities exist in environments as diverse as subzero snowfields to deep ocean thermal vents. Halophiles evolved to live at extreme salinities turn pink the evaporation pans used to win salt from seawater. Sulfur-oxidizing bacteria create very acidic conditions (pH < 1) by producing sulfuric acid as an end product of their metabolism, while other microorganisms survive the opposite end of the pH scale. Given these examples, it should not be surprising that microorganisms have been implicated in the accelerated corrosion and cracking of a correspondingly wide range of industrial systems. For example, the involvement of thermophilic SRB in the severe intergranular pitting of 304L stainless steel condenser tubes in a geothermal electrical power plant operating at >100 °C (> 210 °F) has been reported (Ref 16). In another example, microbiological activity and chloride concentrated under scale deposits were blamed for the wormhole pitting of carbon steel piping used to transport a slurry of magnesium hydroxide and alumina at pH 10.5 (Ref 17).

Whatever the environmental conditions, microorganisms need water, a source of energy to drive their metabolism, and nutrients to provide essential building materials (carbon, nitrogen, phosphorus, trace metals, etc.) for cell renewal and growth. An understanding of these factors can sometimes help in failure investigations. Energy may be derived from sunlight through photosynthesis or from chemical reactions. The importance of photosynthetic metabolism is limited in the context of this article to above-ground facilities or submerged structures that receive sunlight. For closed systems and buried facilities, microbial metabolism is based on energy derived from oxidation reduction (redox) reactions. Under aerobic conditions, reduction of oxygen to water complements the metabolic oxidation of organic nutrients to carbon dioxide. Under anaerobic conditions, electron acceptors other than oxygen can be used. Figure 1 illustrates the range of pH and redox potential where anaerobic forms of microbial metabolism tend to be found (Ref 18).

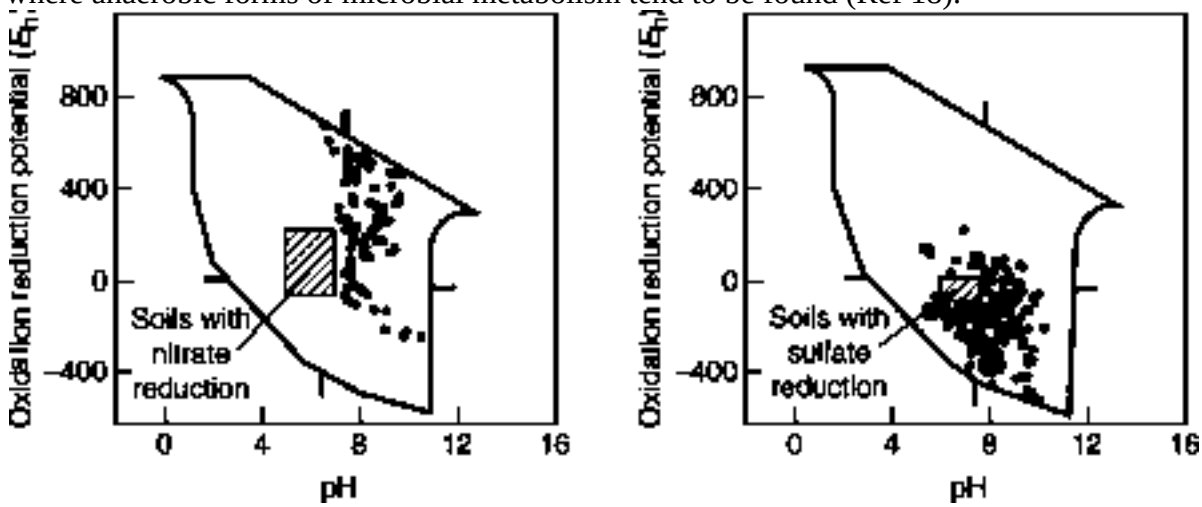


Fig. 1 The pH and oxidation reduction potential for growth of anaerobic bacteria able to reduce nitrate or sulfate (dots in plots) and for soils dominated by the microbial metabolism (boxes). Aerobic bacteria grow over a wide range of pH at E_h > 300 mV (normal hydrogen electrode). Source: Ref 19

Whatever the metabolism, electrochemical reactions catalyzed by enzymes provide energy for cell growth. Many of these reactions are not important under abiotic conditions, because they are kinetically slow in the absence of organisms. By promoting these reactions, microbes produce metabolites and conditions not found under abiotic conditions. In some cases, electrons released by the oxidation of metals are used directly in microbial metabolism. In other cases, it is the chemicals and conditions created by microbial activity that promote MIC. Secondary effects can also be important. These include such things as the biodegradation of lubricants and protective coatings designed to prevent wear or corrosion in an operating system, or the alteration of flow regimes and heat-transfer coefficients due to the biological fouling of metal surfaces.

Given the potential impact of MIC on a wide range of industrial operations, it is not surprising that microbiological effects are of significant concern in failure analysis and prevention. Microbially induced corrosion problems afflict water-handling operations and manufacturing processes in oil and gas production, pipelining, refining, petrochemical synthesis, power production, fermentation, waste water treatment, drinking water supply, pulp and paper making, and other industrial sectors. Microbially induced corrosion is also a concern whenever metals are exposed directly to the environment in applications including marine or buried piping, storage tanks, ships, nuclear waste containers, pilings, marine platforms, and so on.

References cited in this section

1. G.G. Geesey and G.A. Cragnolino, A Review of the Potential for Microbially Influenced Corrosion of High-Level Nuclear Waste Containers in an Unsaturated Repository Site, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion*, P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995, (New Orleans), National Association of Corrosion Engineers International, 1995, p 76/1–76/20

2. B. Little and P. Wagner, An Overview of Microbiologically Influenced Corrosion of Metals and Alloys Used in the Storage of Nuclear Wastes, *Can. J. Microbiol.,* Vol 42, 1996, p 367–374

3. S.C. Dexter and J.P. LaFontaine, Effect of Natural Marine Biofilms on Galvanic Corrosion, *Corrosion,* Vol 54 (No. 11), 1998, p 851–861

4. J. Been and J.S. Grauman, Titanium and Titanium Alloys, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

5. G. Kobrin, *A Practical Manual on Microbiologically Influenced Corrosion,* National Association of Corrosion Engineers, 1993

6. J.F.D. Stott, What Progress in the Understanding of Microbially Induced Corrosion Has Been Made in the Last 25 Years? A Personal Viewpoint, *Corros. Sci.,* Vol 35, 1993, p 667–673

7. S.W. Boronstein, *Microbiologically Influenced Corrosion Handbook,* Industrial Press Inc., 1994

8. T.R. Jack and D.W.S. Westlake, Control in Industrial Settings, *Sulfate-Reducing Bacteria,* Vol 8, *Biotechnology Handbooks,* L.L. Barton, Ed., Plenum Press, London, 1995, p 264–292

9. W.A. Hamilton, and W. Lee, Biocorrosion, *Sulfate-Reducing Bacteria,* Vol 8, *Biotechnology Handbooks,* L.L. Barton, Ed., Plenum Press, London, 1995, p 243–264

10. T.R. Jack, Monitoring Microbial Fouling and Corrosion Problems in Industrial Systems, *Corros. Rev.,* Vol 17 (No. 1), 1999, p 1–31

11. W.P. Iverson, Research on the Mechanisms of Anaerobic Corrosion, *Int. Biodeterior. Biodegrad.,* Vol 47 (No. 2), 2001, p 63–70

12. H.A. Videla, Microbially Induced Corrosion: An Updated Overview, *Int. Biodeterior. Biodegrad.,* Vol 48, 2001, p 176–201

13. *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

14. *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

15. B.N. Herbert and F.D.J. Stott, The Effects of Pressure and Temperature on Bacteria in Oilfield Water Injection Systems, *Microbial Corrosion,* The Metals Society, London, 1983, p 7–17

16. R. Torres-Sanchez, M.C. Gonzalez, H. Martinez, and L. Martinez-Gomez, Microbiologically Influenced Corrosion in a Geothermal Electric Power Plant, *Mater. Perform.,* Vol 40 (No. 11), 2001, p 46–48

17. S.R. Freeman and H. Thielsch, Microbiological Activity and Chlorides Cause Carbon Steel Pipe Pitting, *Mater. Perform.,* Vol 35 (No. 3), 1996, p 65–66

18. D.A. Jones and P.S. Amy, Related Electrochemical Characteristics of Microbial Metabolism and Iron Corrosion, *Ind. Eng. Chem. Res.,* Vol 39, 2000, p 575–582

19. J.E. Zajic, *Microbial Biogeochemistry,* Academic Press, New York, 1969

Thomas R. Jack, NOVA Chemicals Ltd.

Microbial Involvement in Corrosion

Corrosion involves the oxidation of metal atoms, *M*, in an anodic region with a loss of electrons to a complementary reduction reaction in a cathodic region elsewhere on the metal surface (Fig. 2). In the reduction reaction, an electron acceptor, *X*, receives the electrons given up by the oxidation process. The driving force for corrosion is the free energy released by the overall chemical reaction resulting from combined cathodic and anodic half-reactions:

Anodic reaction	$M \rightarrow M^{n+} + n$ electrons
Cathodic reaction	$mX + n$ electrons $\rightarrow mX^{n/m-}$
Overall corrosion reaction	$M + mX \rightarrow M^{n+} + mX^{n/m-}$

(Eq 1)

where *n*, *p*, and *m* are integer values chosen to charge balance the equations based on the number of electrons given up by the metal being oxidized or taken up by the electron acceptor being reduced.

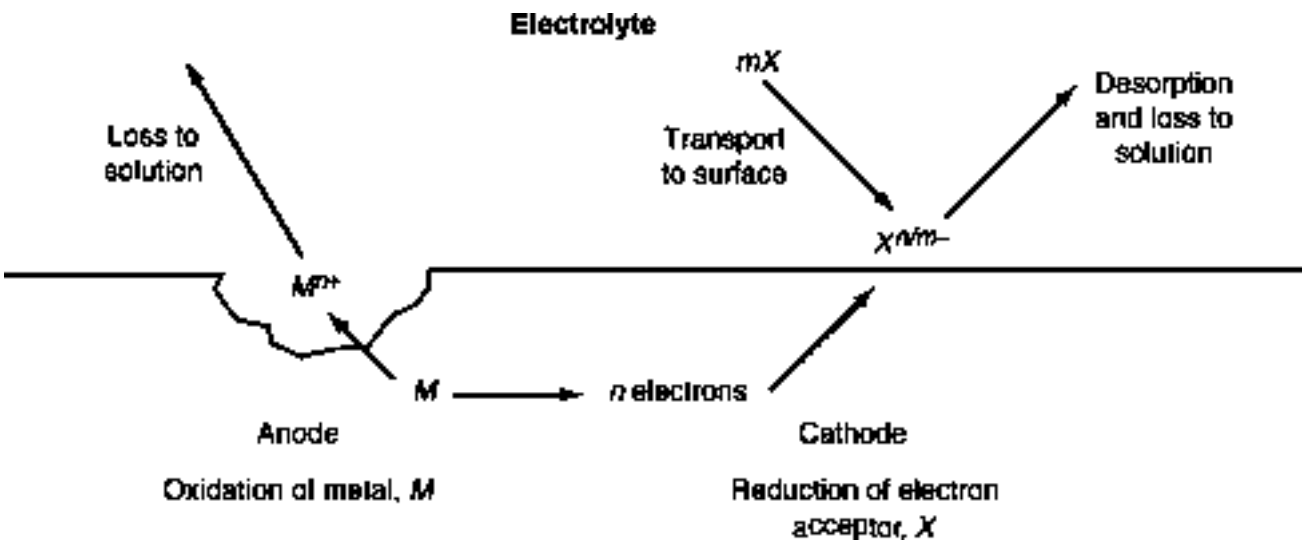


Fig. 2 Schematic diagram of a generic corrosion cell showing anodic oxidation of the metal (*M*) complemented by cathodic reduction of an electron acceptor (*X*). The corrosion rate can be controlled by the rate of arrival of *X* at the cathodic surface, a buildup of metal ions, M^+, at the anode, or a buildup of reduced oxidant, $X^{n/m-}$, at the cathode.

Because metals are good electrical conductors, the cathodic and anodic processes can occur at different locations on a metal surface exposed to a common electrolyte. Where anodes and cathodes frequently change location, general corrosion is seen, but when the anode becomes focused in one location, pitting results. In pitting, the rate of penetration is, to a large measure, a function of the relative size of the cathodic and anodic areas involved in the corrosion cell and is usually much higher than that seen for general corrosion.

The rate of metal loss is determined by the slowest, or rate-determining, step in the corrosion process. For example, where the availability of electron acceptor *X* is limiting, the corrosion rate depends on the transport kinetics for the arrival of *X* at the metal surface from the surrounding solution (Fig. 2). Where the availability of *X* is not a constraint, other processes may be rate limiting. For example, a buildup of primary reaction products such as M^{n+} on the anodic surface can polarize the anode, stifle the rate of metal oxidation, and limit the overall rate of corrosion by anodic polarization. Similarly, a buildup of primary cathodic reaction products, $X^{n/m-}$, can slow the overall process through cathodic polarization. A common example of this occurs under anaerobic conditions, where $X = H^+$. In this case, reduction of protons from the electrolyte leads to formation of atomic

hydrogen on the metal surface. To escape the surface, atoms of hydrogen must combine to form molecular hydrogen that can then be lost to solution or enter the steel matrix, as shown in Fig. 3.

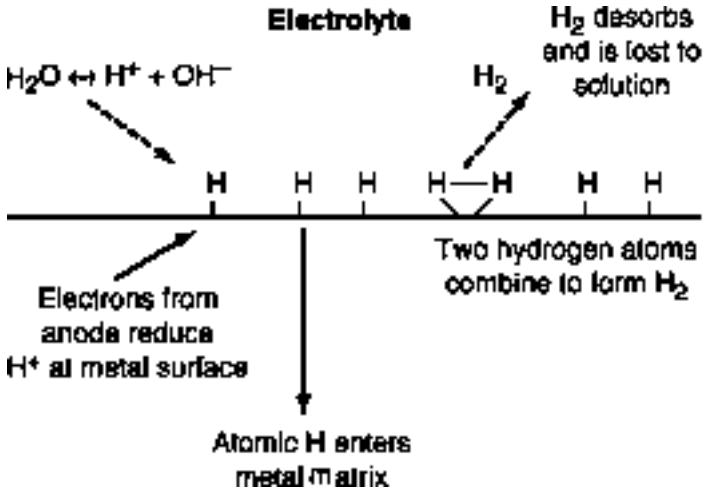


Fig. 3 Schematic diagram of the cathodic surface of an anaerobic corrosion cell in which H^+ is reduced to H on metal surface. The atomic hydrogen formed escapes the cathodic surface by combining to form molecular hydrogen, $2H \rightarrow H_2$, that desorbs from the surface or by entry as atomic hydrogen into the metal matrix.

Byproducts of the various reactions involved may limit the corrosion process by altering the environment at the metal surface (Eq 2–6). For example, reduction of protons at the metal surface (Eq 3) raises the local pH. This can lead to the formation of insoluble deposits. The formation of ferrous carbonate on a steel surface by the sequence of events shown in Eq 2 to 6 provides one example. Precipitation of calcium or magnesium ions from the electrolyte as insoluble carbonates that can passivate a metal surface is another:

$$Fe \rightarrow Fe^{2+} + 2 \text{ electrons} \quad \text{(Eq 2)}$$

$$2H^+ + 2 \text{ electrons} \rightarrow 2H^{\cdot} \text{ (cathodic hydrogen)} \quad \text{(Eq 3)}$$

$$H_2O \leftrightarrow H^+ + OH^- \quad \text{(Eq 4)}$$

$$OH^- + HCO_3^- \rightarrow CO_3^{2-} \quad \text{(Eq 5)}$$

$$CO_3^{2-} + Fe^{2+} \rightarrow FeCO_3 \quad \text{(Eq 6)}$$

Microorganisms accelerate corrosion by changing the nature or kinetics of the rate-controlling reaction or process. They can be directly involved in the electron transfer processes in the electrochemical cell represented by Eq 1 or be less directly involved through a number of mechanisms, including depolarization of the anode or cathode, disruption of passivating films, or rapid regeneration or provision of the electron acceptor, *X*.

Direct Involvement in the Corrosion Process

Corrosion Mechanisms Involving SRB. Perhaps the best-known mechanism of MIC involves corrosion cells generated and sustained on steel surfaces by the action of anaerobic SRB. These organisms reduce sulfate to sulfide in their metabolism and are commonly found in mixed microbial communities present in soils and natural waters. In industrial systems, biodegradable materials, such as some of the hydrocarbons found in oil and gas operations (Ref 20) or susceptible components of coating materials (Ref 21, 22), can provide a source of nutrients for microbial growth. Cathodic hydrogen formed on a metal surface (Eq 3) by active corrosion or by cathodic protection (CP) can specifically promote growth of organisms, including SRB that are able to use hydrogen in their metabolism (Ref 23). Severe corrosion cells develop as sulfide, produced by the microbial reduction of sulfate, combines with ferrous ions, released by the corrosion process, to produce insoluble black iron sulfides:

$$4Fe + SO_4^{2-} + 8H^+ \rightarrow FeS + 3Fe^{2+} + 4H_2O \quad \text{(Eq 7)}$$

Various versions of this MIC mechanism have been suggested (Ref 9). Figure 4 illustrates a plausible mechanism base on a galvanic couple formed between iron and iron sulfide sustained and extended by the active involvement of SRB. The way in which electrons are transferred from iron sulfide to the SRB, for example, is not well resolved. It may occur directly or via formation of cathodic hydrogen, as shown in Fig. 4, or by another reaction involving reduction of H_2S (Ref 6):

$$H_2S + \text{electron} \rightarrow HS^- + H^\cdot \qquad \text{(Eq 8)}$$

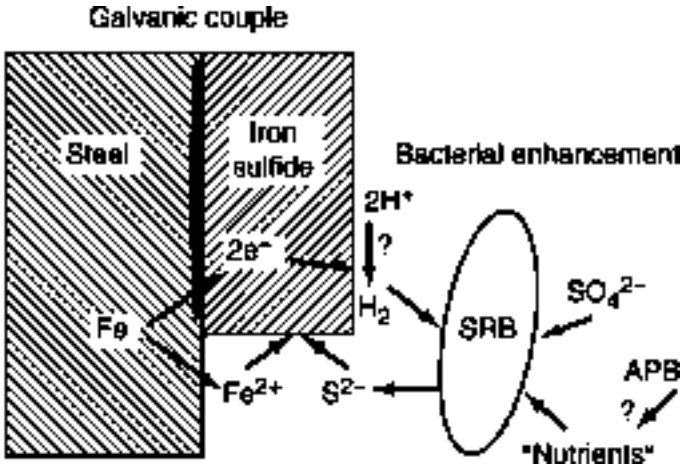


Fig. 4 Schematic diagram of the mechanism in an FeS corrosion cell created by the action of SRB. Iron sulfide sets up a galvanic couple with steel, sustained and extended by the further action of SRB. The bacteria use electrons from the corrosion process, possibly in the form of cathodic hydrogen, to reduce soluble sulfate. Enzyme assays for sulfate reductase and for hydrogenase have been developed as commercially available kits for assessing SRB activity in corrosion sites. The presence of specific iron sulfides is also used to identify this form of MIC. Acid-producing bacteria (APB) may have a role in providing nutrients to SRB, as suggested, and are often found in association.

Typical rates of metal loss for unprotected line pipe steel in an SRB/FeS corrosion scenario are 0.2 mm/year (0.008 in./year) for general corrosion and 0.7 mm/year (0.028 in./year) for pitting corrosion (Ref 24), but the corrosion rate observed depends on the concentration of FeS involved (Fig. 5). Consistent with the importance of this corrosion process in industrial facilities, commercial test kits have been developed for enumerating or assessing the activity of SRB in operating systems (Ref 25).

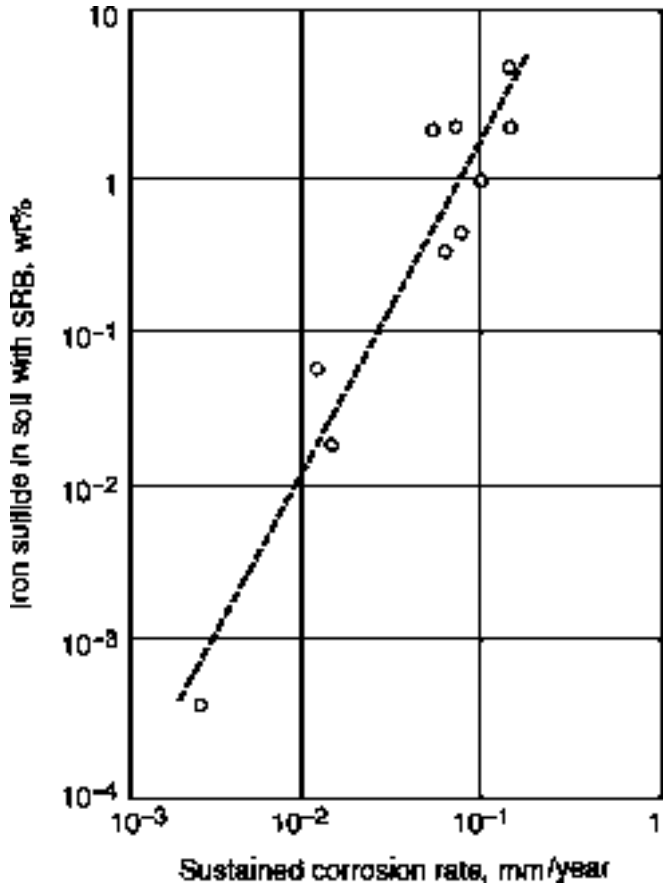


Fig. 5 Rate of corrosion of unprotected steel in biologically active soil as a function of iron sulfide present

Formation of MIC corrosion cells involving iron sulfide need not be limited to the SRB. Reference 26 describes the corrosive action of bacteria isolated from oil production facilities that are able to reduce thiosulfate to sulfide. These organic-acid-producing organisms could generate a corrosion cell based on iron sulfide but would not show up in SRB assays based on sulfate reduction.

Corrosion by Microbially Produced Manganese (IV). Recent work has shown that microbially produced MnO_2 can corrode (Eq 9) steel, provided that contact is made with the metal surface (Ref 27):

$$Fe + MnO_2 + 4H^+ \rightarrow Fe^{2+} + Mn^{2+} + 2H_2O \qquad \text{(Eq 9)}$$

Although the scenario is somewhat more complex than suggested by Eq 9 and 10 there is evidence that marine biofilms help to sustain the corrosion process by the rapid regeneration of MnO_2 (Eq 10) (Ref 28). Manganese-oxidizing organisms have been implicated in the MIC of 304L stainless steel welds (Ref 29). Formation of manganic oxide ennobles the potential of stainless steel in natural waters (Ref 30). Ennoblement shifts the potential above the repassivation potential to the pitting potential for stainless steel and furnishes the increased cathodic current density needed to propagate nucleated corrosion sites:

$$O_2 + Mn^{2+} \xrightarrow[\text{microbial catalysis}]{} MnO_2 \qquad \text{(Eq 10)}$$

Other electron acceptors involved in microbial metabolism include oxygen, nitrate, iron (III), and carbon dioxide. All of these species could theoretically be involved as the electron acceptor, *X*, in the corrosion of metals (Eq 1) but evidence for direct MIC based on other electron acceptors is limited. A more general discussion of possible links between electrochemical reactions mediated in microbial metabolism and corrosion is provided in Ref 18.

Indirect Involvement in the Corrosion Process

Microorganisms can influence the corrosion process by a number of less direct mechanisms.

Depolarization Mechanisms. As previously noted, a buildup of hydrogen on the cathodic surface can stifle the corrosion process through cathodic polarization (Fig. 3). Microorganisms with hydrogenase enzymes are able to

use hydrogen and have been widely cited as accelerating anaerobic corrosion through cathodic depolarization. Even though this concept has been challenged (Ref 6), commercial kits for hydrogenase activity are available for assessing MIC in practical applications (Ref 25).

Metabolites, such as organic acids produced by acid-producing bacteria (APB), may alleviate anodic polarization. Organic acids can form soluble chemical complexes with metal ions released by the corrosion process, reducing the buildup of M^{n+} on anodic surfaces. Figure 6 illustrates the damaging effect of trace levels of acetate (the conjugate base of the organic acid acetic acid) in the corrosion of steel exposed to carbon-dioxide-saturated brine. Acid-producing bacteria and organic acids at the levels shown in Fig. 6 have been identified in external corrosion sites in an extensive Gas Research Institute (GRI) program on pipeline MIC. Commercial kits and guides were produced for enumerating APB in field samples and for identifying related MIC sites (Ref 32, 33). Fungal production of organic acids has been implicated in the corrosion of posttensioned cables used in construction (Ref 34).

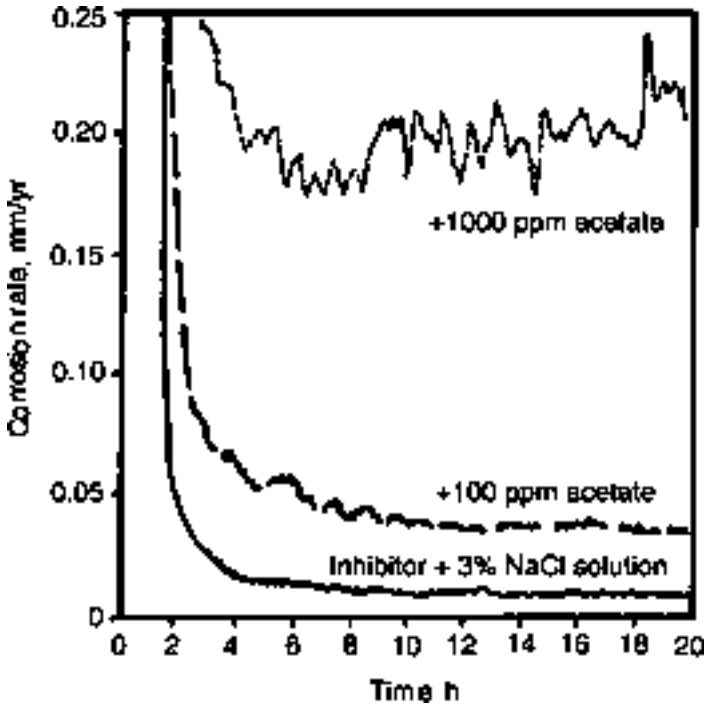


Fig. 6 Initial corrosion rates for steel exposed to carbonated 3% NaCl solution with 0, 100, and 1000 ppm acetate of 4, 9, and 18 mm/year (0.2, 0.4, and 0.7 in./year), respectively. This order is sustained after addition of 25 ppm of a corrosion inhibitor, despite a significant reduction in the corrosion rate in all cases. Source: Ref 31, reproduced with permission of National Association of Corrosion Engineers International

Production of Corrosive Metabolites. Other microbial metabolites have been recognized as agents of MIC. The many scenarios involving various sulfur species that are produced and consumed by microorganisms have been reviewed (Ref 35), and as-yet poorly described microbially produced phosphorus compounds in MIC have been implicated (Ref 11). It has been suggested that hydrogen peroxide produced in biofilms plays a role in the crevice corrosion of 304 and 316L stainless steel (Ref 36).

Alteration of Surface Environments to Create Concentration Cells. In many industrial water systems, colonization of metal surfaces by sessile microorganisms creates a new environment on the surface that may favor corrosion. During the colonization process, microbes from the aqueous phase quickly adhere to exposed steel surfaces and proceed to generate a slime made up of excreted extracellular polysaccharides (EPS). The slime envelops the organisms and creates a unique milieu for further development of microbial communities and possible MIC. The slime holds water, collects nutrients, and recruits other microbes as well as particulates and other debris from the liquid phase. Organisms in the biofilm are protected from direct exposure to biocides at dosages that would readily kill planktonic organisms freely suspended in the water phase. The metal surface is similarly shielded from doses of corrosion inhibitors added to the system.

Corrosion commonly occurs due to concentration cells established under the biofilm, where oxygen levels are depleted and metabolite (e.g., organic acid) concentrations may be high. Slime-forming organisms are a recommended target of monitoring programs for cooling water systems (Ref 37) as are iron- and manganese-oxidizing bacteria capable of laying down deposits of metal oxides on affected surfaces. Mobile anions such as chloride are attracted to the corroding surface under a biofilm and associated deposits by the release of positively charged metal ions. Concentration of chloride can accelerate the corrosion process further, especially in systems that rely on a passivating film on the metal surface to protect the underlying metal matrix from corrosion.

Alteration of Passivating Layers. Some metals are protected from corrosion by the formation of a passivating film of metal oxide or other insoluble deposits. It has been suggested that microbial acceleration of corrosion on copper is the result of incorporation of EPS into the copper oxide on the metal surface (Ref 38). This disrupts the passivating film and is accompanied by a reduction in pH due to metabolic activity. In a more complex example, 316L stainless steel exposed to a flowing microbial culture of *Citrobacter freundii* in the laboratory was selectively colonized at grain boundaries (Ref 39, 40). This led to the local depletion of chromium and iron content relative to nickel in the subsurface region of the superficial oxide. A coculture of *C. freundii* and the SRB, *Desulfovibrio gigas,* similarly selectively colonized grain-boundary regions and enhanced the local depletion of iron and the accumulation of sulfur. The authors of the study suggested that these changes weakened the oxide layer, predisposing the metal to a higher frequency of pitting due to attack by chloride. It has been noted that microorganisms can also produce organic acids and create conditions for formation of HCl or metal chlorides, including $FeCl_3$, under biofilms (Ref 2). These agents cause the passivating layer on stainless steel to fail, allowing pitting and crevice corrosion to proceed.

Massive failures due to pitting in water system components, such as utility condensers made of stainless steel, were identified as an MIC problem in the 1970s (Ref 6). Corrosion damage was ascribed to metastable sulfur oxyanions produced by SRB in the reduction of sulfate or formed in the oxidation of biogenic iron sulfide to elemental sulfur in oxygenated water (Ref 41). Sulfur oxyanions, such as thiosulfate and tetrathionate, have been shown to reduce the pitting potential of stainless steel and promote the localized corrosion of both stainless steels and nickel-base alloys (e.g., Inconels) with or without the presence of chloride (Ref 6, 41). One proposed mechanism involves the release of elemental sulfur by the disproportionation or reduction of thiosulfate anions at the metal surface. The elemental sulfur enhances the anodic dissolution process and accelerates pitting. Active SRB help to inhibit repassivation of the metal surface where chloride is present as a pitting agent.

Microbial Involvement in Environmentally Assisted Cracking

Environmentally assisted cracking (EAC) of susceptible materials under stress can be exacerbated by microbial activity. Potentially affected forms of EAC include sulfide-stress cracking (SSC), hydrogen-induced cracking (HIC), stress-oriented hydrogen-induced cracking (SOHIC), near-neutral pH stress-corrosion cracking (SCC), and corrosion fatigue. Accelerated corrosion and enhanced hydrogen uptake can increase the probability of cracking as well as subsequent crack growth rates.

Microorganisms can influence hydrogen uptake by a metal directly through the microbial production of hydrogen. Alternately, microbial activity can promote entry of cathodic hydrogen into the metal matrix through production of recombination poisons that prevent the formation and escape of molecular hydrogen from the metal surface (Fig. 3). Inside the steel, hydrogen concentrates in areas of stress, such as the plastic zone at the tip of a growing crack, or in areas of imperfection in the metal matrix, such as inclusions or grain boundaries. In the former location, hydrogen causes embrittlement of the metal and facilitates the cracking process in, for example, SCC (Ref 42). At imperfections, formation of molecular hydrogen can force voids to form in a metal matrix through a buildup of gas pressure. This can lead to macroscopic blistering and cracking within the steel matrix in the form of HIC or SOHIC. In general, these effects are of greater concern in higher-strength steels (Ref 43).

Only a modest amount of work has been done on the involvement of microorganisms in EAC relative to the potential importance of this topic. The HIC of high-tensile-strength hard-drawn steel wire used for prestressing concrete pipe on exposure to laboratory cultures of the hydrogen-producing anaerobe, *Clostridium acetobutylicum,* has been demonstrated (Ref 44). While this simple experiment demonstrates a principle, the

real-world situation is likely to be more complicated due to competing effects, such as passivation of the metal surface and consumption of hydrogen by the mixed microbial population likely to be present.

Metabolites, such as hydrogen sulfide or thiosulfate, can act as poisons for the hydrogen recombination reaction and drive cathodic hydrogen into steel (Fig. 3). Sulfide can accelerate crack growth rates even at trace concentrations. Corrosion fatigue crack growth rates (per stress cycle) in RQT 701 steel exposed to seawater have been shown to increase with increasing levels of sulfide exposure in either biotic or abiotic environments (Ref 45). It has also been noted that hydrogen permeation through steel under CP in artificial seawater was greater for a mixed culture of bacteria producing 160 ppm of sulfide than it was for the same concentration of sulfide without bacteria. Attempts to demonstrate these effects more generally led to the conclusion that the microbial scenarios may involve inhibition of hydrogen permeation and crack growth as well as enhancement. Complex lab results have been reported on the enhanced permeation of hydrogen through carbon steel exposed to cultures of SRB isolated from oilfield waters (Ref 46). It was suggested that an initial enhancement in hydrogen flux caused by the bacteria was stifled for a time by formation of a protective iron sulfide (mackinawite) on the steel surface. Accelerated hydrogen permeation was later reestablished, an effect that the researchers attributed to production of additional sulfide and organic acids by the organisms disrupting the protective film.

In addition to corrosion and hydrogen effects, the buildup of certain microbial metabolites in an environment may also foster crack growth by EAC. One example is the production of high levels of CO_2 needed to sustain near-neutral pH SCC (Ref 42). This CO_2 may arise in SCC sites on buried pipelines through biological activity in the surrounding environment or be formed more directly by biodegradation of susceptible components in protective coatings or other materials used in construction of a facility (Ref 21).

Metabolic intermediates in the microbial oxidation and reduction of sulfur can also cause trouble (Ref 35). Metastable sulfur oxyanions (e.g., thiosulfate and tetrathionate) can promote severe intergranular stress-corrosion cracking (IGSCC) of sensitized austenitic stainless steels and nickel-base Inconel alloys at very low concentrations in acidic to slightly acidic solution (Ref 41). As noted previously, metastable sulfur oxyanions can be produced by microbial activity or by secondary oxidation of biogenically produced iron sulfides. One suggested mechanism for their role in IGSCC involves disproportionation of the oxyanion to release elemental sulfur in the acidified crack tip. This would enhance anodic dissolution of chromium-depleted grain boundaries, enabling accelerated cracking. In this case, promotion of crack growth results from enhanced corrosion at the crack tip rather than a hydrogen embrittlement effect.

General souring of an operating environment through the biological formation of hydrogen sulfide is a serious concern in the oil and gas sector. Significant concentrations of hydrogen sulfide require production facilities to be constructed for “sour service” to avoid integrity problems. Souring of oil and gas reservoirs during prolonged production or of gas storage caverns over time can threaten production infrastructure designed to handle “sweet” sulfide-free crude oil or gas. Consequently, control of in situ souring has attracted considerable effort (Ref 47).

References cited in this section

2. B. Little and P. Wagner, An Overview of Microbiologically Influenced Corrosion of Metals and Alloys Used in the Storage of Nuclear Wastes, *Can. J. Microbiol.*, Vol 42, 1996, p 367–374

6. J.F.D. Stott, What Progress in the Understanding of Microbially Induced Corrosion Has Been Made in the Last 25 Years? A Personal Viewpoint, *Corros. Sci.*, Vol 35, 1993, p 667–673

9. W.A. Hamilton, and W. Lee, Biocorrosion, *Sulfate-Reducing Bacteria*, Vol 8, *Biotechnology Handbooks*, L.L. Barton, Ed., Plenum Press, London, 1995, p 243–264

11. W.P. Iverson, Research on the Mechanisms of Anaerobic Corrosion, *Int. Biodeterior. Biodegrad.*, Vol 47 (No. 2), 2001, p 63–70

18. D.A. Jones and P.S. Amy, Related Electrochemical Characteristics of Microbial Metabolism and Iron Corrosion, *Ind. Eng. Chem. Res.*, Vol 39, 2000, p 575–582

20. T.R. Jack, M.J. Wilmott, and R.L. Sutherby, Indicator Minerals Formed during External Corrosion of Line Pipe, *Mater. Perform.,* Vol 34 (No. 11), 1995, p 19–22

21. T.R. Jack, M.M. Francis, and R.G. Worthingham, External Corrosion of Line Pipe, Part II: Laboratory Study of Cathodic Protection in the Presence of Sulfate-Reducing Bacteria, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 339–350

22. R.G. Worthingham, T.R. Jack, and V. Ward, External Corrosion of Line Pipe, Part I: Identification of Bacterial Corrosion in the Field, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 330–339

23. J.A. Hardy, Utilization of Cathodic Hydrogen by Sulphate-Reducing Bacteria, *Br. Corros. J.,* Vol 18, 1983, p 190–193

24. T.R. Jack, Chapter 6 in *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

25. P.J.B. Scott and M. Davies, Survey of Field Kits for Sulfate-Reducing Bacteria, *Mater. Perform.,* Vol 31, 1992, p 64–68

26. J.-L. Crolet and M.F. Magot, Non-SRB Sulfidogenic Bacteria in Oilfield Production Facilities, *Mater. Perform.,* Vol 35 (No. 3), 1996, p 60–64

27. B.H. Olesen, P.H. Nielsen, and Z. Lewandowski, Effect of Biomineralized Manganese on the Corrosion Behaviour of C1008 Mild Steel, *Corrosion,* Vol 56 (No. 1), 2000, p 80–89

28. D.T. Ruppel, S.C. Dexter, and G.W. Luther III, Role of Manganese Dioxide in Corrosion in the Presence of Natural Biofilms, *Corrosion,* Vol 57 (No. 10), 2001, p 863–873

29. C. Zhigang, P. Gumpel, and M. Kaesser, Microbiologically Influenced Corrosion of Weld on Stainless Steel 304L, *Chin. J. Mech. Eng.* (*Engl. Ed.*), Vol 11 (No. 2) 1998, p 159–164

30. W.H. Dickinson and Z. Lewandowski, Electrochemical Concepts and Techniques in the Study of Stainless Steel Ennoblement, *Biodegrad.,* Vol 9, 1998, p 11–21

31. B. Hedges, D. Paisley, and R. Woollam, The Corrosion Inhibitor Model, *Mater. Perform.,* Vol 40 (No. 6), 2001, p 40–44

32. "Microbiologically Influenced Corrosion (MIC): Methods of Detection in the Field," GRI Report GRI-88/0113, Gas Research Institute, 1991

33. "Microbiologically Influenced Corrosion (MIC) II: Internal MIC and Testing of Mitigation Measures," GRI Report GRI-92/0005, Gas Research Institute, 1992

34. B. Little, R. Staehle, and R. Davis, Fungal Influenced Corrosion of Post-Tensioned Cables, *Int. Biodeterior Biodegrad.,* Vol 47, 2001, p 71–77

35. B.J. Little, R.I. Ray, and R.K. Pope, Relationship Between Corrosion and the Biological Sulfur Cycle: A Review, *Corrosion* Vol 56 (No. 4), 2000, p 433–443

36. H. Amaya and H. Miyuki, Mechanism of Microbially Influenced Corrosion on Stainless Steels in Natural Seawater, *J. Jpn. Inst. Met.,* Vol 58 (No. 7), 1994, p 775–781

37. H.M. Herro and R.D. Port, *The Nalco Guide to Cooling Water System Failure Analysis,* Nalco Chemical Company, McGraw-Hill, Inc., 1993

38. B.J. Webster, S.E. Werner, D.B. Wells, and P.J. Bremer, Microbiologically Influenced Corrosion of Copper in Potable Water Systems—pH Effects, *Corrosion,* Vol 56 (No. 9), 2000, p 942–950

39. G.G. Geesey, R.J. Gillis, R. Avci, D. Daly, M. Hamilton, P. Shope, and G. Harkin, Influence of Surface Features on Bacterial Colonization and Subsequent Substratum Chemical Changes of 316L Stainless Steel, *Corros. Sci.,* Vol 38 (No. 1), 1996, p 73–95

40. G.G. Geesey, R.J. Gillis, H.-J. Zhang, and P.J. Bremer, Influence of Surface Features on Microbial Colonization and Susceptibility of Stainless Steels Used in the Food Processing Industry, *Proc. IIW Asian Pacific Welding Congress,* 4–9 Feb 1996 (Auckland, New Zealand), 1996, p 693–707

41. G. Cragnolino and O.H. Tuovinen, The Role of Sulphate-Reducing and Sulphur-Oxidizing Bacteria in the Localized Corrosion of Iron-Base Alloys—A Review, *Int. Biodeterior.,* Vol 20 (No. 1), 1984, p 9–26

42. R.N. Parkins, W.K. Blanchard, Jr., and B.S. Delanty, Transgranular Stress Corrosion Cracking of High-Pressure Pipelines in Contact with Solutions of Near Neutral pH, *Corrosion,* Vol 50 (No. 5), 1994, p 394–408

43. *Hydrogen Transport and Cracking in Metals,* A. Turnbull, Ed., The Institute of Materials, London, 1995

44. T. Ford and R. Mitchell, Microbiological Involvement in Environmental Cracking of High Strength Steels, *Microbially Influenced Corrosion and Biodeterioration,* N.J. Dowling, M.W. Mittleman, and J.C. Danko, Ed., University of Tennessee, 1990, p 3–94 to 3–97

45. R.G.J. Edyvean, J. Benson, C.J. Thomas, J.B. Beach, and H.A. Videla, "Biological Influences on Hydrogen Effects in Steel in Seawater," Paper 206, presented at Corrosion '97, National Association of Corrosion Engineers, 1997

46. H. Liu, L. Xu, and J. Zeng, Role of Corrosion Products in Biofilms in Microbiologically Induced Corrosion of Steel, *Br. Corros. J.,* Vol 35 (No. 2), 2000, p 131–135

47. M. Nemati, G.E. Jenneman, and G. Voordouw, Mechanistic Study of Microbial Control of Hydrogen Sulfide Production in Oil Reservoirs, *Biotechnol. Bioeng.,* Vol. 74 (No. 5), 2001, p 424–434

Biological Corrosion Failures

Thomas R. Jack, NOVA Chemicals Ltd.

Degradation of Protective Systems

Microbiological influence on the integrity of metal systems need not involve direct mediation of corrosion or cracking processes. In some situations, microbiological activity predisposes metals to integrity problems through degradation of protective systems, including corrosion inhibitors, coatings, CP, or lubricants.

Corrosion inhibitors are chemicals introduced into industrial systems to reduce the rate of metal loss. In some cases, these chemicals act as nutrients for MIC problems and are destroyed through biodegradation. In a long-term audit of water recycle systems in a large oilfield, failure costs were found to be inversely related to the residual concentrations of treatment chemicals used to control corrosion, as one would hope. The chemical treatment package included a filming amine corrosion inhibitor and ammonium bisulfite as an oxygen

scavenger. Closer inspection of the data, however, showed that the residual concentrations of the oxygen scavenger correlated with measured corrosion rates, indicating that the ammonium bisulfite was actually being used as a nutrient by the SRB responsible for the corrosion observed. The microbial population had to be independently controlled by targeted biocide addition (Ref 48). This problem of conflicting effects by different additives is a common issue in the design of cost-effective chemical treatment programs.

As environmental concerns push treatment chemical suppliers to less toxic, biodegradable materials (see Ref 49 for an example), competition between biocides intended to reduce microbial populations and corrosion inhibitors that may feed them will become more common.

Protective Coatings. Buried structures and interior surfaces of tanks and piping are often protected by the application of protective coatings. While these coatings have improved over the years, many older coatings were susceptible to biodegradation in service. Microbial attack can be seen, for example, in some polyvinyl chloride (PVC) coatings. While the polymer itself remains relatively inert due to its large molecular size, biological degradation of plasticizers used to make the material flexible can occur (Ref 50). Loss of plasticizer in older PVC tape coatings on pipelines has led to embrittlement and coating failure in service. In a related example, the adhesive used to affix polyethylene tape coatings to line pipe has been found to be a source of nutrients for microbial sulfate reduction in external corrosion sites (Ref 21). Biodegradation of adhesive may contribute to the loss of adhesion seen when these coatings are exposed to biologically active soils (Ref 51). Coating failures of this sort block CP and have led to corrosion and SCC problems on operating pipeline systems (Ref 22, 42).

The modern trend to more environmentally friendly, safer materials also affects the formulation of paints and coatings used to protect industrial systems. This has raised concern that these changes will coincidentally result in a loss of resistance to microbial degradation for products in service (Ref 52). A conflicting strategy of adding more biocides to new formulations has been proposed.

Cathodic Protection. Buried or submerged structures and some interior surfaces in processing facilities are protected by imposition of an electrical potential that prevents oxidation of metal. This can be achieved through the use of sacrificial anodes or by an impressed current system. In either case, maintenance of a protective potential requires the provision of an adequate current density. Production of iron sulfides by active SRB can compromise CP potentials by draining current from exposed steel surfaces. Figure 7 shows the increase in current demand in a laboratory soil box with time as SRB active in the soil produce increasing quantities of electrically conducting iron sulfides around steel coupons at a fixed potential. An increased current demand has also been seen for stainless steel surfaces when a steady-state biofilm is present (Ref 53) and in lab experiments involving an anaerobic slime-forming APB (*Vibrio natriegens*) in pure culture (Ref 54). In the latter case, production of organic acids and exopolymers by *V. natriegens* affected the formation of calcerous scales that would normally seal the cathodically protected surface, reducing current demand. In locations where increased current demand cannot be met, the protective potential is lost. In the field, affected locations may show up as a local dip in close-interval-potential surveys.

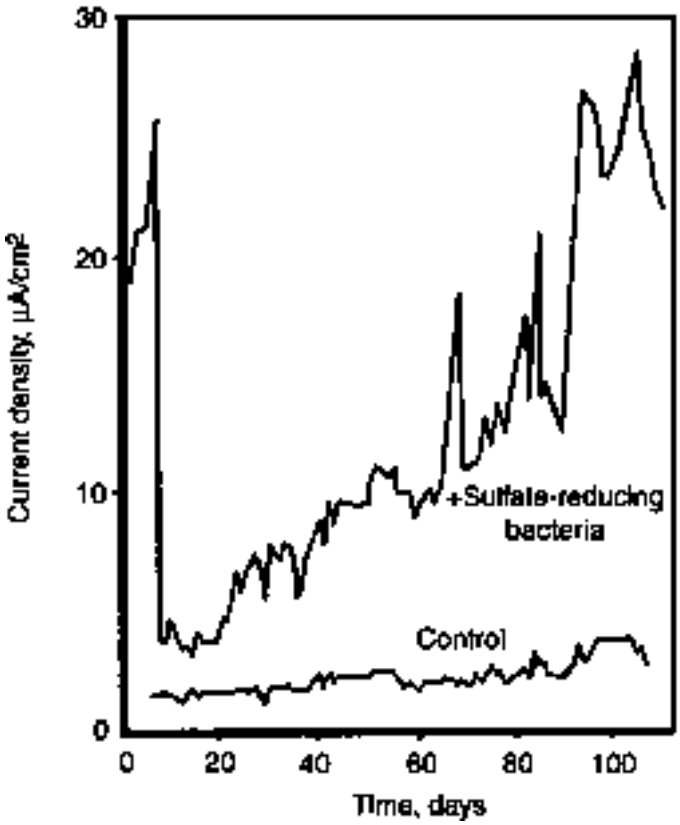


Fig. 7 Current demand needed to sustain a set CP potential increases with time for steel in soil as SRB produce increasing quantities of iron sulfide.

Lubricants protect metal components from corrosion and wear in service. In some cases, these lubricants contain biodegradable hydrocarbons that support microbial growth where water is available and other conditions permit. Microbial growth is, in turn, often accompanied by the production of surfactants that stabilize water/oil emulsions. These extend the interface available for microbial activity and can be quite stubborn to break. The combined effect of biodegradation and emulsification can compromise the performance of lubricants, leading to failure of metal components (Ref 55). A case study involving the failure of bearings in a marine engine is given in Ref 56. Commercial kits are available to track microbial growth in lubricants and identify impending problems in large marine engines as well as other applications.

References cited in this section

21. T.R. Jack, M.M. Francis, and R.G. Worthingham, External Corrosion of Line Pipe, Part II: Laboratory Study of Cathodic Protection in the Presence of Sulfate-Reducing Bacteria, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 339–350

22. R.G. Worthingham, T.R. Jack, and V. Ward, External Corrosion of Line Pipe, Part I: Identification of Bacterial Corrosion in the Field, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 330–339

42. R.N. Parkins, W.K. Blanchard, Jr., and B.S. Delanty, Transgranular Stress Corrosion Cracking of High-Pressure Pipelines in Contact with Solutions of Near Neutral pH, *Corrosion,* Vol 50 (No. 5), 1994, p 394–408

48. T.R. Jack, E. Rogoz, B.J. Bramhill, and P.R. Roberge, Characterization of Sulfate-Reducing Bacteria in Heavy Oil Waterflood Operations, *Microbially Influenced Corrosion Testing,* ASTM STP 1232, J.R. Kearns, and B.J. Little, Ed., American Society for Testing of Materials, 1994, p 108–117

49. C. Chandler, Biodegradable Volatile Corrosion Inhibitors for Offshore and Onshore Installations, *Mater. Perform.,* Vol 40 (No. 2), 2001, p 48–52

50. E.S. Pankhurst, Protective Coatings and Wrappings for Buried Pipelines: Microbiological Aspects, *J. Oil Colour Chem. Assoc.,* Vol 56, 1973, p 373–381

51. T.R. Jack, G. Van Boven, M. Wilmott, and R.G. Worthingham, Evaluating Performance of Coatings Exposed to Biologically Active Soils, *Mater. Perform.,* Vol 35 (No. 3), 1996, p 39–45

52. J.B. Santon, Paint and Coatings—New Trends Respond to Environmental Concerns, *Mater. Perform.,* Vol 39 (No. 9), 2000, p 22–25

53. M.F.L. de Mele, S.G.G. de Saravia, and H.A. Videla, An Overview on Biofilms and Calcerous Deposits Interrelationships on Cathodically Protected Steel Surfaces, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 50/1–50/8

54. J.G. Guezennec, Cathodic Protection and Microbially Induced Corrosion, *Int. Biodeterior. Biodegrad.,* 1994, p 275–288

55. R. Stuart, Something Nasty in the Crank Case, *Ind. Lubr. Tribol.,* Vol 46 (No. 2), 1994, p 5–7

56. I.B. Beech, S.A. Campbell, and F.C. Walsh, Marine Microbial Corrosion, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001, p 11.3–11.14

Biological Corrosion Failures

Thomas R. Jack, NOVA Chemicals Ltd.

Failure Analysis

Although MIC is acknowledged to occur on a wide range of metals, most reported failure analyses have focused on iron, copper, aluminum, and their alloys. This is not surprising, given the importance of these metals in industrial applications.

Steel and Iron

Industry Experience. Failure analysis for steel and iron can begin with an assessment of the susceptibility of an operating system to MIC, based on industry experience. It appears that almost any system with free water present can host microbial activity, provided there is a source of nutrients and an absence of toxic material. In processing facilities, an increase in pressure drop or loss of flow through piping systems (Fig. 8) or a marked decline in the efficiency of heat exchangers are indicators of possible biological fouling. A history of failures due to unexpectedly high corrosion penetration rates may indicate MIC. Favored locations for failures include low spots where water can collect in piping, dead-end or shut-in piping, tank bottoms at the interface between liquid hydrocarbons and underlying water, or areas heavily fouled with surface deposits (Fig. 9). Microbially induced corrosion problems are found in refineries, cooling water systems, sprinkler water systems, oilfield water-handling systems, oil pipelines, and pulp and paper mills but are not recognized as a major issue in steam systems (Ref 57).

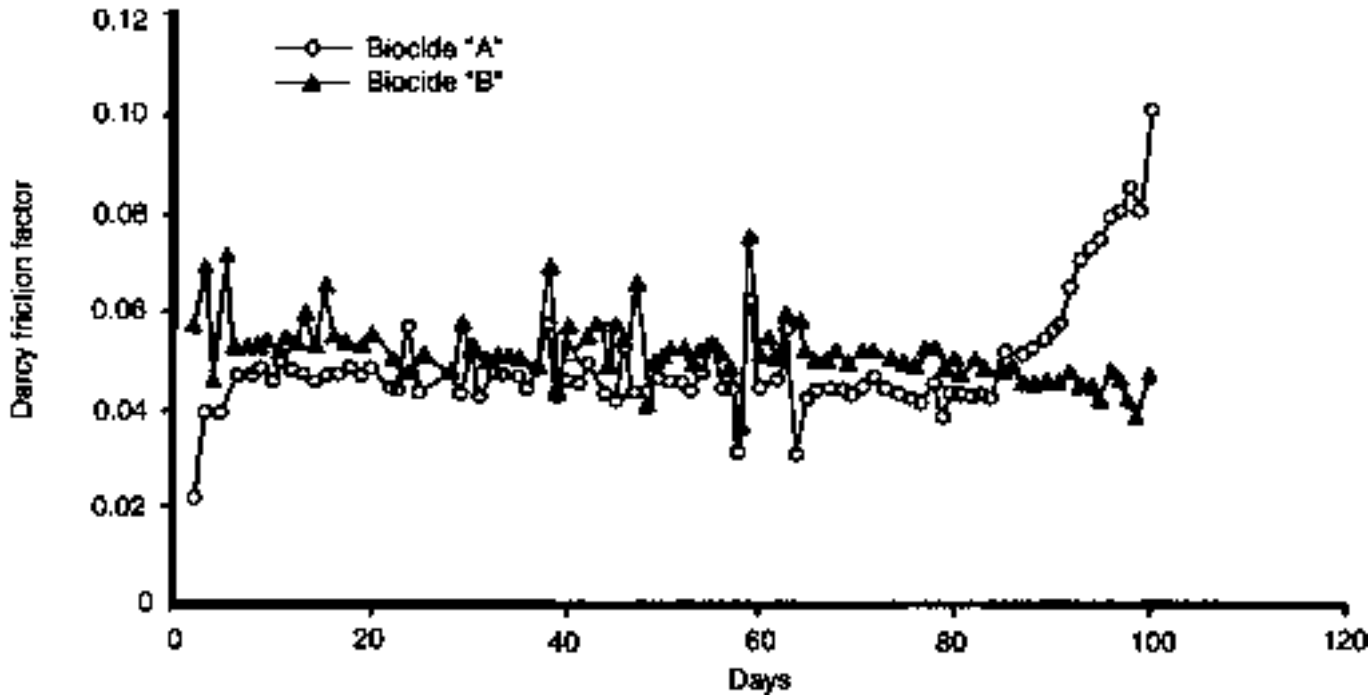


Fig. 8 Friction factors for flow through instrumented sidestreams under biocide treatment in an oilfield-produced water system can indicate biofouling. In this case, biocide "A" loses control of microbial fouling after 80 days, relative to biocide "B," in a comparison carried out in a field performance evaluation facility equipped with parallel test lines (see Fig. 14).

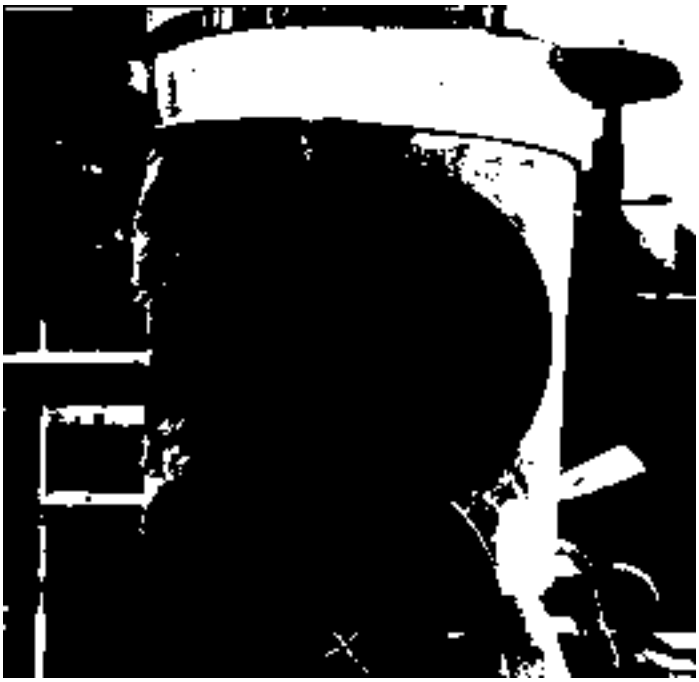


Fig. 9 Dead-end piping reveals extensive black deposits through an open flange in refinery piping.

For external corrosion, past experience of susceptible metal surfaces exposed to a wide range of natural environments has led to a general understanding of where corrosive conditions, including MIC, can be expected to occur (Ref 58, 59).

For soils, corrosion rates for steel and iron have been correlated with the pH, oxidation reduction potential, resistivity, and water content of the soil as well as with the type of soil. Very dry soils that lack the electrolyte needed to support active corrosion cells are benign. Moist, aerobic soils, where oxygen can readily reach exposed steel, show corrosion rates typically in the range of 0.04 to 0.2 mm/year (0.002 to 0.008 in./year). Anaerobic soil environments, where oxygen is absent and where SRB have not precipitated extensive iron

sulfide deposits, show intermediate corrosion rates of 0.002 to 0.01 mm/year (0.00008 to 0.0004 in./year). Where the focused action of SRB has exposed unprotected steel to extensive iron sulfide deposits, very high corrosion rates, >0.2 mm/year (>0.008 in./year), can be seen.

Attempts have been made to identify factors in soil environments that favor MIC by SRB for buried pipe (Ref 60, 61). Table 1 summarizes conditions that correlate with elevated SRB populations. Based on correlations of this sort, predictive models have been developed to prioritize maintenance activities in particular areas.

Table 1 Factors correlating with sulfate- reducing bacteria (SRB) numbers for buried pipeline sites

Factor	**Correlation coefficient**	Range
Bacterial numbers (acid-producing bacteria)	0.829	**10^3-10^8 cells/g wet soil**
Total organic carbon in groundwater	0.645	**0.05–1.2%**
Soil resistivity	-0.642	**500–30,000 Ω · cm**
Soil water content	0.626	**5–36%**
Soil oxidation reduction potential	-0.545	**-316 to 384 mV (CSE)**
Sulfate in groundwater	0.455	**0.3–200 mg/g wet soil**
Clay	**0.407**	**N/A**

Note: CSE, copper-copper sulfate electrode. Source: Ref 60

Anaerobic sediments in freshwater or marine environments generally host anaerobic microbial populations, including SRB that can create MIC problems. Guides have been developed at the University of Manchester Institute of Science and Technology to assess the potential risk of MIC due to SRB for sheet piling in sediments in ports and harbors and for design of CP systems on subsea pipelines (Ref 62).

Identification of MIC Sites. Possible MIC sites are often tentatively identified by the rapidity, severity, and localized nature of the corrosion. Confirmation is based on analysis of metallurgical damage, microorganisms, corrosion products, deposits, and environmental factors. There is no single approach to this problem, but a number of guides have been published (Ref 5, 7, 32, 33, 37, 63).

Guides for identification of internal and external MIC were developed through a major GRI program for natural gas pipelines. The guides provide a scorecard for the identification of MIC sites on mild steel, based on three types of evidence (Ref 32, 33):

- *Metallurgical:* appearance of the corrosion damage
- *Biological:* cell counts for SRB, APB, and general bacterial population
- *Chemical:* identification of corrosion products and other deposits

The metallurgical evidence is given the heaviest weighting, sufficient to pronounce MIC as being “very likely” in the absence of any other evidence from biological or chemical tests. Key features of the metallurgical fingerprint ascribed to MIC include corrosion pits composed of several smaller pits with hemispherical or cuplike appearance, striations in the direction of rolling in the steel pit bottom, and tunneling into the sides of the corrosion pit (Fig. 10). The guides provide appropriate photographs and methods for identification of these features. Biological analyses assess the number of viable SRB, APB, and anaerobic bacteria present. Chemical analysis of corrosion products has limited application in the scorecard and is more fully addressed later. Commercial kits have been developed for the analysis of the biological and chemical criteria used in the guides. Use of these tools is not confined to pipeline applications or soil environments. They are generally useful wherever the corrosion of carbon steel is observed (Ref 64).

Fig. 10 Microbially induced corrosion showing striations in the bottom and tunneling into the walls of overlapping pits, as described in Ref 32

A complementary approach to use of an MIC scorecard is based on the analysis of corrosion products and other deposits found on the corroded surface (Ref 20). This approach was developed for pipeline facilities suffering external corrosion (Ref 22). Qualitative chemical analysis can be done in the field to identify the presence of iron (II), iron (III), calcium, sulfide, and carbonate. Addition of acid to a sample can detect carbonate (by extensive bubble formation) and sulfide (by the odor of hydrogen sulfide or by its reaction with a color indicator such as lead acetate). Iron (II) and (III) can be detected in the resulting acid solution using standard ferricyanide and thiocyanate tests. Addition of a solution of oxalate to the test solution after buffering with excess sodium acetate yields a white precipitate if calcium is present. These qualitative chemical assays have been made available commercially in the form of a test kit through the GRI program noted previously. X-ray diffraction (XRD) analysis of crystalline material in corrosion products and associated deposits can provide more detailed insight (Ref 24). A summary of corrosion products indicating different corrosion scenarios is given in Table 2.

Table 2 Indicator minerals found as corrosion products in various corrosion scenarios seen in pipeline excavations and laboratory soil box tests

Corrosion scenario	**Corrosion products (color, chemistry, mineral form)**	Corrosion rate, mm/year
Simple corrosion processes		
Abiotic aerobic corrosion (O_2 is the electron acceptor, *X*, in cathodic reaction, Eq 2)	Yellow/orange/brown/black iron (III) oxides, including lepidocrosite, goethite, magnetite, maghemite, hematite	**0.04–0.2**
Abiotic anaerobic corrosion (H^+ as *X* in cathodic reaction, Eq 2)	Pasty or dispersed white iron (II) carbonate (siderite)	**0.002–0.01**
Anaerobic MIC (SRB with biotic iron sulfide as *X*, Eq 2)	Black, finely divided iron (II) sulfides, including amorphous iron sulfide, mackinawite, greigite	**0.2 general** **0.7 pitting**
Secondary transformations involving MIC		
Aerobic → anaerobic MIC (SRB/“FeS”)	Iron (II) sulfides, including marcasite and pyrite	**…**

Anaerobic MIC (SRB/"FeS") → aerobic	**Elemental sulfur, iron (III) oxides + residual anaerobic corrosion products**	**2–5[(a)]**

(a) This very high corrosion rate may not be sustained beyond the period of secondary oxidation of the anaerobic site.

The presence of a hard, white, calcium carbonate scale is good evidence that effective CP potentials were achieved on metal surfaces where it is found. Identification of the very severe corrosion scenario due to iron sulfide corrosion cells developed by SRB can be made based on analysis of corrosion products, as noted in Table 2. The observation of iron (II) sulfide in association with severe pitting or areas of significant metal loss in contact with dense, black corrosion products is definitive for this scenario. In some failure sites, extensive iron (II) sulfide present as a finely divided, black solid causes the general blackening of soil and corrosion deposits (Fig. 11). The additional presence of iron (III) suggests secondary oxidation has occurred, with potentially very serious corrosion consequences (Table 2) (Ref 24). In extreme cases, overlapping corrosion pits can lead to extensive areas of metal loss sufficient to cause pressurized piping or vessels to fail through rupture. This type of failure presents an inherently greater risk than formation of a leak due to perforation of the steel by isolated pitting.

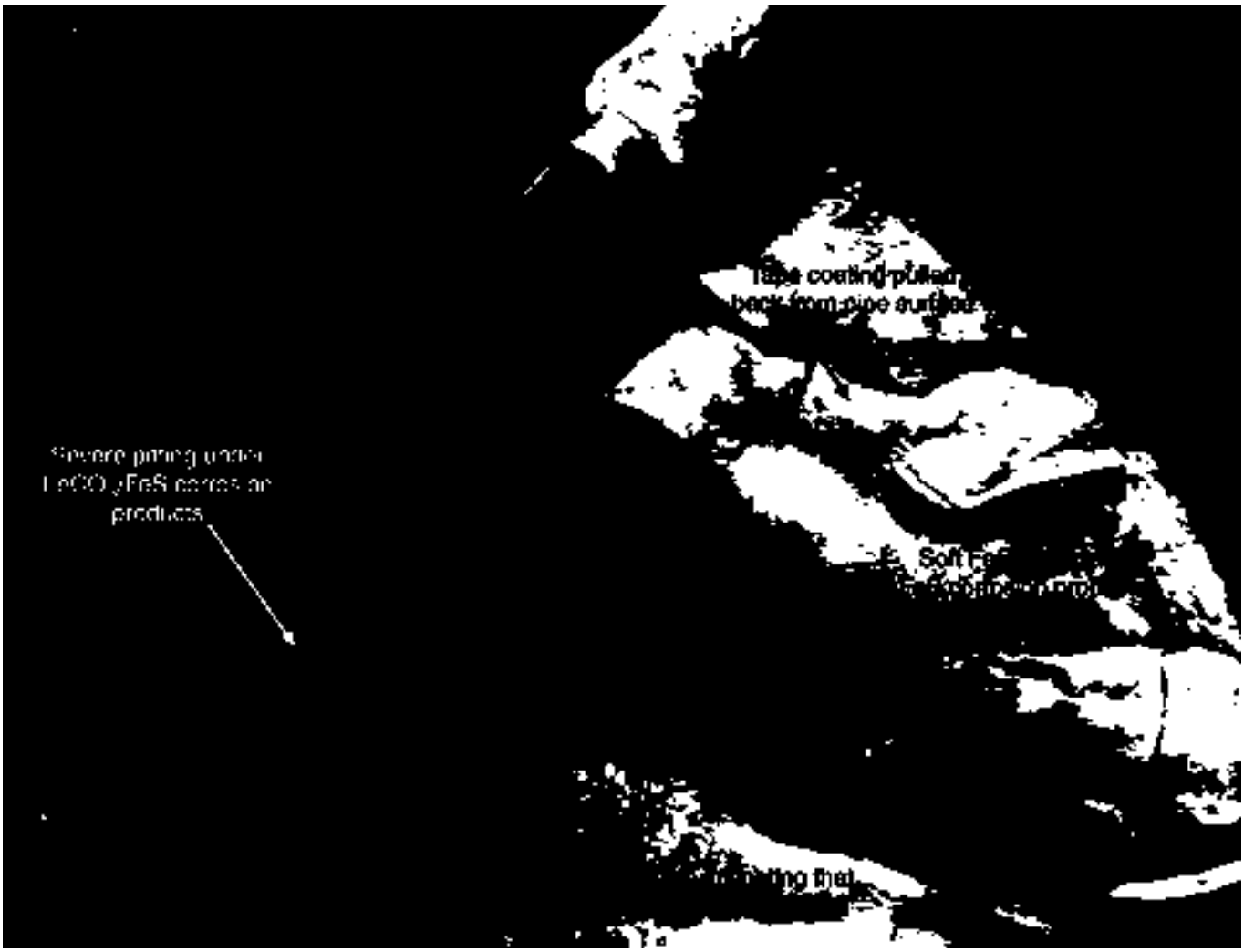


Fig. 11 FeS/SRB corrosion under disbonded polyolefin tape coating on a buried pipeline in a wet anaerobic location shows soft, black corrosion products rich in iron sulfides overlying severe pitting

Table 2 does not include corrosion scenarios due to APB. This possibility can be assessed using the methods described in the MIC field guides described previously or by analysis of deposits for organic acids.

A cryptic guide to the identification of corrosion in oil and gas operations offers two MIC scenarios (Ref 65). One, based on SRB, is identified by:

- Slope-walled pits within pits

- Attached SRB
- A source of sulfate

A second scenario, based on APB, is identified by:

- Steep-sided, sharp-edged, “fibery”-bottomed pits
- Attached APB
- Trace organic materials (presumably organic acids)

Reference 37 describes a wider range of tests and considerations for cooling water systems, and active and passive MIC scenarios are considered. In active scenarios, the organisms participate directly in corrosion processes. Microorganisms involved include SRB and acid producers. In this case, sulfur-oxidizing aerobic bacteria, such as *Thiobacilli*, that produce very acidic conditions and denitrifiers capable of reducing the pH to 3 are considered as well as the organic APB. Consideration of aerobic organisms is consistent with highly oxygenated environments found in cooling water circuits. Passive corrosion scenarios include underdeposit attack due to concentration cells set up under biofilms by slime-forming microbes or iron- and manganese-oxidizing bacteria (referred to as “metal depositors” by the researchers). Other organisms considered include algae that produce dense, thick mats of biomass on sunlit surfaces in cooling towers. These can foster underdeposit attack due to concentration cells and produce high levels of oxygen as well as nutrients for other organisms. In one case, ammonia released by the decay of algal biomass was blamed for SCC in a brass condenser.
The following four factors in the identification of corrosion as MIC were looked for (Ref 37):

- Presence of microorganisms or their byproducts
- Microbiologically unique corrosion morphology
- Specific corrosion products and deposits
- Compatible environmental conditions

The use of these factors for diagnosis of MIC scenarios in cooling water systems is addressed in Table 3.

Table 3 Factors for the diagnosis of MIC scenarios in cooling water systems

Microorganism (metabolite)	**Corrosion morphology**	Specific corrosion products and deposits
Active MIC		
Sulfate-reducing bacteria (sulfide)	Clustered hemispherical pits on stainless steel, Carpenter 20, aluminum, carbon steel. Rare on titanium. Copper poorly defined Very irregular pit surface in less noble metals	**Metal sulfides present** **Voluminous, brown, friable tubercles of iron (III) oxides over pit**
Acid producers (lower-pH organic acids for acid-producing bacteria)[(a)]	Corrosion is localized, moderate Striations in steel under tubercles, as for preferential acid dissolution of microstructure in rolling direction	**None stated**
Passive MIC		
Slimers (gelatinous mass with high microbial numbers)	**General corrosive attack under slime** **Pitting if SRB present**	**Rusting may color surfaces brown**

(a) Acid producers are often associated with SRB but outnumber them in this case. Organisms such as *Clostridia*, *Thiobacillus*, and *Nitrobacter* are cited as potential acid producers.
Source: Ref 37

Corrosion-Resistant Alloys of Steel

Microbially induced corrosion on stainless steel leads mostly to pitting or crevice corrosion failures. Statistical analysis of corrosion failures seen in once-through cooling systems on the Rhine River show that a disproportionate number occur in stainless steel systems relative to carbon steel. Estimates of MIC as a percent of total corrosion failures for stainless steel systems may be as high as 20% but are probably <10% (Ref 66).

Weldments in Stainless Steel. Most failures are associated with welds, because areas of joining tend to be inherently more susceptible to corrosive attack than the base material. Studies have shown that heat-tinted zones are especially vulnerable. These zones are created in a welding process where material above the scaling temperature is contacted by air. The result is a migration of chromium into the surface scale, leaving underlying material depleted and susceptible to corrosion. Removing the heat-tinted scale and underlying surface by pickling, electrochemical cleaning, or mechanical grinding prevents corrosion damage with or without bacteria (Ref 67, 68) for 304L, 308L, or 316L material. It has been suggested that pickling is the most effective approach (Ref 69). Heat tinting can be avoided by use of an effective inert gas blanket in the welding procedure.

Microbially induced corrosion is perhaps the only mechanism that can perforate stainless steel piping in neutral aqueous service, such as river water cooling, in a matter of months (Ref 70). It is suggested that this may be due to the ratio of cathodic to anodic areas, where a single phase, for example, ferrite, is preferentially attacked relative to a large area of less susceptible material. Penetration rates of 17 mm/year (0.7 in./year) in United Numbering System (UNS) 30800 welds have been reported. Perforation of 316L stainless steel weldments in piping with 5.5 mm (0.2 in.) wall thickness in industrial water systems under intermittent flow in four months has been described (Ref 70).

Preferential attack of some sort is a common feature of MIC case studies (Ref 63, 70, 71, 72, 73). Corrosion is often focused on the weld material or at the fusion line for the weld (Fig. 12a). Pit surfaces are often described as dendritic, consistent with preferential corrosion (Ref 63, 71, 72, 73), but the preferential attack of a single phase need not be a feature of MIC. From detailed study of UNS 30800 weld specimens, it has been concluded that either ferrite or austenite can be preferentially attacked, or they may corrode together, depending on a number of possible conditions (Ref 70). Abiotic attack by $FeCl_3$ solutions has been found to give similar effects to MIC, based on comparison of chemically degraded specimens with samples from identified MIC sites (Ref 70, 74). Preferential attack can also occur in a single phase, due to cold work effects on microstructure (Ref 70, 74).

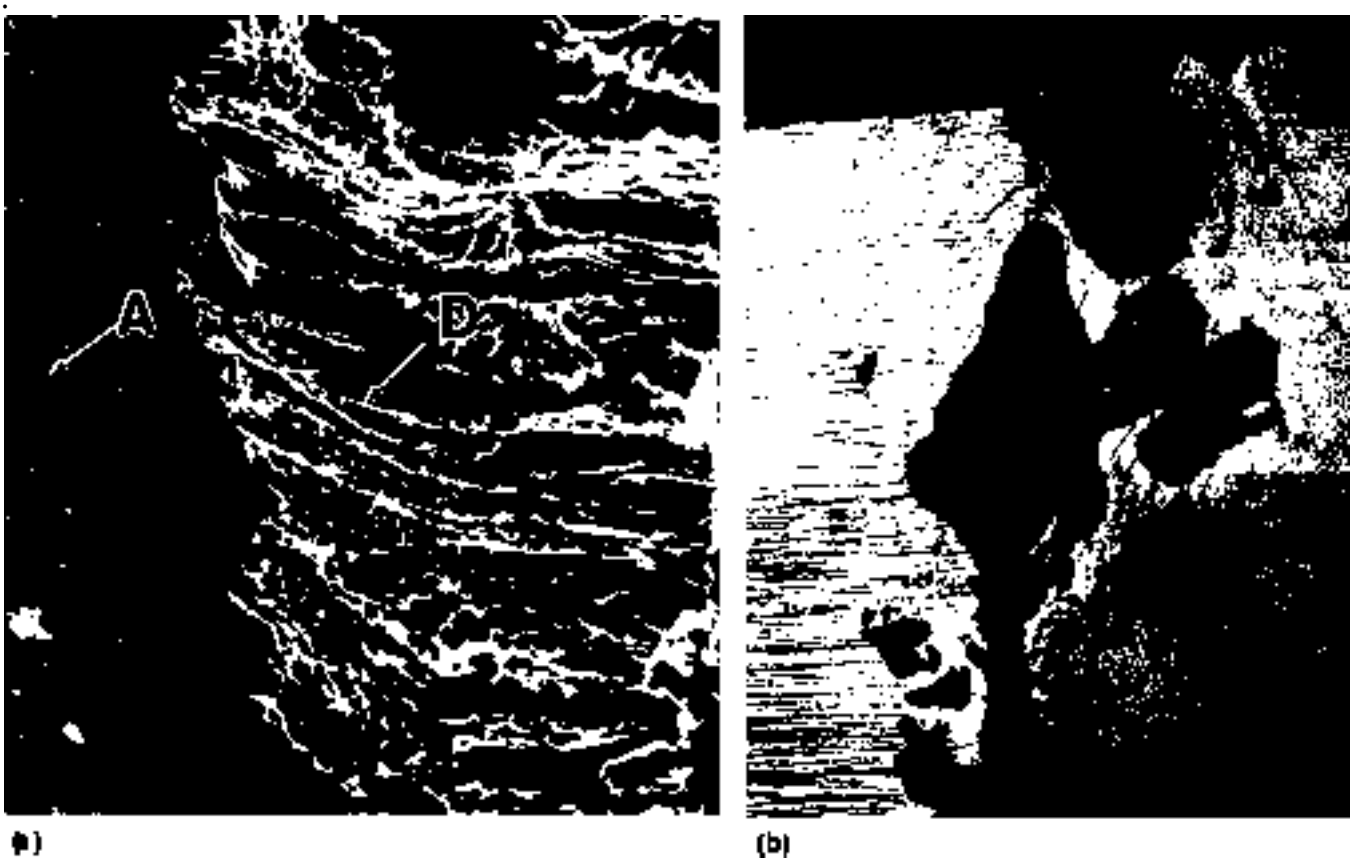


Fig. 12 MIC of stainless steel weldments. (a) MIC showing a surface view of interdendritic attack at the fusion line of a stainless steel weldment. "A," nondentritite; "D," dendrite. (b) Cross section of MIC at a stainless steel weldment showing extensive corrosion of weld metal and fusion line, with a relatively small opening at the bold surface. Source: Ref 71. Reproduced with permission of National Association of Corrosion Engineers International

As previously noted, microbially induced intergranular pitting and IGSCC can occur in sensitized stainless steels, where low chromium content at grain boundaries allows preferential dissolution (Ref 41). However, it has been found that transgranular pitting due to MIC in the heat-affected zone in socket-welded specimens of 304 stainless steel exposed to flowing lakewater (500 to 600 ppm chloride) over 6 to 18 months was not focused in sensitized areas (Ref 74). Instead, pitting occurred along deformation lines left by cold working of the metal during manufacturing. Annealing the material at 1150 °C (2100 °F) was suggested as a way to remove these features and increase resistance of the material to MIC. No pitting was seen in the base metal for either furnace-sensitized or girth-welded specimens of 304 or 316 stainless steel after similar exposure (Ref 74).
A number of trends seem apparent, based on past failure analyses:

- Microbially induced corrosion is often associated with stagnant, untreated water being left in piping over extended periods (Ref 63, 64, 65, 66, 67, 68, 69, 70, 71, 72). It has been suggested that intermittent flow or low flow rates are most damaging (Ref 75).
- Damage often occurs at many welds in an affected section of piping. In one power plant cooling system using lake water, radiography indicated that 50% of the welds in 316L piping showed indications of deep MIC pitting (Ref 72).
- Pitting seems more prevalent in the bottom third of the pipe (Ref 63).
- Low pH or high chloride concentrations in the pit environment enhance attack (Ref 76).

It has been noted that higher alloying in weld combinations seems to improve resistance to MIC (Ref 76). This observation is supported by a systematic laboratory study of stainless steel (304, 316L, and 317L) and Ni-Cr-Mo (alloy 625) alloy weldments cleaned of surface thermal oxides (Ref 77); however, later work on as-received welds showed that thermal oxides produced during the welding process can obscure this dependence. In all cases, exposure of specimens to lake water augmented by active SRB reduced the polarization resistance of the alloys relative to sterile controls. This was true even for alloys with 9% Mo content; however, no documented corrosion failures due to MIC in alloys with 6% Mo or more could be found to support the idea that elevated molybdenum content can provide added resistance to corrosion damage (Ref 2).
It was recommended that failure analysis for stainless steel cooling systems include biological analysis of associated water and deposits, chemical analysis of water, and radiography of welds (Ref 63). Microbially induced corrosion in pitting at weldments in stainless steel was identified by (Ref 74):

- The combination of bacteria present and morphology of pits
- Corrosion features with small surface openings leading to bulbous cavities in the steel matrix at welds
- The absence of other agents that could account for the attack

While SRB in mixed populations are a favorite for laboratory studies, and sulfides are often found in associated deposits on affected metal surfaces in the field, the microbiology found in case studies tends to be complex. A wide range of organisms can be present, especially in cooling systems drawing on natural waters. All sorts of bacteria were found to be present in once-through cooling systems using untreated river water, including sulfur oxidizers, iron oxidizers, iron reducers, SRB, nitrogen oxidizers, and denitrifiers (Ref 66). Aerobes, anaerobes, SRB, and APB were reported to be present in slimes and nodules on the metal surface (Ref 74). The presence of iron oxidizers and slimers for MIC problems in UNS S30800 stainless steel welds has been cited (Ref 69), while *Gallionella* in characteristic MIC pits has been specifically identified (Ref 63). Enhanced numbers of manganese-oxidizing bacteria have been noted in deposits formed on corroded welds in 304L stainless steel specimens exposed to Lake of Constance water in lab studies (Ref 29).
Surface deposits in nine case studies (Ref 63, 66, 67, 68, 72) contained iron, silicon, and sulfur in >75% of the samples analyzed (Fig. 13). Manganese, chromium, and aluminum were also frequently found (>50% of samples), but more soluble ions, such as chloride and potassium, were detected in less than half the samples.

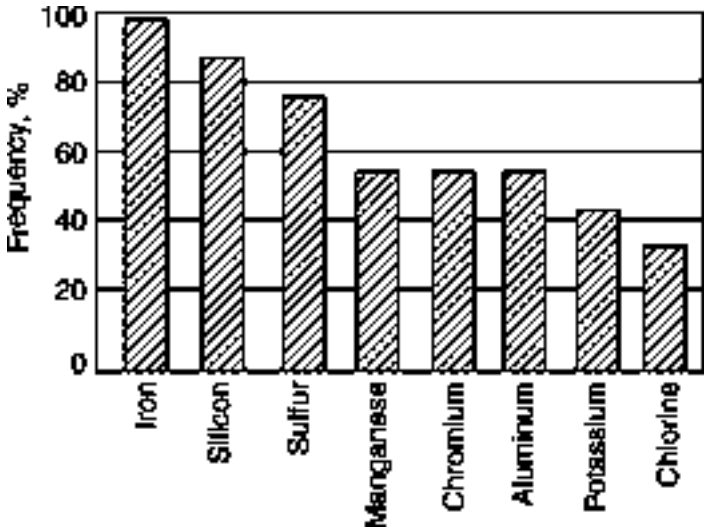


Fig. 13 Most commonly found elements in nine deposits from MIC sites in stainless steel cooling water systems (expressed as percent of deposits showing element). Source: Ref 63, 66, 71

At least two sorts of surface deposits were reported. Most of the surface of service water piping receiving lake water was covered by a tightly packed, black, slimy deposit that had a high content of manganese and iron, with trace sulfide, silicon, and aluminum present (Ref 72). No corrosion was reported under these deposits. Rust-colored deposits found in a small area (6.5 cm^2) at the weld were rich in chromium and iron, with sulfur, chlorine, aluminum, and silicon in smaller concentrations. This rust-colored deposit covered the opening of an extensive corrosion cavity in the underlying metal (Fig. 12b). This is a unique form of pitting associated with MIC in weldments in stainless steel. The cavity openings are often associated with rust-colored stains on the surface metal or with rust-colored deposits rich in iron and manganese (Ref 63, 66).

Radiography or destructive testing of field specimens reveals the large cavities to be a series of pits branching off one another to give a bulbous and irregular void volume sometimes associated with tunneling in the direction of rolling along stringers of ferrite or austenite (Ref 63, 74). This form of pitting is focused on weld metal or the fusion line, with wall perforation occurring through a second small opening on the opposite metal surface. The frequent observation of sulfide in associated surface deposits (Fig. 13) implies that SRB are commonly involved at some stage, but iron-oxidizing bacteria, particularly *Gallionella*, found in the pits have come to be associated with this corrosion morphology. *Gallionella* oxidize Fe^{2+} to Fe^{3+} in their metabolism, leading to the formation of characteristic rust-colored deposits. The acidity of the hydrated ferric ion produced decreases the pH of the local environment. Whether *Gallionella* initiate pitting or are attracted to the anodic area by the release of ferrous ions through a previously existing anaerobic corrosion process is not clear. The latter seems more likely. Once iron-oxidizing organisms are established, reduction in the pH of the corrosion pit and concentration cells established by the buildup of iron (III) oxide deposits help to drive the corrosion process.

Literature reports identify several possible MIC scenarios on stainless steel weldments. Table 4 summarizes the organisms and features that may be useful in failure analysis.

Table 4 MIC scenarios that may play a role in the corrosion of weldments in stainless steel

MIC by	Mechanism	Indicators	Ref
Manganese oxidizers	Ennoblement of stainless steel potential due to MnO_2	Elevated manganese-oxidizing organisms, manganese, and possibly chloride in deposits	**29, 30**
SRB primary	Sulfides, SRB facilitate chloride attack in anaerobic systems	Dark-colored corrosion products, with iron sulfide, chloride, and a high ratio of Fe^{2+}/Fe^{3+}; near-neutral pH	**71**
SRB secondary oxidation	Pitting stabilized by thiosulfate formed by oxidation of sulfides	Cyclic anaerobic, aerobic conditions; surface of corrosion products in pit oxidized red, orange, or brown	**71**
Iron-oxidizing	**Decrease of pH by oxidation of**	**Red/orange corrosion products rich in Fe^{3+};**	**63**

bacteria	**Fe^{2+} to Fe^{3+} in pits**	**iron-oxidizing organisms such as *Gallionella;* pH acidic**	

SRB, sulfate-reducing bacteria

Stainless Steels. Corrosion-resistant alloys can suffer MIC failures in the body of the material not associated with welds. Numerous reports of pitting and crevice corrosion due to MIC have been noted for austenitic (304, 304L, and 316L) stainless steels (Ref 2).

Microbially induced corrosion has been described in underdeposit corrosion in heat exchangers cooled with lake water in tubes made of American Iron and Steel Institute (AISI) type 304L stainless steel and nickel alloys (UNS N08800, N08025, and N08028) (Ref 78). Pitting was found under calcite (calcium carbonate) deposits after one year for N08800 and 304L, while deep pits were found after three years for molybdenum-containing alloys N08025 and N08028. Failure analysis found that microorganisms played a key role in the degradation process. Anaerobic methanogens promoted deposition of calcium carbonate, setting the stage for crevice corrosion. Oxygen introduced with flow on startup then oxidized biogenic sulfides produced by SRB during periods of stagnation, when anaerobic conditions prevailed. Formation of thiosulfate through oxidation stabilized metastable pitting in the affected alloys, promoting corrosion even at low chloride levels. It has been suggested that the same mechanism was responsible for perforation of 316 stainless steel weldments (Ref 71).

Elemental mapping of deposits in and around the pits that formed under the calcite deposits showed:

- High levels of nickel, iron, and sulfur around the pit
- Chromium as the major metal component in the pit
- Chlorine at low levels or not at all

The calcite scale overlying corrosion stains on the metal surface was etched and contained enhanced levels of iron, nickel, and chromium derived from the corrosion process (Ref 78).

An example of MIC in martensitic stainless steel under severe service conditions has been described (Ref 79). Rapid corrosion under a thick, slimy, jellylike deposit on stainless steel (UNS S40300) drive chain systems in clarifiers in a wastewater treatment plant led to 40% metal loss in the first year of operation. The steel had been tempered for wear resistance at the cost of reduced corrosion resistance. Components made of 304L substituted into service in the unit corroded at half the rate of 403 but were subject to unacceptable levels of wear. The gray outer layer of the slime was rich in silicon and oxygen, with carbon, sulfur, and chlorine present in decreasing amounts. The dark-black inner layer showed major amounts of sulfur, chromium, and oxygen. Sulfate-reducing bacteria, APB, slime formers, and pseudomonads were all identified in samples of water and slime. High levels of biogenic hydrogen sulfide were generally present in the water phase, but the atmosphere over the fluid in the clarifiers was aerobic. This led to proliferation of sulfide- and sulfur-oxidizing organisms (*Thiothrix* and *Beggiatoa*) at the interface, producing very acidic local conditions for exposed components of the chain system. The high chloride content of the wastewater (up to 200 ppm) was also an issue. Given the extreme operating environment created, in part, by microbial activity, finding materials able to offer a reasonable service life at an acceptable cost remains a challenge.

Copper and Its Alloys

Copper is widely used in a variety of applications, because it is relatively low-cost, strong, and corrosion resistant. In addition, it conducts heat and electricity well and is readily formed, machined, and joined. Copper and brasses (copper and zinc alloys) are used, for example, in electrical wiring, water piping, architectural applications, heat exchangers, condensers, bearings, and valves, while bronzes (silicon, tin, and aluminum-amended copper) are used in bearings, impellers, pumps, screens, and special-purpose tubing. Cupronickel alloys are used in heat exchanger tubing and ships.

Copper is more-or-less susceptible to MIC in all its forms; however, copper-base alloys do show significant resistance to biofouling (Ref 80). Copper-nickel alloys, in particular, are used in marine applications, because of their resistance to both fouling and corrosion. Resistance to biofouling appears to be based on the slow release of toxic copper ions.

Elevated production of slime by microbial biofilms that develop on copper and its alloys suggests that EPS acts as a binding agent for copper ions that would otherwise inhibit microbial growth (Ref 81). The EPS exopolymers tend to be acidic and contain functional groups that bind metal ions. This capability has been

linked to the formation of copper concentration cells under biofilms, the transport of metal ions away from the corroding surface, and variations in potential on the underlying surface (Ref 82). Incorporation of EPS in the oxide surface film that normally protects copper and its alloys from corrosive attack is a key reason for the occurrence of MIC in potable water systems (Ref 38). While EPS may bind the metal ions released by corrosion, it also allows an influx of mobile anions to balance the buildup of positive charge at the corroding surface. Chloride and other aggressive anions can be concentrated under biofilms, leading to disruption of passive films that would otherwise protect the metal surface.

Concentration cells set up under biofilms and production of corrosive metabolites, such as reduced pH, organic acids, and anaerobic sulfide, can lead to corrosion of copper and copper alloys. Copper is susceptible to pitting, especially in acidic media where oxidants are present. In alkaline media, production of ammonia or ammonium salts can promote SCC.

Sulfide production by SRB is one of the most potent MIC scenarios. Sulfate-reducing bacteria can become established during long periods of stagnation or in periods of intermittent flow in water systems (Ref 75). Copper sulfides deposited in the protective film by bacterial activity are rapidly oxidized when a flow of oxygenated water is introduced into the system, resulting in exposure and corrosion of the underlying metal. Failure analysis can be based on identification of specific copper sulfides formed as corrosion products. These include digenite (Cu_9S_5), spionkopite ($Cu_{39}S_{28}$), chalcocite (Cu_2S), and covellite (CuS). It is thought that digenite, formed initially, undergoes subsequent transformation to chalcocite, the most characteristic corrosion product for sulfate-reducing bacteria MIC scenarios (Ref 80). Djurleite ($Cu_{1.96}S$) formed in a SRB scenario may deposit as a passivating film, but this is likely to lack the mechanical stability needed to provide lasting protection to the metal surface in most industrial situations (Ref 2). The biogenic origin of the sulfide in MIC scenarios on copper has been supported by stable isotope analysis showing enrichment of ^{32}S in the corrosion products relative to the sulfate present in solution (Ref 35).

Pitting morphology can also provide a key to the identification of MIC in failure analysis. Two forms of pitting have been identified with MIC in potable water systems (Ref 80). One of these has been dubbed “pepper-pot pitting.” In this form of damage, a conical cap of gray corrosion products (copper sulfate and cupric oxide) overlies a cluster of pits. Sulfides are present in some pits, and biofilms rich in EPS are invariably seen. In the second morphology identified with MIC, hemispherical pits are filled with crystalline cuprous oxide, with chloride often present at the metal surface. A tubercle over these pits is composed of pale-green basic copper sulfate carbonate ($Cu(OH)_x(SO4)_y$) and blue $Cu(OH)_2 \cdot CaCO_3$. Tubercles are covered by an outer layer of black cupric oxide, in some cases. Biofilm materials are seen associated with the original metal surface and with the tubercle.

Example 1: MIC of Brass Piping. A failure of buried brass (92% Cu, 8% Zn) piping used to carry drinking water in wet clay soil after less than two years service was attributed to MIC (Ref 83). Excavation showed the presence of soil blackened by deposition of sulfides and high numbers of SRB around the pipe. The external pipe surface showed a loss of zinc consistent with selective leaching of this metal from the alloy, and the groundwater contained appreciable levels of chloride that may have assisted in the breakdown of the passivating film protecting the metal surface.

A comparison was done of the corrosion failure of power station condenser tubing cooled by seawater for two copper alloys, an aluminum brass alloyed with arsenic (UNS C68700, ASTM B111, or CuZn20Al Deutsche Industrie-Normen (DIN) 17660), and a cupronickel 70-30 alloy with iron added (C71500, ASTM B111, or CuNi30Fe DIN17665) (Ref 84). Both kinds of tube had identical dimensions and had seen similar service, with failure occurring by perforation from internal pitting under sediment deposits. The presence of $Cu_{2-x}S$ in the corrosion products implicated MIC by SRB as the cause of failure, but this occurred more rapidly in the cupronickel alloy. The copper sulfide was close to stoichiometric Cu_2S on the aluminum brass but enriched in sulfur on the cupronickel tube. In the latter, large spherical pits were seen, with perforation taking the form of large, round holes. In the aluminum brass, big elliptic pits were seen, with small holes perforating the tube wall. These differences were attributed to differences in the pitting mechanism for the two alloys.

Alloy Cu-10%Ni (UNS C70600) is used extensively for condenser tubing in seawater applications, because it offers good corrosion resistance at reasonable cost. In polluted, brackish waters, however, severe localized corrosion has led to failures within three years of service. An investigation of MIC in these systems used on-line monitoring techniques and found elevated numbers of SRB in both the water phase (10^7 cells mL^{-1}) and in a surface biofilm (10^5 cells mL^{-1}) (Ref 85). A chlorination treatment intended to control the microbial problem destabilized the protective oxide film on the metal surface and made matters worse.

Example 2: SCC of Admiralty Brass Condenser Tubes. Microbes initiated SCC failures in admiralty brass condenser tubes in a nuclear plant cooled by freshwater (Ref 86). About 2500 tubes had to be replaced over a span of six years' operation. Analyses were carried out for microorganisms, water chemistry (for both intake and outfall), and corrosion products in the operating system and on test coupons exposed to the operating environment. Nitrate-reducing bacteria from the lakewater used in cooling were found to produce high levels of ammonia (5.8 mg/L) when established in biofilms. Ammonia levels at the metal surface were 300 times higher than background levels in the lake water. Copper amine complexes were identified in the surface deposits, and nitrate and oxygen in the incoming water were considered as accelerating factors for the cracking process.

Aluminum and Its Alloys

Aluminum is the third most abundant metal in the crust of the earth and second only to iron in commercial importance. It is nonmagnetic but electrically conducting and shows high thermal conductivity and reflectance. Its low density gives it a notable strength-to-weight ratio, and strength can be improved by alloying. It finds use in many household applications as well as structural roles, especially where weight is an issue. Incorporation of ceramic particles has created aluminum composite materials with enhanced stiffness, strength, and wear resistance that find application in markets such as the automotive industry.

Aluminum is an active metal that owes its corrosion resistance to the formation of a protective oxide film. The metal and its alloys remain passive in the pH range circa 5 to 8.5 but corrode in more acidic or more alkaline environments consistent with the amphoteric nature of aluminum. In a few instances where the oxide is insoluble or where it is maintained by strongly oxidizing conditions, corrosion is limited. Under dilute aqueous conditions where MIC is found, this is not the case. In general, the presence of chloride facilitates the anodic reaction by formation of soluble aluminum chloride species that hydrolyze with the release of H^+ to lower the pH at the metal surface.

Aluminum and its alloys show good corrosion resistance in pure freshwater or seawater or in dry, well-drained soils, where the pH of the environment is near neutral and oxidation reduction potentials are likely to favor the protective oxide layer on the surface of the metal. Corrosion is more likely to be a problem in polluted waters or wet, marshy soils (Ref 87).

Biological attack usually results in the formation of soft tubercles of aluminum hydroxide over anodic pits. Suggested mechanisms include the creation of oxygen concentration cells under microbial colonies in generally aerobic environments and the acceleration of the cathodic hydrogen reaction by anaerobic microorganisms in anaerobic environments. A wide range of microbes has been implicated. Microbially induced corrosion by SRB causes the formation of discrete hemispherical pits on aluminum in cooling water systems (Ref 37). Organic acids produced by many kinds of microorganisms (pH 3 to 4) greatly accelerate corrosion under biofilms. In the case of *Cladosporium resinae,* pit initiation occurs at points where the fungus adheres to the metal surface. The corrosion process can be closely related to the enzymatic activity of the microorganism (Ref 88). Identification of this form of MIC in failure analysis is based on observation of fungal mycelia associated with pitting on the metal surface.

The microstructure and processing of metal specimens has been related to the damage caused by exposure of the aluminum aerospace alloy 2219-T87 to microbial cultures developed from environmental and life support waters from the U. S. aerospace program (Ref 89). These waters contained pseudomonads as well as SRB, *Klebsiella* and *Enterobacter* bacterial species. Enhanced corrosion damage was attributed to the accelerated evolution of hydrogen at the cathode of the corrosion cell, but evidence of organic acids was also noted. Welded areas were more susceptible to MIC, and anodized samples showed corrosion rates several orders of magnitude higher than corresponding sterile controls.

The tensile stress-strain response of aluminum alloy 6061 (UNS A96061) and an alumina-particle-reinforced 6061 composite was altered adversely by exposure of material specimens to a culture of the marine bacterium *Pseudomonas* NCMB 2021 (Ref 90). The ultimate strength and strain to failure was especially reduced in the case of the composite material. In both materials, localized pitting resulting from the biofilm-forming tendencies of the microorganism was identified as the primary cause for the reduction in mechanical properties. Boron-carbide-reinforced aluminum composites are also susceptible to MIC (Ref 2).

Example 3: MIC of Aluminum Alloy Aircraft Fuel Tanks. An early case of MIC on aluminum was identified in aircraft wing tank failures in the 1950s. In this case, water condensed into the fuel tanks during flight led to microbial growth on the jet fuel. Pitting attack occurred under microbial deposits on the metal surface in the

water phase or at the water-fuel interface. The scope of MIC attack on aluminum components used structurally in seven different military and civil aircraft has recently been investigated (Ref 91). A wide range of microorganisms has been isolated from various locations on the aircraft, including 158 bacteria, 36 yeasts, and 14 fungi, all of which were heterotrophic facultative aerobes or anaerobes. Only one SRB was detected. Exposure of aluminum 7075 to cultures of these isolates showed that 27 bacterial isolates and 3 fungi could seriously corrode the aluminum alloy over several weeks. Corrosive organisms included all the species of staphylococci, enterococci, and micrococci isolated as well as selected species of isolated bacilli, coryneforms, aspergillus, and penicillium. This work indicates a widespread potential for MIC wherever conditions permit microbial growth on aluminum or its alloys.

References cited in this section

2. B. Little and P. Wagner, An Overview of Microbiologically Influenced Corrosion of Metals and Alloys Used in the Storage of Nuclear Wastes, *Can. J. Microbiol.*, Vol 42, 1996, p 367–374

5. G. Kobrin, *A Practical Manual on Microbiologically Influenced Corrosion*, National Association of Corrosion Engineers, 1993

7. S.W. Boronstein, *Microbiologically Influenced Corrosion Handbook*, Industrial Press Inc., 1994

20. T.R. Jack, M.J. Wilmott, and R.L. Sutherby, Indicator Minerals Formed during External Corrosion of Line Pipe, *Mater. Perform.*, Vol 34 (No. 11), 1995, p 19–22

22. R.G. Worthingham, T.R. Jack, and V. Ward, External Corrosion of Line Pipe, Part I: Identification of Bacterial Corrosion in the Field, *Biologically Induced Corrosion*, S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 330–339

24. T.R. Jack, Chapter 6 in *A Practical Manual on Microbiologically Influenced Corrosion*, Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

29. C. Zhigang, P. Gumpel, and M. Kaesser, Microbiologically Influenced Corrosion of Weld on Stainless Steel 304L, *Chin. J. Mech. Eng.* (*Engl. Ed.*), Vol 11 (No. 2) 1998, p 159–164

32. "Microbiologically Influenced Corrosion (MIC): Methods of Detection in the Field," GRI Report GRI-88/0113, Gas Research Institute, 1991

33. "Microbiologically Influenced Corrosion (MIC) II: Internal MIC and Testing of Mitigation Measures," GRI Report GRI-92/0005, Gas Research Institute, 1992

35. B.J. Little, R.I. Ray, and R.K. Pope, Relationship Between Corrosion and the Biological Sulfur Cycle: A Review, *Corrosion* Vol 56 (No. 4), 2000, p 433–443

37. H.M. Herro and R.D. Port, *The Nalco Guide to Cooling Water System Failure Analysis*, Nalco Chemical Company, McGraw-Hill, Inc., 1993

38. B.J. Webster, S.E. Werner, D.B. Wells, and P.J. Bremer, Microbiologically Influenced Corrosion of Copper in Potable Water Systems—pH Effects, *Corrosion*, Vol 56 (No. 9), 2000, p 942–950

41. G. Cragnolino and O.H. Tuovinen, The Role of Sulphate-Reducing and Sulphur-Oxidizing Bacteria in the Localized Corrosion of Iron-Base Alloys—A Review, *Int. Biodeterior.*, Vol 20 (No. 1), 1984, p 9–26

57. O. Jonas, Corrosion and Water Chemistry Problems in Steam Systems—Root Causes and Solutions, *Mater. Perform.*, Vol 40 (No. 12), 2001, p 42–47

58. I. Matsushima, Carbon Steel—Corrosion by Soils, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000, p 555–560

59. M.J. Wilmott and T.R. Jack, Corrosion by Soils, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000, p 329–348

60. S.Y. Li, Y. Kim, K. Jeon, Y. Kho, and T. Kang, Microbiologically Influenced Corrosion of Carbon Steel Exposed to Anaerobic Soil, *Corrosion,* Vol 57 (No. 9), 2001, p 815–828

61. F. Kajiyama, Y. Koyama, and Y. Nakamura, "Corrosion of Ductile Iron Pipes Buried in Soils. A Summary of Research Activities Performed during the Last Decade," Paper 438, Proc. 13th Int. Corrosion Congress, (Claytion, Australia) ACA Inc., 1996

62. R.A. King and J.D.A. Miller. "An Outline of 150 Man-Years of Microbial Corrosion Studies at UMIST," Paper 368, presented at Corrosion '87, National Association of Corrosion Engineers International, 1987

63. S.W. Boronstein and P.B. Lindsay, "Planning and Conducting Microbiologically Influenced Corrosion Failure Analysis," Paper 381, presented at Corrosion '87, National Association of Corrosion Engineers, 1987

64. R.T. Huang, B.L. McFarland, and R.Z. Hodgman, "Microbial Influenced Corrosion in Cargo Oil Tanks of Crude Oil Tankers," Paper 535, presented at Corrosion '97, National Association of Corrosion Engineers, 1997

65. *Corrosion Recognition and Control,* Petrolite Corporation, Treatolite Division, St. Louis, MO 1992

66. M.H.W. Renner, "Corrosion Engineering Aspects Regarding MIC Related Failures on Stainless Steel," Paper 285, presented at Corrosion '98, National Association of Corrosion Engineers International, 1998

67. P. Li, R.A. Buchanan, C.D. Lunden, P. Angell, A.H. Tuthill, and R.E. Avery, Laboratory Study of Microbially Influenced Corrosion of Stainless Steel Weldments in Simulated Tennessee River Water, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 40/1 to 40/14

68. P. Li, R.A. Buchanan, C.D. Lunden, D.E. Sachs, A.H. Tuthill, R.E. Avery, and J. Bullen, Corrosion Studies of Stainless Steel Weldments in a Microorganism-Containing Arizona Ground Water System, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 41/1–41/13

69. P. Gümpel, K. Neumann, T. Volimer, and W. Racky, Effect of Surface Conditions on Microbially Influenced Corrosion of Stainless Steels, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 26/1–26/10

70. Y. Chung, H.J. Mantle, and G.E. Lakso, *Proc. 1995 Int. Conf., on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 13-1–13-13

71. F. Kajiyama, Role of Sulfate-Reducing Bacteria in Corrosion of Type 316 Stainless Steel, *Corros. Eng.,* Vol 46, 1997, p 377–387

72. D.W. Brinkley III and A.A. Moccari, MIC Causes Pipe Weld Joint Problems, *Mater. Perform.,* Vol 39 (No. 6), 2000, p 68–70

73. Y. Kikuchi, F. Matsuda, K. Tomoto, and M. Naihimura, Microbially Influenced Corrosion of Stainless Steel Welds in Sewage Treatment Plant Water: Laboratory Measurements, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 34/1–34/7

74. A.A. Stein, "Metallurgical Factors Affecting the Resistance of 300 Series Stainless Steels to Microbiologically Influenced Corrosion," Paper 107, presented at Corrosion '91, National Association of Corrosion Engineers, 1991

75. G. Licina, MIC in the Power Industry, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

76. N.J.E. Dowling, C. Lundin, D. Sachs, J.A. Bullen, and D.C. White, Principles of the Selection of Effective and Economic Corrosion-Resistant Alloys in Contact with Biologically Active Environments, *Proc. Mat. Res. Soc. Symp.,* Vol 294, Materials Research Society, 1993, P 361–367

77. T.A. Petersen and S.R. Taylor, The Effects of Sulfate Reducing Bacteria on Stainless Steel and Ni-Cr-Mo Alloy Weldments, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers, 1995, p 56/1–56/20

78. A.M. Brennenstuhl, T.S. Gendron, and R. Cleland, Mechanisms of Underdeposit Corrosion in Freshwater Cooled Austenitic Alloy Heat Exchangers, *Corros. Sci.,* Vol 35, 1993, p 699–711

79. S. Basu, J. McAuley, N. White, S. Lamb, and P. Pitt, Microbiologically Influenced Corrosion in a Wastewater Treatment Plant, *Mater. Perform.,* Vol 40 (No. 7), 2001, p 52–57

80. C.A.C. Sequeira, Corrosion of Copper and Copper Alloys, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

81. W.R. Fischer, D. Wagner, H. Peinemann, P. Arens, and G.J. Tuschewitzki, The Influence of Corrosion on Biological Parameters Relevant for Microbially Influenced Corrosion of Copper, *Proc., 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 21/1–21/14

82. J.D. Gu, T.E. Ford, and R. Mitchell, Microbiological Corrosion of Metals, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000, p 915–927

83. I. Alanis, L. Berardo, N. De Cristofaro, C. Moina, and C. Valentini, A Case of Localized Corrosion in Underground Brass Pipes; *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 102–108

84. P. Yoffe, Microbiologically Influenced Failures of Cupronickel and Aluminum Brass Condenser Tubes: Comparison of Real Cases, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion*, P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 69/1–69/5

85. M. de Romero, Z. Duque, O. de Rincon, O. Perez, I. Araujo, and A. Martinez, Online Monitoring Systems of Microbiologically Influenced Corrosion on Cu-10% Ni Alloy in Chlorinated, Brackish Water, *Corrosion*, Vol 56 (No. 8), p 867–876

86. T.S. Rao and K.V.K. Nair, Microbiologically Influenced Stress Corrosion Cracking Failure of Admiralty Brass Condenser Tubes in a Nuclear Power Plant Cooled by Freshwater, *Corros. Sci.* Vol 40 (No. 11), 1998, p 1821–1836

87. E. Ghali, Aluminum and Aluminum Alloys, *Uhlig's Corrosion Handbook*, 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

88. H.A. Videla, The Action of *Cladosporium resinae* Growth on the Electrochemical Behaviour of Aluminum, *Biologically Induced Corrosion*, S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986 p 215–222

89. D.W. Walsh and A.G. Demmons, Microbiologically Influenced Corrosion of Aluminum Aerospace Alloys, *Proc. 5th Int. Conf. on Trends in Welding Research*, (Pine Mountain, GA), ASM International, 1999, p 904–909

90. R.U. Vaidya, D.P. Butt, L.E. Hersman, and A.K. Zurek, Effect of Microbiologically Influenced Corrosion on the Tensile Stress-Strain Response of Aluminum and Alumina-Particle Reinforced Aluminum Composite, *Corrosion*, Vol 53 (No. 2), 1997, p 136–141

91. A. Hagenauer, R. Hilpert, and T. Hack, Microbiological Investigations of Corrosion Damage in Aircraft, *Werkst. Korros.*, Vol 45, 1994, p 355–360

Thomas R. Jack, NOVA Chemicals Ltd.

Monitoring Industrial Systems

Monitoring methods for MIC have recently been reviewed (Ref 10, 92, 93). Monitoring is generally applicable to closed industrial systems where internal fouling and corrosion are a concern rather than to external surfaces exposed to the natural environment. The focus of monitoring is on the surface-bound sessile microbial population associated with events on the metal surface rather than on freely suspended planktonic organisms. Monitoring can be done to identify MIC as a factor in failure analysis or can be instituted as a routine procedure in susceptible systems to catch MIC problems in an early stage of development and assess the cost-effectiveness of various management and control strategies.

Practical aspects of designing a corrosion monitoring program have been recently reviewed (Ref 94). Monitoring systems should consist of both probes that respond to short-term events and sampling coupons that provide a time-averaged view of system conditions. These devices should be mounted in adjacent fittings, so that information obtained can be integrated in interpretation; however, locating these devices correctly can be a challenge. Probes and coupons must be mounted in locations that reflect worst-case conditions for system

failure. Dead legs and low spots are often targeted in systems with suspended solids or mixed fluid flow, because this is where water and sediments collect (Fig. 9). Monitoring devices should contact the water phase and be flush-mounted to avoid fluid-flow effects, including erosion. They should also be located where they will not be damaged or interfere with process flow or cleaning operations.

Probes and Condition Monitoring

Various probes are available to monitor conditions relevant to MIC in industrial systems.
Fouling can be assessed by monitoring the pressure drop across a length of tubing with a known flow rate or by assessing the heat exchange efficiency from flow rate and temperature readings upstream and downstream of a heat exchanger. These measurements can be made on actual operating units on line but are also done using instrumented pipe loops and model heat exchanger units run in parallel to system flow. Figure 14 shows one such test unit with five parallel, instrumented pipe runs. Water flow from the target system is diverted through this unit, so that conditions are representative of the actual operating system. Measurements of friction factors and heat exchange efficiencies can indicate fouling. While deposits of any sort can affect flow and heat transfer in an operating industrial system, biofilms are especially effective. A 165 μm thick biofilm shows 100 times the relative roughness of a calcite scale and a thermal conductivity close to that of water—almost 100 times less than carbon steel (Ref 10). The assumption is generally made that a susceptible system showing extensive fouling is prone to MIC. Systems showing the effects of extensive fouling are operating inefficiently and may warrant remedial action, in any case.

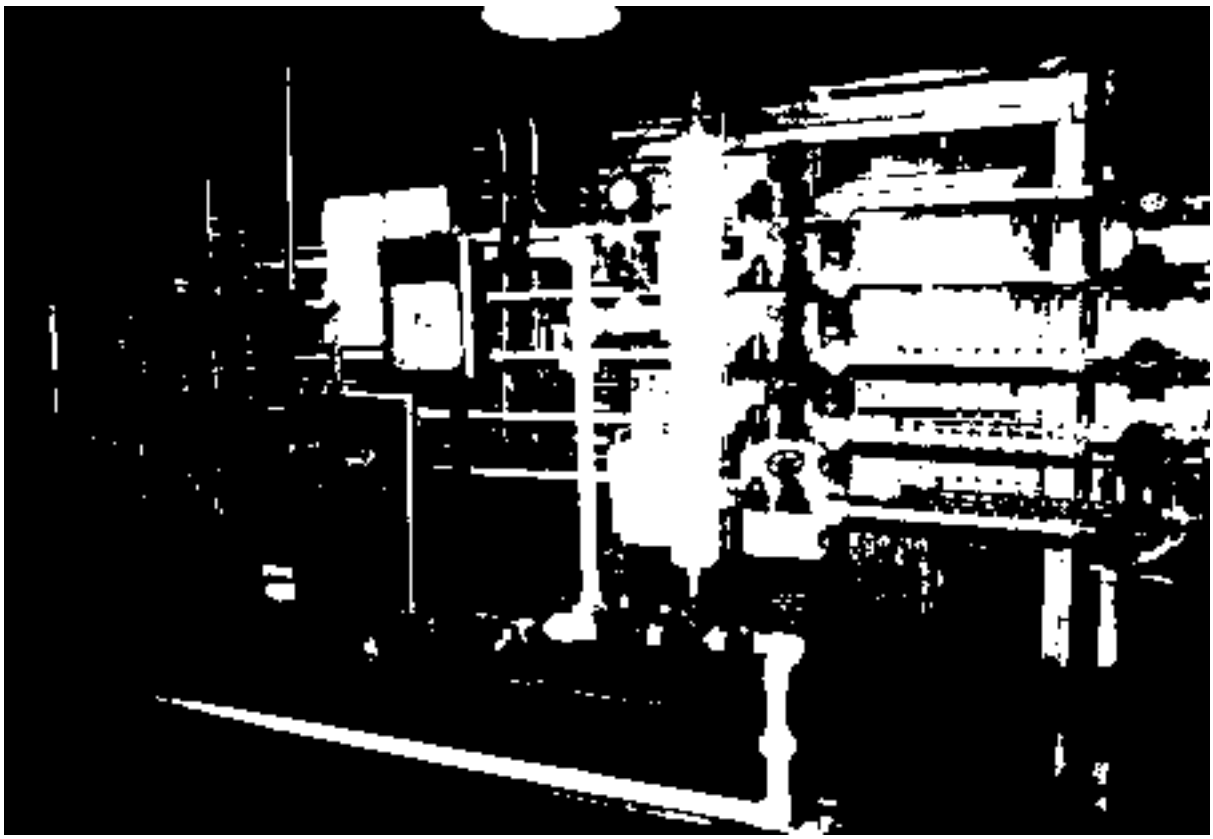


Fig. 14 Instrumented sidestreams in a biocide field test facility, each equipped with chemical injection pumps, pressure-drop measurement, corrosion probes, flush-mounted sampling, and weight-loss studs and flow meters

Electrochemical measurements can be used to track biofilm development, the appearance of specific metabolites, and corrosion.
Biological activity can be detected using various probes. In one design, the probe consists of a stack of corrosion-resistent stainless steel or titanium discs electrically isolated into two sets. An electrochemical potential is applied between the sets for a certain period each day, and the current needed to sustain the potential is measured. The rest of the time, the external potential is turned off, and the current flux between the

sets of discs is monitored. Because biofilm activity is based on electrochemical reactions, changes in biological activity on the disc surfaces are reflected as changes in current.

Another approach is based on the use of small silver, silver chloride, or silver sulfide electrodes capable of detecting chlorides or sulfides by changes in the potential between the Ag/AgCl or Ag/Ag_2S electrode pairs (Ref 93).

Corrosion Rates. Galvanic probes consist of a metal of interest in contact with a more noble metal. This sets up a galvanic couple that can be monitored readily. Changes in conditions within a system, such as addition of corrosion inhibitor or biofilm formation, are reflected by changes in the galvanic corrosion cell (Ref 93).

More conventional approaches include electrical resistance probes, linear polarization resistance (LPR), electrochemical noise monitoring, and electrochemical impedance spectroscopy (EIS). The principles and operation of these techniques can be found elsewhere (Ref 95, 96, 97, 98). An annotated summary of various methods is presented in Table 5.

Table 5 Electrochemical monitoring methods used for MIC

Measurement	Remarks
Corrosion potential	**Easy to do, difficult to interpret without other information**
Redox potential	**Limited value—not selective and not directly related to corrosion; provides insight into aerobic/anaerobic status of system**
Linear polarization resistance	**Gives trend information for uniform corrosion; sensitive to fouling by nonconducting material such as hydrocarbons**
Electrochemical impedance spectroscopy	**For uniform corrosion rate and information on various corrosion-related processes; requires significant expertise to interpret; small amplitude polarization does not disturb microbial numbers or activity (Ref 99)**
Electrochemical relaxation	**Fast transient techniques yield information for uniform corrosion.**
Electrochemical noise	**No external perturbation; indicates localized corrosion events and rates; well suited to MIC**
Electrical resistance	**Resistance measurement reflects loss of conductor cross section by pitting or general corrosion; robust, and suited to systems with multiphase flow; fouling by electrically conducting iron sulfides can cause erroneous readings**

Source: Ref 98

Literature has questioned the value of expensive probe-based monitoring, pointing out that instantaneous corrosion readings by electrical resistance or LPR probes are not particularly efficient at detecting the localized corrosion failures typical of MIC (Ref 92). Electrochemical noise and EIS are deemed more useful, presumably for the insight these techniques can provide, but the need for a high level of expertise in interpretation is noted. In general, carefully located weight-loss coupons are considered to be more effective (Ref 92).

Sampling and Inspection

Probe measurements alone will not indicate that development of a biofilm is responsible for fouling or that MIC is the mechanism of metal loss. This requires further sampling and analysis of conditions usually achieved by installation of removable coupons in the target system. Sampling provides a chance to assess the nature of deposits, microbial populations, weight-loss rates, and metallurgical damage directly; however, such analysis can be time-consuming and expensive. Where special pressure fittings are required to insert and withdraw samples, significant expense may also be incurred. Installation is best done during construction of a pressure system, to avoid the difficulty of hot taps and other issues related to later installation (Ref 94). Safety concerns and handling issues must also be considered, depending on the nature of the system. Fortunately, many target facilities, such as cooling water systems, are relatively accessible, free of contaminants, and operate at low pressures.

Analysis of fluid samples provides information on the chemical and microbiological composition of the liquid phase; however, this may not represent the environment under biofilms or other deposits on the metal surface. Samples scraped from the interior of an operating system or withdrawn intact on removable coupons or studs provide better insight into processes and conditions that contribute to corrosion. Examples of removable studs

can be seen in the gallery of instrumented pipe runs in the field test facility shown in Fig. 14. Analysis of surface deposits and corrosion products can be carried out using quantitative chemical analysis or more sophisticated techniques such as XRD, as described previously, while organic acids and similar entities can be quantified using standard gas chromatographic methods.

Tared coupons made of the same metal as the target system can provide time-averaged corrosion rates, but, for a range of reasons, these rates often do not match the instantaneous rates obtained by electrochemical probes. In extensive monitoring of six water return systems in a large oilfield waterflood operation, corrosion rates given by electrochemical probes, when integrated over time, were proportional to average corrosion rates obtained by weight-loss coupons but were 4 times higher (Ref 48). Corrosion rates from intrusive coupons mounted in an oilfield test loop gave pit depths significantly greater than seen on the pipe (Ref 100). This reinforces the point that location and mounting of weight-loss coupons is a key issue. Such discrepancies in monitoring results led to the observation, "Numbers are incidental; trends are everything. The most important element in a monitoring program is early detection of changes in the system" (Ref 92).

Biological assays can be performed on liquid samples or on suspensions of solid deposits to identify and enumerate viable microorganisms, quantify metabolic or specific enzyme activity, or determine the concentration of key metabolites.

Table 6 documents methods used to detect and describe microorganisms in terms of total cells present, viable cell numbers, and metabolic activity. Methods where commercial kits are available are noted. These kits are inexpensive and easy to use, even on site. Growth-based assays employ selective media to discriminate broad categories of microorganisms.

Table 6 Inspection, growth, and activity assays for microbial populations

Assay	**Method**	Comments
Microorganisms Cell numbers Specific organisms	*Microscopic examination:* Cell numbers are obtained with a Petroff-Hausser counting chamber. Fluorescent dyes can light up specific microbes.	**Requires a microscope with high magnification, phase contrast, and ultraviolet fluorescence, as appropriate. Cell counts are straightforward, but particulates and fluorescent materials may interfere.**
Viable cell counts SRB APB Other organisms	*Most probable number:* Sample is diluted into a series of tubes of specific growth medium. Cell numbers based on growth at various dilutions	**Commercially available kits can be inoculated in the field, but growth takes days to weeks.**
	Dip slides: A tab coated with growth medium is immersed, then incubated 2 to 5 days. Microbial colonies are counted visually.	**Crude numbers of bacteria, yeast, or fungi can be estimated in contaminated crankcase oil or fuels, for example. Commercial kits require minimal cost and expertise.**
Identification of microorganisms	*Fatty acid methyl ester analysis:* Methyl esters of fatty acids from field sample are analyzed by gas chromatograph.	**Fatty acid composition fingerprints specific organisms. The technique is available commercially.**
	Probes based on nucleic-acid base sequences: Probes bind to DNA or RNA for specific proteins or organisms.	**Probes require special expertise and lab facilities.**
	Reverse sample genome probing: DNA from MIC microbes is spotted on a master filter. Labeled DNA from field samples will bind to that DNA, if the reference organism is present.	**The assay answers the question, "Is this organism present?" It can also be used to track changes in a population after chemical treatment, etc., but requires special expertise and equipment.**
Biomass	*Protein, lipopolysaccharide, nucleic acid analysis:* Established methods, widely available	**Concentration of cell constituents correlates to level of organisms present in field sample.**
Metabolic activity	*Adenosine triphosphate (ATP):*	**ATP is the "energy currency" of**

	Fluorometer and supplies commercially available for field use	**microbial metabolism. Its concentration reflects the level of activity in a sample.**
Sulfate reduction	^{35}S *sulfate reduction:* Radioactively labeled sulfate is incubated in field sample. H_2S formed is liberated by acid addition, trapped on a zinc acetate wick, and measured by scintillation counter.	**A specialized lab technique useful in discovering nutrient sources for MIC and in quickly screening biocide activity in samples taken from the target system**
Enzyme activity	***Hydrogenase assay or sulfate reductase assay:* Measures enzyme activity in a field sample as a rate of reaction**	**Commercial kits are available. Hydrogenase activity may be related to MIC, while sulfate reductase assesses the presence of SRB.**

Table 6 also identifies assays that can be used to establish the presence of biomass in a field sample, identify the organisms present, and assess the activity of enzymes, such as hydrogenase, that are thought to accelerate corrosion through cathodic depolarization. More sophisticated assays based on nucleic acid sequences are also described. These are powerful techniques for characterizing microbial populations but remain in the domain of specialists with appropriate laboratory facilities, despite the number of practical applications published. A problem in the application of these techniques is the separation of sufficient DNA or RNA from field samples. A special technique (the polymerase chain reaction, PCR) is often used to amplify the material present to generate useful amounts for analysis. Unfortunately, PCR does not uniformly amplify all the material present. This can lead to a distorted picture of the makeup of the original population in the field sample.

In practice, a monitoring program for MIC in an industrial system can be expected to detect the presence of microorganisms and assess the number of SRB and APB in the population as well as the level of enzyme activity for hydrogenase or sulfate reductase by using inexpensive, commercially available kits designed for on-site use. A round-robin test comparing a number of these kits has been published (Ref 25). The information obtained can be used, along with other information from corrosion probes and coupon deposit analysis and damage assessment, to assess the susceptibility of an industrial system to MIC. These same assays are also useful to assess the performance of various control strategies such as biocide treatments.

References cited in this section

10. T.R. Jack, Monitoring Microbial Fouling and Corrosion Problems in Industrial Systems, *Corros. Rev.*, Vol 17 (No. 1), 1999, p 1–31

25. P.J.B. Scott and M. Davies, Survey of Field Kits for Sulfate-Reducing Bacteria, *Mater. Perform.*, Vol 31, 1992, p 64–68

48. T.R. Jack, E. Rogoz, B.J. Bramhill, and P.R. Roberge, Characterization of Sulfate-Reducing Bacteria in Heavy Oil Waterflood Operations, *Microbially Influenced Corrosion Testing*, ASTM STP 1232, J.R. Kearns, and B.J. Little, Ed., American Society for Testing of Materials, 1994, p 108–117

92. P.J.B. Scott, Microbiologically Influenced Corrosion Monitoring: Real World Failures and How to Avoid Them, *Mater. Perform.*, Vol 39 (No. 1), 2000, p 54–59

93. T.P. Zintel, G.J. Licina, and T.R. Jack, Techniques for MIC Monitoring, *A Practical Manual on Microbiologically Influenced Corrosion*, Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001, p 10.1–10.19

94. D.E. Powell, D.I. Ma'Ruf, and I.Y. Rahman, Practical Considerations in Establishing Corrosion Monitoring for Upstream Oil and Gas Gathering Systems, *Mater. Perform.*, Vol 40 (No. 8), 2001, p 50–54

95. F. Mansfeld and B.J. Little, A Critical Review of the Application of Electrochemical Techniques to the Study of MIC, *Microbially Influenced Corrosion and Biodeterioration*, N.J. Dowling, M.W. Mittleman, and J.C. Danko, Ed., University of Tennessee, 1990, p 5–33 to 5–39

96. F. Mansfeld and B.J. Little, A Technical Review of Electrochemical Techniques Applied to Microbially Influenced Corrosion, *Corros. Sci.*, Vol 32, 1991, p 247–272

97. S.C. Dexter, D.J. Duquette, O.W. Siebert, and H. Videla, Use and Limitations of Electrochemical Techniques for Investigating Microbiological Corrosion, *Corrosion*, Vol 47, 1991, p 308–318

98. G. Schmitt, Sophisticated Electrochemical Methods for MIC Investigation and Monitoring, *Mater. Corros.*, Vol 48, 1997, p 586–601

99. M.J. Franklin, D.E. Nivens, J.B. Guckert, and D.C. White, Effect of Electrochemical Impedance Spectroscopy on Microbial Biofilm Cell Numbers, Viability, and Activity, *Corrosion*, Vol 47 (No. 7), 1991, p 519–522

100. M. Attard, S. Papavinasam, and R.W. Revie, Comparison of Pitting Corrosion Rates of Coupons and Pipe, *Mater. Perform.*, Vol 39 (No. 10), 2000, p 58–61

Prevention and Control Strategies

Control of MIC in industrial systems has been reviewed (Ref 8). Table 7 summarizes prevalent integrity-protection strategies by facility type. In general, metal surfaces exposed to air are protected by the selection of appropriate materials and the use of protective coatings. Those in contact with soil or natural waters are primarily protected by coatings and CP. The internal surface of metal structures can be protected by a combination of materials selection, water removal (for nonaqueous systems), cleaning, and chemical treatments, with coatings and CP being applied in special cases.

Table 7 Prevention of corrosion in industrial facilities

Facility	Prevention approach					
	Materials selection	**Coatings**	**Cathodic protection**	**Water removal**	**Cleaning**	Chemical treatment
Upstream oil and gas	…	…	…	…	…	**…**
Well casings	X	…	X	…	…	**X**
Piping	…	…	X	…	X	**X**
Tanks, batteries	…	X	X	…	X	**X**
External pipelines	…	X	X	…	…	**…**
Internal pipelines	…	…	…	…	…	**…**
Gas	…	X	…	X	X	**…**
Oil	…	…	…	X	X	**X**
Storage tanks	X	X	X	X	X	**X**
Cooling systems (cooling towers, heat exchangers)	**X**	**…**	**X**	**…**	**X**	**X**

Materials Selection. Most metals and alloys are more or less susceptible to MIC, with the possible exception of beryllium and titanium. In practice, materials selection is largely controlled by cost, with the consequence that cheaper materials, such as carbon steel, are often extended, through the use of protective coatings and CP, into

applications where MIC can be a problem. In some cases, nonmetallic materials have been substituted for metallic ones to avoid corrosion or EAC; however, non-metallic materials are not without problems of their own, including microbial degradation (Ref 13, 14).

Protective coatings can be used to prevent MIC on metal surfaces exposed to soil or natural waters or on internal surfaces in contact with operating fluids. A wide range of commercial products is available, but without exception, correct selection and application are critical to successful performance.

Some coatings provide barrier protection by sealing the surface from contact with the local environment. These coatings offer a high electrical resistance and extremely low water permeability. Isolation of the steel surface provides effective protection for the underlying metal provided that the coating has been applied without defect, the coating material does not degrade in service, and the coating remains strongly bonded. In practice, some barrier coatings on pipelines have shown a tendency to lose adhesion and disbond in service (Ref 22). Where this allows an aqueous environment to reach the metal surface under the disbondment, MIC and other types of corrosion and EAC have been seen. Coating suppliers and applicators are aware of these past problems and have worked hard to develop products to avoid them.

Another kind of external coating that has enjoyed success in service is one that fails in a manner that allows CP to reach the underlying metal surface. In practice, CP is often used in conjunction with protective coatings, especially on buried or submerged structures. Coatings that resist cathodic disbondment but are permeable enough to allow CP penetration are referred to as being CP compatible. This approach has had a good record of preventing pipeline failures, for example, even where coating disbondment, defects, or damage have occurred.

Cathodic protection can prevent all kinds of corrosion, including MIC, if adequate potentials are sustained. Cathodic protection is generally applied to the external surfaces of structures buried underground or submerged in water but has also been applied to the inside of tanks, clarifiers, and heat exchangers, in some circumstances (Ref 101).

The interrelationship between CP and MIC in marine systems has been reviewed (Ref 102). In the presence of anaerobic bacteria, including SRB, applied CP potentials must be more negative than usual to achieve good protection for exposed steel. Potentials of -950 mV ($Cu/CuSO_4$) or polarization potentials of -200 mV are generally accepted as being protective. These potentials do not affect bacterial attachment to protected surfaces (Ref 54). Applied current densities up to 100 $\mu A/cm^2$ fail to remove attached biofilms from stainless steel surfaces. It was further noted that CP seemed unable to stop localized corrosion initiation when a stable biofilm of SRB was already established on stainless steel in a chloride-containing medium (Ref 53). Potentials sufficiently negative to drive hydrogen evolution at the metal surface cause a significant increase in pH that may discourage bacterial attachment and activity (Ref 103); however, such potentials are neither practical nor prudent for field systems, where undesirable hydrogen effects and coating damage are likely to result.

System Design and Operation. For water-handling systems, microbial problems are often associated with deposits formed by the settling of suspended solids or a buildup of metal oxides or scale as well as the growth of biofilms on susceptible surfaces. The key issue in preventing problems is good housekeeping, as indicated in the quote, “Keep the system clean. To the best of our knowledge, MIC has never been seen on a clean metal surface,” (Ref 104). Excluding water and other contaminants from nonaqueous systems prevents MIC. Natural gas transmission systems that enforce tight specifications on incoming gas quality do not suffer internal corrosion problems.

In nonaqueous industrial systems where water is present, minimizing the water content and not allowing standing water to collect greatly reduces problems. This can be done by a combination of good system design and good housekeeping. Tanks and storage vessels for fuels and other hydrocarbons should be designed to permit frequent drainage of collected water at the bottom of the vessel (Ref 64). For large engines (in ocean-going ships, for example), breaking emulsions and separating water from crankcase oils can be done by circulation of the oil through a centrifuge (Ref 55). Prevention and timely removal of deposits and scales can also be beneficial. For pipelines, cleaning devices called “pigs” can be run through the line pipe, pushed by the fluid flow. These tools can remove deposits, corrosion tubercles, and water collected at low spots (Ref 105, 106, 107, 108), provided that the facilities have been designed for pigging and are equipped with appropriate launchers and receivers.

Even with good housekeeping practices, chemical treatments including biocides are often employed to ensure system integrity.

Chemical Treatments. Chemical treatment involves the planned addition of a site-specific combination of chemicals to an operating system. The treatment may include agents designed to inhibit corrosion, scavenge

oxygen, alter pH, control scaling, suspend solids, control microbial growth or activity, and so on. Designing cost-effective synergistic combinations of treatment chemicals requires careful consideration. Cost of the chemicals along with the cost of installing, monitoring, controlling, and maintaining injection systems can be significant, especially for distributed applications in widespread facilities such as oilfield operations. Risk analysis can be used to target treatments for maximum benefit (Ref 109). Selection of agents is generally done on the basis of past experience, but recommendations of the vendor must be verified by monitoring.

Biocides are generally added to chemical treatment packages to control MIC or biological fouling. Good discussions of the application of biocides can be found in recent articles focused on cooling water systems (Ref 110), water-handling systems (Ref 48), and oil and gas operations (Ref 109, 111, 112).

Biocidal agents used in industrial applications include strong oxidants (such as chlorine or ozone), reactive aldehydes (such as glutaraldehyde), quaternary ammonium salts (such as cocodiamine), and more exotic chemicals. Some are administered in pure form (e.g., injection of gaseous chlorine), while others are formulated in a solution or carrier. Surfactants or emulsifiers may be added to penetrate or disperse deposits and biofilms. In practice, application is often combined with pigging or another form of mechanical cleaning to enable the maximum kill of organisms present during disruption and dispersal of surface deposits and sediment (Ref 10). In cold climates, alcohols may be added to product formulations to prevent freezing.

Prescreening of potential biocides is commonly done before implementation to ensure site-specific performance. Even so, the efficacy of a biocide product can decline in service as organisms sheltered from its reach by biofilm slime or deposits multiply. A buildup of resistance often occurs more quickly when biocides are injected at a low continuous rate. Higher-concentration scheduled batch treatments tend to be more effective in the long term and can be cheaper. Using combinations of biocides at a lower concentration than they would be used individually can extend the range of treatment and reduce cost (Ref 113). Switching biocide products periodically can restore effective control in a system developing resistance to the continuous use of one product (Ref 8).

Biocidal agents that come to be recognized as posing an unacceptable risk to the environment, workers, or the public continue to be phased out. For example, the International Maritime Organization has recently banned the use of tributyltin as an antifoulant in marine paints (Ref 114). Chromates, mercurials, copper salts, and chlorinated products provide other examples of agents that have been similarly removed from many past applications. Design of a biocide treatment must now also include the safe disposal of treated fluids after use (Ref 115). These trends continue to challenge those involved in biocide treatment to come up with new products and approaches that are cost-effective and acceptable.

References cited in this section

8. T.R. Jack and D.W.S. Westlake, Control in Industrial Settings, *Sulfate-Reducing Bacteria,* Vol 8, *Biotechnology Handbooks,* L.L. Barton, Ed., Plenum Press, London, 1995, p 264–292

10. T.R. Jack, Monitoring Microbial Fouling and Corrosion Problems in Industrial Systems, *Corros. Rev.,* Vol 17 (No. 1), 1999, p 1–31

13. *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

14. *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

22. R.G. Worthingham, T.R. Jack, and V. Ward, External Corrosion of Line Pipe, Part I: Identification of Bacterial Corrosion in the Field, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 330–339

48. T.R. Jack, E. Rogoz, B.J. Bramhill, and P.R. Roberge, Characterization of Sulfate-Reducing Bacteria in Heavy Oil Waterflood Operations, *Microbially Influenced Corrosion Testing,* ASTM STP 1232, J.R. Kearns, and B.J. Little, Ed., American Society for Testing of Materials, 1994, p 108–117

53. M.F.L. de Mele, S.G.G. de Saravia, and H.A. Videla, An Overview on Biofilms and Calcerous Deposits Interrelationships on Cathodically Protected Steel Surfaces, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 50/1–50/8

54. J.G. Guezennec, Cathodic Protection and Microbially Induced Corrosion, *Int. Biodeterior. Biodegrad.,* 1994, p 275–288

55. R. Stuart, Something Nasty in the Crank Case, *Ind. Lubr. Tribol.,* Vol 46 (No. 2), 1994, p 5–7

64. R.T. Huang, B.L. McFarland, and R.Z. Hodgman, "Microbial Influenced Corrosion in Cargo Oil Tanks of Crude Oil Tankers," Paper 535, presented at Corrosion '97, National Association of Corrosion Engineers, 1997

101. L. Manian and J.W. Hutt, Cathodic Protection of a Biotreatment Tankage Facility, *Mater. Perform.,* Vol 40 (No. 2), 2001, p 26–29

102. B.J. Little and P.A. Wagner, The Interrelationship Between Marine Biofouling, Cathodic Protection and Microbiologically Influenced Corrosion, *Mater. Sci. Forum,* Vol 192–194, 1995, p 433–446

103. P. Watkins, S. Cheung, and I. Beech, The Influence of Cathodic Protection on the Attachment of Sulphate-Reducing Bacteria to Mild Steel Surfaces, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 12/1–12/13

104. A.J. Freedman, "Microbiologically Influenced Corrosion: A State-of-the-Art Technology," Paper 635, presented at Corrosion '93, National Association of Corrosion Engineers, 1993

105. J.S. Smart and G.L. Smith, Pigging and Chemical Treatment of Pipelines, *Pipes Pipelines Int.,* May-June, 1998

106. J.S. Smart and G.L. Smith, Pigging and Chemical Treatment of Pipelines, Part II, *Pipes Pipelines Int.,* July–Aug 1998, p 31–36

107. J.S. Smart and G.L. Smith, Pigging and Chemical Treatment of Pipelines, *Proc. 1998 Pipeline Pigging Conf.,* (Houston), Clarion Technical Conferences, Pipes & Pipelines International, 1998

108. T.E. Cloete, P.J. Allison, and W.I.J. Poulton, The Detection and Mitigation of MIC, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001, p. 11.15–11.21

109. S. Ciaraldi, A. Abdallah, A.M. Attia, H.A. El-Leil, S. Bedair, M. Konecki, and A. Lain, Rehabilitation of GUPCO's Massive and Aged Pipeline Infrastructure in the Gulf of Suez, *Pipes Pipelines Int.,* Jan-Feb 2000, p 5–17

110. S.G. Choudhary, Emerging Microbial Control Issues in Cooling Water Systems, *Hydrocarbon Process.,* May 1998, p 91–102

111. D.M. Brandon, J.P. Fillo, A.E. Morris, and J.M. Evans, “Biocide and Corrosion Inhibition Use in the Oil and Gas Industry: Effectiveness and Potential Environmental Impacts,” SPE 29735, Society of Petroleum Engineers, 1995

112. D.H. Pope, T.P. Zintel, H. Aldrich, and D. Duquette, Efficacy of Biocides and Corrosion Inhibitors in the Control of Microbiologically Influenced Corrosion, *Mater. Perform.,* Vol 29 (No. 12), 1990, p 49–55

113. L. Hongfang, X. Liming, Z. Jiashen, and L. Jing, New Bactericide for Biocide-Resistant Sulfate-Reducing Bacteria, *Mater. Perform.,* Vol 39 (No. 4), 2000, p 52–55

114. M. Reisch, Maritime Dilemma. Scramble is on for Best Biocide to Replace Tributyltin in Marine Antifoulant Coatings, *Chem. Eng. News,* 12 Nov 2001, p 16

115. E.Y. Chen and R.B. Chen, “Wet Lay-up Fluid: Can It Be Both Corrosion Preventive and Environmentally Friendly?,” SPE 37891, Society of Petroleum Engineers, 1997

Conclusions

The role of the microorganisms in MIC is to alter conditions on the metal surface so as to trigger known corrosion mechanisms that would not occur in the absence of the microbes (Ref 69). This suggests that few MIC problems are definitively identified by inspection of metallurgical features in damaged areas. In many cases, it is only the rate and severity of localized damage, along with the absence of any other plausible cause, that triggers consideration of MIC as a possible mechanism.

Because microorganisms have many direct and indirect ways of contributing to corrosion and EAC, examination of the damage to metal structures alone is usually insufficient to identify a MIC problem. Similarly, identification of the presence of microorganisms alone does not signify MIC. Microorganisms are found in many water-bearing systems without being the cause of integrity damage. Failure analysis must take a wider view to identify those combinations of damage, deposits, water chemistry, physical conditions, and microbial activity that indicate MIC.

It is hoped that the information presented here provides sufficient background to enable a wider investigation and to allow some basis for comparison to past experience in failure analysis. A brief overview of management and mitigation tools has been provided to guide the response to identification of MIC as a potential issue in an operating system.

Reference cited in this section

69. P. Gümpel, K. Neumann, T. Volimer, and W. Racky, Effect of Surface Conditions on Microbially Influenced Corrosion of Stainless Steels, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 26/1–26/10

References

1. G.G. Geesey and G.A. Cragnolino, A Review of the Potential for Microbially Influenced Corrosion of High-Level Nuclear Waste Containers in an Unsaturated Repository Site, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995, (New Orleans), National Association of Corrosion Engineers International, 1995, p 76/1–76/20

2. B. Little and P. Wagner, An Overview of Microbiologically Influenced Corrosion of Metals and Alloys Used in the Storage of Nuclear Wastes, *Can. J. Microbiol.,* Vol 42, 1996, p 367–374

3. S.C. Dexter and J.P. LaFontaine, Effect of Natural Marine Biofilms on Galvanic Corrosion, *Corrosion,* Vol 54 (No. 11), 1998, p 851–861

4. J. Been and J.S. Grauman, Titanium and Titanium Alloys, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

5. G. Kobrin, *A Practical Manual on Microbiologically Influenced Corrosion,* National Association of Corrosion Engineers, 1993

6. J.F.D. Stott, What Progress in the Understanding of Microbially Induced Corrosion Has Been Made in the Last 25 Years? A Personal Viewpoint, *Corros. Sci.,* Vol 35, 1993, p 667–673

7. S.W. Boronstein, *Microbiologically Influenced Corrosion Handbook,* Industrial Press Inc., 1994

8. T.R. Jack and D.W.S. Westlake, Control in Industrial Settings, *Sulfate-Reducing Bacteria,* Vol 8, *Biotechnology Handbooks,* L.L. Barton, Ed., Plenum Press, London, 1995, p 264–292

9. W.A. Hamilton, and W. Lee, Biocorrosion, *Sulfate-Reducing Bacteria,* Vol 8, *Biotechnology Handbooks,* L.L. Barton, Ed., Plenum Press, London, 1995, p 243–264

10. T.R. Jack, Monitoring Microbial Fouling and Corrosion Problems in Industrial Systems, *Corros. Rev.,* Vol 17 (No. 1), 1999, p 1–31

11. W.P. Iverson, Research on the Mechanisms of Anaerobic Corrosion, *Int. Biodeterior. Biodegrad.,* Vol 47 (No. 2), 2001, p 63–70

12. H.A. Videla, Microbially Induced Corrosion: An Updated Overview, *Int. Biodeterior. Biodegrad.,* Vol 48, 2001, p 176–201

13. *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

14. *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

15. B.N. Herbert and F.D.J. Stott, The Effects of Pressure and Temperature on Bacteria in Oilfield Water Injection Systems, *Microbial Corrosion,* The Metals Society, London, 1983, p 7–17

16. R. Torres-Sanchez, M.C. Gonzalez, H. Martinez, and L. Martinez-Gomez, Microbiologically Influenced Corrosion in a Geothermal Electric Power Plant, *Mater. Perform.,* Vol 40 (No. 11), 2001, p 46–48

17. S.R. Freeman and H. Thielsch, Microbiological Activity and Chlorides Cause Carbon Steel Pipe Pitting, *Mater. Perform.,* Vol 35 (No. 3), 1996, p 65–66

18. D.A. Jones and P.S. Amy, Related Electrochemical Characteristics of Microbial Metabolism and Iron Corrosion, *Ind. Eng. Chem. Res.,* Vol 39, 2000, p 575–582

19. J.E. Zajic, *Microbial Biogeochemistry,* Academic Press, New York, 1969

20. T.R. Jack, M.J. Wilmott, and R.L. Sutherby, Indicator Minerals Formed during External Corrosion of Line Pipe, *Mater. Perform.,* Vol 34 (No. 11), 1995, p 19–22

21. T.R. Jack, M.M. Francis, and R.G. Worthingham, External Corrosion of Line Pipe, Part II: Laboratory Study of Cathodic Protection in the Presence of Sulfate-Reducing Bacteria, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 339–350

22. R.G. Worthingham, T.R. Jack, and V. Ward, External Corrosion of Line Pipe, Part I: Identification of Bacterial Corrosion in the Field, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 330–339

23. J.A. Hardy, Utilization of Cathodic Hydrogen by Sulphate-Reducing Bacteria, *Br. Corros. J.,* Vol 18, 1983, p 190–193

24. T.R. Jack, Chapter 6 in *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

25. P.J.B. Scott and M. Davies, Survey of Field Kits for Sulfate-Reducing Bacteria, *Mater. Perform.,* Vol 31, 1992, p 64–68

26. J.-L. Crolet and M.F. Magot, Non-SRB Sulfidogenic Bacteria in Oilfield Production Facilities, *Mater. Perform.,* Vol 35 (No. 3), 1996, p 60–64

27. B.H. Olesen, P.H. Nielsen, and Z. Lewandowski, Effect of Biomineralized Manganese on the Corrosion Behaviour of C1008 Mild Steel, *Corrosion,* Vol 56 (No. 1), 2000, p 80–89

28. D.T. Ruppel, S.C. Dexter, and G.W. Luther III, Role of Manganese Dioxide in Corrosion in the Presence of Natural Biofilms, *Corrosion,* Vol 57 (No. 10), 2001, p 863–873

29. C. Zhigang, P. Gumpel, and M. Kaesser, Microbiologically Influenced Corrosion of Weld on Stainless Steel 304L, *Chin. J. Mech. Eng.* (*Engl. Ed.*), Vol 11 (No. 2) 1998, p 159–164

30. W.H. Dickinson and Z. Lewandowski, Electrochemical Concepts and Techniques in the Study of Stainless Steel Ennoblement, *Biodegrad.,* Vol 9, 1998, p 11–21

31. B. Hedges, D. Paisley, and R. Woollam, The Corrosion Inhibitor Model, *Mater. Perform.,* Vol 40 (No. 6), 2001, p 40–44

32. "Microbiologically Influenced Corrosion (MIC): Methods of Detection in the Field," GRI Report GRI-88/0113, Gas Research Institute, 1991

33. "Microbiologically Influenced Corrosion (MIC) II: Internal MIC and Testing of Mitigation Measures," GRI Report GRI-92/0005, Gas Research Institute, 1992

34. B. Little, R. Staehle, and R. Davis, Fungal Influenced Corrosion of Post-Tensioned Cables, *Int. Biodeterior Biodegrad.,* Vol 47, 2001, p 71–77

35. B.J. Little, R.I. Ray, and R.K. Pope, Relationship Between Corrosion and the Biological Sulfur Cycle: A Review, *Corrosion* Vol 56 (No. 4), 2000, p 433–443

36. H. Amaya and H. Miyuki, Mechanism of Microbially Influenced Corrosion on Stainless Steels in Natural Seawater, *J. Jpn. Inst. Met.,* Vol 58 (No. 7), 1994, p 775–781

37. H.M. Herro and R.D. Port, *The Nalco Guide to Cooling Water System Failure Analysis,* Nalco Chemical Company, McGraw-Hill, Inc., 1993

38. B.J. Webster, S.E. Werner, D.B. Wells, and P.J. Bremer, Microbiologically Influenced Corrosion of Copper in Potable Water Systems—pH Effects, *Corrosion,* Vol 56 (No. 9), 2000, p 942–950

39. G.G. Geesey, R.J. Gillis, R. Avci, D. Daly, M. Hamilton, P. Shope, and G. Harkin, Influence of Surface Features on Bacterial Colonization and Subsequent Substratum Chemical Changes of 316L Stainless Steel, *Corros. Sci.,* Vol 38 (No. 1), 1996, p 73–95

40. G.G. Geesey, R.J. Gillis, H.-J. Zhang, and P.J. Bremer, Influence of Surface Features on Microbial Colonization and Susceptibility of Stainless Steels Used in the Food Processing Industry, *Proc. IIW Asian Pacific Welding Congress,* 4–9 Feb 1996 (Auckland, New Zealand), 1996, p 693–707

41. G. Cragnolino and O.H. Tuovinen, The Role of Sulphate-Reducing and Sulphur-Oxidizing Bacteria in the Localized Corrosion of Iron-Base Alloys—A Review, *Int. Biodeterior.,* Vol 20 (No. 1), 1984, p 9–26

42. R.N. Parkins, W.K. Blanchard, Jr., and B.S. Delanty, Transgranular Stress Corrosion Cracking of High-Pressure Pipelines in Contact with Solutions of Near Neutral pH, *Corrosion,* Vol 50 (No. 5), 1994, p 394–408

43. *Hydrogen Transport and Cracking in Metals,* A. Turnbull, Ed., The Institute of Materials, London, 1995

44. T. Ford and R. Mitchell, Microbiological Involvement in Environmental Cracking of High Strength Steels, *Microbially Influenced Corrosion and Biodeterioration,* N.J. Dowling, M.W. Mittleman, and J.C. Danko, Ed., University of Tennessee, 1990, p 3–94 to 3–97

45. R.G.J. Edyvean, J. Benson, C.J. Thomas, J.B. Beach, and H.A. Videla, "Biological Influences on Hydrogen Effects in Steel in Seawater," Paper 206, presented at Corrosion '97, National Association of Corrosion Engineers, 1997

46. H. Liu, L. Xu, and J. Zeng, Role of Corrosion Products in Biofilms in Microbiologically Induced Corrosion of Steel, *Br. Corros. J.,* Vol 35 (No. 2), 2000, p 131–135

47. M. Nemati, G.E. Jenneman, and G. Voordouw, Mechanistic Study of Microbial Control of Hydrogen Sulfide Production in Oil Reservoirs, *Biotechnol. Bioeng.,* Vol. 74 (No. 5), 2001, p 424–434

48. T.R. Jack, E. Rogoz, B.J. Bramhill, and P.R. Roberge, Characterization of Sulfate-Reducing Bacteria in Heavy Oil Waterflood Operations, *Microbially Influenced Corrosion Testing,* ASTM STP 1232, J.R. Kearns, and B.J. Little, Ed., American Society for Testing of Materials, 1994, p 108–117

49. C. Chandler, Biodegradable Volatile Corrosion Inhibitors for Offshore and Onshore Installations, *Mater. Perform.,* Vol 40 (No. 2), 2001, p 48–52

50. E.S. Pankhurst, Protective Coatings and Wrappings for Buried Pipelines: Microbiological Aspects, *J. Oil Colour Chem. Assoc.,* Vol 56, 1973, p 373–381

51. T.R. Jack, G. Van Boven, M. Wilmott, and R.G. Worthingham, Evaluating Performance of Coatings Exposed to Biologically Active Soils, *Mater. Perform.,* Vol 35 (No. 3), 1996, p 39–45

52. J.B. Santon, Paint and Coatings—New Trends Respond to Environmental Concerns, *Mater. Perform.,* Vol 39 (No. 9), 2000, p 22–25

53. M.F.L. de Mele, S.G.G. de Saravia, and H.A. Videla, An Overview on Biofilms and Calcerous Deposits Interrelationships on Cathodically Protected Steel Surfaces, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 50/1–50/8

54. J.G. Guezennec, Cathodic Protection and Microbially Induced Corrosion, *Int. Biodeterior. Biodegrad.,* 1994, p 275–288

55. R. Stuart, Something Nasty in the Crank Case, *Ind. Lubr. Tribol.,* Vol 46 (No. 2), 1994, p 5–7

56. I.B. Beech, S.A. Campbell, and F.C. Walsh, Marine Microbial Corrosion, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001, p 11.3–11.14

57. O. Jonas, Corrosion and Water Chemistry Problems in Steam Systems—Root Causes and Solutions, *Mater. Perform.,* Vol 40 (No. 12), 2001, p 42–47

58. I. Matsushima, Carbon Steel—Corrosion by Soils, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000, p 555–560

59. M.J. Wilmott and T.R. Jack, Corrosion by Soils, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000, p 329–348

60. S.Y. Li, Y. Kim, K. Jeon, Y. Kho, and T. Kang, Microbiologically Influenced Corrosion of Carbon Steel Exposed to Anaerobic Soil, *Corrosion,* Vol 57 (No. 9), 2001, p 815–828

61. F. Kajiyama, Y. Koyama, and Y. Nakamura, “Corrosion of Ductile Iron Pipes Buried in Soils. A Summary of Research Activities Performed during the Last Decade,” Paper 438, Proc. 13th Int. Corrosion Congress, (Claytion, Australia) ACA Inc., 1996

62. R.A. King and J.D.A. Miller. “An Outline of 150 Man-Years of Microbial Corrosion Studies at UMIST,” Paper 368, presented at Corrosion '87, National Association of Corrosion Engineers International, 1987

63. S.W. Boronstein and P.B. Lindsay, “Planning and Conducting Microbiologically Influenced Corrosion Failure Analysis,” Paper 381, presented at Corrosion '87, National Association of Corrosion Engineers, 1987

64. R.T. Huang, B.L. McFarland, and R.Z. Hodgman, “Microbial Influenced Corrosion in Cargo Oil Tanks of Crude Oil Tankers,” Paper 535, presented at Corrosion '97, National Association of Corrosion Engineers, 1997

65. *Corrosion Recognition and Control,* Petrolite Corporation, Treatolite Division, St. Louis, MO 1992

66. M.H.W. Renner, “Corrosion Engineering Aspects Regarding MIC Related Failures on Stainless Steel,” Paper 285, presented at Corrosion '98, National Association of Corrosion Engineers International, 1998

67. P. Li, R.A. Buchanan, C.D. Lunden, P. Angell, A.H. Tuthill, and R.E. Avery, Laboratory Study of Microbially Influenced Corrosion of Stainless Steel Weldments in Simulated Tennessee River Water, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 40/1 to 40/14

68. P. Li, R.A. Buchanan, C.D. Lunden, D.E. Sachs, A.H. Tuthill, R.E. Avery, and J. Bullen, Corrosion Studies of Stainless Steel Weldments in a Microorganism-Containing Arizona Ground Water System, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 41/1–41/13

69. P. Gümpel, K. Neumann, T. Volimer, and W. Racky, Effect of Surface Conditions on Microbially Influenced Corrosion of Stainless Steels, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 26/1–26/10

70. Y. Chung, H.J. Mantle, and G.E. Lakso, *Proc. 1995 Int. Conf., on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 13-1–13-13

71. F. Kajiyama, Role of Sulfate-Reducing Bacteria in Corrosion of Type 316 Stainless Steel, *Corros. Eng.,* Vol 46, 1997, p 377–387

72. D.W. Brinkley III and A.A. Moccari, MIC Causes Pipe Weld Joint Problems, *Mater. Perform.,* Vol 39 (No. 6), 2000, p 68–70

73. Y. Kikuchi, F. Matsuda, K. Tomoto, and M. Naihimura, Microbially Influenced Corrosion of Stainless Steel Welds in Sewage Treatment Plant Water: Laboratory Measurements, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 34/1–34/7

74. A.A. Stein, “Metallurgical Factors Affecting the Resistance of 300 Series Stainless Steels to Microbiologically Influenced Corrosion,” Paper 107, presented at Corrosion '91, National Association of Corrosion Engineers, 1991

75. G. Licina, MIC in the Power Industry, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001

76. N.J.E. Dowling, C. Lundin, D. Sachs, J.A. Bullen, and D.C. White, Principles of the Selection of Effective and Economic Corrosion-Resistant Alloys in Contact with Biologically Active Environments, *Proc. Mat. Res. Soc. Symp.,* Vol 294, Materials Research Society, 1993, P 361–367

77. T.A. Petersen and S.R. Taylor, The Effects of Sulfate Reducing Bacteria on Stainless Steel and Ni-Cr-Mo Alloy Weldments, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W.

Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers, 1995, p 56/1–56/20

78. A.M. Brennenstuhl, T.S. Gendron, and R. Cleland, Mechanisms of Underdeposit Corrosion in Freshwater Cooled Austenitic Alloy Heat Exchangers, *Corros. Sci.,* Vol 35, 1993, p 699–711

79. S. Basu, J. McAuley, N. White, S. Lamb, and P. Pitt, Microbiologically Influenced Corrosion in a Wastewater Treatment Plant, *Mater. Perform.,* Vol 40 (No. 7), 2001, p 52–57

80. C.A.C. Sequeira, Corrosion of Copper and Copper Alloys, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

81. W.R. Fischer, D. Wagner, H. Peinemann, P. Arens, and G.J. Tuschewitzki, The Influence of Corrosion on Biological Parameters Relevant for Microbially Influenced Corrosion of Copper, *Proc., 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 21/1–21/14

82. J.D. Gu, T.E. Ford, and R. Mitchell, Microbiological Corrosion of Metals, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000, p 915–927

83. I. Alanis, L. Berardo, N. De Cristofaro, C. Moina, and C. Valentini, A Case of Localized Corrosion in Underground Brass Pipes; *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986, p 102–108

84. P. Yoffe, Microbiologically Influenced Failures of Cupronickel and Aluminum Brass Condenser Tubes: Comparison of Real Cases, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 69/1–69/5

85. M. de Romero, Z. Duque, O. de Rincon, O. Perez, I. Araujo, and A. Martinez, Online Monitoring Systems of Microbiologically Influenced Corrosion on Cu-10% Ni Alloy in Chlorinated, Brackish Water, *Corrosion,* Vol 56 (No. 8), p 867–876

86. T.S. Rao and K.V.K. Nair, Microbiologically Influenced Stress Corrosion Cracking Failure of Admiralty Brass Condenser Tubes in a Nuclear Power Plant Cooled by Freshwater, *Corros. Sci.* Vol 40 (No. 11), 1998, p 1821–1836

87. E. Ghali, Aluminum and Aluminum Alloys, *Uhlig's Corrosion Handbook,* 2nd ed., R.W. Revie, Ed., John Wiley & Sons, Inc., 2000

88. H.A. Videla, The Action of *Cladosporium resinae* Growth on the Electrochemical Behaviour of Aluminum, *Biologically Induced Corrosion,* S.C. Dexter, Ed., National Association of Corrosion Engineers, 1986 p 215–222

89. D.W. Walsh and A.G. Demmons, Microbiologically Influenced Corrosion of Aluminum Aerospace Alloys, *Proc. 5th Int. Conf. on Trends in Welding Research,* (Pine Mountain, GA), ASM International, 1999, p 904–909

90. R.U. Vaidya, D.P. Butt, L.E. Hersman, and A.K. Zurek, Effect of Microbiologically Influenced Corrosion on the Tensile Stress-Strain Response of Aluminum and Alumina-Particle Reinforced Aluminum Composite, *Corrosion,* Vol 53 (No. 2), 1997, p 136–141

91. A. Hagenauer, R. Hilpert, and T. Hack, Microbiological Investigations of Corrosion Damage in Aircraft, *Werkst. Korros.,* Vol 45, 1994, p 355–360

92. P.J.B. Scott, Microbiologically Influenced Corrosion Monitoring: Real World Failures and How to Avoid Them, *Mater. Perform.,* Vol 39 (No. 1), 2000, p 54–59

93. T.P. Zintel, G.J. Licina, and T.R. Jack, Techniques for MIC Monitoring, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001, p 10.1–10.19

94. D.E. Powell, D.I. Ma'Ruf, and I.Y. Rahman, Practical Considerations in Establishing Corrosion Monitoring for Upstream Oil and Gas Gathering Systems, *Mater. Perform.,* Vol 40 (No. 8), 2001, p 50–54

95. F. Mansfeld and B.J. Little, A Critical Review of the Application of Electrochemical Techniques to the Study of MIC, *Microbially Influenced Corrosion and Biodeterioration,* N.J. Dowling, M.W. Mittleman, and J.C. Danko, Ed., University of Tennessee, 1990, p 5–33 to 5–39

96. F. Mansfeld and B.J. Little, A Technical Review of Electrochemical Techniques Applied to Microbially Influenced Corrosion, *Corros. Sci.,* Vol 32, 1991, p 247–272

97. S.C. Dexter, D.J. Duquette, O.W. Siebert, and H. Videla, Use and Limitations of Electrochemical Techniques for Investigating Microbiological Corrosion, *Corrosion,* Vol 47, 1991, p 308–318

98. G. Schmitt, Sophisticated Electrochemical Methods for MIC Investigation and Monitoring, *Mater. Corros.,* Vol 48, 1997, p 586–601

99. M.J. Franklin, D.E. Nivens, J.B. Guckert, and D.C. White, Effect of Electrochemical Impedance Spectroscopy on Microbial Biofilm Cell Numbers, Viability, and Activity, *Corrosion,* Vol 47 (No. 7), 1991, p 519–522

100. M. Attard, S. Papavinasam, and R.W. Revie, Comparison of Pitting Corrosion Rates of Coupons and Pipe, *Mater. Perform.,* Vol 39 (No. 10), 2000, p 58–61

101. L. Manian and J.W. Hutt, Cathodic Protection of a Biotreatment Tankage Facility, *Mater. Perform.,* Vol 40 (No. 2), 2001, p 26–29

102. B.J. Little and P.A. Wagner, The Interrelationship Between Marine Biofouling, Cathodic Protection and Microbiologically Influenced Corrosion, *Mater. Sci. Forum,* Vol 192–194, 1995, p 433–446

103. P. Watkins, S. Cheung, and I. Beech, The Influence of Cathodic Protection on the Attachment of Sulphate-Reducing Bacteria to Mild Steel Surfaces, *Proc. 1995 Int. Conf. on Microbially Influenced Corrosion,* P. Angell, S.W. Boronstein, R.A. Buchanan, S.C. Dexter, N.J.E. Dowling, B.J. Little, C.D. Lundin, M.B. McNeil, D.H. Pope, R.E. Tatnall, D.C. White, and H.G. Ziegenfuss, Ed., 8–10 May 1995 (New Orleans), National Association of Corrosion Engineers International, 1995, p 12/1–12/13

104. A.J. Freedman, “Microbiologically Influenced Corrosion: A State-of-the-Art Technology,” Paper 635, presented at Corrosion '93, National Association of Corrosion Engineers, 1993

105. J.S. Smart and G.L. Smith, Pigging and Chemical Treatment of Pipelines, *Pipes Pipelines Int.,* May-June, 1998

106. J.S. Smart and G.L. Smith, Pigging and Chemical Treatment of Pipelines, Part II, *Pipes Pipelines Int.,* July–Aug 1998, p 31–36

107. J.S. Smart and G.L. Smith, Pigging and Chemical Treatment of Pipelines, *Proc. 1998 Pipeline Pigging Conf.,* (Houston), Clarion Technical Conferences, Pipes & Pipelines International, 1998

108. T.E. Cloete, P.J. Allison, and W.I.J. Poulton, The Detection and Mitigation of MIC, *A Practical Manual on Microbiologically Influenced Corrosion,* Vol 2, J.G. Stoecker II, Ed., National Association of Corrosion Engineers International, 2001, p. 11.15–11.21

109. S. Ciaraldi, A. Abdallah, A.M. Attia, H.A. El-Leil, S. Bedair, M. Konecki, and A. Lain, Rehabilitation of GUPCO's Massive and Aged Pipeline Infrastructure in the Gulf of Suez, *Pipes Pipelines Int.,* Jan-Feb 2000, p 5–17

110. S.G. Choudhary, Emerging Microbial Control Issues in Cooling Water Systems, *Hydrocarbon Process.,* May 1998, p 91–102

111. D.M. Brandon, J.P. Fillo, A.E. Morris, and J.M. Evans, "Biocide and Corrosion Inhibition Use in the Oil and Gas Industry: Effectiveness and Potential Environmental Impacts," SPE 29735, Society of Petroleum Engineers, 1995

112. D.H. Pope, T.P. Zintel, H. Aldrich, and D. Duquette, Efficacy of Biocides and Corrosion Inhibitors in the Control of Microbiologically Influenced Corrosion, *Mater. Perform.,* Vol 29 (No. 12), 1990, p 49–55

113. L. Hongfang, X. Liming, Z. Jiashen, and L. Jing, New Bactericide for Biocide-Resistant Sulfate-Reducing Bacteria, *Mater. Perform.,* Vol 39 (No. 4), 2000, p 52–55

114. M. Reisch, Maritime Dilemma. Scramble is on for Best Biocide to Replace Tributyltin in Marine Antifoulant Coatings, *Chem. Eng. News,* 12 Nov 2001, p 16

115. E.Y. Chen and R.B. Chen, "Wet Lay-up Fluid: Can It Be Both Corrosion Preventive and Environmentally Friendly?," SPE 37891, Society of Petroleum Engineers, 1997

Fundamentals of Wear Failures

Raymond G. Bayer, Tribology Consultant

Introduction

WEAR is a persistent service condition in many engineering applications with important economic and technical consequences. In terms of economics, the cost of abrasion wear has been estimated as ranging from 1 to 4% of the gross national product of an industrialized nation. The effect of abrasion is particularly evident in the industrial areas of agriculture, mining, mineral processing, and earth moving. Likewise, wear is a critical concern in many types of machine components; in fact, it is often a major factor in defining or limiting the suitable lifetime of a component. An important example is the wear of dies and molds.

Wear generally is manifested by a change in appearance and profile of a surface. Some examples illustrating these types of changes are shown in Fig. 1. Wear results from contact between a surface and a body or substance that is moving relative to it. Wear is progressive in that it increases with usage or increasing amounts of motion, and it ultimately results in the loss of material from a surface or the transfer of material between surfaces. Wear failures occur because of the sensitivity of a material or system to the surface changes caused by wear. Typically, it is the geometrical or profile aspects of these changes, such as a dimensional change, a change in shape, or residual thickness of a coating, that cause failure. However, a change in appearance and the nature of the wear damage also can be causes for failure. An example of the former would be situations where marring is a concern, such as with optical scanner windows, lens, and decorative finishes. Examples of the latter include valves, which can fail because of galling, and structural components, where cracks caused by wear can reduce fatigue life (Ref 1, 2). In addition to these differences, the same amount or degree of wear may or may not cause a wear failure; it is a function of the application. For example, dimensional changes in the range of several centimeters may not cause wear failure on excavator bucket teeth, but wear of a few micrometers might cause failure in some electromechanical devices. As a consequence of these differences, there is no universal wear condition that can be used to define failure. The specific nature of the failure condition generally is an important factor in resolving or avoiding wear failures. It can affect not only the solutions to a wear problem but also the details of the approaches used to obtain a solution. While this is the case, there are some general considerations and approaches that can be of use in resolving or avoiding wear problems.

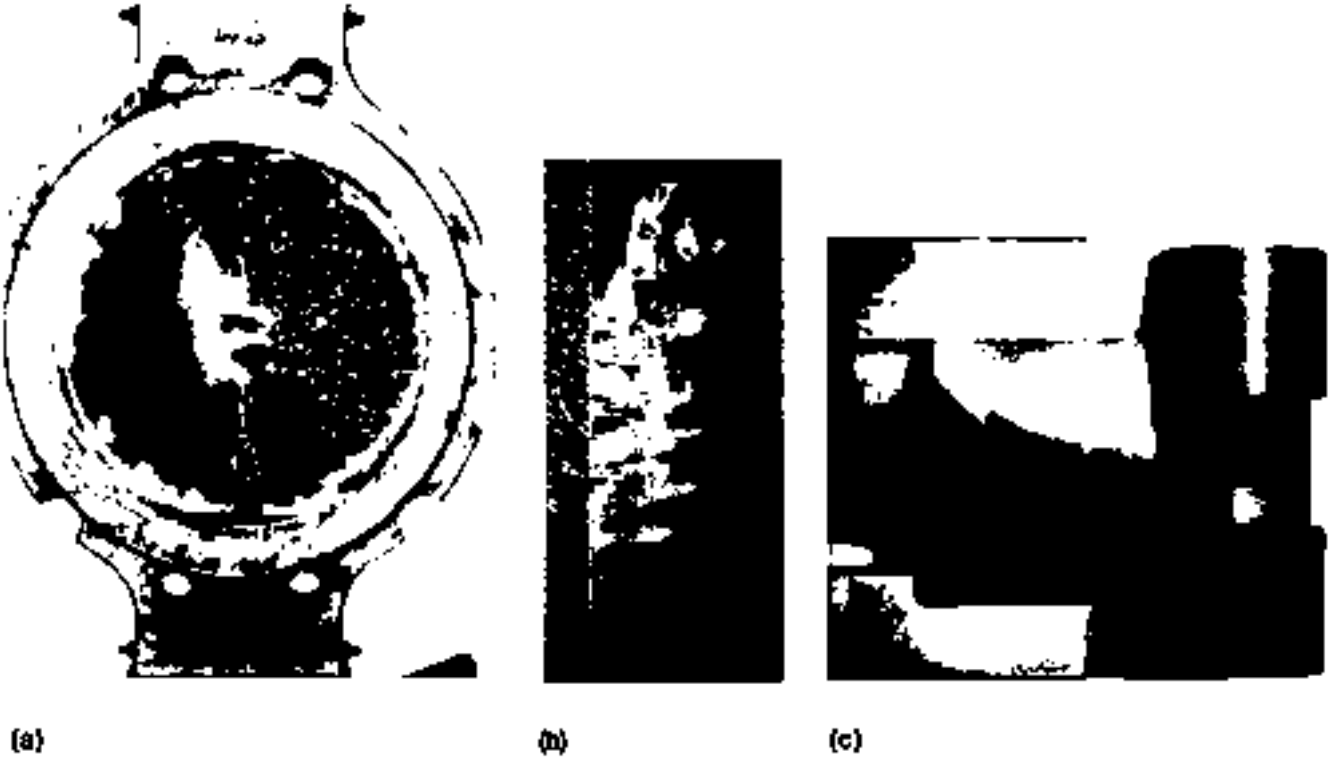


Fig. 1 Examples of wear showing loss of material, changes in dimension, and changes in appearance. (a) Erosion damage on a butterfly valve component. (b) Fretting damage on a friction band. (c) Sliding wear on a cam follower

The types of activities generally required for the resolutions of wear problems are:

- Examining and characterizing the tribosystem
- Characterizing and modeling the wear process
- Obtaining and evaluating wear data
- Evaluating and verifying the solution

While these activities roughly follow the sequence in the list, they generally are interwoven, and the overall approach is somewhat iterative in practice. For example, some modeling considerations might influence the details of the examination of the tribosystem, or failure during the verification can lead to additional tribosystem examination and modeling. Brief descriptions of the need for and the nature of these types of activities are presented in the following sections. More detailed treatment of these activities can be found in Ref 3 and 4.

References cited in this section

1. K. Budinski, Incipient Galling of Metals, *Proc. Intl. Conf. On Wear of Materials,* ASME, 1981, p 171–185

2. C. Lutynski, G. Simansky, and A.J. McEvily, Fretting Fatigue of Ti-6Al-4V Alloy, *Materials Evaluation Under Fretting Conditions,* STP 780, ASTM, 1982, p 150–164

3. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994

4. R.G. Bayer, *Wear Analysis for Engineers,* HNB Publishing, New York, 2002

Raymond G. Bayer, Tribology Consultant

Examination and Characterization of the Tribosystem

Wear is a system characteristic or phenomenon; it is not a materials property. Materials wear differently in different wear situations, and different materials wear differently in the same situation. Therefore, it is necessary to examine and characterize a number of different parameters, not simply the worn part. A tribosystem consists of all those elements that influence the wear process. The basic elements of a tribosystem are:

- Contacting materials
- Geometrical parameters (shape, size, roughness)
- Relative motion
- Loading
- Type of lubrication
- Environment

This tribosystem concept can be extended to include those elements or factors that affect the fundamental ones listed. In practice, it generally is appropriate to think of the tribosystem as at least extending to the mechanism or device in which the wear occurs. The tribological aspect number (TAN) is a method for characterizing tribosystems (Ref 5). This system is useful in evaluating the relevance of data and determining the most appropriate simulation test. The wear situation is described in terms of the contact velocity, contact area, contact pressure, and entry angle.

The purpose of the examination and characterization is to be able to define the tribosystem at the point of contact or wear site. It is necessary to define to some degree all of the basic parameters for that contact situation. The degree that is necessary generally is not the same for all the basic parameters. It depends on a number of factors. Fundamentally, it depends on the potential sensitivity of the wear to the various parameters in the particular service environment. It is influenced also by the detail that is needed to obtain a solution and the type of solution that is acceptable. For some engineering situations, a very crude description might be sufficient, such as describing the tribosystem as being a lightly loaded, lubricated contact at low sliding speed in an ambient room environment. However, greater detail is always desirable and generally necessary. Typically, magnitudes of most parameters, material identification, and details about the nature of the contact geometry, loading, and relative motion, are necessary.

The examination of the tribosystem should include also the inspection and measurement of the wear scars. The shape, morphology, and location of the wear scars provide important information generally needed to characterize the tribosystem and the wear process. Quantifying the amount of wear, particularly in terms of depth, generally is useful as well. The magnitude of the wear can support the characterization of the wear behavior and aid in the identification of a solution when used in conjunction with various models and analytical relationships. A generally good practice in examining wear scars is to examine them using several different methods, such as visual, low-power optical, and scanning electron microscopy (SEM). In many situations, magnifications between 30 and a few hundred are most useful.

In addition to these methods for examining wear scars, a variety of other methods are often used. These procedures can and often do include methods to characterize materials, measure dimensions and surface roughnesses, determine loads, determine contact stresses, and determine environmental conditions. Some of these techniques are discussed in other articles in this Volume, as well as in Ref 3, 4, 6.

In general, the amount of wear or root cause that results in the failure should be identified and defined. A criterion for acceptable wear also should be identified. Both pieces of information generally are important in developing an economical and practical solution. These factors should be determined as part of the examination and characterization process.

References cited in this section

3. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994

4. R.G. Bayer, *Wear Analysis for Engineers,* HNB Publishing, New York, 2002

5. R.M. Voitik, Realizing Bench Test Solutions to Field Tribology Problems by Utilizing Tribological Aspect Numbers, *Tribology: Wear Test Selection for Design and Application,* STP 1199, A.W. Ruff and R.G. Bayer, Ed., ASTM, 1993

6. P.J. Blau, Ed., *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992

Raymond G. Bayer, Tribology Consultant

Characterization and Modeling of the Wear Situation

Wear behavior is different in different service environments. The significance of factors and parameters tends to change with the type of wear situation. Significant wear phenomena also tend to be different. As a consequence, some form of characterization of the wear process is not only helpful but also generally needed to resolve or avoid future wear failures. The characterization enables the sorting and identification of appropriate information on wear behavior, model development, and selection of wear data. The most useful method of characterization is to classify the situation in terms of broad types of wear and then refine these in terms of specific operational features. The term *abrasive wear* is used to describe situations where the principal cause of the wear is scratching or cutting by abrasive particles. The term *nonabrasive wear* is used to describe all other wear situations involving contact between two solid bodies (for example, sliding). The term *erosion* is used where the wear is caused by a fluid, a stream of particles, or bubbles (in the case of cavitation), not by contact between two solid bodies. The operational classification for nonabrasive wear situations (Table 1) is directly based on the characterization of the elements of the tribosystem at the contact. The nominal type of motion, that is rolling, sliding and impact, and in the case of sliding, lubricated or nonlubricated wear, often are considered as major subcategories of nonabrasive wear situations. Some subcategories for abrasive wear and erosion situations, and operational elements that are often significant, are given in Tables 2 and 3.

Table 1 Operational attributes of nonabrasive wear

Attribute	Variations
Motion	
Rolling	With slip
	Without slip
Impact	With slip
	Without slip
Sliding	Unidirectional or reciprocating
	High speed or low speed
	Fretting or gross sliding
Lubrication	Lubricated or not
	Type of lubricant
	Lubricant
Load	Heavy or light
	Constant or variable

	Impact or not
Contact geometry	Point, line, or area
	Conforming or nonconforming
Contact stress	Above or below yield
Environment	Hostile or nonhostile
	High temperature or low temperature
	With or without abrasive particles
	pH level
Materials	Type-to-type or dissimilar

Table 2 Subcategories of abrasive wear

Attribute	**Variations**
Number of bodies	One-body (abrasive surface) or two-body (loose particles between surfaces)
Stress level	High stress or low stress
Surface alteration	Scratching, polishing, or gouging
Presence of fluid	Dry abrasion or slurry abrasion (particles in fluid)
Relative hardness of particles to surface	Surface harder or softer than particles

Table 3 Subcategories of erosion

Type	**Variables**
Particle erosion	Grazing or high angle of incidence
	High or low temperature
Cavitation erosion	Grazing or high angle of incidence
	High or low temperature
	Corrosive or noncorrosive fluid
Slurry erosion	Grazing or high angle of incidence
	High or low temperature
	Corrosive or noncorrosive fluid
Droplet erosion	Grazing or high angle of incidence
	High or low temperature
	Corrosive or noncorrosive fluid
Jet erosion	Grazing or high angle of incidence
	High or low temperature
	Corrosive or noncorrosive fluid

It often is desirable and helpful to augment this type of classification with a characterization of the principal wear mechanism or mechanisms involved. Generally, for engineering purposes, characterization in terms of broad generic types of wear mechanisms is adequate, and often, the most useful. Identification and characterization in terms of specific mechanisms generally are not as useful or needed since they can change with materials and other tribosystem parameters. Wear mechanisms generally can be grouped into six generic types:

- Adhesive mechanisms
- Single-cycle deformation mechanisms
- Repeated-cycle deformation mechanisms
- Chemical mechanisms
- Thermal mechanisms
- Tribofilm mechanisms

More than one type of mechanism can be involved in a wear situation. Also, these individual mechanisms can interact sequentially to form a more complex wear process. However, one mechanism generally is the controlling and primary mechanism. The relative importance or occurrence of individual mechanisms can change with changes in tribosystem parameters. Therefore, materials can exhibit transitions in wear behavior as a result of changes in other operational parameters, such as load, velocity, and friction (Ref 7). The wear map for unlubricated sliding between steels (Fig. 2) illustrates such transitions (Ref 8).

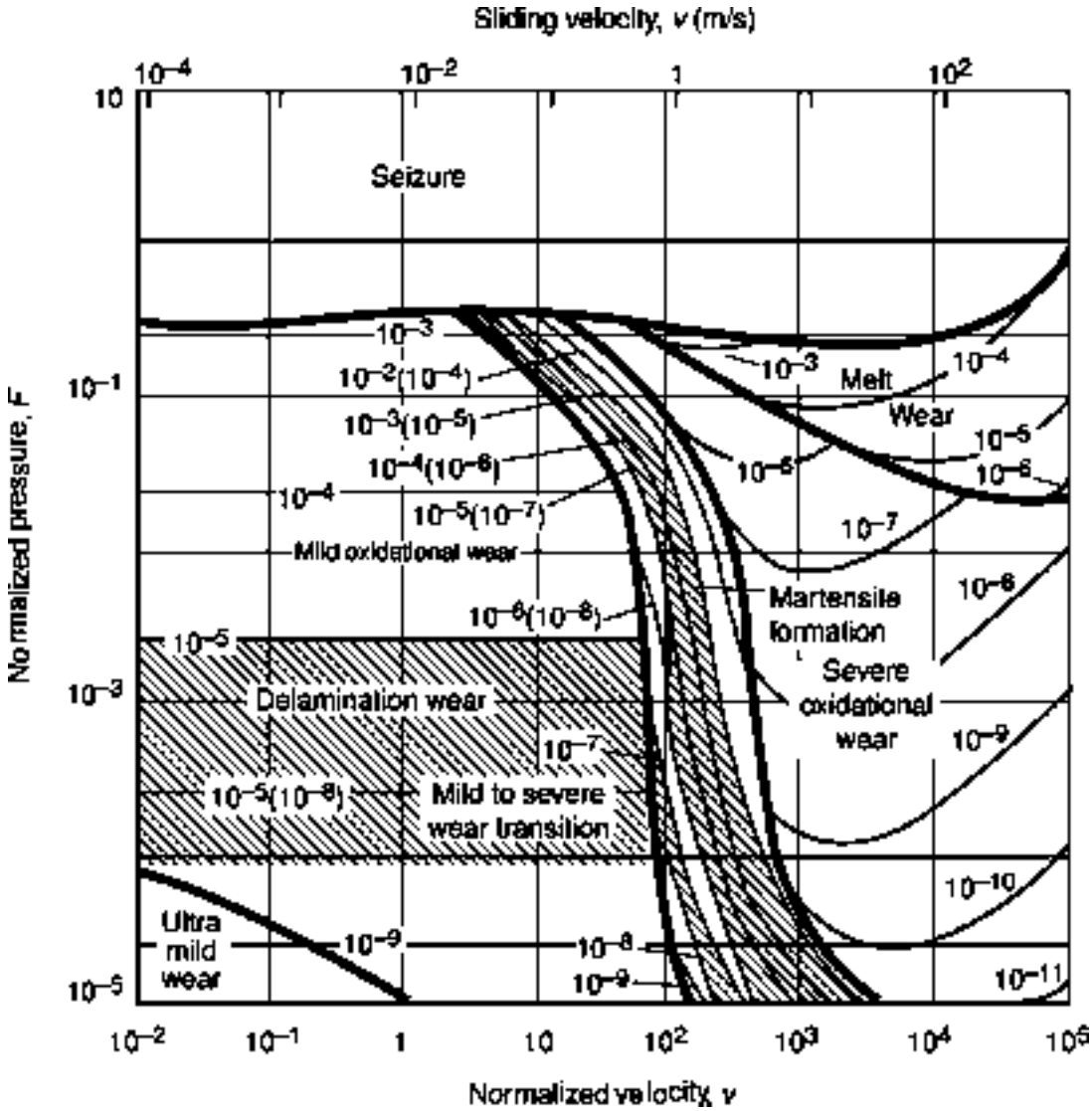


Fig. 2 Wear-mechanism map for unlubricated sliding of a steel couple. The normalized pressure is the contact pressure divided by hardness. The normalized velocity is the velocity multiplied by the ratio of the radius of the contact to the thermal diffusivity. The contour lines are lines of constant normalized wear rate. Pin-on-disk configuration. Source: Ref 8.

Except for adhesive and tribofilm mechanisms, these generic types are possible with all types of relative motion involving one or two solid surfaces. Adhesive and tribofilm mechanisms generally are limited to sliding contact between two solid surfaces. Adhesive wear mechanisms are those involving adhesion and transfer of material. Single-cycle deformation mechanisms are mechanical processes, which occur as a result of a single engagement, such as plastic deformation or cutting. Repeated cyclic deformation mechanisms are mechanical processes, which require repeated engagements, such as fatigue or ratcheting. Chemical or oxidative mechanisms are wear processes in which the rate-controlling parameters are those associated with a chemical reaction at the surface, such as oxidation or corrosion. Similarly, thermal mechanisms are wear processes where the rate-controlling mechanism is associated with temperature. Tribofilm mechanisms are those that involve the formation of layers of wear debris on and between surfaces and the loss of material from these layers. Tribofilm mechanisms tend to be significant in unlubricated sliding situations, particularly between polymers and metals.

Wear also can be classified in relative terms as mild and severe. This classification is based on the nature of the wear, not the amount. The differentiation is primarily in terms of the features of the wear scars and secondarily by wear rate. Coarse features and high wear rates are characteristics of severe wear. Examples of mild and severe sliding wear are shown in Fig. 3. Most types of wear mechanisms have mild and severe forms, and most materials can exhibit both types of behavior. Transitions between these two states are often sharp and can be associated with most parameters of the tribosystem; speed, temperature, load, and lubrication are the most common. Severe wear behavior usually cannot be tolerated. Consequently, if the wear observed is severe, the minimum corrective action required to solve the wear problem is to make changes (by selecting more wear-resistant materials or by modifying other parameters, for example, shape and load) to achieve mild wear behavior. Because of the typical large difference in wear rates between these states, many times such changes are adequate, particularly in applications that are less sensitive to wear.

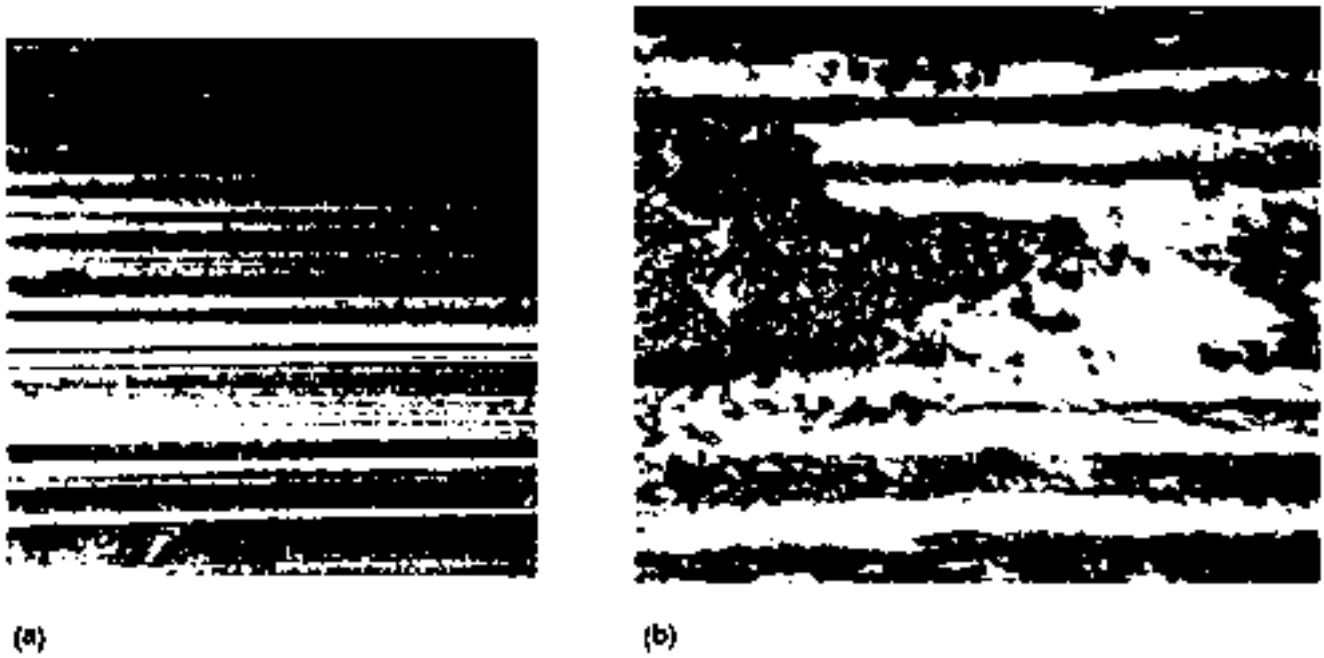


Fig. 3 Examples of mild and severe wear morphology. (a) A lubricated sliding wear scar on steel in the mild wear regime. (b) The appearance of the same type of scar in the severe wear regime

Generally, by characterizing the wear situation in these manners, there is sufficient information to formulate a model and to assess the significance of various parameters, including materials, which can be changed to resolve the wear problem.

Modeling is an essential element in resolving wear failures. The characterizations of the tribosystem and the wear situation provide the basis for models. Simple phenomenological models sometimes are adequate. However, more complex models, using analytical relationships for wear behavior, always are superior and sometimes necessary. Through the modeling activity, reasons for the failure and measures needed to resolve the wear problem both are identified. Models are particularly useful when they involve analytical relations to identify the significance of parameters and to assess the significance of proposed changes. The incorporation of analytical relationships generally facilitates the identification of causes and minimizes the amount of testing associated with obtaining a solution.

An effective model for a wear situation may be as simple as concluding that, because of too much clearance, fretting occurred. Based on this model, two solutions can be proposed. One solution is to select materials that are more resistant to fretting wear. The second is to change tolerances to reduce the clearance. This method is an example of a simple phenomenological model. A more elaborate model would be describing the wear situation as unlubricated mild sliding, impact, or rolling wear, and using analytical models for these wear situations to develop relationships between wear and design parameters. These relationships would then be used to evaluate the effects of changing different design parameters to achieve adequate wear behavior. Case studies, which further illustrate this type of modeling and the use of these models, can be found in Ref 3 and 4.

References cited in this section

3. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994

4. R.G. Bayer, *Wear Analysis for Engineers,* HNB Publishing, New York, 2002

7. P.J. Blau, *Friction and Wear Transitions of Materials,* Noyes Publications, Park Ridge, NJ, 1989

8. S.C. Lim and M.F. Ashby, *Acta Metall.,* Vol 35 (No. 1), 1987, p 1–24

Obtaining and Evaluating Wear Data

No matter which type of model is developed, the resolution of the problem involves the use of material wear data. For phenomenological models, this is generally in the form of rankings or relative wear performance for different materials obtained from tests or prior experience. For models involving analytical relationships, this generally is in the form of values for empirical wear coefficients, associated with the underlying wear equations. All equations proposed for wear have such coefficients (Ref 9). Because of the wide range of wear behavior possible and the system nature of wear, the applicability of any data generally depends on how closely the conditions of the test used to obtain the data simulate those of the wear situation (Ref 5, 10). If the test produces a poor simulation, then the data are less applicable. This concept applies to wear data obtained from wear tests done for a specific situation, from material suppliers, or from the technical literature. Generally, the best simulation and, hence, the more relevant data are obtainable with wear tests done for the specific situation. However, this is usually the least feasible in many engineering environments. On the other hand, vendor and technical literature data generally are based on tests that provide less simulation, and their relevance needs to be assessed in terms of the simulation that they provide. Some form of extrapolation is generally involved with the use of such data. Often it is the trends in such data that are important. The key to the simulation is to ensure that the appropriate wear mechanisms occur in both the test and the application.

A testing approach that may be used in conjunction with a phenomenological model is testing that simulates the wear situation and establishes a correlation. This approach is most often used for materials solutions to wear problems, but it can be used for other parameters, such as roughness, load, or lubrication. Such a test can then be used for rankings. With the use of a reference or central material, an estimate of the improvement in the wear situation can be obtained (Ref 9).

The usefulness, methodology, and general types of laboratory, bench, and component tests are summarized in Ref 11.

References cited in this section

5. R.M. Voitik, Realizing Bench Test Solutions to Field Tribology Problems by Utilizing Tribological Aspect Numbers, *Tribology: Wear Test Selection for Design and Application,* STP 1199, A.W. Ruff and R.G. Bayer, Ed., ASTM, 1993

9. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994, Section C2

10. A.W. Ruff and R.G. Bayer, Ed., *Tribology: Wear Test Selection for Design and Application,* STP 1199, ASTM, 1993

11. M. Anderson and F.E. Schmidt, Jr., Wear and Lubricant Testing, Chapter 25, *ASTM Manual on Fuels, Lubricants, and Standards: Application and Interpretation,* ASTM, 2002

Evaluation and Verification of Solutions

The development of a model and possible solutions to resolve or avoid a wear problem typically involves making assumptions, interpolating and extrapolating data, and often working with incomplete information about the tribosystem or conditions of operation. Because of the complex nature of wear behavior, it is important to provide some structure and control to the normal engineering practice of evaluating and verifying solutions and designs. Generally, this structure and control involve the characterization and control of hardware and operating conditions and identifying a procedure for monitoring wear. Ideally this procedure should involve the measurement of wear as a function of usage or exposure. This often facilitates, simplifies, and shortens the verification by allowing the determination of stable wear behavior and projecting improvement and life, as illustrated in Fig. 4. Frequently, the relationship between usage and a measure of wear often used in engineering, such as depth or width, is nonlinear; that is, doubling the amount of usage does not result in doubling the wear depth but results in something less than double. Because of the precision and variation typical of these measurements in engineering situations, it generally is adequate and preferable to select measurement intervals on a log or semi-log basis, such as decades or half decades. An example would be measurements after one, ten, one hundred, five hundred, and a thousand hours of operation, rather than after one hundred, two hundred, three hundred hours, and so on.

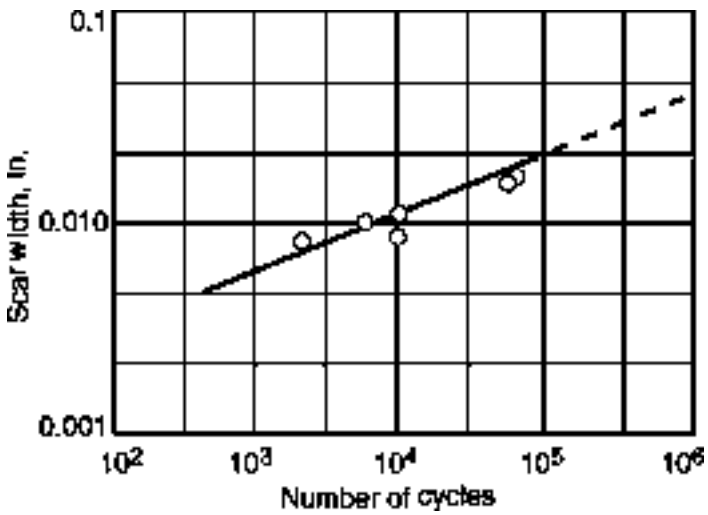


Fig. 4 Example of a log-log plot of wear data used to identify stable wear behavior and projecting wear. A straight line in this type of plot is indicative of stable behavior.

The characterization of the hardware and operating conditions serves two purposes. One, it insures that the conditions and materials are as they should be. The other purpose is that in the event that wear behavior is unsatisfactory, this information is needed for the continuation of the process to resolve the wear problem. All the tribosystem elements should be characterized. However, the principal parameters to characterize in most cases are material conditions, dimensions, roughness, and adjustments. If multiple tests can be done, or if different conditions can be included in a single test, it is useful and desirable to sort the hardware into different categories, particularly ones representing extremes, to determine the range of wear behavior that would be associated with these differences. This concept is particularly true when such differences are large and may be significant. The same approach should be used for operating conditions, when there are pronounced differences over the range of these parameters.

Avoiding Wear Failures

The ideal way of addressing a wear failure is by designing the mechanism initially to give adequate life (Ref 3). This approach is sometimes referred to as *wear design.* This method is similar to that of resolving a wear failure and becomes identical to it once there are some testing results. With design, the initial step is not the examination of hardware, as is the case with a failure, but developing various scenarios about the operation of the device and what constitutes a wear failure. As a minimum, the initial design should follow good tribological design practices and be selected to ensure that severe wear is avoided. For some applications, this approach may be adequate. For applications that are more sensitive to wear, analytical relationships for detailing the wear process may then be used to further refine and evaluate designs. Once hardware is built and some testing done, worn parts are available for examination, and the activities are identical to those that started with a wear failure. It is essential to determine the root cause of the wear failure.

Wear failures and problems tend to be very individualistic in terms of wear behavior and what constitutes a failure. Because of this tendency, there is no one type of solution or model that can be used as the basis for a solution. There are, however, some common elements in approaches that are used to resolve or avoid wear problems, and these have been described. Combined, they provide a general methodology for approaching wear failures. A key aspect in resolving wear failures is to recognize that wear is a system property or characteristic and not a materials property. While material changes are often involved in solutions to wear problems, this is not the only way of resolving such problems. Often, changes in other parameters are adequate by themselves or required in conjunction with a material change. Very often a key to the resolution of a wear problem is by detailed examination of worn parts and studying the operation of the device. While not always necessary, the use of analytical relationships for analyzing wear failures and developing solutions is feasible, generally worthwhile, and recommended (Ref 4). Further information about techniques used to investigate wear failures, as well as information about failure modes and wear behavior in different situations, is presented in other articles of this section. Several of the references (Ref 3, 4, 5) provide additional information about wear behavior, testing, and the use of analytical relationships for wear that are useful in resolving and avoiding wear failures.

References cited in this section

3. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994

4. R.G. Bayer, *Wear Analysis for Engineers,* HNB Publishing, New York, 2002

5. R.M. Voitik, Realizing Bench Test Solutions to Field Tribology Problems by Utilizing Tribological Aspect Numbers, *Tribology: Wear Test Selection for Design and Application,* STP 1199, A.W. Ruff and R.G. Bayer, Ed., ASTM, 1993

Fundamentals of Wear Failures

Raymond G. Bayer, Tribology Consultant

References

1. K. Budinski, Incipient Galling of Metals, *Proc. Intl. Conf. On Wear of Materials,* ASME, 1981, p 171–185

2. C. Lutynski, G. Simansky, and A.J. McEvily, Fretting Fatigue of Ti-6Al-4V Alloy, *Materials Evaluation Under Fretting Conditions,* STP 780, ASTM, 1982, p 150–164

3. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994

4. R.G. Bayer, *Wear Analysis for Engineers,* HNB Publishing, New York, 2002

5. R.M. Voitik, Realizing Bench Test Solutions to Field Tribology Problems by Utilizing Tribological Aspect Numbers, *Tribology: Wear Test Selection for Design and Application,* STP 1199, A.W. Ruff and R.G. Bayer, Ed., ASTM, 1993

6. P.J. Blau, Ed., *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992

7. P.J. Blau, *Friction and Wear Transitions of Materials,* Noyes Publications, Park Ridge, NJ, 1989

8. S.C. Lim and M.F. Ashby, *Acta Metall.,* Vol 35 (No. 1), 1987, p 1–24

9. R.G. Bayer, *Mechanical Wear Prediction and Prevention,* Marcel Dekker, 1994, Section C2

10. A.W. Ruff and R.G. Bayer, Ed., *Tribology: Wear Test Selection for Design and Application,* STP 1199, ASTM, 1993

11. M. Anderson and F.E. Schmidt, Jr., Wear and Lubricant Testing, Chapter 25, *ASTM Manual on Fuels, Lubricants, and Standards: Application and Interpretation,* ASTM, 2002

Abrasive Wear Failures

Jeffrey A. Hawk and Rick D. Wilson, Albany Research Center; Daniel R. Danks, Danks Tribological Services; Matthew T. Kiser, Caterpillar Inc.

Abrasive Wear Mechanisms

Of all the areas where abrasive wear is a problem, probably the most severe environment is in the excavation, earth moving, mining, and minerals processing industries, where component deterioration occurs in a wide variety of equipment, such as bulldozer blades, excavator teeth, rock drill bits, crushers, slushers, ball mills and rod mills, chutes, slurry pumps, and cyclones. However, abrasive wear is not limited to these activities. Abrasion presents problems in many wear environments at one point or another, even though it may not be the primary wear mechanism to begin with. In any tribosystem where dust and wear debris are not, or cannot be, controlled and/or excluded, abrasive wear is eventually a major problem. The wear of parts, the cost of repair and replacement of these parts, and the associated downtime related to these activities result in significant costs to many industries.

The individual factors that influence abrasive wear behavior are shown in Table 2. For both the abrasive and wear material, the majority of factors that affect abrasive wear behavior are related to their respective mechanical properties. Also of importance is the mechanical aspect of the abrasive/wear material interaction. Chemical processes, however, are also important, that is, corrosion or oxidation, because they directly influence the rate of wear of a material in the environment of interest.

Table 2 Factors influencing abrasive wear behavior

Abrasive properties	Particle size

	Particle shape Hardness Yield strength Fracture properties Concentration
Contact conditions	Force/impact level Velocity Impact/impingement angle Sliding/rolling Temperature Wet/dry pH
Wear material properties	Hardness Yield strength Elastic modulus Ductility Toughness Work-hardening characteristics Fracture toughness Microstructure Corrosion resistance

The influence of the parameters listed in Table 2 can be explained by their effect on the mechanism by which material is removed from a worn surface. The simplest model of abrasive wear is one in which rigidly supported hard particles indent and are forced across the surface of the wear material. Depending on the properties of the abrasive and wear materials, one of several wear mechanisms (Fig. 4) may occur (Ref 7, 10):

- *Plowing* occurs when material is displaced to the side, away from the wear particles, resulting in the formation of grooves that do not involve direct material removal. The displaced material forms ridges adjacent to grooves, which may be removed by subsequent passage of abrasive particles.
- *Cutting* occurs when material is separated from the surface in the form of primary debris, or microchips, with little or no material displaced to the sides of the grooves. This mechanism closely resembles conventional machining.
- *Fragmentation* occurs when material is separated from a surface by a cutting process and the indenting abrasive causes localized fracture of the wear material. These cracks then freely propagate locally around the wear groove, resulting in additional material removal by spalling.

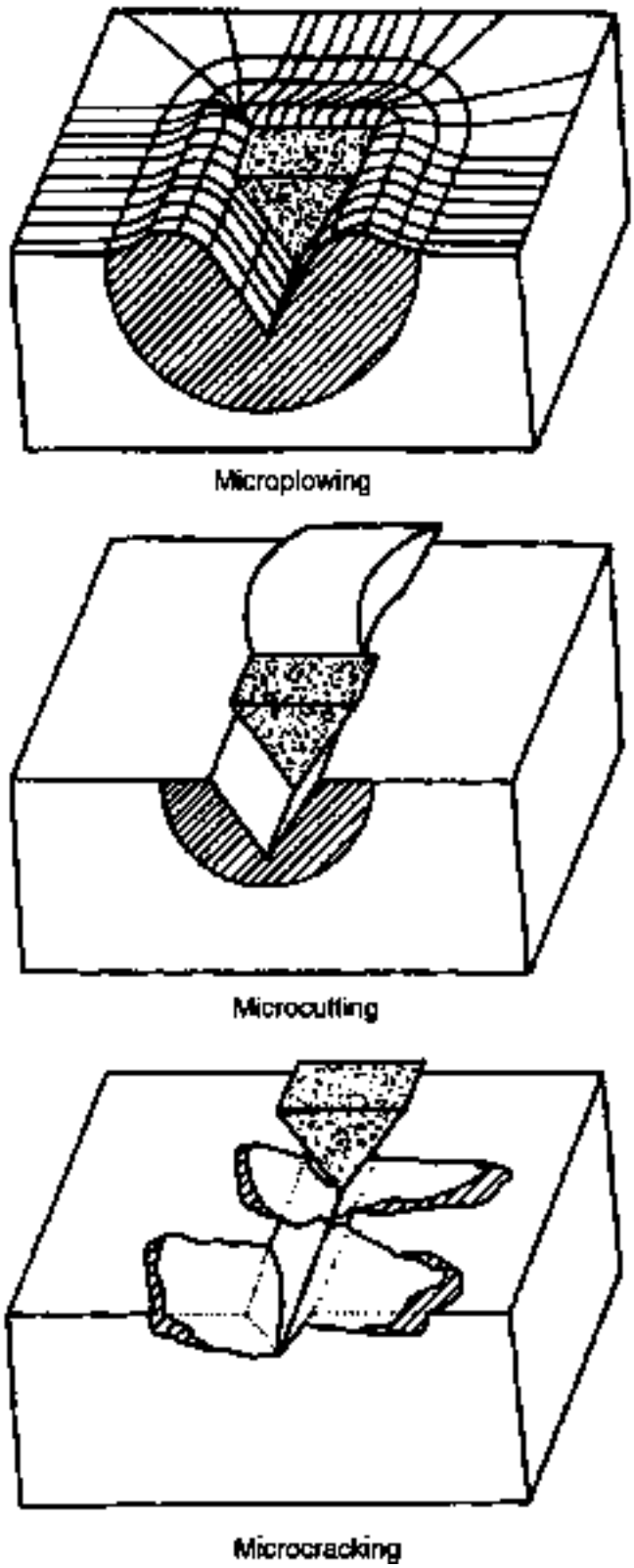


Fig. 4 Microscopic mechanisms of material removal between abrasive particles and the surface of materials. Source: Ref 6

The plowing and cutting mechanisms involve predominately plastic deformation of the wear material, while the third mechanism also involves fracture. Thus, the dominant mechanisms that occur for a particular operating condition are influenced to a great extent by the plastic deformation and fracture behavior of the wear material. Materials that exhibit high fracture resistance and ductility with relatively low yield strength are more likely to be abraded by plowing. Conversely, materials with high yield strength and with low ductility and fracture resistance abrade through fragmentation (Ref 7, 10, 11).

Additional wear mechanisms can operate in materials that exhibit a duplex microstructure or that are made up of two or more component phases (e.g., a composite-type material, such as a high-chromium white iron, or a composite drill bit, such as diamond or tungsten carbide inserts in a steel matrix), the individual components of which vary in their mechanical properties. Under some abrasive wear conditions, removal of the softer phase (usually the matrix) can occur by one or more of the previously mentioned mechanisms. This process then leaves the harder phase unsupported, in which case it may either become detached from the wear surface by a pull-out mechanism or be more susceptible to wear by fragmentation. In addition, the presence of an interface between the various components of the composite may promote cracking and fragmentation of the harder phase, particularly under impact-abrasion conditions.

The rate of material removal (or its wear rate) for any of the previously mentioned processes is influenced by the extent of indentation of the wear material surface by the abrasive particle. This depth of indentation, for a given load, is a function of the hardness of the wear material and the shape of the abrasive particle. Angular particles indent the wear surface to a greater extent than rounded particles, leading to higher wear rates. In addition, angular particles are more efficient in cutting and machining (Ref 11).

The hardness of the wear material, or more particularly, the hardness of the worn surface, is an important parameter in determining the resistance of a material to abrasion. An increase in the surface hardness of the wear material reduces the depth of penetration by the abrasive particle, leading to lower wear rates. However, an increase in the hardness of a material is also accompanied by an attendant reduction in its ductility, resulting in a change in the abrasion mechanism, for example, from predominately plowing and/or cutting to fragmentation (Ref 12).

For mechanisms involving predominately plastic deformation, that is, plowing and cutting, the wear rate can be expressed through the following parameters (Ref 7, 9): the probability of wear debris formation, the proportion of plowing and cutting processes, the abrasive particle shape and size, the applied stress, and the hardness of the wear surface.

For brittle materials (e.g., ceramics), a transition from a purely cutting mechanism to one that also involves fragmentation occurs when the nature of contact changes from elastic-plastic indentation to Hertzian fracture (Ref 11). The conditions under which this transition occurs are dependent on the size and shape of the abrasive particles, the applied stress, and the hardness of the wear surface. In addition, the fracture resistance of the wear material, as measured by the fracture toughness, is also important. Decreasing the hardness of the wear material, and increasing the fracture toughness, increases the critical abrasive size at which the transition to fragmentation occurs. Thus, reducing the hardness of brittle materials, or alternatively, increasing their fracture toughness, leads to lower wear rates (Ref 7, 10).

Abrasive wear mechanisms involving plastic deformation, cutting, and fragmentation occur predominantly in materials with relatively high elastic modulus, that is, metals, ceramics, and rigid polymers. As the elastic modulus decreases, the nature of the abrasive/wear material contact changes, with localized elastic deformation becoming more significant. The probability of wear occurring by plastic deformation mechanisms decreases, such that for elastomers, cutting mechanisms can occur only with contact against sharp abrasive particles. For contact against blunt abrasive particles, the two main wear mechanisms are tensile tearing and fatigue.

For the abrasive wear of polymeric materials, the following material parameters are important (Ref 10): elastic modulus and resilience, friction coefficient, tensile strength and tear resistance, elongation at break, and hardness. The wear behavior of elastomers is particularly sensitive to abrasive impingement angle, because this influences the dominant modes of deformation and hence, the wear mechanisms. At low impingement angles, tensile tearing occurs, and the tear resistance of the material is important. At impingement angles close to 90°, the behavior of the elastomer is essentially elastic, and resilience is a major factor in determining wear resistance. For abrasive wear conditions in which significant energy levels are dissipated in the abrasive/wear material contact, elastomeric linings are usually designed such that impingement angles are as close to 90° as possible.

Contact Pressure. In general, higher contact pressures between the abrasive particle and the wear surface in abrasive wear situations cause higher wear rates. As the force applied to an abrasive particle increases, the contact pressure between the abrasive and component wear surface also increases. As the contact pressure nears and exceeds the yield strength of the wear surface in the contact zone, the depth of abrasive penetration increases. For a given length of relative motion between the abrasive and the wear surface, deeper penetration generally removes more material or causes damage to a larger volume of material. The contact pressure/wear rate relationship has been discussed in more detail by several authors (Ref 13, 14, 15, 16, 17, 18, 19).

One of the more general (and generally accepted) wear equations was developed by Archard in 1953 (Ref 20):

$$W = ks\left(\frac{P}{P_m}\right)$$

where W is volume of worn (removed or disturbed) material, K is a constant related to probability of surface contact and debris formation, s is sliding distance, P is applied load, and p_m is flow pressure (related to hardness) of wearing surface.

The constant K is usually treated as a material property and is determined empirically. Typical values for various materials have been established through extensive wear testing (Ref 20). Sliding distance is one part of the volume aspect of affected material. Applied load combined with the flow pressure/hardness provide a measure of the depth of penetration of abrasive particles and supply the second volume dimension.

Abrasive Characteristics. Abrasive particle size has a significant effect on material wear, with the greatest effect being for nonmetals (i.e., ceramics and polymers) (Ref 21). In nonmetals, the effect of particle size is associated with changes in the predominant mechanism of material removal. Ceramics undergo a transition to fragmentation above a critical abrasive particle size, whereas elastomers undergo a transition from elastic behavior to either tearing or fatigue.

In metals, the effect of abrasive particle size is minimal for particle sizes >100 μm. Below this particle size, the wear rate decreases rapidly with decreasing particle size (Ref 11). This particle size effect is usually attributed to the nature of the abrasive/wear material contact, with decreasing size favoring elastic rather than plastic contact.

Abrasive hardness, or the ratio of hardness (Vickers) of the wear material to the hardness (Vickers) of the abrasive (H/H_a), is a critical parameter in abrasive wear. It is well known that the abrasive wear rate decreases as the hardness of the worn surface approaches that of the abrasive (Ref 10, 14, 15, 16, 19). When the hardness of the worn material exceeds that of the abrasive, the wear rate decreases rapidly. Figure 5 shows this particular effect for metals and ceramics.

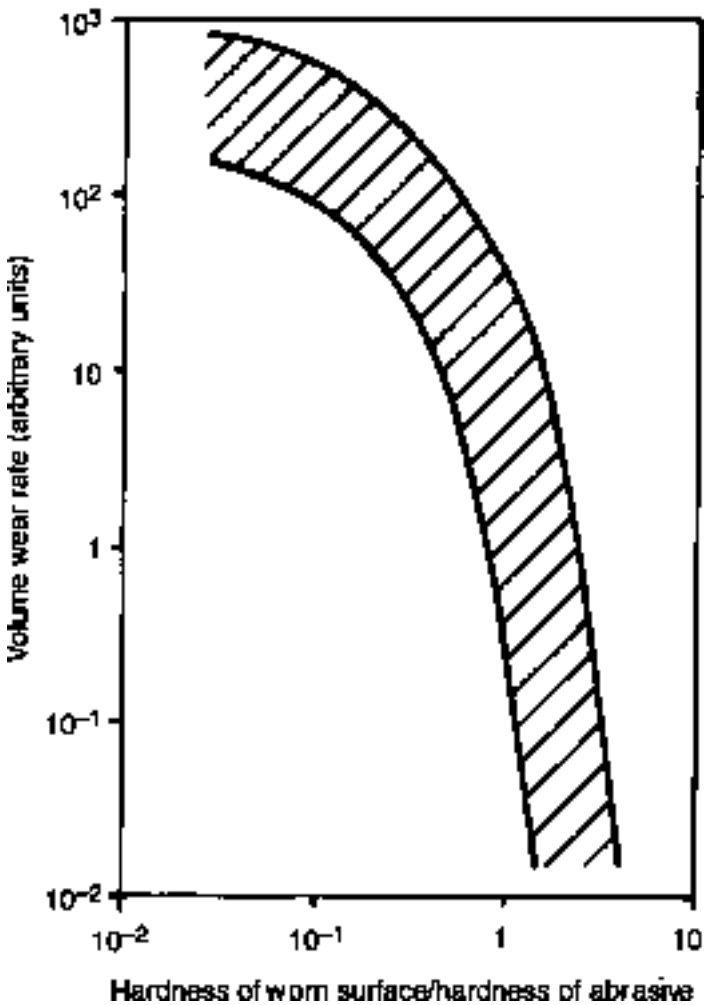


Fig. 5 Effect of abrasive hardness on wear behavior of metals and ceramics. Source: Ref 7

This ratio effect of H/H_a on abrasive wear results from a change in the nature of the contact mechanics. At H/H_a ratios between 0.6 and 0.8, the contact conditions give rise to extensive plastic deformation. At higher H/H_a ratios, the nature of the contact becomes essentially elastic (Ref 10, 11, 22). As a result, wear rates decrease, unless material is removed by mechanisms other than cutting and plowing, for example, fragmentation.

As the hardness of the worn surface approaches that of the abrasive, plastic flow of the abrasive may occur, leading to a reduction in the cutting ability of the abrasive particle. In addition, it is possible to fracture the abrasive when the plastic zone in the abrasive particle reaches a critical size (Ref 10, 19). This effect for the abrasive particle is analogous to the transition from purely cutting to fragmentation in the abrasion of brittle materials.

The effect of H/H_a on wear behavior is also influenced by the size and compressive strength of the abrasive particles. Coarse abrasives are more likely to fracture than fine abrasives, partly due to a decrease in tensile strength with increasing particle size. In addition, fracture of the abrasive may regenerate sharp facets and produce loose abrasive fragments, which in turn increase wear rates. Loading conditions also have an effect, because increasing the contact stress between abrasive and wear material increases the probability of fracture of the abrasive particles.

Contact Conditions. It is difficult to assess the effects of individual contact conditions on the wear interactions among abrasive and wear material, because their effect is synergistic in nature. Force, impact level, velocity, and impingement angle combine to influence the wear rate of the material.

Increasing the contact stress between the abrasive and the wear surface results in greater indentation depths and an increased tendency for fracture and fragmentation in both brittle wear materials and abrasives (Ref 10, 19). This generally leads to increased wear rates, although exceptions may occur if the abrasivity of particular minerals decreases with failure of the abrasive particle. The effect of nominal contact stress on the relative abrasion rating of a range of metallic materials is shown in Fig. 6.

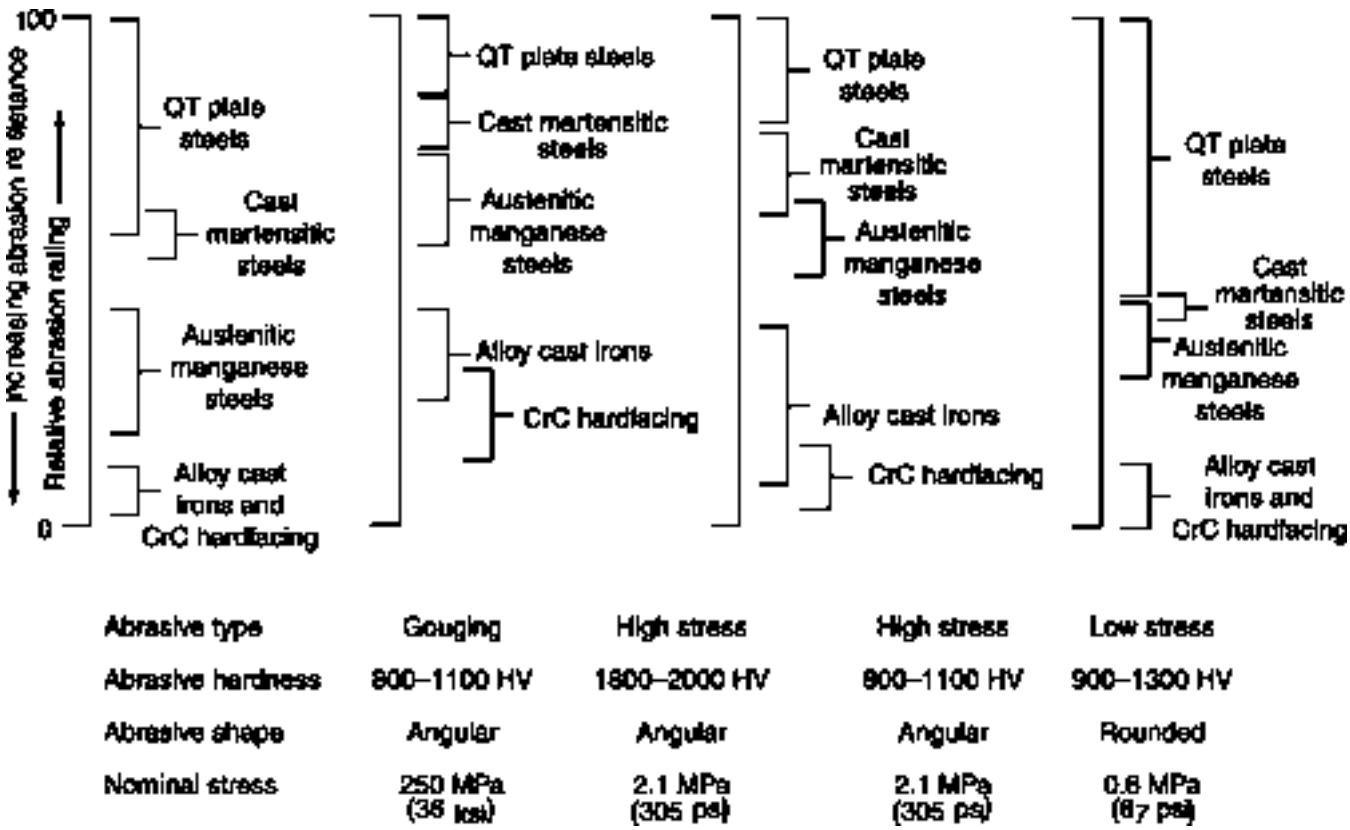


Fig. 6 Effect of nominal contact stress on relative abrasion rating of metallic wear materials. QT, quenched and tempered. Source: Ref 5

The general effect of increasing nominal contact stress levels is to increase wear rates; however, the occurrence of significant impact during the abrasive/wear material contact may also accelerate the rate of material removal. This acceleration results from the increased amount of kinetic energy dissipated during contact. This energy

may arise either from the moving abrasive particles or from the moving wear surfaces, as in the case of impact crushers. For brittle materials, increased fragmentation occurs under impact conditions, whereas elastomers may only suffer an increase in cutting and tearing modes, associated with insufficient elastic recovery.

The nominal force level may also determine the level of constraint experienced by abrasive particles. Under low-stress abrasive conditions, the abrasive particles can be free to rotate as they move across the surface of the wear material and are less likely to indent and scratch the surface. The tendency for the abrasive particle to rotate also depends on the abrasive particle shape, with angular particles being more likely to slide rather than roll. Increasing the nominal force acts to constrain the abrasive particle in its orientation to the wear surface, thereby increasing the wear rate.

The effect of velocity on wear behavior is associated with the dissipation of kinetic energy during abrasive/wear material contact. For a large number of abrasive wear environments, the velocity of abrasive particles is relatively low (<10 m/s) and therefore of little importance. However, in slurry pumps, that is, the transport of abrasive particles in slurry or pneumatic form, the effect of velocity is significant. Under these conditions, the wear rate is proportional to velocity ($W \alpha V^n$, where W is the wear rate, V is the velocity of the abrasive, and n is a constant for the abrasive/wear material synergism). The value of n falls in the range of 2 to 3, although the exact value for any particular condition is dependent on the properties of both the abrasive and wear material and on the angle of impingement (Ref 10, 22, 23). For softer abrasives, n tends to increase with decreasing abrasive particle size. For brittle materials and high impingement angles, the value of n tends toward the higher extreme.

In erosion, a change in abrasive impact velocity can lead to a change in the dominant wear mechanism (Ref 24), resulting in a change in wear rate. Higher values for n are associated with increased cutting and fragmentation. Figure 7 depicts the effect of erosive particle velocity on the wear of a range of materials.

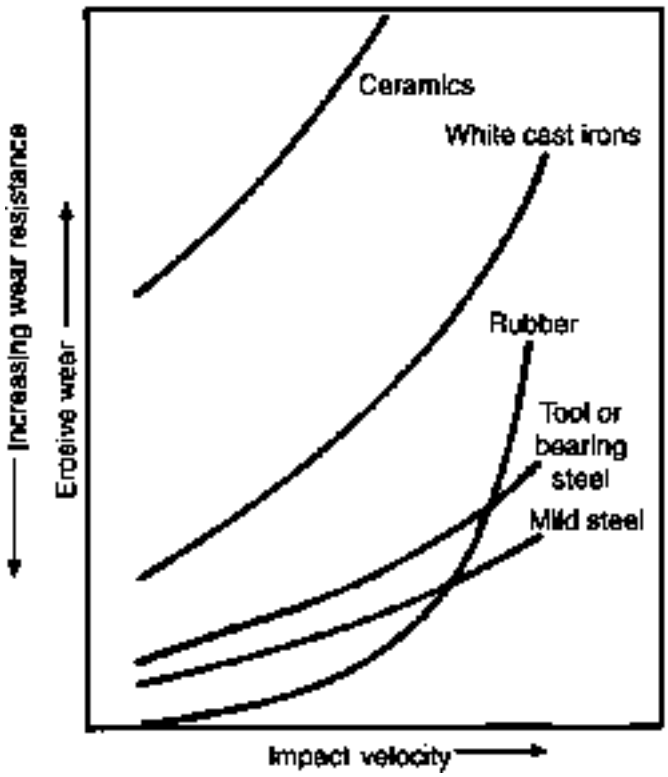


Fig. 7 Effect of impact velocity on erosive wear. Source: Ref 12

The influence of impingement angle on wear depends on the properties of the wear material and is associated with changes to the dominant wear mechanism. At high impingement angles (60 to 90°), brittle materials typically experience elevated wear rates, resulting from increased fragmentation and spalling. Conversely, elastomers are more effective under these conditions, because much of the impact energy can be dissipated through elastic deformation.

At low impingement angles (10 to 30°), elastomers cut and tear more readily, leading to increases in wear rates. Hard, brittle materials usually perform better under these conditions. For materials intermediate in their

mechanical properties, for example, some metals, the effect of impingement angle depends on the ductility of the material (Ref 24). The effect of impingement angle on the erosive wear behavior of a number of materials is shown schematically in Fig. 8.

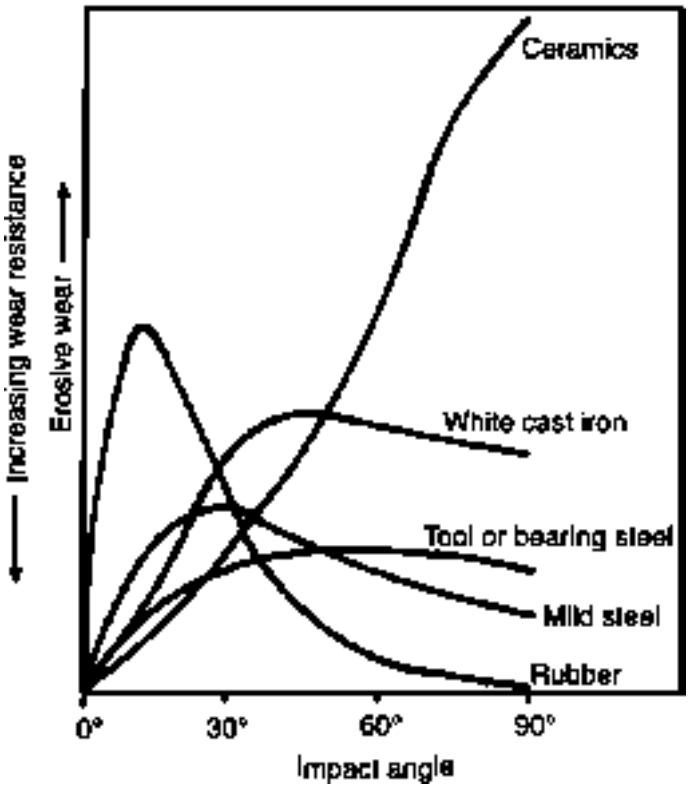


Fig. 8 Effect of impingement angle on erosive wear. Source: Ref 12

Properties of the Wear Material. The properties of the wearing material that influence wear behavior are grouped into the following categories: mechanical properties, microstructure effects, and other properties (corrosion resistance, friction, thermal effects). Because abrasive wear is primarily a mechanical process, particularly in the absence of corrosive environments, mechanical properties are of major importance, whereas the role of microstructure depends on the severity of the wear environment. The following discussion is patterned after the approach of Mutton (Ref 10) and Moore (Ref 11).

Mechanical Properties. The resistance to indentation (hardness) is an important variable in determining abrasion resistance. Laboratory wear tests indicate that the abrasive wear resistance for particular material types increases with increasing hardness (Fig. 9). However, large differences in abrasion resistance can also occur at similar hardness levels. These differences arise from variations in the plastic flow characteristics of the various materials, which in turn influence the predominant wear mechanism (Ref 7). The general relationships between abrasion resistance and hardness for plowing and fragmentation are similar to those shown in Fig. 10 for pure metals and ceramics.

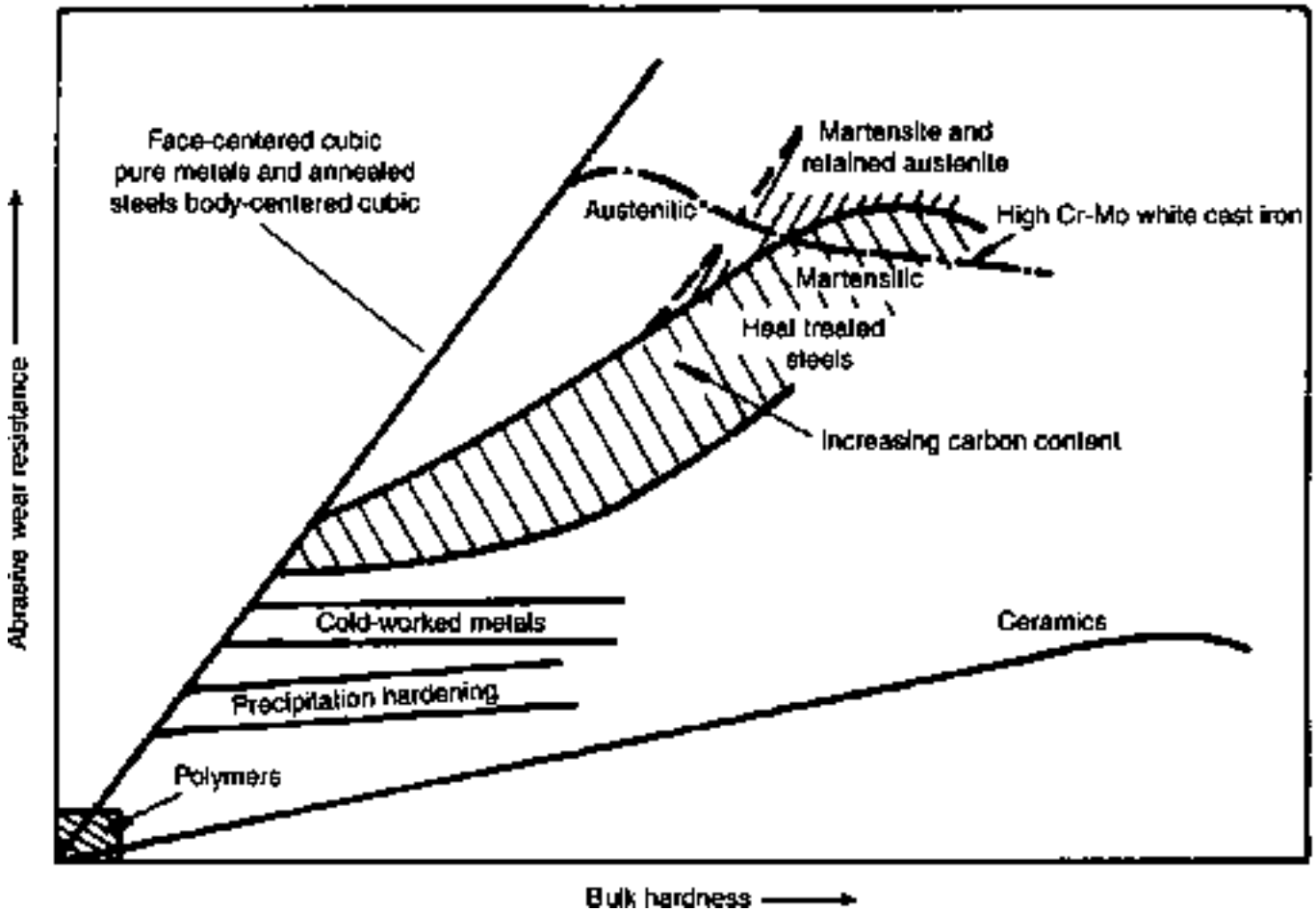


Fig. 9 Abrasion resistance versus hardness for various material types in high-stress pin abrasion tests (silicon carbide abrasive). Source: Ref 6

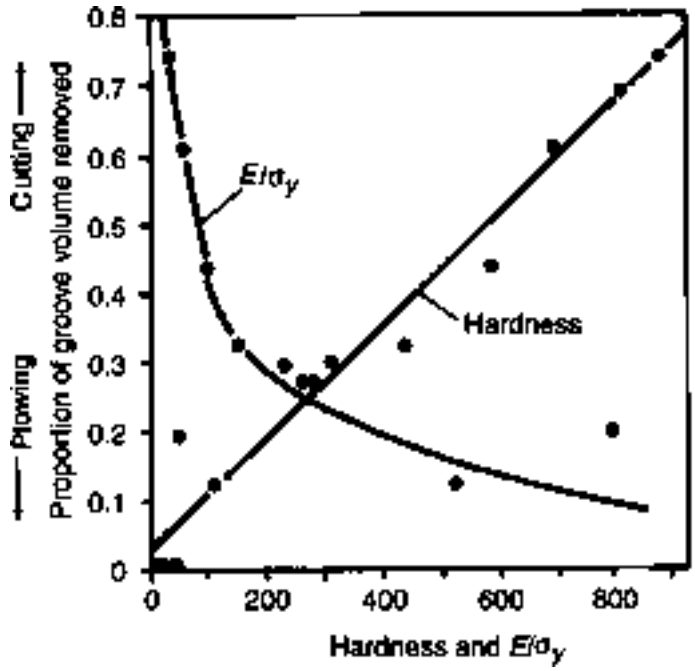


Fig. 10 Influence of hardness and E/σ_y on dominant wear mechanism. Source: Ref 5

In the absence of fragmentation, plowing and cutting involve plastic flow of the wear material either in front of or to the sides of the indenting abrasive particles. The plastic flow behavior of the material can be characterized by the ratio E/σ_y where E is the elastic modulus and σ_y is the flow stress of the non-work-hardened material.

Decreasing E/σ_y favors a cutting mechanism, in which a high proportion of material is removed as microchips. Increasing the hardness of the wear material has a similar effect, as shown in Fig. 10.

Although Fig. 10 indicates that changing E/σ_y can significantly affect the dominant wear mechanism, its impact on wear resistance is not as clearly established. For a constant value of E, a decrease in σ_y favors plowing. This should lead to lower material wear rates. However, decreasing σ_y also decreases hardness, which favors greater depths of indentation and a reduction in the wear resistance of the material. For materials of equivalent hardness, the general trend is for wear resistance to increase with higher E/σ_y values (Ref 11).

The wear material properties discussed thus far (i.e., hardness, elastic modulus, and flow strength) relate to material behavior at relatively low strains. For example, the plastic strain produced by indentation hardness testing of metals is typically between 8 and 10%. During the abrasive wear process, the extent of plastic strain experienced by a worn surface may approach these values. For these cases, the behavior of the wear material at high plastic strains is important, and here, high values of the work-hardening coefficient and ductility are favorable.

Tensile strength and ductility are also important mechanical property parameters in a wear environment, especially for polymers, which undergo elongations-to-fracture approaching 500%. Laboratory wear studies indicate that the wear resistance of both rigid and ductile polymers is roughly proportional to the work of rupture, which is defined as the product of the rupture stress and elongation-to-fracture.

The abrasion resistance of polymers also increases with increasing indentation resistance, although this relationship is not as well defined for polymers as it is for metals and ceramics. This is due in part to the viscoelastic nature of polymer deformation behavior. In addition, the absolute hardness levels for polymers are much lower than those for metals and ceramics.

For elastomeric materials such as natural rubbers and polyurethanes, the deformation behavior during abrasive wear may be entirely elastic, in which case resilience or hysteresis loss is important. High resilience favors increased abrasion resistance, while increasing hysteresis losses under high impact levels result in material breakdown, exacerbated by heat buildup.

For materials that undergo wear by fragmentation, abrasion resistance varies with fracture toughness. However, the contribution from microcracking and fragmentation depends on the severity of the wear environment, in particular the applied load and abrasive particle size.

For metals and ceramics, the general form of the relationship between wear resistance, hardness, and fracture toughness (K_c) is illustrated in Fig. 11 (Ref 25). For low-fracture-toughness materials (>14 MPa$\sqrt{m}$, or 13 ksi$\sqrt{in}$), wear resistance increases with increasing fracture toughness, despite a marked decrease in hardness. The increase in the wear resistance of the material results from a reduction in the fragmentation contribution to the total wear rate. For materials with high fracture toughness, wear occurs by cutting and plowing only. Hence, the wear rate of the material is controlled by the indentation resistance. In this case, wear resistance decreases with decreasing hardness.

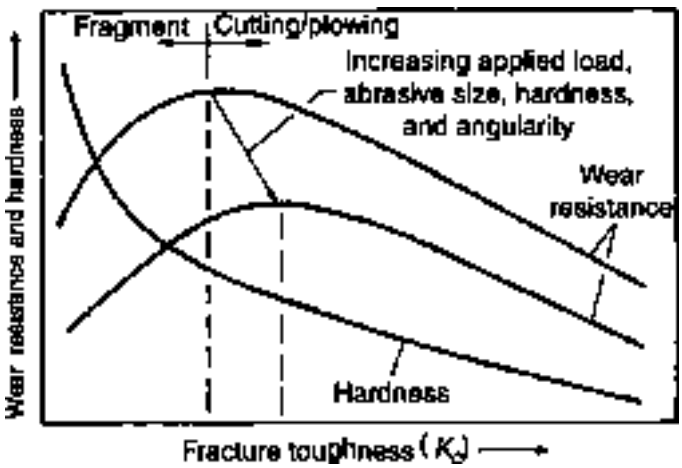


Fig. 11 Schematic relationship between wear resistance, hardness, and fracture toughness. Source: Ref 6

The transition from fracture (fragmentation) to plastic deformation (cutting and plowing) is dependent on the critical groove size (p_{crit}) and the contact stress. These parameters, in turn, also depend on the ratio of fracture

toughness to hardness (K_c/H). From Fig. 11, maximum abrasion resistance is obtained by optimizing the ratio K_c/H while maintaining the indentation hardness or groove size slightly below that at which the transition to fragmentation occurs (Ref 11).

Increasing the severity of the wear environment, by increasing either the applied load or the size, hardness, or angularity of the abrasive particles, shifts the transition from fragmentation to cutting/plowing to higher values of fracture toughness (Fig. 11). This results in a decrease in the maximum wear resistance.

Microstructure. Microstructural effects on abrasive wear depend on the overall magnitude or scale of the wear environment. For high-impact loads, large abrasive particle size, and so on, the size of the microstructural features is generally much smaller than that of the abrasive wear damage. In this case, the role of microstructure on wear resistance is limited to its effect on bulk mechanical properties. Abrasive wear rates, particularly for materials that exhibit homogeneous microstructures, may correlate with bulk hardness (Ref 11). Figure 12 shows the effect of microstructure and hardness on the high-stress abrasion resistance of steels. Over a limited hardness range (200 to 300 HV, or 2.0 to 3.0 GPa), relative abrasion resistance decreases in the order (Ref 11): austenite > bainite > pearlite > martensite.

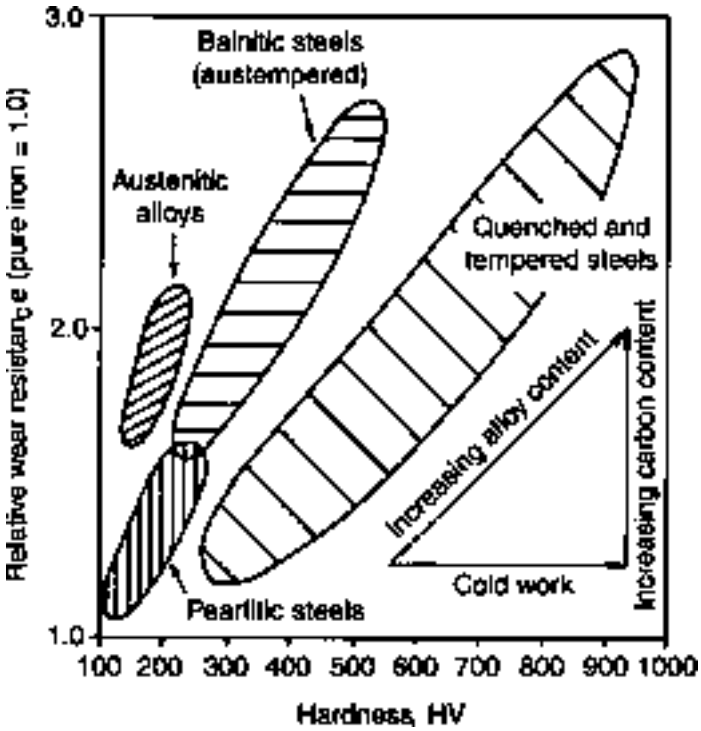


Fig. 12 Effect of microstructure and hardness on the abrasion resistance of steels: high-stress abrasion, alumina abrasive. Source: Ref 7

This effect can be attributed to the influence of microstructure on plastic deformation behavior. However, bainitic and martensitic microstructures offer greater abrasion resistance at higher hardness levels (>400 HV, or 4.0 GPa).

At lower loads, and with smaller abrasive particle sizes, microstructural features are more effective as discrete components, and the mechanical properties of individual phases assume increased importance. This occurs when the size of the microstructural features is approximately equal to, or larger than, that of the abrasive. Under these conditions, materials with duplex microstructures, for example, alloy white irons and some ceramics, are more sensitive to microstructural effects. In such materials, size, spacing, and volume fraction of the harder phase, as well as the mechanical properties of both hard and soft (matrix) phases, can significantly affect wear behavior. Thus, the abrasion resistance of the material either increases, decreases, or remains the same by increasing the volume fraction of the harder phase. The net effect depends on the various contributions of the plowing, cutting, fragmentation, and pull-out mechanisms to the total wear rate. Figure 13 shows schematically this behavior for alloy white irons and alloy carbide materials. The transition from increasing to decreasing abrasion resistance with increasing carbide volume fraction is associated with the relative contributions of fragmentation of the carbides and plowing/cutting of the matrix to the total wear. Carbide hardness, and size relative to that of the wear damage, also influences these trends (Ref 11, 26).

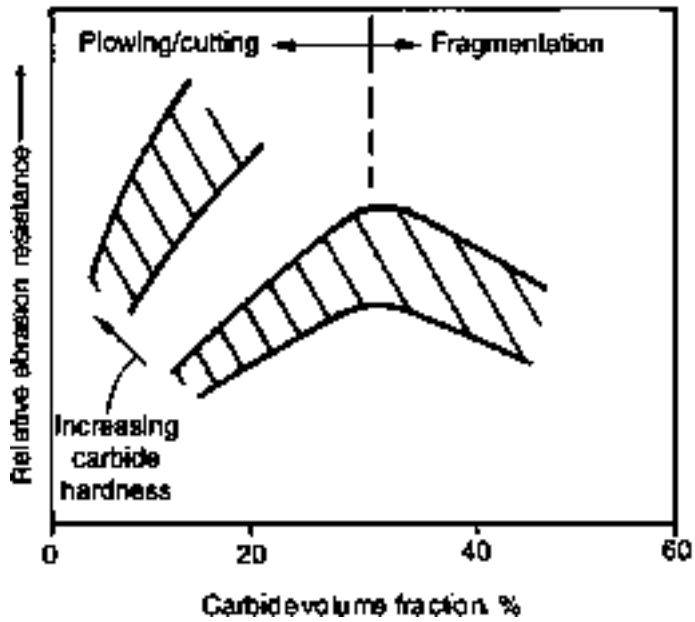


Fig. 13 Influence of microstructure on abrasive wear behavior of alloy white irons and chromium carbide materials. Source: Ref 5

Other Properties. Wear processes that involve plastic deformation (i.e., cutting and plowing) also generate significant increases in temperature at the wearing surface. These increases occur as a result of frictional energy dissipation at the interface between the wearing surfaces. Under these circumstances, the wear behavior is influenced by characteristics such as the friction coefficient and thermal conductivity of the two wear bodies. In materials that exhibit low values of thermal conductivity, one of two additional effects may arise, depending on other material properties. Localized thermal expansion in ceramics may result in additional material loss by spalling; conversely, localized temperature increases in polymers may give rise to melting. Materials that exhibit low friction coefficients under dry sliding conditions (e.g., some polymers and finely ground ceramics) suffer less degradation through frictional temperature rise effects.

References cited in this section

5. H. Czichos, Systems Approach to Wear Problems, *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Ed., American Society of Mechanical Engineers, 1980, p 17–34

6. J.D. Gates and G.J. Gore, Wear of Metals: Philosophies and Practicalities, *Mater. Forum,* Vol 19, 1995, p 53–89

7. K.-H. Zum Gahr, *Microstructure and Wear of Materials,* Elsevier Science Publishers, 1987

9. T.E. Norman, Wear in Ore Processing Machinery, *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Ed., American Society of Mechanical Engineers, 1980, p 1009–1051

10. P.J. Mutton, *Abrasion Resistant Materials for the Australian Minerals Industry,* Vol 1, Australian Minerals Industries Research Association Limited, 1988

11. M.A. Moore, Abrasive Wear, *Fundamentals of Friction and Wear of Materials,* D.A. Rigney, Ed., American Society for Metals, 1981, p 73–118

12. M.J. Murray, P.J. Mutton, and J.D. Watson, Abrasive Wear Mechanisms in Steels, *Wear of Materials,* W.A. Glaeser, K.C. Ludema, and S.K. Rhee, Ed., American Society of Mechanical Engineers, 1979, p 257–265

13. D.R. Danks and P. Clayton, Comparison of the Wear Process for Eutectoid Rail Steels: Field and Laboratory Tests, *Wear,* Vol 120, 1987, p 233–250

14. R.C.D. Richardson, The Maximum Hardness of Strained Surfaces and the Abrasive Wear of Metals and Alloys, *Wear,* Vol 10, 1967, p 353–382

15. R.C.D. Richardson, The Wear of Metals by Hard Abrasives, *Wear,* Vol 10, 1967, p 291–309

16. M.M. Khruschov, Resistance of Metals to Wear by Abrasion as Related to Hardness, *J. Mech. Eng.,* 1957, p 655–659

17. M.M. Khruschov and M.A. Babichev, Issledovaniya iznashivaniya metallov, *Izdat. Akad. Nauk SSSR,* 1960, p 889–893: Research on Wear of Metals, N.E.L. translation

18. M.A. Moore, The Relationship Between the Abrasive Wear Resistance, Hardness and Microstructure of Ferritic Materials, *Wear,* Vol 28, 1974, p 59–68

19. R.C.D. Richardson, The Wear of Metals by Relatively Soft Abrasives, *Wear,* Vol 11, 1968, p 245–275

20. J.F. Archard, Contact and Rubbing of Flat Surfaces, *J. Appl. Phys.,* Vol 24, 1953, p 981–988

21. A. Misra and I. Finnie, On the Size Effect in Abrasive and Erosive Wear, *Wear,* Vol 65, 1981, p 359–373

22. I.M. Hutchings, *Tribology: Friction and Wear of Engineering Materials,* CRC Press, 1992

23. T.H. Kosel, Solid Particle Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* P.J. Blau, Ed., ASM International, 1992, p 199–213

24. A.R. Lansdown and A.L. Price, *Materials to Resist Wear, A Guide to Their Selection and Use,* Pergamon Press, 1986

25. K.-H. Zum Gahr, Relation Between Abrasive Wear Rate and the Fracture Toughness of Metallic Materials, *Z. Metallkd.,* Vol 69, 1978, p 643–650

26. K.-H. Zum Gahr and D.V. Doane, Optimizing Fracture Toughness and Abrasion Resistance in White Cast Irons, *Metall. Trans. A,* Vol 11, 1980, p 613–620

Jeffrey A. Hawk and Rick D. Wilson, Albany Research Center; Daniel R. Danks, Danks Tribological Services; Matthew T. Kiser, Caterpillar Inc.

Wear Failure Analysis

Table 3 presents the procedural sequence used in general failure analysis studies. Failure analysis studies can be broken down into the following major categories through the collection of documentary evidence: service conditions; materials, handling, storage, and identification; interviews; field investigation; laboratory analysis; and dimensional analysis (Ref 2). The same general procedure is used in analyzing wear failures.

Table 3 General fracture failure analysis procedural sequence

Step 1: Determine prior history
Review documentary evidence Test certificates Mechanical/wear test data Pertinent specifications Correspondence Interviews Depositions and interrogatories Review service parameters Design or intended operating parameters Actual service conditions Temperature data (magnitude and range) Environmental conditions Service loads/stresses
Step 2: Clean failed parts or specimens
Step 3: Perform nondestructive tests
Macroscopic examination of wear/fracture surface Presence of color or texture Temper colors Oxidation Corrosion products Contaminants Presence of distinguishing surface features Shear lips Beach marks Chevron markings Gross plasticity Large voids or exogenous inclusions Secondary cracks

Direction of propagation Fracture origin Detection of surface and subsurface defects Magnaflux Dye penetrant Ultrasonics Hardness measurements Macroscopic Microscopic Chemical and microstructural analysis Macrochemistry (“wet” chemical analysis) Spectrographic analysis X-ray diffraction Electron beam microprobe analysis
Step 4: Perform destructive tests
Metallographic (light microscopy, scanning electron microscopy, transmission electron microscopy) Macroscopic Microscopic Structure Grain size Cleanliness Mechanical tests Tensile Impact Fracture toughness Corrosion tests Wet chemical analysis
Step 5: Perform stress analysis
Step 6: Store and preserve parts or specimens

Source: Ref 1

Wear is a systems problem (Ref 5, 27). The process of accurately detailing an abrasive wear failure consists of a series of defined tasks undertaken by the failure analysis specialist. Each task is designed to obtain specific information from the failed components and system. These tasks are generic to most failure analysis investigations and can be summarized in the following list (Ref 28):

1. Identify the actual materials used in the worn part, noting also the operating environment, abrasive causing the material loss, the wear debris, and any lubricant used (as needed).
2. Identify the specific wear mechanism, or combination of wear mechanisms, that caused the loss of material or change in the surface dimensions: adhesive wear, abrasive wear, corrosive wear, surface fatigue, erosive wear, and so on.
3. Determine the change in the overall dimensions to the surface configuration between the worn surface and the original surface.
4. Determine the relative motions involved in the tribosystem that caused the loss of material, for example, abrasive wear, including the direction and velocity of the relative motion.
5. Determine the force or contact pressure between mating surfaces or between the worn surfaces, that is, the counterfaces and the abrasive particles, on both the macroscopic and the microscopic level.
6. Determine the wear rate by calculating the material loss over some unit of time or distance.
7. Determine the coefficient of friction (if possible).
8. Identify the type of lubricant used and its effectiveness in slowing down material loss, for example, was the grease, oil, modified surface, naturally occurring oxide layer, adsorbed film, or intentional foreign material beneficial or not.
9. Establish whether the observed wear is normal or abnormal for the particular application.
10. Devise a solution, if required.

In most failures, some or even most of this information is simply not available. The more information that can be accurately determined, however, the better the chances of making a successful determination of cause and therefore, chance of remediation. These steps are used in each of the example failure analyses in this article.

In order to perform effective analyses of abrasive wear failures, it is not enough to have a broad-based knowledge of the mechanical situation found in the wear environment. A good background in metallurgy, ceramic science, or materials science is also necessary. In addition, access to investigative diagnostic tools is required for examination of worn surfaces and structures on both a macroscopic and microscopic level (Ref 29). This includes methods such as optical and electron microscopy, electron spectroscopy for chemical analysis, x-ray diffraction, low-energy electron diffraction, and Auger electron spectroscopy. Above all, a well-documented plan of investigation is needed before any analysis is begun.

Once a plan of investigation is decided on, it is then necessary to develop a case history of the failure, documenting as much as possible all aspects of the problem. This may not be easy to do, because in most cases, detailed operating records are not kept. However, it is important to collect as much information about the operating history of the component as possible in this portion of the investigation. More importantly, it is necessary to determine at this stage whether the component and associated parts were operating properly, that is, as they were intended. If they were not, then it is easy to correct the problem by specifying the correct operating procedures and then to monitor subsequent operational situations to see if the same problem occurs.

However, in many cases, operation of the component and associated parts was within limits. In these instances, a thorough examination of the parts is necessary, using some or all of the analytical tools mentioned previously. It is important, however, to examine the worn surfaces of the part of the failed component prior to its disassembly. Doing this provides a broad operational overview of the failure of the component, or more precisely, how all the parts fit together and operated. At this point in the analysis process, all operational questions need to be answered. Once a record is obtained, no matter how incomplete, the component can be broken down into its constituent parts, as was done in the case studies investigated in this article. Then the wear surfaces can be analyzed for manufacturing and/or material defects or any other unusual occurrences, and a preliminary determination as to the cause of the wear failure expounded. This may or may not lead to remediation procedures. If it does, these steps can be investigated on subsequent components to see if the failure was properly resolved (such as a change in material used for the parts or a redesign of the part to eliminate a possible design feature leading to accelerated wear). Of course, the material or design requirements

to solve an abrasive wear failure, or any failure for that matter, may prove to be too expensive, leading to other solution approaches, such as the complete redesign of the component and the associated parts that failed.

References cited in this section

1. E. Rabinowicz, *Friction and Wear of Materials,* 2nd ed., John Wiley & Sons, Inc., 1995, p 5–7

2. V.J. Colangelo and F.A. Heiser, *Analysis of Metallurgical Failures,* 2nd ed., John Wiley & Sons, Inc., 1987, p 1–10

5. H. Czichos, Systems Approach to Wear Problems, *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Ed., American Society of Mechanical Engineers, 1980, p 17–34

27. H. Czichos, *Tribology,* Elsevier Science Publishing Co., 1978, p 14–23

28. R.C. Tucker, Wear Failures, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 145–162

29. K.C. Ludema, *Friction, Wear, Lubrication: A Textbook in Tribology,* CRC Press, 1996, p 205–211

Abrasive Wear Failures

Jeffrey A. Hawk and Rick D. Wilson, Albany Research Center; Daniel R. Danks, Danks Tribological Services; Matthew T. Kiser, Caterpillar Inc.

Examples of Abrasive Wear

Example 1: Jaw-Type Rock Crusher Wear. Although rock crushing and mineral comminution components are expected to lose surface material during their operating lifetime, this loss of material still results in reduced and/or eliminated component function and leads directly to failure. In this type of system, mineral ore flows down a feed chute into the upper portion of the crushing zone, which consists of two plates, one stationary and one moving. The chunks of rock enter at the top and are reduced in size each time the jaws cycle toward each other. The mineral then moves through the crushing zone until it reaches the desired size at the bottom, where the crushed pieces exit through the gap at the bottom of the plate assembly each time the plates separate to accept new rock.

Figure 14 shows the jaw crusher wear plates after processing a quantity of mineral ore. In this case, the crusher plates are of the size and type outlined in ASTM G 89. The plates were quenched-and-tempered, low-alloy steel (~0.30 wt% C at 514 HB hardness), with an elliptical motion of the movable plate relative to the fixed plate (Ref 30). In this case, the entering rock is 50 to 75 mm (2 to 3 in.) in size. On exiting the jaw crusher plate assembly, the mineral is approximately less than 6 mm (0.2 in.) in size. Notice that for this jaw crusher configuration, the stationary plate absorbs the most severe gouging-abrasive wear (right side of the plate). Table 4 summarizes the damage characteristics on both the stationary and moveable plates.

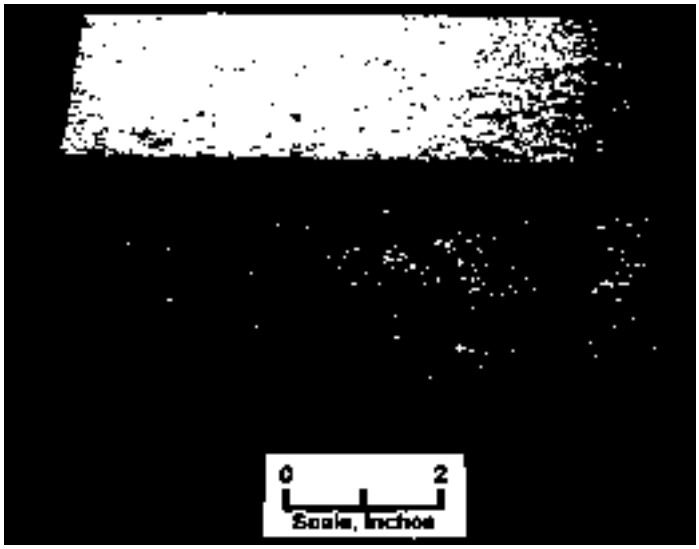


Fig. 14 Photograph of the movable (top) and stationary (bottom) wear plates from the jaw crusher. The feed moves across the plates from left to right, so the most severe wear occurs at the exit area toward the right.

Table 4 Jaw crusher plate damage summary

Plate		Location (from top)		Damage type	Typical event size		Relative material loss	Displacement direction
		mm	in.		mm	in.		
Moveable	Zone 1	0–50	0–2	Blunt with flow	12	0.5	None	Up
	Zone 2	50–100	2–4	Mixed rm	25	1.0	Minor	Up
	Zone 3	100–175	4–7	Sharp and deep, rm	25	1.0	Moderate	Up
	Zone 4	175–210	7–8.3	Small, direct penetration	6	0.25	Minor	Up
Stationary	Zone 1	0–50	0–2	Blunt, direct penetration	12	0.5	None	None
	Zone 2	50–100	2–4	Sharp, direct penetration	18	0.75	Moderate	Up
	Zone 3	100–210	4–8.3	Sharp, rm	12	0.5	Majority	Up

rm, relative movement direction; that is, direction the surface metal is flowing

Example 2: Electronic Circuit Board Drill Wear. Very small drill bits are used for drilling holes in electronic printed circuit boards (PCBs). To be economical, the drilling process must be completed quickly, because of the large number of holes in each board. This high-speed drilling operation thus requires automatic drilling machines capable of identifying hole location, starting and stopping quickly, and changing worn drill bits as needed. The maximum drilling rate for the system is the rate of maximum drill bit breakage. Generally, the optimal drilling conditions are determined by pushing the drilling conditions to their limit, that is, until drill bit failure. The failure of the drill bits is then analyzed.

Commercial drill bits vary in size, but popular sizes for drilling holes in PCBs are on the order of 0.343 to 0.457 mm (0.0135 to 0.0180 in.) in diameter by 6.35 mm (0.25 in.) in length. In this case, the drill bits are made of cobalt with approximately 50 vol% of uniformly distributed 0.1 μm sintered WC particles in the cobalt matrix. Vickers hardness of the drill bits was measured at 1589 HV. The PCBs are composite materials, with layers of fiberglass epoxy resin sandwiched between copper layers. A PCB contains as many as 14 layers of copper and fiberglass-resin layers. The glass fibers in the fiberglass have a hardness of 500 HV, while the copper has a hardness of 40 HV.

The drilling process is computer operated and numerically controlled. Typical operating conditions are as follows: the feed for drilling PCB holes is approximately 2000 mm (80 in.)/min. The speed corresponds to the number of drill bit rotations per minute, and for this application, it was between 80,000 and 100,000 rpm. The hit rate, or the number of holes drilled in a particular time interval, was 80 holes/min. After 1500 hits, the drill bits were resharpened. The material removal rate (MRR) is 4.8 mm^3/s and is given by the following equation (Ref 31):

$$MRR = nf_r \frac{p\,d^2}{4000}$$

where d is the diameter in millimeters, n is the rotational speed in revolutions per second, and f_r is the feed per revolution in millimeters.

In this study, a sharp, new 0.343 mm (0.0135 in.) diameter drill bit was loaded into a tool holder and operated at the appropriate drilling conditions to drill a hole in a PCB. Figure 15 shows the results of this operation: a hole 0.3404 mm (0.0134 in.) in diameter. Notice that the hole has sharp edges and round, straight internal dimensions. The drill bit made chips of fiberglass and copper foil while cutting through the PCB. The glass fibers break by brittle fracture on impact with the cutting edge of the drill bit. Intense stress develops under the surface of the glass fiber where it makes contact with the bit, generating a plastic zone and a median vent crack in the plane of the stress field. Lateral cracks also develop and curve back to the fiber surface, liberating wear debris of glass fibers and epoxy. When the drill bit cuts through the layers of copper, ordinary ductile metal chips form by shearing. The chips tend to flow up the rake face of the drill bit and deposit in the flute, causing copper buildup on the edge.

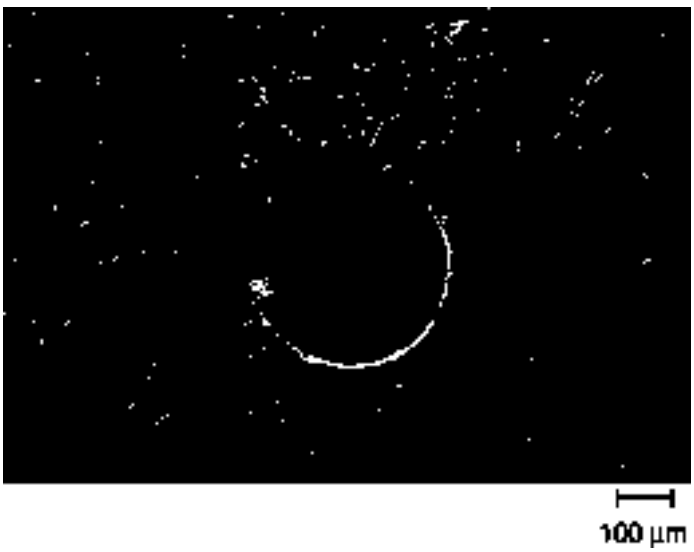


Fig. 15 A scanning electron micrograph of the details of a hole drilled with a new, sharp drill bit. Note the clean hole with only a minor amount of damage to the hole periphery.

One way a drill bit fails is by catastrophic breakage from, for example, too high a load applied to the bit. A dull drill bit is susceptible to the buildup of wear debris on the flute. As wear debris accumulates on the bit, the friction force between the bit and hole increases. The increased friction generates additional heat, exacerbating wear debris buildup on the bit cutting edge. This reduces bit strength and increases hole size, a direct result of a change in bit diameter due to thermal expansion. The buildup of wear debris also rounds the cutting edge. Thus, in order to maintain cutting efficiency, the cutting force must be increased, which leads to increased shear stress on the drill bit shank. When the applied shear stress exceeds that of the bit material, it fractures, usually in the region between the spiral flutes and the shank.

Another way that drill bits ultimately fail is by accelerated wear. Drill bits typically have a break-in period, where sharp burs on the drill bit are worn away. This break-in period is followed by steady-state wear when the majority of the holes are drilled with uniformly acceptable characteristics (Fig. 15). The majority of the tool lifetime is spent in the steady-state wear regime. Taylor's tool lifetime equation can be used to predict this time (Ref 32):

$$VT^n = C$$

where V is the cutting speed, T is the tool life, C is a material-dependent constant, and n is a constant that depends on tool material and cutting conditions. Eventually, the sharpness of the bit degrades to a point where edge wear-debris accumulation becomes a problem. At this point, the drill bit has effectively failed, because the characteristics of the hole have significantly changed (Fig. 16). Examining Fig. 16, it is clear that the worn drill bit has slid away from the intended hole location by 0.3810 mm (0.015 in.), approximately one drill bit diameter. The margins of the hole are also rounded, and it is now oversized by 12.6%.

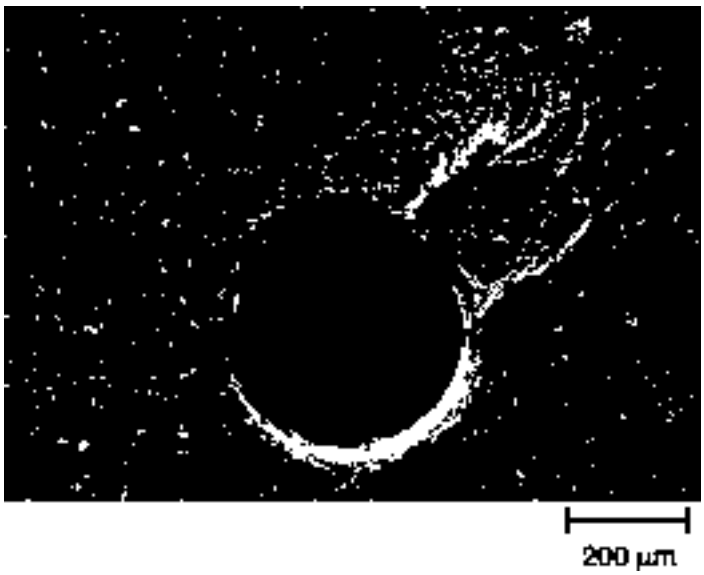


Fig. 16 A scanning electron micrograph of a hole drilled with a worn drill bit. Note that the final position of the hole is not where the drilling started (i.e., the drill wandered across the surface before "biting") and the ragged nature of the periphery of the hole.

The microstructure of these drill bits has been optimized, so that further improvements in hardness and wear resistance may not be possible. However, some things can be done to increase the productivity of the drilling operation by properly determining drill bit life and changing them at the beginning of the accelerated wear phase of the tool. One way to do this is by measuring the drilling force on the chuck, using a three-axis force dynamometer. When the cutting force deviates from the steady-state constant force condition, the bit must be resharpened or replaced. Feeds and speeds of the drilling operation in combination with changing the cutting angle of the drill bits may also lower the number of fractured bits.

Example 3: Grinding Plate Wear Failure Analysis. In this example, a 230 mm (9 in.) diameter disk attrition mill was scheduled to grind 6.35 mm (0.25 in.) diameter quartz particles to a 0.075 mm (0.003 in.) diameter powder. Due to severe wear on the grinding plates, however, the unit was unable to complete the task of grinding the rock. The unit was disassembled and an analysis performed to determine the cause of plate failure. Several steps were used to analyze the disk attrition mill failure, which included both an analysis of the grinding plates and the rock to be comminuted. Historical information was collected about the disk attrition mill and used to compare the present operating conditions with previous conditions where acceptable product was produced. Several wear mechanisms were identified and used to make recommendations for future improvements.

Grinding Plates. The disk attrition mill is shown in Fig. 17. It consists of a heavy gray cast iron frame, a gravity feeder port, a runner, and a heavy-duty motor. The frame and gravity feeder weighed over 200 kg (440 lb) and, in some areas, was over 25 mm (1 in.) thick. Gray cast iron is considered ideal for the machine frame and base, because it absorbs vibration and promotes smooth, quiet operation during grinding. To obtain the operating speed of 200 rpm, a gear system is used to transmit the torque from the 2 hp motor. The runner consisted of a 50 mm (2 in.) diameter shaft and two gray cast iron grinding plates (i.e., the parts used to grind the mineral).

Fig. 17 Photograph of the attrition mill, showing placement of the attrition mill wear plates

In Fig. 18, a new, unused grinding plate (left) is a 230 mm (9 in.) diameter and 25 mm (1 in.) thick disk of gray cast iron with a hardness of 338 HV and 0.1 mm (0.004 in.) graphite flakes, which are an excellent absorber of vibrations during mineral grinding. Radiating outward from the center are six cutting flutes that are used to shear and fracture the quartz particles. The 32 mm (1.25 in.) wear ring on the outside of the plate grinds the broken rock fragments to their final size. The procedure used to grind quartz in the disk attrition mill included: (1) setting the space between the gray cast iron plates to zero clearance, (2) turning on the unit, and (3) feeding in quartz between the plates. The quartz was precrushed to 6.35 mm (0.25 in.) diameter and then fed directly into the feed hopper at a rate of 0.5 kg (1 lb)/min. The process continued until approximately 4.5 kg (10 lb) of the 75 μm product was deposited in the collection tray. The product was removed from the collection tray and the process repeated. Eventually, the machine failed due to grinding plate wear.

Fig. 18 Macrographs of new (left) and worn (right) attrition mill wear plates

Characterization of the Quartz Feed. Quartz from western Oregon was the primary feed material used in the attrition mill. This rock has a Mohs hardness of 7, corresponding to a Vickers hardness of between 700 and 800 kg/mm^2. The quartz mineral possesses a rupture strength of 70 MPa (10 ksi) in the transverse direction. On slow cooling, quartz forms a tetrahedral silicate structure, with silicon at the center and oxygen atoms occupying the corners. Quartz fractures conchoidally along crystallographic planes when fractured. The sharp, angular shape of the quartz particles has a significant effect on the wear behavior of the grinding plates in the attrition mill, with quartz attrition leading to failure of the part in the following way.

When the quartz particles are pulverized by the rubbing action between the rotating plate and the stationary plate, particles become embedded in the two grinding plates and abrade through, scratching the wear ring surfaces at the edge of the plate. Initially, the clearance between the plates is set to zero, but as the quartz particles flow into the space between the plates, they wedge them apart. The particles wedged between the

plates can see an elastic force equal to 4.05 kN, leading to compressive stresses on the particles of 0.12 MPa (17 psi), sufficient to overcome the transverse rupture strength of 0.07 MPa (10 psi) for the quartz.
Additionally, the quartz particles are twice as hard as the cast iron grinding plates and cause severe wear by two distinct wear mechanisms, illustrated in Fig. 19. Because the quartz particles are sharp and angular, gouging abrasion was the primary mechanism of material removal (Fig. 20) at the wear ring on the edge of the grinding plate. The forces generated during gouging abrasion are higher than the yield strength of the cast iron, and these areas plastically deform under the compressive force of the quartz. The angle that the cutting edge of the particle makes with the wear surface is often referred to as the rake angle in cutting during machining operations. The plowing mechanism of material removal is illustrated in Fig. 21. In this case, a particle is oriented such that a “blunt” edge is presented to the grinding plate, and material is pushed to the sides of the particle as it passes through the lathe, piling up material at the periphery of the wear groove. In the machining wear mechanism, also illustrated in Fig. 21, particles are oriented so a “sharp” edge is presented to the grinding plate, and these particles cut into the plate and remove a chip from the surface. The chips are discontinuous, due to the flake graphite incorporated in the gray cast iron. The gouging, cutting, and plowing process continues until the thickness of the plate has been reduced to the point where no adjustments can be made to zero, the clearance between the stationary and moving plates.

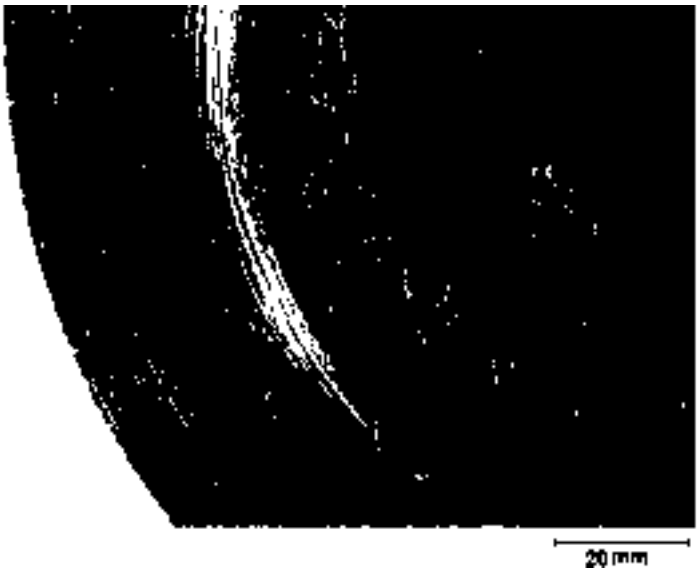


Fig. 19 An optical macrograph of a segment of the attrition mill wear plate showing gouging abrasion near the inner ring and grinding abrasion near the outer ring

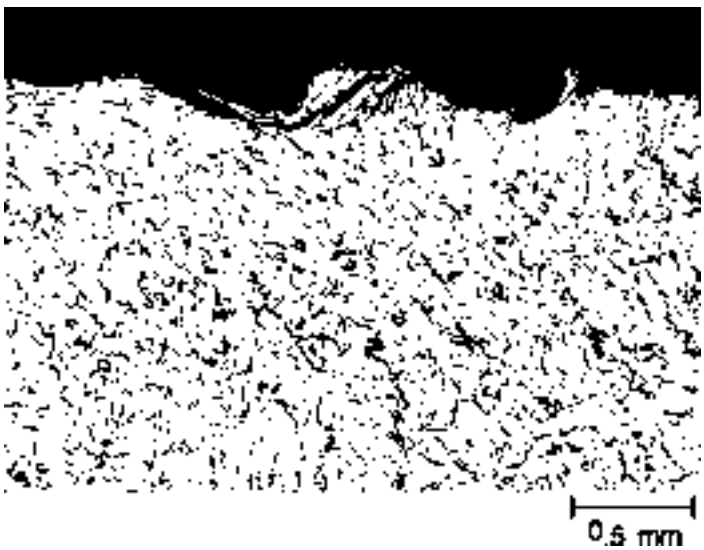


Fig. 20 Optical micrograph of a cross section at the plate ring showing gouging abrasion. Notice the deep depressions with deformed material and cracks.

Fig. 21 Optical micrograph of a cross section of the outer section of the attrition mill wear plate showing no gouging abrasion and little plastic deformation. Along the upper wear surface are shallow wear grooves from grinding abrasion.

From the failure analysis study on the wear plates, suggestions for improving the wear life can be made. Evidence from the wear patterns on the face of the grinding plates suggest that modifications to plate design may reduce overall wear and improve mill efficiency. The cutting flutes on the grinding plates show uniform gouging wear on both sides of the cutting edges, as illustrated in Fig. 19. This means that the cutting edge of the grinding plate, per se, is not involved in the actual shearing of the quartz particles but only in crushing them. The clearance between the flutes could, therefore, be reduced, providing more shearing action to the particles before they reach the inner wear ring.

Several material changes could also be made that might extend the life of the grinding plates. Lower cost is the main advantage of using gray cast iron for the wear material of the grinding plates. A negative consequence of using gray iron is that it has a high wear rate and can contaminate the product with the substantial iron debris as it wears. The ratio of iron to quartz is approximately 0.5. Because no indications of severe impact wear were noted, and because fracture and spalling were not observed, a harder plate material such as white cast iron or a work-hardening manganese steel could be used. Either material would lead to a larger value of the hardness ratio, possibly leading to lower wear. The manganese steel has a second attractive property: good high-temperature strength. High-temperature strength is important, because the temperature between the grinding plates during operation can reach 900 °C (1650 °F). These high temperatures can significantly reduce the strength and wear resistance of the gray cast iron grinding plates.

Example 4: Impact Wear of Disk Cutters.* Steel disks are used on large tunnel-boring machines to continuously cut away at rock faces. Although material is worn from the disks through indentation and scratching wear processes (Fig. 22), one of the predominant material removal mechanisms is material fracture near the edges of the disk cutting face (Fig. 23).

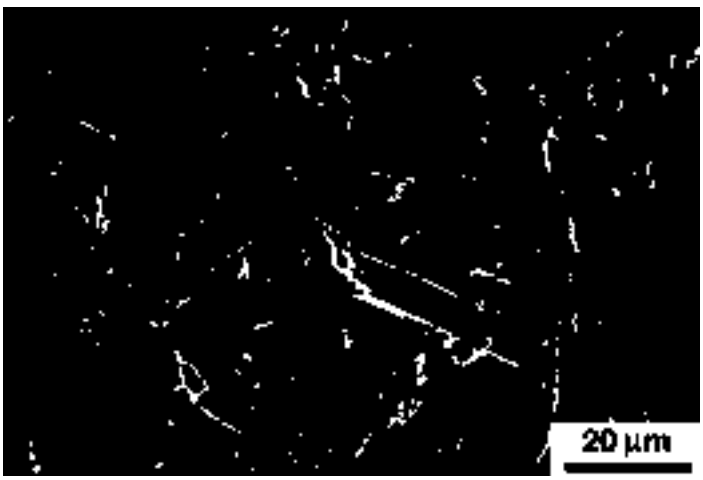


Fig. 22 Scanning electron micrograph showing the surface wear on the disk cutter, which shows elements of indentation/impact and scratching or gouging

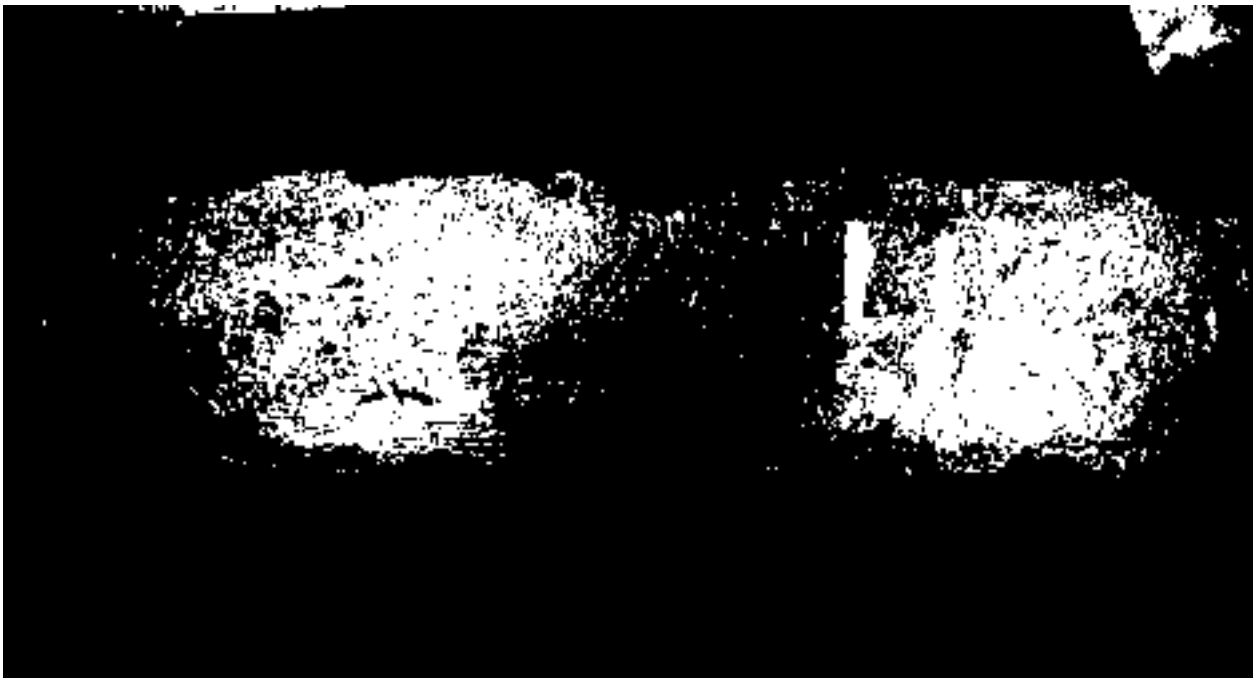


Fig. 23 Disk cutter, comprised of H13 tool steel, showing material fracture at edge of contact surface

Because of the dominance of the fracture process on the material removal rate, it was decided that a tougher material should be tried to provide increased fracture resistance. The original disk material was an H13 tool steel. The proposed replacement alloy was a medium-carbon steel, which had a lower hardness but greater fracture toughness and impact toughness.

Material	**Hardness, HRC**	**Fracture toughness, (K_{Ic})**		**Charpy impact energy at 24 °C (75 °F)**	
MPa $\sqrt{m}$	**ksi $\sqrt{in}$**	**J**	**ft · lbf**		
H13 tool steel	55–57	14–22[(a)]	13–20[(a)]	3–4	2–3
Medium-carbon steel	49–51	130	118	28–33	21–24

(a) Depending on direction of crack growth

Higher-toughness disks were placed on the head of a tunnel-boring machine, along with those made of H13. Despite having significantly less toughness than the medium-carbon alloy steel, the H13 disks exhibited what could be equated to 1.7 times longer life.

Examination of the cutting edge of the tougher steel disk revealed even larger fracture areas, similar in appearance to spalls (Fig. 24). As with many field-return worn components, further surface examination at higher magnifications revealed no useful information, due to continued wear and oxidation. Instead, nital-etched metallographic cross sections were prepared through the wear surfaces of disks made of each material. Figure 25 shows a cross section taken near the edge of the contact surface of an H13 disk. This image shows that plastic flow due to the high contact pressures has led to mushrooming at the edge of the disk cutting face. Cracking is subsequently initiating from these mushroomed regions. Figure 26 shows similar cracking in the lower-hardness alloy steel. The cracked zone is significantly larger and thus would lead to more extensive material removal, once crack growth is completed. Closer examination of these cracked regions revealed that cracks are typically following white-etching shear localization bands (Fig. 27). These bands likely form in a single impact event and are believed to be nano-crystalline ferrite, with all other elements trapped in solution (Ref 33, 34, 35). Hardness values within the band have been found to approach 1000 HV. Due to the brittleness of this layer, subsequent load cycles lead to cracking and spallation along the shear localization band. The

lower hardness and plastic flow stress of the medium-carbon alloy steel resulted in the formation of larger shear localization bands and thus, larger spalls with more rapid material removal.

Fig. 24 More severe material fracture/spalling on surface of disk made from tougher, medium-carbon alloy steel

Fig. 25 Micrograph showing the edge of the contact face in an H13 disk. Mushrooming is occurring due to contact pressures and resulting metal flow. Cracks resulting from fatigue or fracture at localized shear bands are leading to material removal

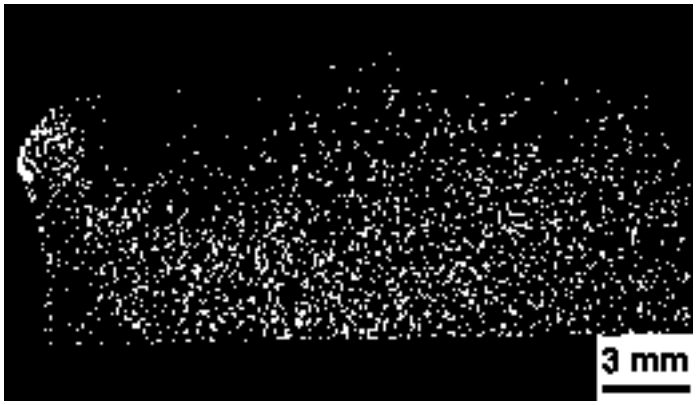


Fig. 26 More extensive cracking observed in tougher but softer medium-carbon steel

Fig. 27 Micrograph showing cracking along white-etching localized shear zone

Example 5: Identification of Abrasive Wear Modes in Martensitic Steels. Wear modes experienced by a given component can be quite diverse, depending on wear environment and component function. The abrasive wear modes experienced by steels in typical applications may be broadly divided into three categories: low-stress scratching, high-stress gouging, and impact or indentation. Scanning electron micrographs of these different wear modes are shown in Fig. 28. Wear rates of a given material depend strongly on wear mode and although a material may perform well in one mode of wear, it may not perform well in another.

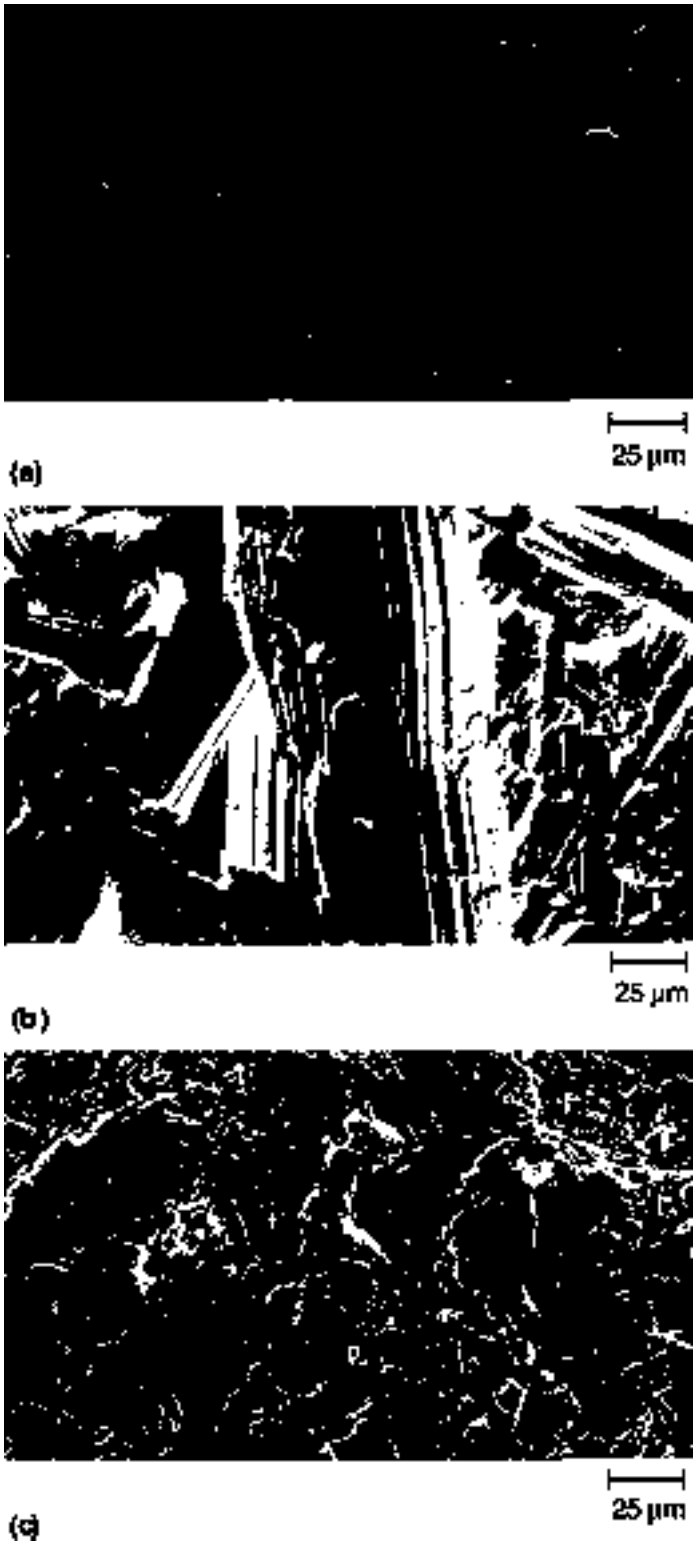


Fig. 28 Scanning electron micrographs showing the three modes of abrasive wear typically found in steels: (a) low-stress scratching, (b) higher-stress gouging, and (c) impact or indentation

For martensitic steels, wear rate is most strongly dependent on material hardness. Figure 29 shows approximate lines of normalized wear rate versus hardness for the different wear modes. Data is plotted against Vickers hardness rather than Rockwell C, because Vickers hardness directly relates to material yield strength, while the Rockwell C scale has no such direct correlation. Figure 29 shows that low-stress scratching has the strongest correlation with hardness. As the severity of wear increases, the dependence on hardness decreases. For gouging and impact wear, other material attributes, such as toughness and work-hardening ability, begin to play a larger roll in dictating wear rate. This difference in dependencies leads to materials that perform well in one wear mode but not in others. Figure 30 shows some materials that change the ranking when wear mode is changed. While the 50 HRC ground-engaging steel performs as expected, given its hardness in low-stress

scratching, as the severity of wear environment increases, it becomes the favored material. In contrast to the 50 HRC steel, the steel composite material performs very well in low-stress scratching (due to the presence of hard phases) but more poorly as the wear severity increased, due to its low material toughness.

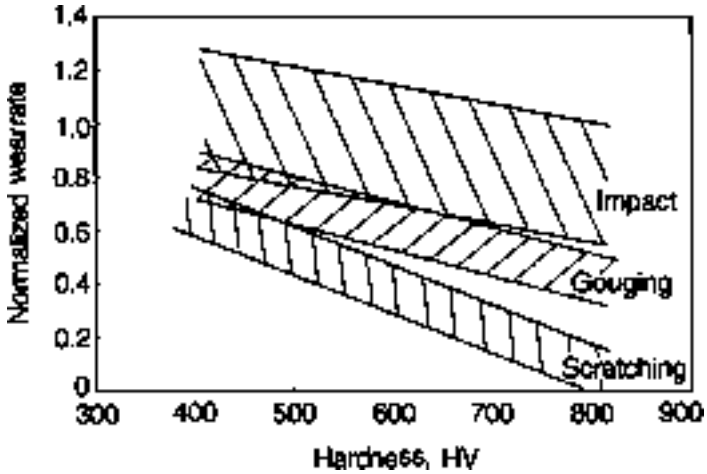


Fig. 29 Bands of normalized wear rate versus hardness for low-stress scratching, high-stress gouging, and impact wear. Low-stress scratching shows the strongest dependence on hardness, while impact abrasion shows the least. The scatter in the impact abrasion data suggests a growing contribution of material attributes other than hardness.

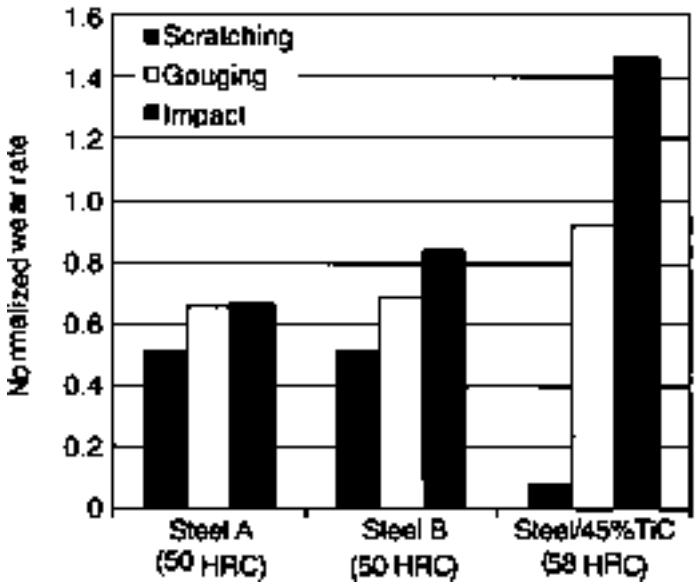


Fig. 30 Correlation of hardness with wear rate for three materials. The two 50 HRC materials both exhibit the same low-stress scratching wear resistance. However, as the wear severity increases, the steel designed for ground-engaging tools (steel A) exhibits moderate improvements in gouging wear and significant improvements in impact abrasion resistance. In contrast, due to the hard phases, the composite material performs better than would be expected, given its bulk hardness; however, due to its low fracture toughness, it performs significantly worse in more severe wear modes.

The dependency of the wear performance on a given material or wear mode emphasizes the need to know the wear mode being experienced by a component. Unfortunately, components returned from field service rarely reveal information as clearly as seen in Fig. 28. Such surface features are frequently damaged by corrosion. Therefore, another means of identifying wear modes is desired. One such method used for martensitic steels is to examine the polished and etched cross section of the wear surfaces. Figure 31 shows cross sections taken through the surface of components experiencing the three wear modes discussed. In low-stress scratching, there is essentially no subsurface microstructural modification of the steel. Even the few microns just below the surface show little or no evidence of plastic deformation (Fig. 31a). As the stress and severity of the abrasive

event increase, microstructural modification becomes more evident. This may be manifested as a deep, plastically deformed layer (evident in structural deformation) or in white-etching layers (Fig. 31b). The white-etching layers are believed to be extremely fine-grained ferrite, with all alloying elements in solution (Ref 33, 34, 35, 36, 37). In most cases, the formation of white-etching layers indicates a higher-stress abrasive event. However, it is also reflective of the material condition. Figure 32 shows two cross sections that were exposed to the same high-stress abrasive conditions. The material exhibiting white-etching layer formation is approximately 20 HRC softer than that with only slight surface plasticity. Impact abrasion is manifested in the appearance of severe plastic deformation at the surface and the presence of white-etching localized shear bands below the surface (Fig. 31c). The shear bands are created in a single impact event and are again the result of very high strains leading to what is believed to be the same structure as observed on the surface of components experiencing gouging wear (Ref 33, 34, 37). Identification of these wear modes may be used to help guide materials selection and processing to improve the wear resistance of components.

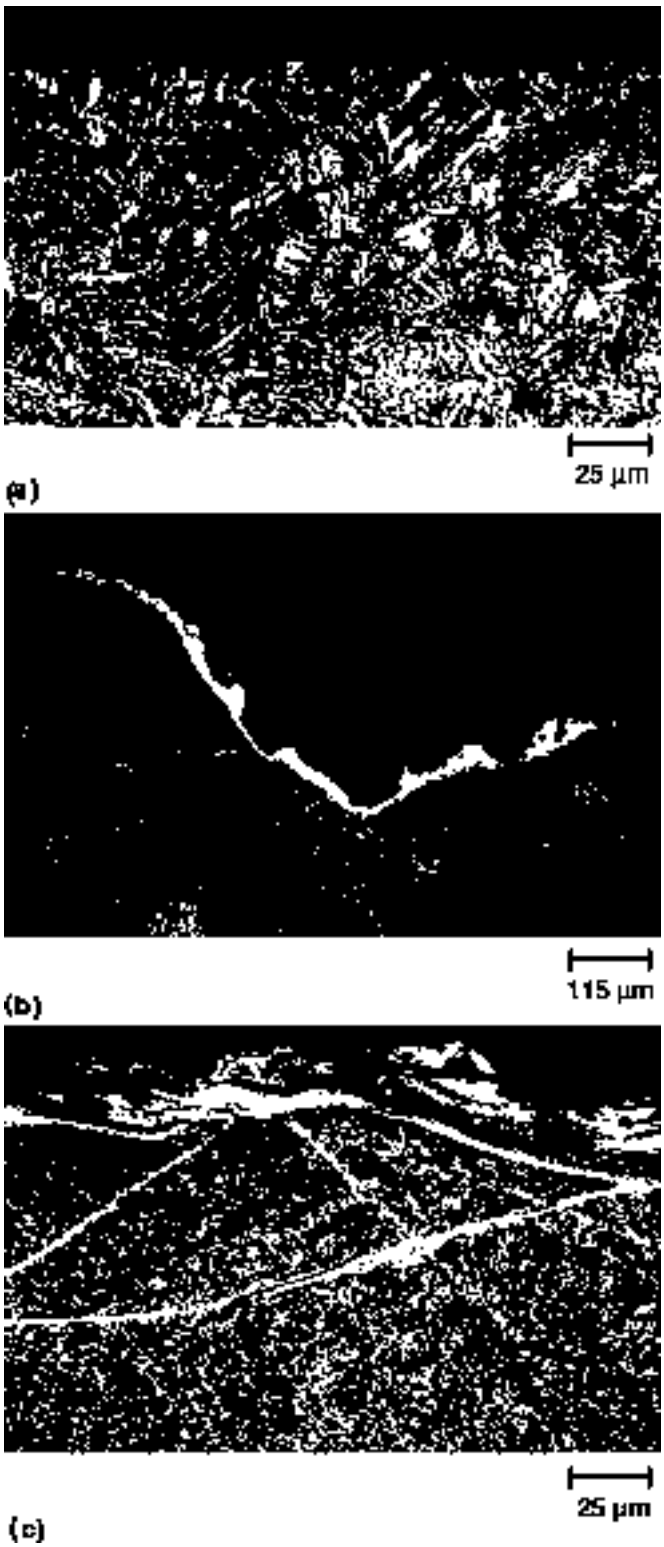


Fig. 31 Metallographic cross sections of the worn surface regions showing (a) no microstructural modification in low-stress scratching, (b) white-etching layers in high-stress gouging, and (c) subsurface white-etching shear bands resulting from impact abrasion

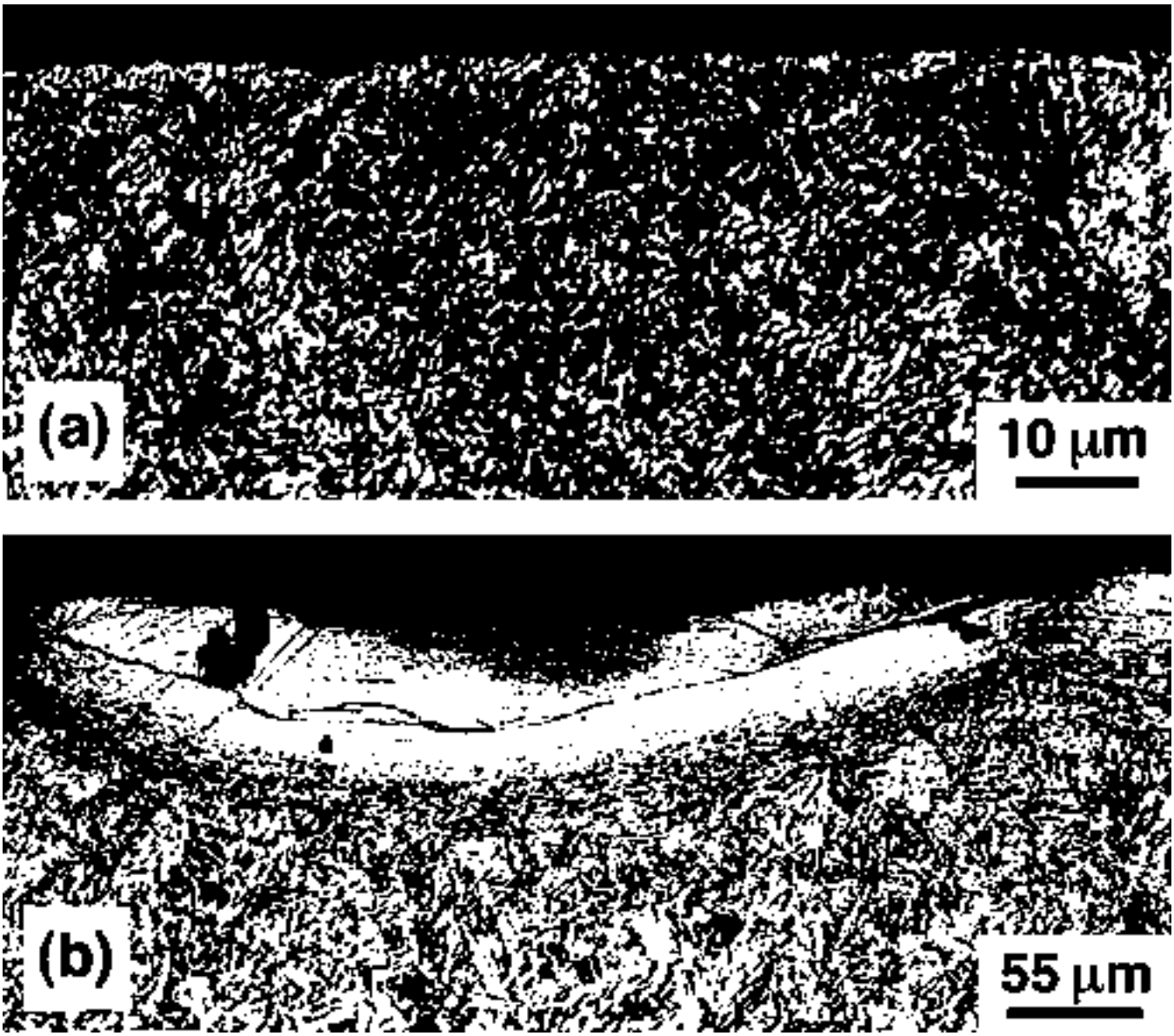


Fig. 32 Metallographic cross sections of wear surfaces of two materials that experienced the same high-stress abrasion. (a) At 60 HRC, the material exhibited no white-etching layers, while (b) at 20 HRC points, softer, significant gouging and white-etching layer formation is observed.

Footnote

* *The authors would like to recognize Donald Sherman of Caterpillar Inc. for his original work in this investigation.

References cited in this section

30. F. Borik and D.L. Sponseller, Gouging Abrasion Test for Materials Used in Ore and Rock Crushing, Part 1: Description of the Test, *J. Mater.*, Vol 6, 1971, p 576–589

31. S. Kalpakjian and S.R. Schmid, *Manufacturing Engineering and Technology*, 4th ed., Prentice-Hall, Inc., 2001, p 627–635

32. S.D. El Wakil, *Process and Design for Manufacturing,* 2nd ed., PWS Publishing Company, 1998, p 349–353

33. B. Zhang, W. Shen, Y. Liu, X. Tang, and Y. Wang, Microstructures of Surface White Layer and Internal White Adiabatic Shear Band, *Wear,* Vol 211, 1997, p 164–168

34. Y.-Y. Yang, H.-S. Fang, Y.-K. Zheng, Z.-G. Yang, and Z.-L. Jiang, The Failure Models Induced by White Layers During Impact Wear, *Wear,* Vol 185, 1995, p 17–22

35. Y.-K. Zheng and Y.-Y. Yang, Wear Characteristics of Large Grinding Balls in an SAG Mill, *Tribol. Trans.,* Vol 40, 1997, p 720–724

36. L. Xu, S. Clough, P. Howard, and D. St. John, Laboratory Assessment of the Effect of White Layers on Wear Resistance for Digger Teeth, *Wear,* Vol 181–183, 1995, p 112–117

37. R. Bulpett, T.S. Eyre, and B. Ralph, The Characterization of White Layers Formed on Digger Teeth, *Wear,* Vol 162–164, 1993, p 1059–1065

Abrasive Wear Failures

Jeffrey A. Hawk and Rick D. Wilson, Albany Research Center; Daniel R. Danks, Danks Tribological Services; Matthew T. Kiser, Caterpillar Inc.

References

1. E. Rabinowicz, *Friction and Wear of Materials,* 2nd ed., John Wiley & Sons, Inc., 1995, p 5–7

2. V.J. Colangelo and F.A. Heiser, *Analysis of Metallurgical Failures,* 2nd ed., John Wiley & Sons, Inc., 1987, p 1–10

3. T.F.J. Quinn, The Classifications, Laws, Mechanisms and Theories of Wear, *Fundamentals of Tribology,* N.P. Suh and N. Saka, Ed., The MIT Press, 1980, p 477–492

4. M.B. Peterson, Classification of Wear Processes, *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Ed., American Society of Mechanical Engineers, 1980, p 9–15

5. H. Czichos, Systems Approach to Wear Problems, *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Ed., American Society of Mechanical Engineers, 1980, p 17–34

6. J.D. Gates and G.J. Gore, Wear of Metals: Philosophies and Practicalities, *Mater. Forum,* Vol 19, 1995, p 53–89

7. K.-H. Zum Gahr, *Microstructure and Wear of Materials,* Elsevier Science Publishers, 1987

8. H.S. Avery, The Measurement of Wear Resistance, *Wear,* Vol 4, 1961, p 427–449

9. T.E. Norman, Wear in Ore Processing Machinery, *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Ed., American Society of Mechanical Engineers, 1980, p 1009–1051

10. P.J. Mutton, *Abrasion Resistant Materials for the Australian Minerals Industry,* Vol 1, Australian Minerals Industries Research Association Limited, 1988

11. M.A. Moore, Abrasive Wear, *Fundamentals of Friction and Wear of Materials,* D.A. Rigney, Ed., American Society for Metals, 1981, p 73–118

12. M.J. Murray, P.J. Mutton, and J.D. Watson, Abrasive Wear Mechanisms in Steels, *Wear of Materials,* W.A. Glaeser, K.C. Ludema, and S.K. Rhee, Ed., American Society of Mechanical Engineers, 1979, p 257–265

13. D.R. Danks and P. Clayton, Comparison of the Wear Process for Eutectoid Rail Steels: Field and Laboratory Tests, *Wear,* Vol 120, 1987, p 233–250

14. R.C.D. Richardson, The Maximum Hardness of Strained Surfaces and the Abrasive Wear of Metals and Alloys, *Wear,* Vol 10, 1967, p 353–382

15. R.C.D. Richardson, The Wear of Metals by Hard Abrasives, *Wear,* Vol 10, 1967, p 291–309

16. M.M. Khruschov, Resistance of Metals to Wear by Abrasion as Related to Hardness, *J. Mech. Eng.,* 1957, p 655–659

17. M.M. Khruschov and M.A. Babichev, Issledovaniya iznashivaniya metallov, *Izdat. Akad. Nauk SSSR,* 1960, p 889–893: Research on Wear of Metals, N.E.L. translation

18. M.A. Moore, The Relationship Between the Abrasive Wear Resistance, Hardness and Microstructure of Ferritic Materials, *Wear,* Vol 28, 1974, p 59–68

19. R.C.D. Richardson, The Wear of Metals by Relatively Soft Abrasives, *Wear,* Vol 11, 1968, p 245–275

20. J.F. Archard, Contact and Rubbing of Flat Surfaces, *J. Appl. Phys.,* Vol 24, 1953, p 981–988

21. A. Misra and I. Finnie, On the Size Effect in Abrasive and Erosive Wear, *Wear,* Vol 65, 1981, p 359–373

22. I.M. Hutchings, *Tribology: Friction and Wear of Engineering Materials,* CRC Press, 1992

23. T.H. Kosel, Solid Particle Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* P.J. Blau, Ed., ASM International, 1992, p 199–213

24. A.R. Lansdown and A.L. Price, *Materials to Resist Wear, A Guide to Their Selection and Use,* Pergamon Press, 1986

25. K.-H. Zum Gahr, Relation Between Abrasive Wear Rate and the Fracture Toughness of Metallic Materials, *Z. Metallkd.,* Vol 69, 1978, p 643–650

26. K.-H. Zum Gahr and D.V. Doane, Optimizing Fracture Toughness and Abrasion Resistance in White Cast Irons, *Metall. Trans. A,* Vol 11, 1980, p 613–620

27. H. Czichos, *Tribology,* Elsevier Science Publishing Co., 1978, p 14–23

28. R.C. Tucker, Wear Failures, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., American Society for Metals, 1986, p 145–162

29. K.C. Ludema, *Friction, Wear, Lubrication: A Textbook in Tribology,* CRC Press, 1996, p 205–211

30. F. Borik and D.L. Sponseller, Gouging Abrasion Test for Materials Used in Ore and Rock Crushing, Part 1: Description of the Test, *J. Mater.,* Vol 6, 1971, p 576–589

31. S. Kalpakjian and S.R. Schmid, *Manufacturing Engineering and Technology*, 4th ed., Prentice-Hall, Inc., 2001, p 627–635

32. S.D. El Wakil, *Process and Design for Manufacturing*, 2nd ed., PWS Publishing Company, 1998, p 349–353

33. B. Zhang, W. Shen, Y. Liu, X. Tang, and Y. Wang, Microstructures of Surface White Layer and Internal White Adiabatic Shear Band, *Wear*, Vol 211, 1997, p 164–168

34. Y.-Y. Yang, H.-S. Fang, Y.-K. Zheng, Z.-G. Yang, and Z.-L. Jiang, The Failure Models Induced by White Layers During Impact Wear, *Wear*, Vol 185, 1995, p 17–22

35. Y.-K. Zheng and Y.-Y. Yang, Wear Characteristics of Large Grinding Balls in an SAG Mill, *Tribol. Trans.*, Vol 40, 1997, p 720–724

36. L. Xu, S. Clough, P. Howard, and D. St. John, Laboratory Assessment of the Effect of White Layers on Wear Resistance for Digger Teeth, *Wear*, Vol 181–183, 1995, p 112–117

37. R. Bulpett, T.S. Eyre, and B. Ralph, The Characterization of White Layers Formed on Digger Teeth, *Wear*, Vol 162–164, 1993, p 1059–1065

Fretting Wear Failures

Introduction

FRETTING is a wear phenomenon that occurs between two mating surfaces; initially, it is adhesive in nature, and vibration or small-amplitude oscillation is an essential causative factor. Fretting frequently is accompanied by corrosion. Fretting of ferrous materials in air produces a characteristic reddish-brown debris of ferric oxide, which, when mixed with oil or grease, produces debris that is often referred to as “blood,” “cocoa,” or “red mud.” In components that are lubricated so that ordinary corrosion is not likely to occur, the presence of reddish brown debris is indicative of fretting. If a component is not lubricated, the presence of oxide powder may not necessarily signify fretting but, rather, wear. Fretting also occurs in nonoxidizing materials, such as gold, platinum, and cupric oxide. The debris formed by many nonferrous metals is largely unoxidized and is larger in particle size than ferrous materials are. If fretting occurs in an inert or protective atmosphere, little debris is produced.

The general characteristics of fretting wear, with an emphasis on steel, are reviewed in this article. This review is followed by several examples of failures related to fretting wear. Common sites for fretting are in joints that are bolted, keyed, pinned, press fitted, or riveted; in oscillating bearings, splines, couplings, clutches, spindles, and seals; in press fits on shafts; and in universal joints, base plates, shackles, and orthopedic implants. Generally, fretting occurs at contacting surfaces that are intended to be fixed in relation to each other but that actually undergo minute alternating relative motion that is usually produced by vibration. There are exceptions, however, such as contact between balls and raceways in bearings and between mating surfaces in oscillating bearings and flexible couplings.

Fretting Wear

In general, fretting occurs between two tightly fitting surfaces that are subjected to a cyclic, relative motion of extremely small amplitude. For oxidizing metals, one of the immediate consequences of the process in normal atmospheric conditions is the production of oxide debris, hence the term *fretting wear,* or *fretting corrosion,* is applied to the phenomenon. The movement is usually the result of external vibration, but in many cases, it is the consequence of one of the members of the contact being subjected to a cyclic stress (that is, fatigue), which gives rise to another and usually more damaging aspect of fretting, namely the early initiation of fatigue cracks. This characteristic is termed *fretting fatigue,* or *contact fatigue.* Fatigue cracks also may initiate where the contacting surfaces are under a very heavy normal load or where there is a static tensile stress in one of the surfaces. There are cases where the movement is not simply tangential but is complicated by the normal force also oscillating to the extent that the surfaces lose contact in each cycle. This leads to a hammering effect, which is termed *impact fatigue.* In this case, the phase relationship between the two motions can be an important factor.

Other terms that have been used to designate fretting-type damage are *false Brinelling,* which is applied particularly to bearings where the craters caused by the vibration of the ball against the race are circular and resemble Brinell impressions, and “fitting rust,” where gages or shims are clamped together and experience vibration.

The terms *fretting* and *fretting corrosion* are widely accepted now in most languages, but the term *reiboxydation* still is used in Germany, and *corrosion de trepidation* is used occasionally in France.

Fretting was first reported by Eden, Rose, and Cunningham in 1911 (Ref 1), who found that brown oxide debris was formed in the steel grips of their fatigue machine in contact with a steel specimen. It was not until 1927 that the first investigation of the process was conducted and two machines were designed to produce small-

amplitude rotational movement between two annuli in the first case, and an annulus and a flat in the second case (Ref 2). The movement was controlled by a long lever system. Because the resultant debris on steel specimens was the red iron oxide α-Fe_2O_3, which had risen from a chemical reaction with oxygen in the air, Tomlinson (Ref 2) coined the phrase *fretting corrosion.* The investigation also established that the damage could be caused by movements with amplitudes as small as a few millionths of an inch (~125 nm, or 1250 Å) and the important fact that relative movement, which was termed "slip," had to occur.

The next important stage in the development of fretting studies was made in 1941 when fretting damage was produced on the gage length of steel fatigue specimens and a subsequent reduction in fatigue strength was found due to the pitting of the surface of between 13 and 17% (Ref 3). This result was to be expected, but later investigations, particularly that presented in Ref 4, showed that the conjoint action of fretting and fatigue, which is the usual case in practice, was much more dangerous, producing strength-reduction factors (SRFs) of 2 to 5 and even greater. In 1958, studies demonstrated that fretting accelerated the crack initiation process greatly. In normal fatigue, crack initiation may account for 90% of fatigue life, whereas in fretting fatigue, initiation could occur in 55% of the fatigue life.

Fretting Wear in Mechanical Components

Because vibration is one of the main causes of the fretting movement, it follows that the most likely area for it to occur is in machinery. The contacts between hubs, shrink and press fits, and bearing housings on loaded rotating shafts or axles are particularly prone to fretting damage, but because the movement arises from alternating stresses in the shaft surface, the problem is more one of fatigue than wear (Ref 6). However, wear rather than fatigue can be a problem in bearing housings. Thin-shell bearings are used universally in diesel engines, and such bearings involve an interference fit between the bearing and the housing. If the contact pressure is not high enough, movement can occur, giving rise to the fretting damage shown in Fig. 1 (Ref 7).

Fig. 1 Fretting damage incurred in a bearing shell component because of insufficient contact pressure between the shell and housing of a bearing. Source: Ref 7

Flexible couplings and splines, particularly where they form a connection between two shafts and are designed to accommodate some slight misalignment, can suffer severe fretting wear. Examples of such cases are described in Ref 8.

Jointed structures are another source of fretting problems. It has been reported (Ref 9) that there is no such thing as a static joint on an aircraft. An aircraft structure is largely an assemblage of aluminum and steel components riveted together or joined by fasteners. Even in a simple riveted joint, there are at least three possible sites where fretting can occur (Fig. 2):

- Between the riveted panels
- Between the underside of the rivet head and the panel
- Between the shank of the rivet and the rivet holes

Typical fretting damage is shown in Fig. 3.

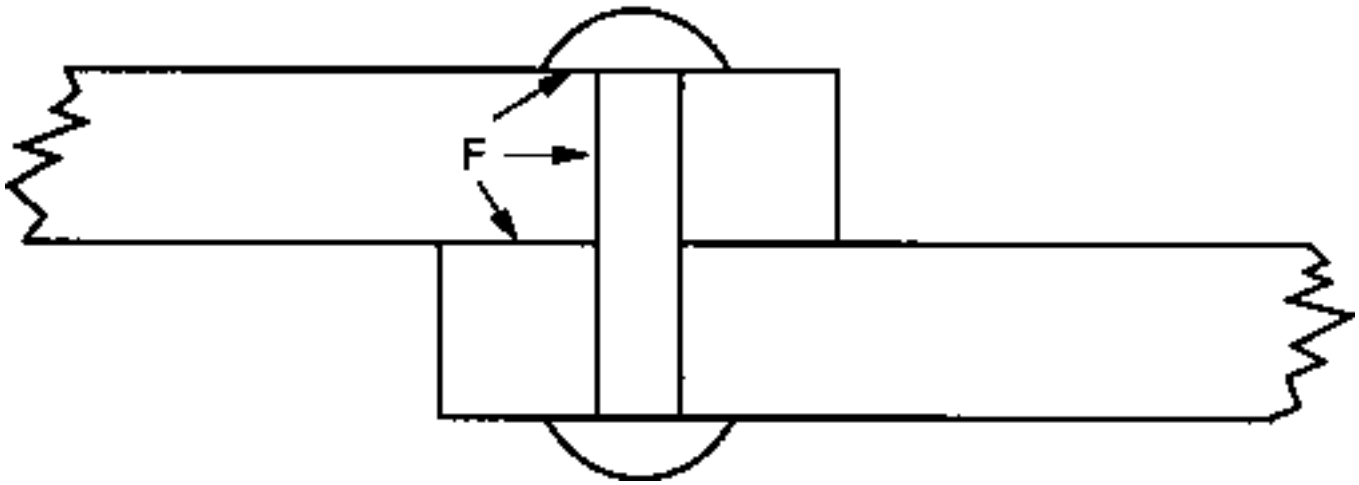


Fig. 2 Locations (indicated by F) prone to fretting in an aluminum-steel riveted joint

Fig. 3 Fatigue crack generated by fretting damage at a rivet hole

Turbines, both steam and gas, operate under conditions that subject these components to fretting damage. There are three main sites:

- The point at which the turbine disk is either shrink fitted onto the driving shaft or attached to the shaft by means of a bolted flange
- The position where the blades are fixed into the disk by either a dovetail or fir-tree type fixture
- The place where snubbers at the outer ends of the blades contact those on adjacent blades

The first two sites are more likely to be areas for fatigue crack initiation, but loss of material by wear of the snubbers will give rise to increased vibration of the blades and the onset of impact fretting. In some turbines, a wire is threaded through holes in the outer ends of the blades to dampen down vibration. In this case, the frictional energy dissipated in the fretting process is the effective agent and may be considered one of the few cases where fretting is beneficial.

The rotors of electrical generators have slots along their lengths in which the windings are located. These are held in place by wedges. Fretting can occur between the wedges and the undercut portion of the winding channel with the development of fatigue cracks in the rotor. These problems in turbines and generators are discussed in Ref 10.

Steel ropes are used widely as mooring ropes, haulage ropes, mining ropes, and on cable cars. Overhead power lines are in effect ropes made up of aluminum wires wound on a steel support rope; more recently, they have been made up entirely of aluminum (the all-aluminum conductor). Any bending movement or alternating tension can result in local movement in the numerous interwire contacts. Most ropes include some form of internal lubrication to lessen the effects of such movement. Again, one of the main dangers is fatigue failure of a wire, which increases the load on the remaining wires. However, in the locked coil rope that is designed with shaped wires to give an almost solid outer structure (Fig. 4), fretting debris with a greater volume than that of the metal from which it was formed can lead to the forcing apart of the wires, with the escape of lubricant and the ingress of a corrosive atmosphere. The accumulation of debris also leads to a reduction in flexibility at this point, and, in the case of a mining rope, can cause jolting when this part of the cable passes on to the winding drum.

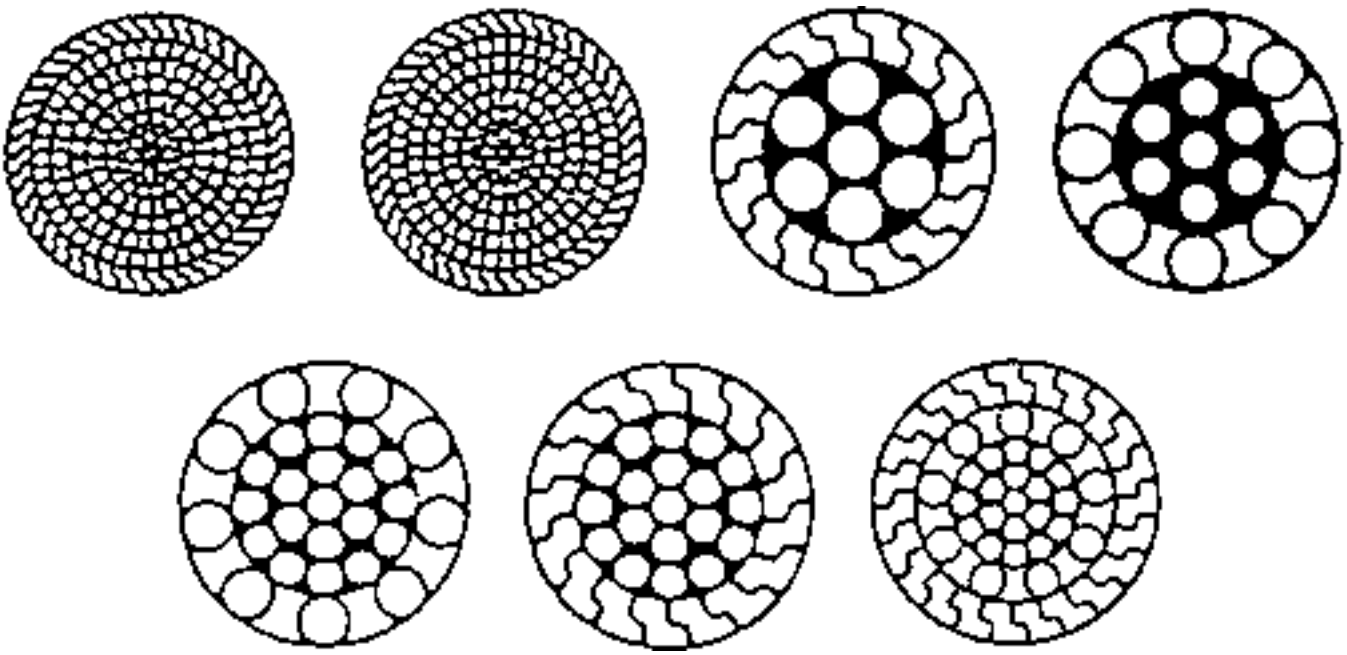


Fig. 4 Cross sections of typical locked-coil track strand wire rope construction.

In steam generators and heat exchangers, flow-induced vibration results in fretting between the tubes and the supports, or baffles through which they pass. In the past, this phenomenon appears to have been a particular problem in the nuclear power industry (Ref 11, 12) and chemical processing plants (Ref 13). A large contribution to the fretting in these industries is from impact fretting, which is discussed in the section “Impact Fretting” of this article.

Another example where movement is an integral part of the modus operandi of the device is the leaf spring used at one time in automobiles and railway rolling stock. It consists of an assemblage of steel plates, whose purpose is to dampen down vibration transmitted from the wheels to the carriage. Accumulation of debris could result in the plates not moving easily over one another and a stiffening of the spring.

The accumulation of debris is also the main problem in the fretting of electrical contacts (Ref 14). Because the oxide debris is nonconducting, interference or distortion of electrical signals can be generated in telephone equipment and also in computer connections. In some countries, aluminum conductors are used in domestic electrical circuits. Here, the movement is due to temperature variations and is, therefore, of extremely low-cycle frequency. However, the buildup of aluminum oxide in connections can cause the formation of electrical arcs, which in extreme instances has led to houses being burned down.

The transportation of goods by road or rail obviously means that these items are subjected to vibration. This fact gave rise to the classic case of fretting in the bearings of automobiles being transported by rail from Detroit to the West Coast of the United States. As the bearings were stationary, fretting occurred between the loaded balls and the race, giving rise to the false Brinelling mentioned previously. This has also been a problem with the transport of stacks of aluminum sheets with highly polished finishes. It also led to the abandonment of using stacked aluminum trays to transport food stuffs. Another example involved the transport of live ammunition in aluminum boxes, because the fretting debris was found to be pyrophoric. In a case investigated by the author, high-purity (HP) aluminum “logs” (cylinders 250 mm, or 10 in., in diameter) were transported by rail to the plant where they were to be extruded as buss bars. Black marks (that is, fretting debris) were noticed on the cylinders (Fig. 5) that gave rise to serious surface tears on extrusion (Fig. 6).

Fig. 5 Typical fretting marks on a high-purity aluminum cylinder

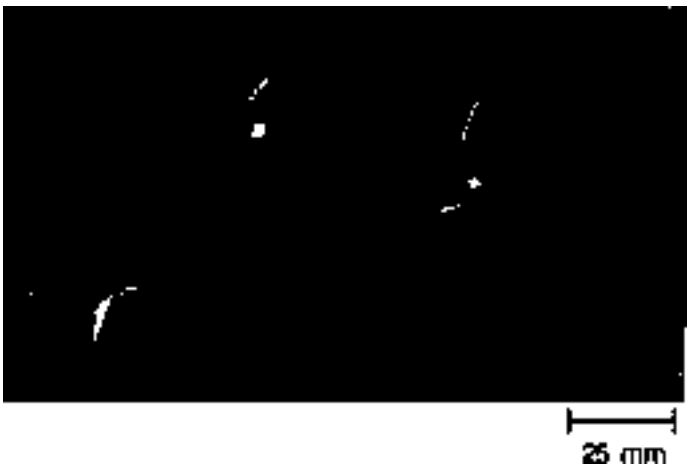


Fig. 6 Surface tears caused by fretting damage when aluminum cylinders in Fig. 5 are extruded

Fretting can occur also in certain orthopedic devices, particularly fracture fixation devices such as bone plates. These devices are used to hold together the parts of fractured bone. They consist of plates of corrosion-resistant materials that are attached to the bone by means of screws of the same material. Fretting can occur between the underside of the screw head and the countersink of the hole in the plate. Fretting in this case results in disruption of otherwise protective films and the continual release of heavy metal ions into the body fluids, with possible toxic reactions. This aspect of fretting in aqueous electrolytes is discussed in detail subsequently. Fatigue failure is also possible (Fig. 7).

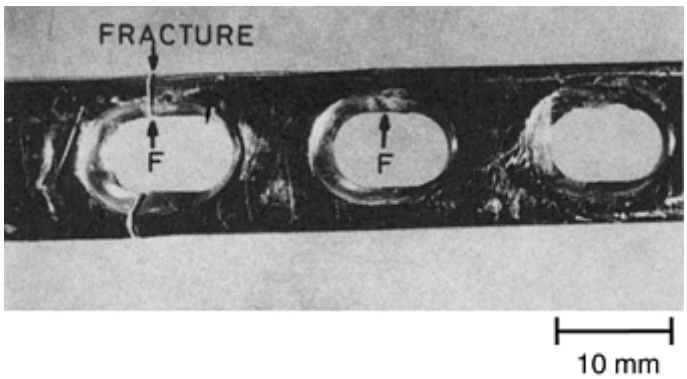


Fig. 7 Fracture in an orthopedic bone plate. A failure was caused by fretting damage (loss of protective oxide layer) in the countersunk portion of the plate.

The possible situations in which fretting can occur are legion, and it is possible in this review to give only some typical examples. In most of the cases described, suitable methods of overcoming the problem have been forthcoming as the result of detailed investigation; these studies are described in detail in the section "Prevention of Fretting Damage" in this article.

At one time, the author was approached by a professor of dentistry from Paris, who was convinced that the spots of damage (caries) that develop between contacting teeth, particularly the molars, were a form of fretting corrosion due to the relative movement occurring during chewing and swallowing, and certainly the examples he produced looked very much like it. This theory extends the possibility of fretting damage to materials other than metals. With increased use of polymers, both simple and reinforced, in addition to ceramics, this is an area where further examples of a somewhat different kind are likely to be encountered. Some research already is being devoted to the topic and is described in the section "Parameters Affecting Fretting" in this article.

Parameters Affecting Fretting

Experimental investigations have concentrated on the effect of specific physical variables (such as amplitude of slip, normal load, frequency of vibration, and the circumstances of the fretting situation, such as type of contact, mode of vibration, and the condition of the surfaces) to indicate how the problem can be overcome in the future design of contacting components. These results are discussed in detail next.

Amplitude of Slip. It established that relative movement is essential for fretting to occur and that extremely small movements of the order of a few nanometers were capable of causing damage (Ref 2). This aspect of fretting—the lower limit of slip—has more recently been the subject of investigation (Ref 15, 16). One of the problems in experimental work has been the control and measurement of small amplitudes of movement. The most satisfactory method that avoids possible losses due to elastic deformation in the supports or actuators is to concentrate on the situation of partial slip. This method is studied best with a contact consisting of a ball (typically, a bearing ball) on a flat, a situation that was analyzed by Mindlin. Figure 8 shows the form of contact of a ball on a flat developed by Mindlin from the Hertzian solution for a stationary contact. On the application of a tangential force or a torsional force, slip occurs over an outer annular region of the circle of contact. If the applied force is oscillating, then fretting occurs in the slip region. This method was used to verify the Mindlin analysis (Ref 17). Later, it was demonstrated that the stress distributions are somewhat distorted, but the basic idea remains the same. The center of the contact region (that is, the nonslip region) remains locked. The

amplitude of movement in the slip region for torsional excitation presumably rises from zero at the inner boundary to some greater value at the outer edge. The dimensions of the damage area closely approach these limits.

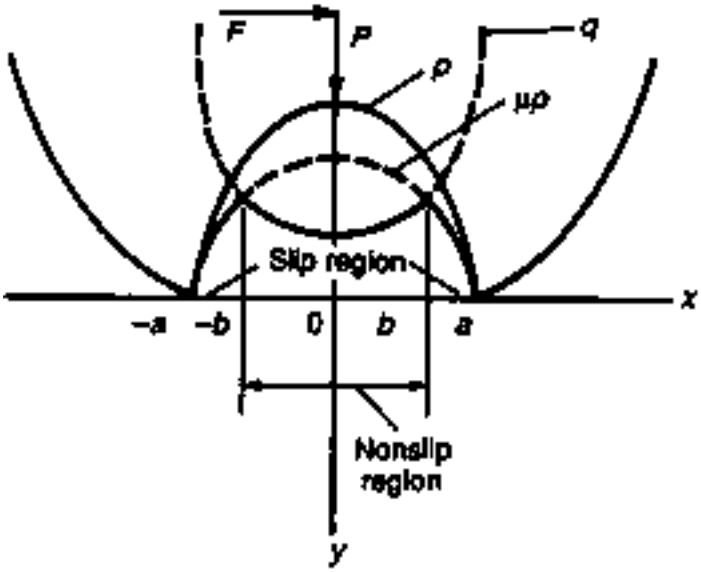


Fig. 8 Mindlin analysis of partial slip for a ball in contact with a flat and subjected to a tangential force

In some investigations (Ref 15, 16), the amplitude of slip at which damage was detected on steel specimens under a load of 88 N (8.8 kgf) was 0.06 μm (2.4 μin.). A narrow annulus of damage was detected at an amplitude of 0.1 μm (4 μin.). As the amplitude was increased, the damage area, characterized by the appearance of the typical reddish-brown oxide, increased in width, with the outer radius remaining constant. This situation has been investigated using a linear vibratory motion (Ref 19), where it was found that the annulus of damage for a fixed amplitude increased in width with the number of fretting cycles, but the outer radius remained constant (Fig. 9).

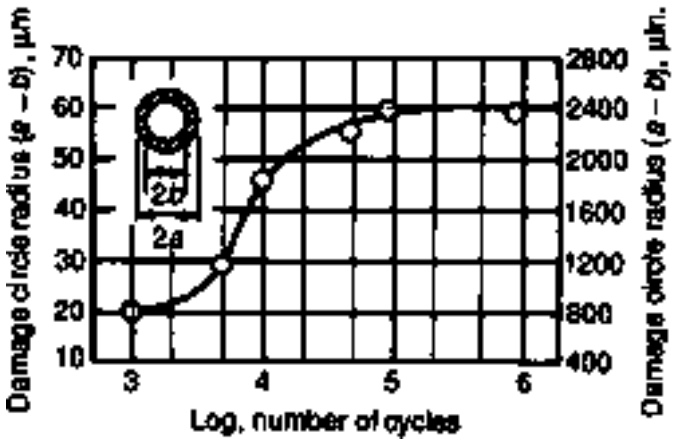


Fig. 9 Relation between diameter of fretting scar and the number of cycles for a ball on a flat under conditions of total slip. *P*, 437 N (43.7 kgf); P_{max}, 1.2 GPa (175 ksi); amplitude of movement, 8 μm (320 μin.); frequency, 50 Hz

In some experiments (Ref 15, 16), the damage was detectable by profilometry of the flat surface at an amplitude of 0.53 μm (21 μin.) and was said to be severe at 2.8 μm (110 μin.). The level of slip amplitude at which slip occurs over all of the contact area depends on such factors as the normal load and the mechanical properties of the materials. In heavier loads used in other experiments (Ref 20), which were sufficient to produce a Brinell impression, the slip annulus was extremely narrow, and it is doubtful whether total slip could ever be achieved under these conditions (Fig. 10). Under more normal conditions (amplitude 8 to 25 μm, or 320 to 1000 μin.; normal load 135 N, or 13.8 kfg), it was found that total slip occurred at an amplitude of 25 μm (1000 μin.) (Ref 19). At this amplitude and higher, the radius of the damage circle increased linearly with the number of cycles.

Fig. 10 Annulus of fretting damage produced by contact of a ball on a flat under heavy normal load

The amplitude of slip where the slip becomes total has been termed the *critical amplitude* Δ. In another study, a crossed-cylinder arrangement was used with steel specimens (Ref 21), and the relation between the critical amplitude and normal load was mapped (Fig. 11). These results are somewhat different from those reported in Ref 15 and 16, where the experimental values agree well with calculations based on elastic theory (Fig. 12). The discrepancy is attributed to plastic deformation occurring in the crossed-cylinder experiments (Ref 22).

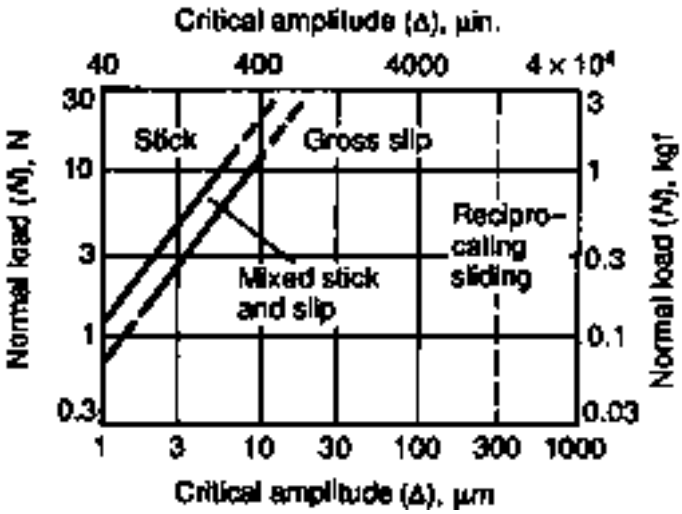


Fig. 11 Plot of normal load versus critical amplitude as a function of normal slip for a crossed steel cylinder arrangement. Source: Ref 21

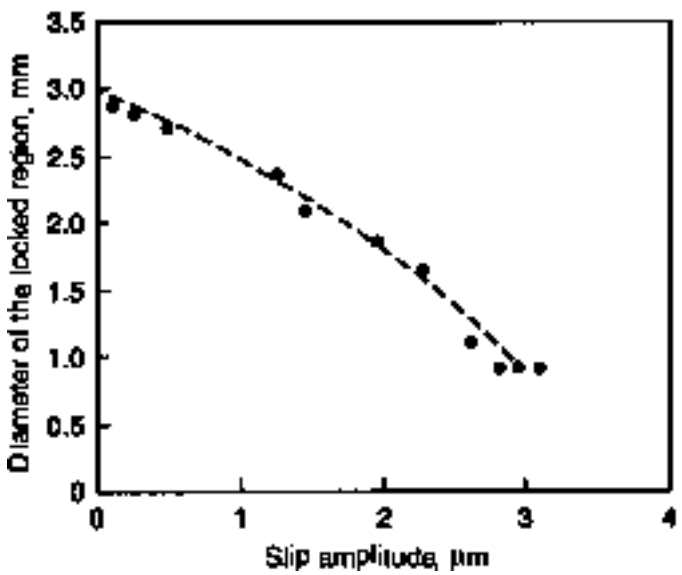


Fig. 12 Comparison of the theoretically predicted diameter of the nonslip region versus slip amplitude based on experimental data

In other investigations, using a variety of different contact geometries and different methods of assessing the wear rate and plotting the results as the specific wear rate (that is, the loss of material per unit of sliding distance per unit of applied normal load), the curves are of the form shown in Fig. 13. The curves are sigmoidal, with the specific wear rate becoming constant at amplitudes >100 μm (>0.004 in.) (that is, identical with unidirectional sliding or reciprocating sliding wear rates), providing a possible upper limit for the slip amplitude for the case of true fretting.

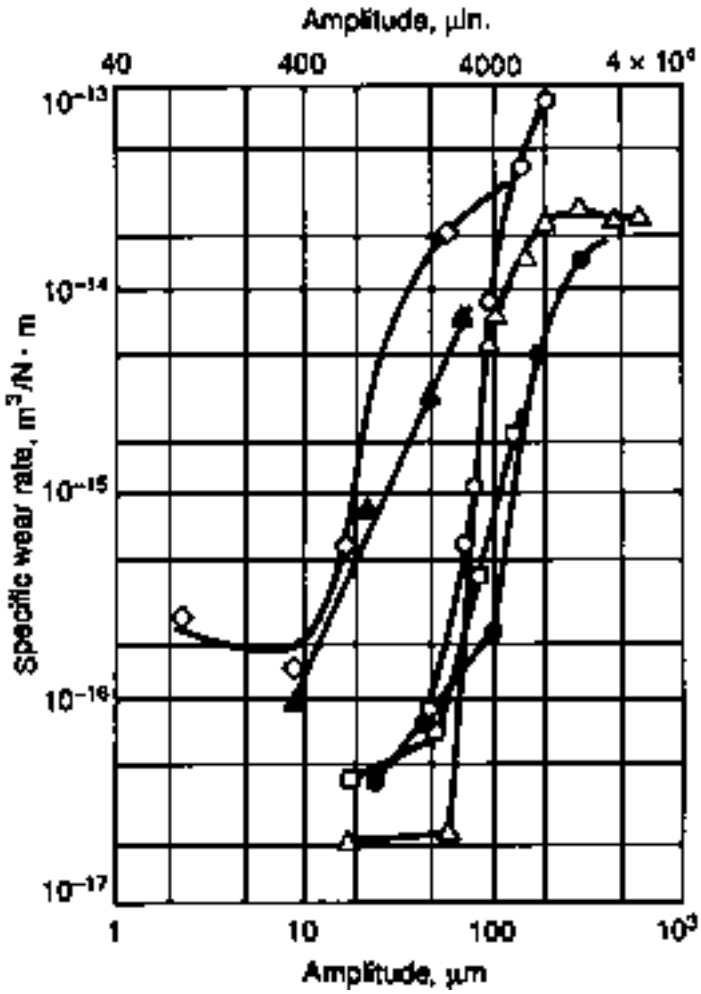


Fig. 13 Plot of specific wear rate versus amplitude of slip. Each core is the result of a separate investigation

The small-amplitude characteristic of fretting is the basis for many of the important features of the process. It means that the relative velocities of the two surfaces are much lower, even at high frequencies, compared with conditions in unidirectional sliding. At a frequency of 25 Hz and an amplitude of 25 μm (1000 μin.), the average velocity is 1.25 mm/s (0.0492 in./s), whereas sliding speeds in bearings are of the order of 1 m · s. Debris tends to stay where it is generated, so that high spots (for example, machining marks) become imprinted on the other surface if it has a higher degree of finish (Fig. 14). The debris may never become exposed to the atmosphere and can become compacted under the compressive load, often changing its appearance, as discussed in the section “Effect of Debris on Wear” in this article.

Fig. 14 Marks produced by a machined surface fretting against a polished flat

Normal Load. When two surfaces are placed in contact and loaded, plastic deformation of contacting high spots (asperities) occurs (Ref 23). However, certain contacts (for example, sphere on flat and cylinder on flat), where the contact area tends to be extremely small, still conform to the elastic theory of Hertz because the overall contact is determined by the elastic stress field developed in the hinterland, or bulk, of the material. In most practical cases of contact between conforming surfaces, plastic deformation is to be expected at contacting high points. The real area of contact, A, is therefore directly proportional to the applied load, P:

$$A = \frac{P}{p_o} \qquad \text{(Eq 1)}$$

where p_0 is the yield pressure and approximately equal to $3Y$, and where Y is the yield stress in tension. Because fretting occurs only at contacting areas, it follows that the amount of wear should be directly proportional to the applied load. This generally is found to be the case if the amplitude of slip is kept constant. In the early stages, there is a running-in period in which metal-to-metal contact predominates (Fig. 15). This contact can be detected by measurement of contact resistance, which remains low, and depends for a given number of cycles on the magnitude of the load (Fig. 16). These observations are discussed in greater detail in the section “Mechanism of Fretting Wear” in this article.

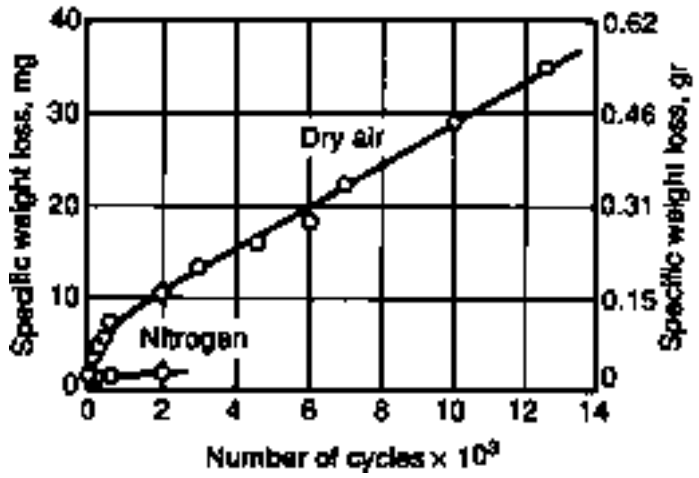


Fig. 15 Plot of fretting wear versus number of cycles for mild steel with 90 μm (0.0036 in.) slip amplitude in both dry air and nitrogen atmospheres. Source: Ref 24

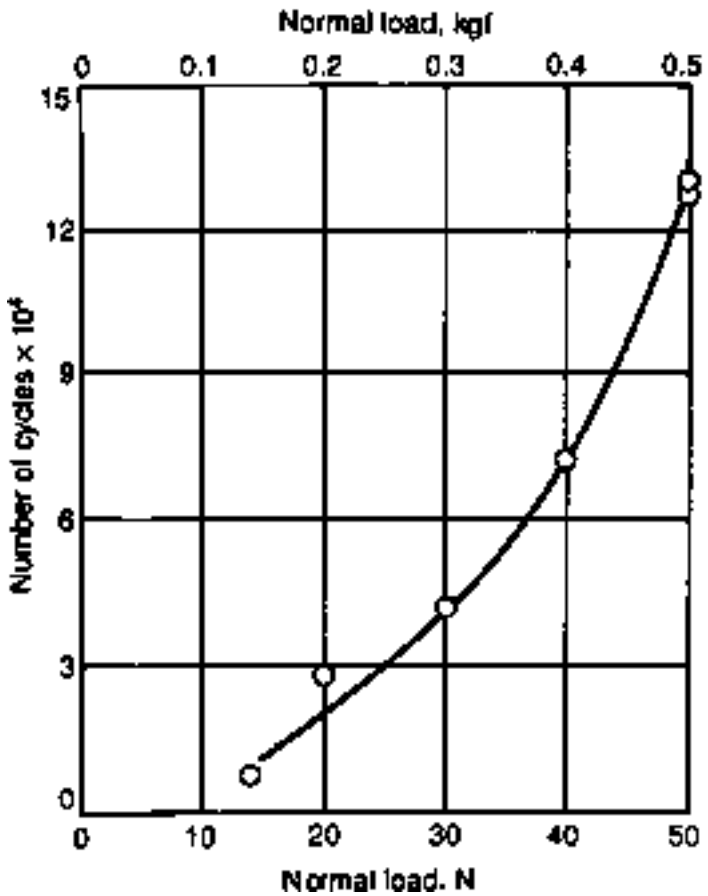


Fig. 16 Relation between initial period of metal-to-metal contact and normal load

Frequency. In the initial study of the effect of frequency (Ref 24), an increase in wear volume at low frequencies was found for a given number of cycles. This rise was related to the formation and disruption of oxide films and was considered in Ref 25. The question was considered more recently and investigated as a means of speeding up fretting wear tests, assuming that high frequencies had no significant effect (Ref 25). The range of frequencies investigated was from 10 to 10^4 Hz, using steel specimens in a crossed-cylinder contact. The conclusion was that, in the partial-slip situation, although there was little effect on measurable wear, the increase in the interfacial strain rate at high frequencies led to increased fatigue damage and increased corrosion due to the rise in temperature. However, in the total-slip situation, there was little effect over the frequency range 100 to 20×10^4 Hz, and these high frequencies could be used in accelerated tests.

Not all movements that can lead to fretting damage take place at anything like the frequencies just mentioned. Relative movement can be due to differential thermal expansion, which, if arising from diurnal changes in

temperature, can have a frequency of only a few cycles a day. The pressurizing of airplane cabins is in a similar frequency range. The movement of mooring cables due to water flow is of considerably higher frequency but is assumed to be ~0.1 Hz. In such circumstances, the environmental and corrosion effects have much greater importance, as can be seen by the extrapolation of the results reported in Ref 24 to lower frequencies (Fig. 17).

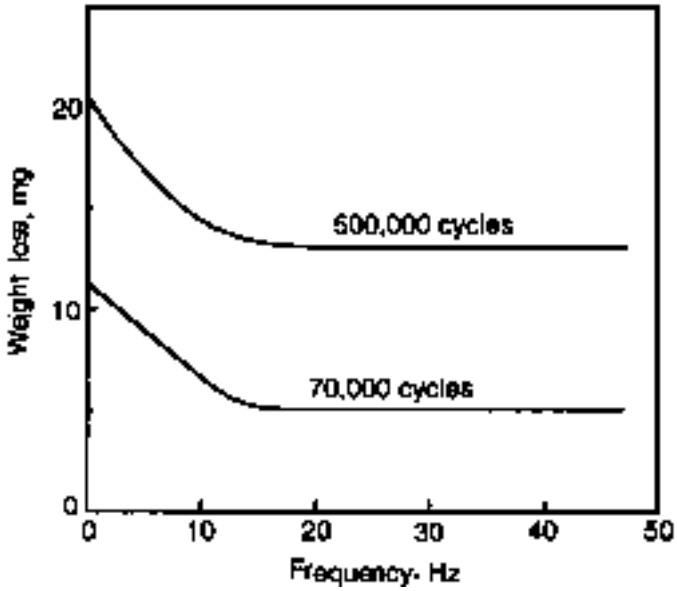


Fig. 17 Plot of wear versus frequency of fretting vibration

The area of low frequency has been investigated mainly in relation to electrical contacts. In many of these investigations, the low-frequency movement (0.003 to 0.03 Hz) has been produced by cyclic heating and cooling of a copper rod. This cycle allows also very low amplitudes to be investigated (Ref 27). Extensive work in the United States (Ref 28) and in Canada (Ref 29) has explored the effect of different contacting surfaces on the contact resistance under fretting conditions. In these studies, the aim was to maintain metal-to-metal contact throughout the working life of the contact. A complication in such experiments is the formation of a friction polymer from carbonaceous impurities in the atmosphere, which can have a profound effect on the contact resistance. For separable connectors, gold plating appears to be the most satisfactory surface, although cheaper coatings are actively being sought (Ref 30) and are discussed further in the section “Prevention of Fretting Damage” in this article.

Type of Contact. As discussed in the section “Effect of Debris on Wear” in this article, the ease with which debris can escape from the contact region is an important factor in the fretting process itself. It also gives rise to further problems by contamination of other surfaces (in a machine, for example).

The original design of an apparatus with torsional vibration of annulus on flat (Ref 2) has much to commend it because no part of the contacting surfaces becomes exposed, and for debris to escape, it must move at right angles to the direction of motion. Another slight advantage is that the amplitude of slip has a small variation from the inner to the outer edge and can therefore be used to investigate the effect of amplitude in one test. With flat contact surfaces, however, the initiation and development of areas of wear damage are sporadic no matter how carefully the surfaces are prepared and the alignment controlled. Contact pressure is usually expressed as the nominal value calculated from the apparent area of contact and the applied load. Measurement of the real area of contact would require contact resistance measurements at the outset of the test, or it could be calculated with a knowledge of the yield pressure (Eq 1). Contacts consisting of geometrical surfaces such as sphere on flat, cylinder on flat, and crossed cylinders allow some initial calculations based on elastic theory of the area of contact and stress distribution therein. However, these quantities are rapidly changed as the wear process proceeds. Also, such contacts increase the possibility for debris to escape. It was shown that the escape of debris in the crossed-cylinder arrangement is greatly influenced by the direction of motion (Ref 31). The arrangement shown in Fig. 18(a) allows the debris to escape by being pushed out by the axial movement of the upper cylinder, leading to more frequent metal-to-metal contact and a higher wear rate than the arrangement shown in Fig. 18(b). However, in experimental investigations, the ease and cost of preparing specimens are

significant factors, and the crossed-cylinder arrangement is one of the most convenient. It also mirrors the situation of contact between wires in steel ropes.

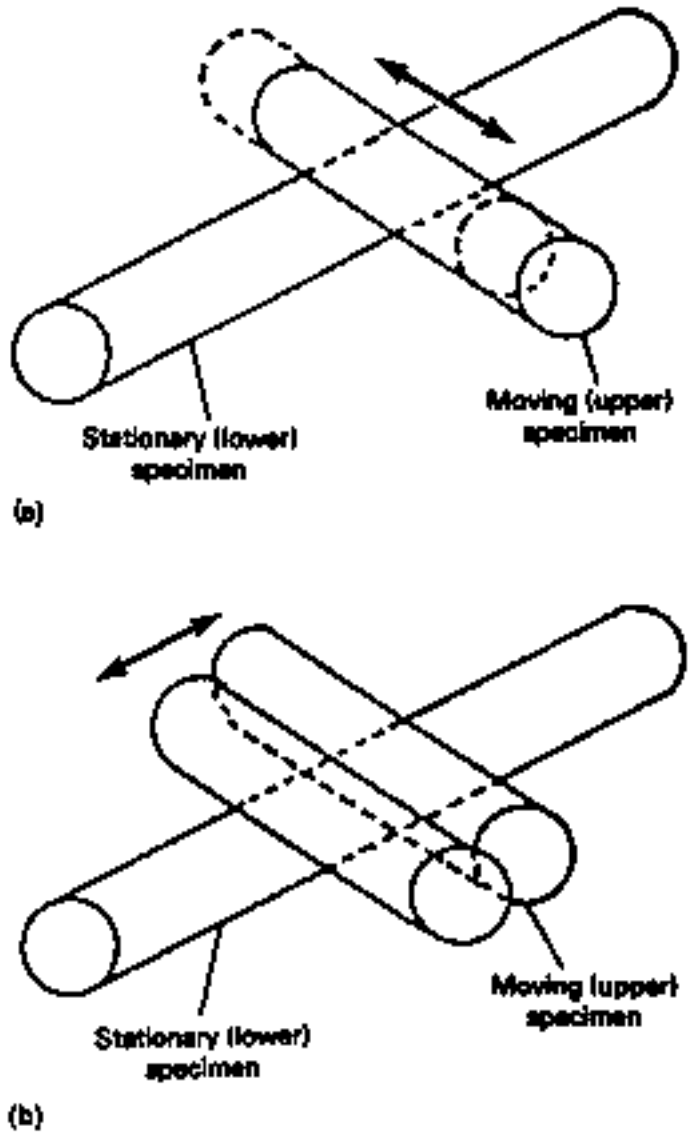


Fig. 18 Two test directions for determining fretting in a crossed-cylinder arrangement. (a) Parallel to axis of lower specimen. (b) At right angles to axis of lower specimen

Type of Vibration. In most experimental investigations, the vibration is simple harmonic motion of constant amplitude. This may be the case in practical cases of rotating shafts, but it is very far from situations in which frequencies may vary and pass through areas of resonance. In addition, aircraft structures will be entirely under the influence of random loading with no fixed frequency or amplitude. The same can be said of ropes and cables in moorings, overhead ropeways, and elevator ropes.

There have been few investigations into this aspect of fretting wear. A detailed analysis of the vibrations in one of the bearings supporting the shaft of an ac/dc generator operating onboard a ship was conducted (Ref 32). Spectra of the relative displacements in the axial, tangential, and radial directions were determined in the frequency ranges 0 to 500 Hz and 0.5 to 10 kHz, with no load and under full load. Although it was not possible to distinguish between elastic movement and rigid body movement, it was assumed that amplitudes >1 μm (>40 μin.) were due to the latter. The amplitude range was from 4.30 to 10^{-4} μm (173 to 0.004 μin.) in die frequency range <250 Hz, with several of the peaks being >0.05 μm (>2 μin.). The tangential component under full load had an upper limit of 8 μm (320 μin.), and combining this with the movements in the axial and radial directions gave a vector sum of 12 μm (480 μin.), at which considerable fretting damage was experienced. The analysis indicated that there was significant relative motion normal to the fretting surface.

This latter observation has been investigated in the construction of a tribometer that allowed frequency variations, phase differences between tangential and normal vibration, and the control of resonance (Ref 33). Three areas were identified:

- Subresonant loading conditions in which the damage was said to resemble typical fretting damage
- Resonant conditions in which the surface damage was significantly less due to the lower dissipation of energy due to friction
- Superresonant operation, which showed tearing of the surface that was associated with a phase shift of 90° between the tangential and normal vibrations

Impact Fretting. In the section "Type of Vibration," the effect of a cyclically varying normal force superimposed on the tangential movement is considered for situations in which the two surfaces remained in constant contact. The circumstances where the two surfaces can become separated during some part of each cycle are somewhat different because of the contribution of impact. As mentioned earlier in the section "Fretting Wear in Mechanical Components," this is a problem frequently encountered in heat exchangers, where pipes or fuel rod containers are restrained by supports or baffles but experience vibration as a result of fluid flow.

Two types of wear can be distinguished in the practical case encountered in nuclear reactors. The first type involves whirling of the tube, resulting in intermittent unidirectional sliding and is obviously not fretting. The small-amplitude movement that results in fretting-type damage is the result of flow-induced vibrations. The important parameters are impact force, clearance between tube and support, support plate thickness, and frequency. The wear rate for stainless steel increases by an order of magnitude for a tenfold increase in impact force, and a similar increase for a fourfold increase in the clearance. Increasing the support plate thickness from 10 to 20 mm (0.4 to 0.8 in.) reduces the wear rate from 0.7 to 0.2 μm (30 to 8 μin.) per 10^6 cycles. At frequencies below 40 Hz, there is no effect of frequency, but at higher frequencies the wear rate is reduced due to the reduction in contact time per cycle (Ref 34). These results were obtained at room temperature, whereas in operation, these devices would be at a higher temperature and thus the wear rates could be profoundly influenced both by temperature and environment; such effects are discussed in the section "Environmental Effects."

Many years of work (Ref 35) indicated that the combination of impact and fretting was more severe than fretting alone, whereas much of the work on tube and tube support vibration suggests that the impact contributes little to the wear rate and that continuous contact is more damaging (Ref 36, 37). An important variable in considering this type of motion is the phase relation between the normal and tangential vibrations. Differences are to be expected between impact at maximum tangential velocity and zero velocity. Also, the time period per cycle that contact is made is of great significance. In Ref 34, the following contact times per cycle are quoted as a function of frequency:

Frequency, Hz	**Contact time per cycle, s**	**Difference between impact at maximum tangential velocity and zero velocity, %**
26	0.022	57
40	0.010	40
60	0.005	30

Rigs have been constructed to study the effects of these phase relationships, particularly under severe environmental conditions (Ref 34).

Surface Finish. Early observations of practical examples of fretting damage suggested that the more highly polished the contacting surfaces are, especially if flat, the worse the damage will be. On polished aluminum where the debris is largely Al_2O_3 (corundum) with a high hardness, it could spread the damage, whereas on a rough surface the debris can escape into the hollows between the real contact areas. There is also the possibility that on a rough surface, if the exciting force is of small amplitude, some of the relative movement will be taken up by elastic deformation of the asperities. One study shows that a rougher surface suffers less damage (Ref 38).

Shot peening is a widely applied surface treatment that produces roughening, work hardening, and a residual compressive stress in the surface. The author showed that shot peening a steel surface reduced the coefficient of friction in fretting (Ref 39), but this could also be due to work hardening and possibly also to the compressive stress. To investigate this phenomenon further, and to try and separate the three effects of shot peening, a series

of tests were carried out on treated cylindrical steel specimens. The residual stress and work hardening could be removed by annealing, the residual stress could be reduced by controlled tensile straining, and the surface roughness could be removed by careful surface polishing (Ref 40). The shot-peened surface of 0.4% C steel had a very low value for fretting wear; however, polishing raised the level of wear to that of the unpeened steel.

Residual Stresses. Because one of the possible mechanisms of fretting damage is surface fatigue, it is to be expected that a residual stress field will have some effect on the wear process. Experiments were performed with a bearing ball on a bearing steel (1.0C–1.5Cr) specimen that was statically bent to give residual stresses over a range of -900 to 900 MPa (-130 to 130 ksi). When the specimen was vibrated with a contact load of 216 N (22 kgf) and an alternating tangential load of ±20 N (±2 kgf) so that partial slip was occasioned, the depth of the wear scar increased with tensile stress but decreased with compressive stress (Fig. 19).

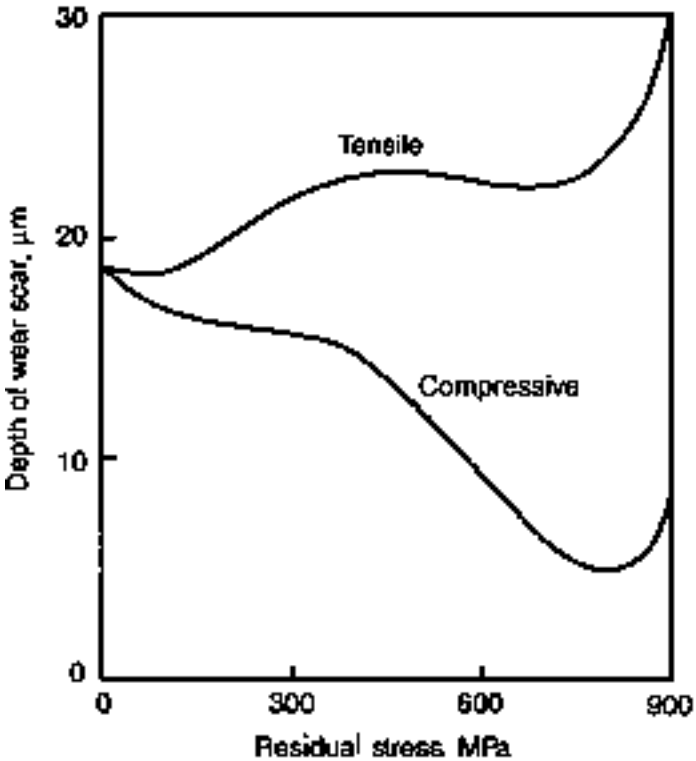


Fig. 19 Plot of depth of wear scar versus compressive and tensile residual stresses in the fretted surface. Source: Ref 41

It was shown earlier that residual stresses can be generated by the fretting action (Ref 42) and this has been confirmed in further work (Ref 43). In fretting, flat steel surfaces with residual compressive stresses up to 735 MPa (107 ksi) were generated; these were most important in a direction at right angles to the direction of motion. The converse of this effect, namely the modification of existing compressive stresses by fretting, has been observed in fretting-fatigue experiments (Ref 44). The stresses fade under normal fatigue but disappear much more rapidly under fretting fatigue (Fig. 20).

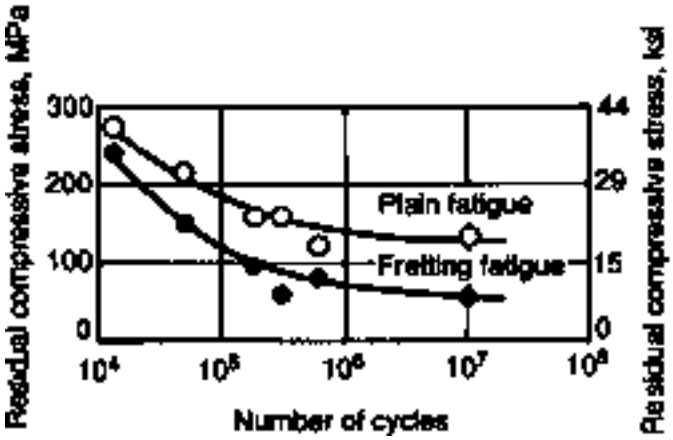


Fig. 20 Fading of surface compressive stress induced with a number of fretting cycles by shot peening

Finally, the presence of a static tensile stress can lead to the propagation of cracks initiated by the fretting (Ref 45).

Materials

If fretting wear is a possible occurrence in the design of a machine or structure, it may be possible to reduce the effect by choosing materials that are more resistant to such damage. This section deals with bulk materials not coated or surface treated; coated and surface-treated materials are dealt with in the section "Prevention of Fretting Damage." In addition, information on the behavior of new materials such as composites and ceramics also must be considered because of their increasing use in engineering applications.

Steels have been most extensively studied, largely because of their ability to change mechanical properties over a wide range by heat treatment without changing their chemical composition. In experiments with a steel ball on a steel flat, in which the hardness was varied from 220 to 850 HV, there was little difference in the wear loss (Ref 46). In crossed-cylinder experiments with three alloys of the same crystal structure, increasing hardness is shown to lead to a decrease in the critical amplitude between partial and total slip (Ref 47). The materials were copper, copper silicon, and austenitic stainless steel, all materials of low stacking fault energy; as a consequence, work hardening occurred at low frequencies, but softening set in at high frequencies because of the temperature rise.

In some investigations, notably those described in Ref 48 and 49, the formation of a white etch-resistant layer has been detected in fretting of carbon steels (Fig. 21). This material is extremely hard and tends to develop cracks. It often is referred to as *martensite,* and some authors have regarded this as evidence that high local temperatures can be developed in fretting. The question of temperature rise is controversial and is discussed more fully subsequently. However, if there is a temperature rise, materials that rely on low-temperature heat treatment for their optimum mechanical properties (for example, most aluminum alloys) are likely to suffer a further type of damage in fretting.

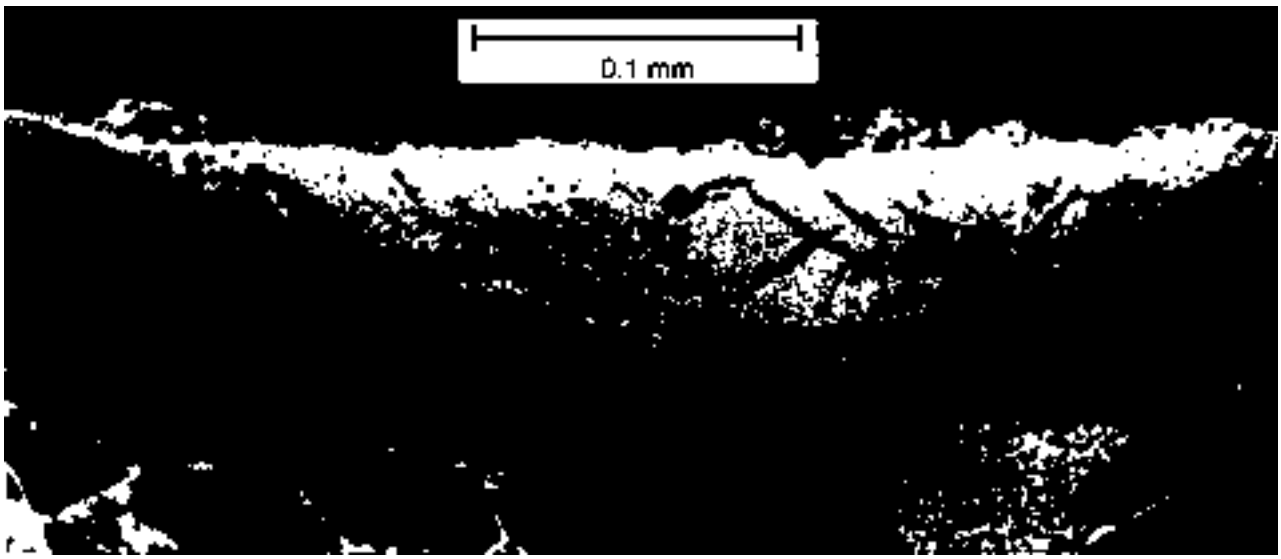


Fig. 21 Cross-sectional view of a white layer of martensite produced by fretting of a carbon steel connecting rod. Axial stress, 0 to 380 MPa (0 to 55 ksi); contact stress, 40 MPa (6 ksi); fretting cycles, 10^5. Sample was nital etched and viewed with scanning electron microscopy (SEM).

One development has been die, rapid quenching of liquid metals to produce amorphous metals. Fretting of two of these materials, Co-4Fe-2Mo-16Si-12B and Fe-4Si-15B, indicates that they are not wear resistant and become embrittled by the fretting action (Ref 50).

Metals and alloys that rely on a protective oxide film for corrosion resistance can suffer considerable fretting damage in environments (for example, ambient air), that would not be considered particularly corrosive. These include the more reactive metals such as aluminum, titanium, zirconium, and chromium when used as elements

to alloy stainless steels. This aspect of fretting is considered in the section “Environmental Effects” in this article.

When dissimilar metals pairs are fretted against each other, their mutual solubility is a factor in the severity of the damage. Nickel, chromium, and iron are damaged severely when fretted against iron or chromium because of high mutual solubility; however, the damage is mild when these metals are fretted against copper or silver because there is no solubility (Ref 51).

In the fretting of polymers against steel, the possibility exists that one material will be transferred to the opposite surface. Damage arises in this case from platelike iron oxide particles adhering to the polymer; thus, the process becomes the fretting of two oxide-coated surfaces. Damage can occur to the polymer by the formation of cylindrical wear particles formed by rolling up of thin detached surface layers, which are extruded in a direction at right angles to the motion (Ref 52, 53).

In an investigation of fretting wear of fiber-reinforced polymers (composites) in contact with an aluminum alloy, weight loss was used to measure the damage (Ref 54). At amplitudes <900 μm (<0.036 in.), the wear was low and showed a slight increase with increasing amplitude. Above 900 μm (0.036 in.), the wear accelerated. Wear rate showed a stepwise increase with frequency, the critical frequency being 35 Hz at room temperature but reduced to 20 Hz at 80 to 100°C (175 to 212°F) due to a local rise in temperature. Unlike metals, which show a linear increase in wear with increasing normal load, these materials show little effect of load up to a critical value at a pressure of 20 MPa (3 ksi), above which a rapid increase occurs. Fiber orientation with respect to the sliding direction is another variable with these materials. Here the specific wear rate was negligible at angles of 0 and 90° between fiber axis and sliding direction, but showed a pronounced peak at 45°, as did the coefficient of friction (Fig. 22).

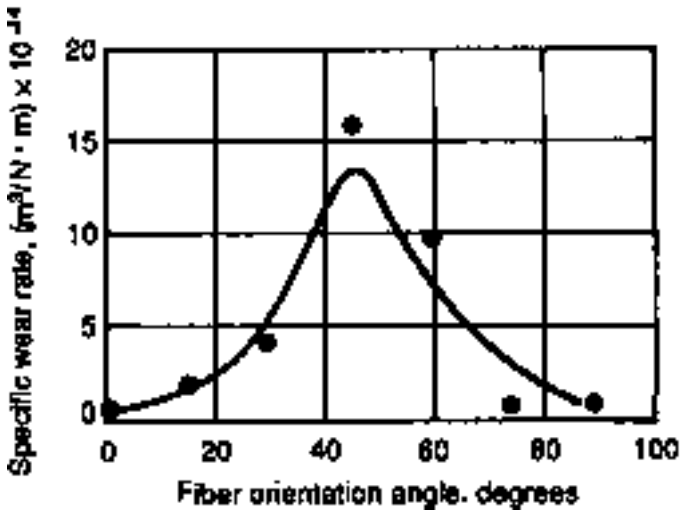


Fig. 22 Plot of fretting wear versus fiber orientation angle in a composite

In experiments involving the fretting of sintered alumina against a wide variety of metallic materials in a crossed-cylinder configuration, no detectable wear was found on the alumina, although metal transfer to its surface occurred (Ref 55). A similar result was found in fretting a steel ball against alumina, silicon nitride, and zirconia in dry air (Ref 56); however, silicon carbide produced equal wear on the ceramic and steel. An increase in relative humidity generally resulted in a decrease in wear rate except in the case of silicon nitride, which exhibited behavior similar to that of silicon carbide.

It appears that when oxides such as alumina and zirconia are fretted, there is little damage. Fretting alumina against alumina produced no measurable wear, although some polishing of the surface was apparent (Ref 31, 57) (Fig. 23).

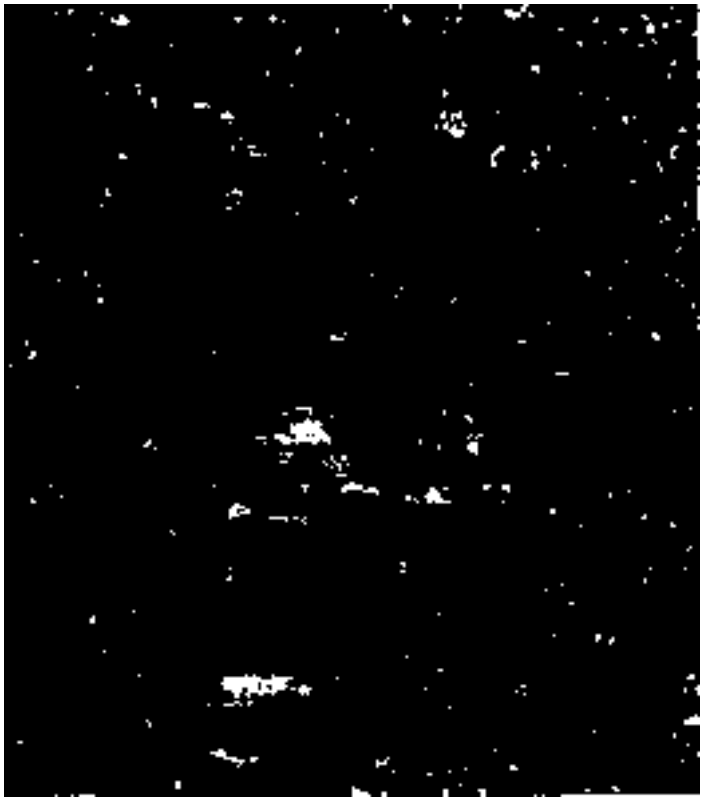


Fig. 23 Fretting mark generated by fretting an alumina cylinder against an alumina flat

In one of the earliest observations of fretting wear (Ref 58), a steel ball was fretted against the underside of a glass microscope slide, and the process was observed under a microscope. This type of experiment has been repeated in Japan, but no quantitative measurements of wear have been made. However, valuable information has been gained on the initiation of cracks in the glass, a typical brittle material, and on the onset of damage to the steel, and is therefore relevant to discussions on fretting mechanisms (Ref 59).

Effect of Debris on Wear

Debris oozing from a joint or connection in machinery is often the first indication that fretting damage is occurring. In recent years, the role of debris in wear processes has been given considerable attention because it can build up between the sliding surfaces and increase the number of interfaces from one to two (Ref 60). This is particularly the case in fretting where escape of the debris is greatly restricted compared with unidirectional sliding. This section focuses on the form, composition, and role of the debris.

It is well established that the debris on mild steel and other steels in air is essentially α-Fe_2O_3, which, when finely divided, is reddish brown in color and nonmagnetic. Its hexagonal crystal structure is identical to that of α-Al_2O_3. Under high-compressive loads, it can become compacted, and its color is then black and iridescent. In this respect, it is identical to the mineral hematite, which in massive form is blackish but is red when powdered. On aluminum, the debris is black and consists of aluminum particles surrounded by Al_2O_3. The aluminum content is 23% and the powder is pyrophoric (Ref 61). In other metals, such as copper and titanium, die debris is the expected oxide. In some cases, depending on the method of its production, the debris can contain some unoxidized metal.

The debris is usually in the form of thin platelets a few microns thick and 50 to 100 μm (0.002 to 0.004 in.) in diameter, but it can be much smaller, as in Fig. 24, where the debris resulted from fretting steel against a polymer. The particle size was shown in this case to be related to the surface energy of the polymer. Sometimes fretting occurs between the walls of fatigue cracks, when spherical and cylindrical shapes of debris may result due to the very small movements and the high compressive stresses during the compression part of the fatigue cycle.

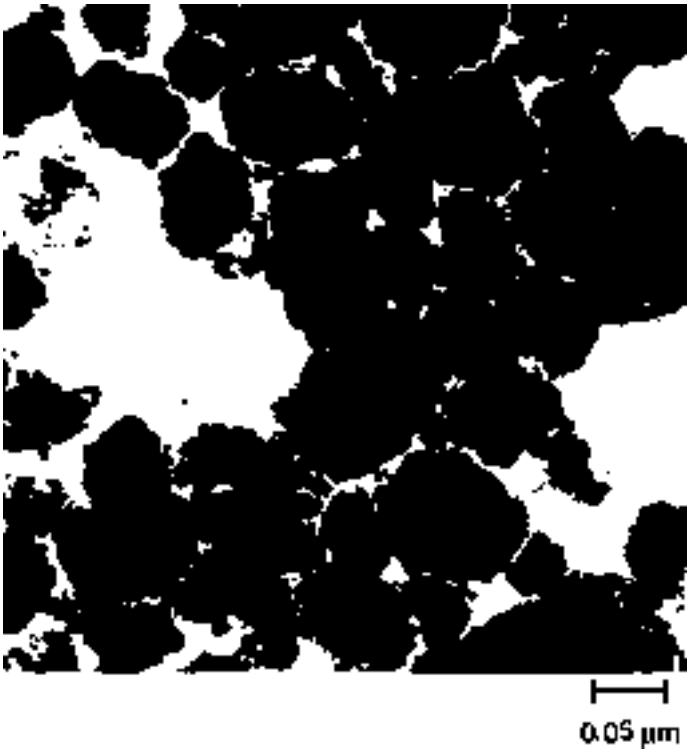


Fig. 24 Oxide debris produced when fretting steel against a polymer

The role of the debris on steel specimens has been demonstrated by performing tests where the fretting surfaces were periodically separated and the debris removed (Ref 62), resulting in an increase in the wear rate. Conversely, the introduction of α-Fe_2O_3 powder between the surfaces reduced the wear rate; the resultant surfaces were hardly damaged and showed signs of slight polishing. In long-term experiments, the terminal stage of fretting was identified as the formation of extensive beds of compacted oxide with a much-reduced wear rate (Ref 63). With very small slip amplitudes of less than 10 µm (400 µin.), the wear ceases because the oxide beds support the load and separate the surfaces (Ref 64). At higher amplitudes of ≥20 µm (≥800 µin.), the wear rate is much reduced; wear continues, however, and occasional metal-to-metal wear still can be detected by contact resistance measurements, even after millions of cycles (Ref 65). These observations have considerable bearing on the ideas that have been put forward for explaining the mechanism of fretting, which are discussed subsequently. The conclusion is, however, that the accumulation of debris between the surfaces in compacted layers is more likely to reduce further wear and that abrasion by the oxide debris is not a significant contribution to the problem.

It is interesting to note that in one of the later investigations (Ref 64), it was stated that the debris had some of the characteristics of a thixotropic liquid, an observation that had been made in 1951 (Ref 58).

Environmental Effects

Obviously, most investigations and reported cases of fretting wear occur in a normal atmosphere where oxygen and water vapor are the main corrosive agents, although there is evidence that nitrogen may be involved in alloys containing nitride-forming elements.

Vacuum. An investigation on the effect of reduced air pressure on the fretting of a carbon steel cylinder on a flat showed that in a vacuum of 1 mPa (1.5×10^{-7} psi), the coefficient of friction was high at 3.0, and, although little measurable wear took place, considerable surface damage occurred due to adhesive transfer of metal, which ceased at pressures above 0.1 Pa (1.5×10^{-5} psi) (Ref 66). As the pressure was increased further, the coefficient of friction fell to a value of ~1.0 at 10 Pa (1.5×10^{-3} psi), whereas the wear remained low up to this pressure. Above 10 Pa (1.5×10^{-3} psi), the coefficient of friction remained constant but the wear rate increased rapidly (Fig. 25). Below this critical pressure, any debris formed was black and identified as Fe_3O_4, but above the critical pressure the debris was red and identified as α-Fe_2O_3. Similar results were obtained with an austenitic stainless steel, except in this case, the plastic deformation at the low pressures resulted in martensite formation and the growth of surface fatigue cracks (Ref 67). The effects of slip amplitude and normal load were

investigated on the initiation and growth of such cracks at a pressure of 4 mPa (6×10^{-7} psi); the tests showed that the number of cracks decreased rapidly at an amplitude ≥50 μm (≥0.002 in.) but that the length of the longest crack increased markedly above this amplitude (Ref 68). The number of cracks increased steadily over the load range 8 to 30 N (0.8 to 3 kgf), whereas the length of the longest cracks was greatest at 8 N (0.8 kgf) but decreased rapidly at loads above 13 N (1.3 kgf). These results have considerable significance in relation to the fatigue aspects of fretting, particularly in connection with space research. Further work on an Fe-Cr-Ni-W amorphous alloy (hardness, 1300 HV) fretting against stainless steel gave similar results, although the coefficient of friction remained low over the whole pressure range, which was attributed to cylindrically shaped debris acting as rollers (Ref 69).

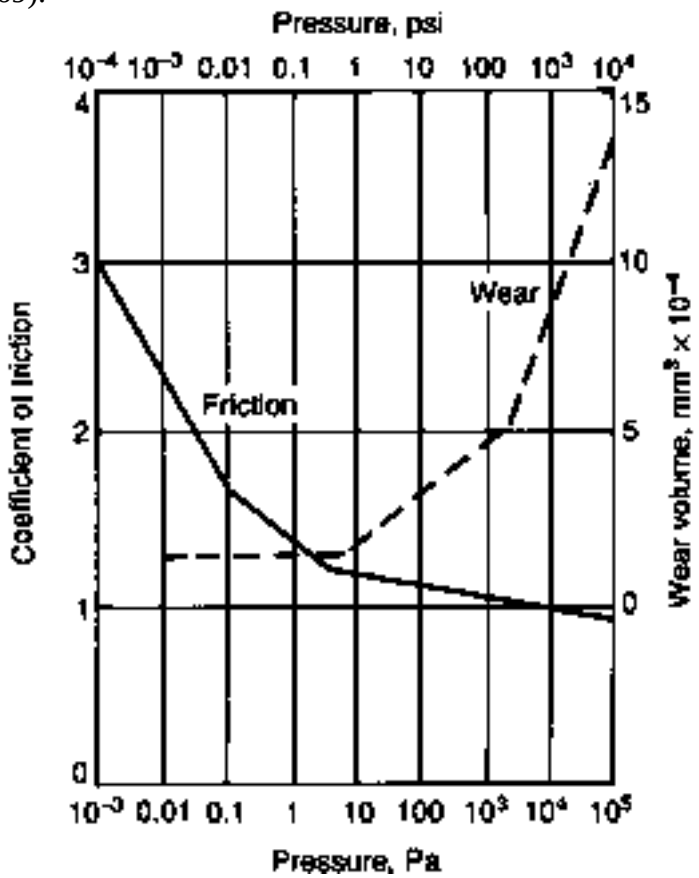


Fig. 25 Effect of atmospheric pressure on the friction and wear of a carbon steel cylinder

The effect of high pressure has not been investigated extensively. The studies that have been conducted on the effects of high pressure have been associated with the nuclear-power industry, specifically relating to problems in the advanced gas-cooled reactor (AGR). In this case, the reactor is cooled by carbon dioxide at a pressure of 4 MPa (40 atm), with temperatures in the region of 600 °C (1110 °F) (see the section "Low and High Temperature" in this article).

Humidity. After oxygen, water vapor is the most significant environmental factor in fretting wear. In a study of a wide variety of pure metals tested in a hemisphere on flat configuration and at a slip amplitude of 80 μm (0.0030 in.), most metals, such as silver, copper, titanium, and iron, showed a peak coefficient of friction value between 10 and 15% relative humidity (RH), with a corresponding peak in wear volume (Ref 70). With chromium, the peaks appeared at 5% RH, but with aluminum and nickel the peaks occurred at very low humidity. At normal atmospheric humidity (~50% RH), all the metals except nickel showed a gradual decrease in fretting wear. This corresponds well with earlier observations on steel specimens (Ref 24). Because aluminum alloys are particularly sensitive to the effects of water vapor in fatigue situations due to the reactivity of an exposed aluminum surface and the liberation of hydrogen, further work has been devoted to the aluminum-zinc-magnesium alloy (Ref 71). Fretting wear was found to be the same in dry air as in dry argon, thus discounting oxygen as an active agent; however, in air of 60% relative humidity, the wear rate was much higher because the wear debris was nonadherent and, thus, metal-to-metal contact was increased. In addition, it was found that softening due to overaging of the surface material in the contact region developed in humid air.

As mentioned previously, water vapor has a considerable effect on the fretting wear of certain ceramics, particularly silicon carbide and silicon nitride (Ref 56). With silicon carbide, alumina, and zirconia, the high rate of wear at 5% RH is greatly reduced at humidities >50%, with a steady decrease in the coefficient of friction. Silicon nitride shows a similar wear rate to silicon carbide at 5% RH, but the rate increases linearly with time at ≥50% RH. This is related to the propensity for silicon nitride to be transformed into silica.

Aqueous Electrolytes. Fretting of corrosion-resistant metals in aqueous environments was investigated originally because of problems associated with orthopedic implants in the human body. In some of these devices, metal-to-metal contacts (for example, screws and bone plates) could suffer fretting due to the activity of the patient. In potentiostatic experiments on an austenitic stainless steel, it was found that the corrosion current increased linearly with the amplitude of slip at a fixed frequency and also was linearly related to frequency at a fixed amplitude (Ref 72). Where potential control was not applied, there were large drops in potential with a slow recovery when the fretting action ceased (Fig 26). In studies in physiological saline (artificial body fluid) of fretting of type 316L stainless steel and Co-Cr-Mo alloy, the latter was found to have the lower wear rate, although the addition of albumin reduced the wear rate of the stainless steel (Ref 73). In tests with a fretting simulator using screws and a bone plate of type 316L stainless steel, the weight loss of the screws was an order of magnitude greater in 0.9% NaCl compared with the same solution to which 10% bovine serum had been added (Ref 74). Measurements of potential drop on fretting obtained in laboratory experiments were compared with in vivo measurements on bone plates fitted in sheep walking on a treadmill. The potential drop was found to be related to the load at low levels of torque to which the screws had been tightened (Ref 75). Further experiments on the simulator revealed that the wear weight loss of plates and screws was highest for stainless steel (2.8 mg, or 0.043 gr), less for wrought Co-Ni-Cr-Mo (0.65 mg, or 0.010 gr) and lowest for Ti-6Al-4V ELI (0.13 mg, or 0.0020 gr) (Ref 76), where ELI is extra-low interstitial. It has since been established that the effect of proteins on fretting corrosion depends on the relation of the pH of the solution to the isoelectric point of the protein. At pH values below the isoelectric point, no effect is seen, but at pH values above the isoelectric point there is a reduction in the wear rate (Ref 77).

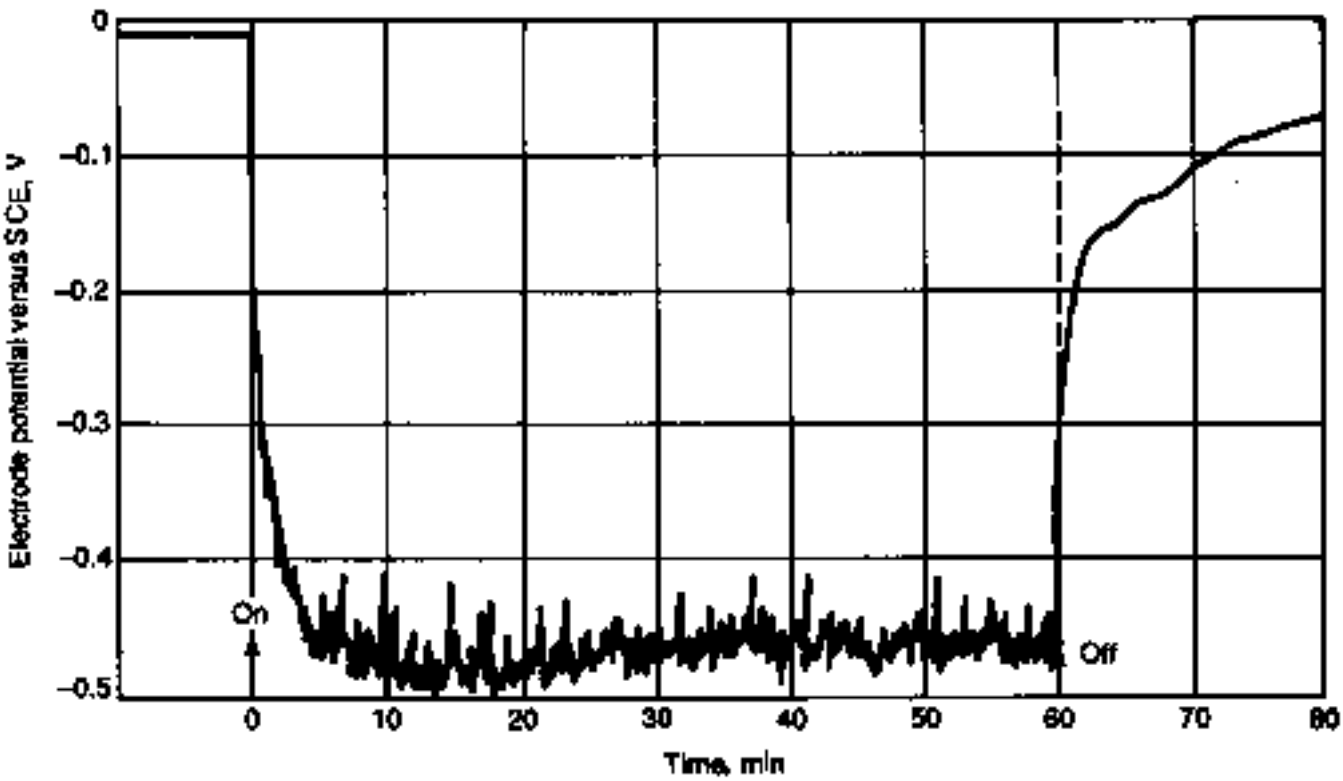


Fig. 26 Potential drop obtained when fretting Ti-6Al-4V (IMI 318) titanium alloy in a saline solution

The investigation of corrosion-resistant materials under fretting conditions in aqueous electrolytes was obviously of considerable interest because the materials rely on a protective oxide film for their corrosion properties. Continual disruption of this film is known to produce high corrosion rates, as confirmed by measurements of weight loss or volume of pits, which equates to rapid wear rates. The development of offshore

oil rigs in the 1980s provided many more examples of fretting corrosion occurring on less-corrosion-resistant materials, such as the weldable steels used in the construction of the rigs and the steel ropes used to moor them in the marine environment. The specific wear rate for a 1.5% Mn steel in air was 1.07 $m^3/N \cdot m \times 10^{-16}$, but in seawater it increased to 10.0 $m^3/N \cdot m \times 10^{-16}$ at 50 Hz and 235.5 $m^3/N \cdot m \times 10^{-16}$ at 1 Hz (Ref 78). Fortunately, application of cathodic protection of -950 mV versus standard calomel electrode (SCE) reduced the wear rate almost to that in air.

Steel ropes are used extensively in marine conditions. The development of the offshore oil industry has increased the use of these ropes as moorings for the oil rigs. The single-strand ropes are made up of layers of 5 mm (0.2 in.) diameter steel wires; successive layers are spiraled in the opposite sense to the adjacent layers. Numerous interwire contacts are referred to as *trellis contacts.* The fretting behavior of these wires in artificial seawater has been investigated extensively (Ref 79). In crossed-cylinder tests with an amplitude of slip of 50 μm (0.002 in.), a normal load of 3 N (0.3 kgf), and a frequency of 30 Hz, it was found that the wear volume after 5×10^6 cycles for the 0.64% C steel was 55×10^{-4} mm^3 (34×10^{-8} $in.^3$) in air, 167×10^{-4} mm^3 (102×10^{-8} $in.^3$) in seawater, but reduced to 4×10^{-4} mm^3 (2×10^{-8} $in.^3$) when a potential of -950 mV versus SCE was applied (Ref 80). In 3.5% NaCl solution the wear was even higher at 214×10^{-4} mm^3 (130×10^{-8} $in.^3$), indicating that the magnesium salts in the seawater were having some inhibiting effect. The effect of coating the wire by hot dip galvanizing was similar to the effect of cathodic protection, reducing the wear to ~10×10^{-4} mm^3 (~6×10^{-8} $in.^3$) in seawater.

More aggressive solutions have not been investigated widely, apart from two examples by the author. Fretting of an austenitic stainless steel in H_2SO_4 produced severe pitting in the contact region (Fig. 27). In some work related to the cooling of suspensions of radioactive solids in concentrated nitric acid, it was found that austenitic stainless steel tubes in a crossed-cylinder configuration when fretted in this mixture developed very strong adhesion, as though the scouring action of the solid matter had produced clean metal surfaces that were able to form cold welds. It is interesting to note that strong adhesion was found when austenitic stainless steel was fretted against itself in liquid sodium at 500 °C (930 °F). The force to part the surfaces was calculated to require a tensile stress of 500 MPa (73 ksi), greater than the ultimate tensile strength (UTS) of the material (Ref 81).

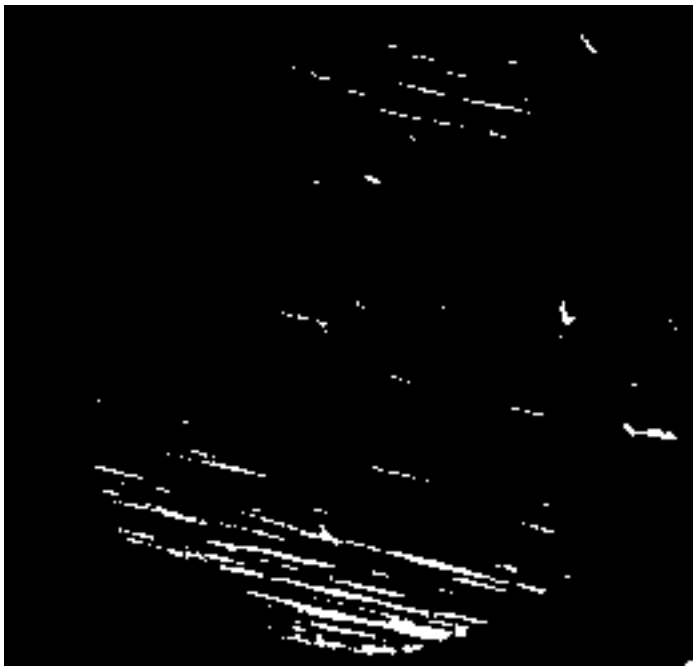


Fig. 27 Local pitting produced when an austenitic stainless steel ball is fretted against an austenitic stainless steel flat in 0.1 N H_2SO_4

Low and High Temperature. The increasing development of space engineering has meant that dynamic systems have to operate not only in vacuum but also at very low temperatures. Fretting of a spherical surface on a flat, both composed of austenitic stainless steel, at 4 K (-452 °F) in liquid helium resulted in high friction but little wear, very similar to the situation in high vacuum (see the section “Vacuum” in this article). The explanation

was similar; that is, oxide films did not grow at this temperature. At 77 K (-320 °F), the friction and wear were the same as at 293 K (-68 °F).

Much more attention has been given to high-temperature fretting in oxidizing atmospheres because this is a situation that is very common in practice (for example, in gas turbine aeroengines and in the AGR device). Pioneer work (Ref 82) and later studies (Ref 83, 84) showed that there was a great reduction in the coefficient of friction and wear when mild steel was fretted in air at temperatures up to 500 °C (930 °F). In argon, however, surface damage was severe. At about this time, the reciprocal sliding of certain alloys at high temperature in an oxidizing atmosphere was investigated, and it was discovered that many alloys containing combinations of iron, nickel, and chromium developed a very smooth oxide, which was termed *glaze oxide,* which had low friction and low wear (Ref 85). A similar type of oxide was found in fretting a nickel-base alloy at 540 °C (1005 °F) (Ref 86). Fretting wear tests at amplitudes of 10 and 40 μm (0.0004 and 0.0016 in.) showed low coefficients of friction at 280 and 540 °C (535 and 1005 °F) compared with room temperature, and corresponding reductions in wear damage (Ref 87). However, if the air pressure is reduced to 1 mPa (l.5 × 10^{-7} psi) severe surface cracking is produced (Ref 88). The alloys that are capable of forming the glaze oxide are those on which an oxide with a spinel structure can develop (for example, mild steel [$FeO \cdot Fe_2O_3$], stainless steels [$FeO \cdot Cr_2O_3$, $NiO \cdot Cr_2O_3$, $NiO \cdot Fe_2O_3$], and nickel-chromium alloys [$NiO \cdot Cr_2O_3$]). If the oxide film is damaged, it is self repairing at the high temperature but does not survive at room temperature (Ref 89). Titanium alloys do not behave in quite the same way; however, as shown in Fig. 28, implantation with bismuth or barium ions results in very low friction and wear results at temperatures between 400 and 600 °C (750 to 1110 °F) (Ref 90). For such films to survive, the substrate must have sufficient creep resistance to withstand the high stresses in the contact regions; otherwise, cracking can occur (Ref 91) (Fig. 29).

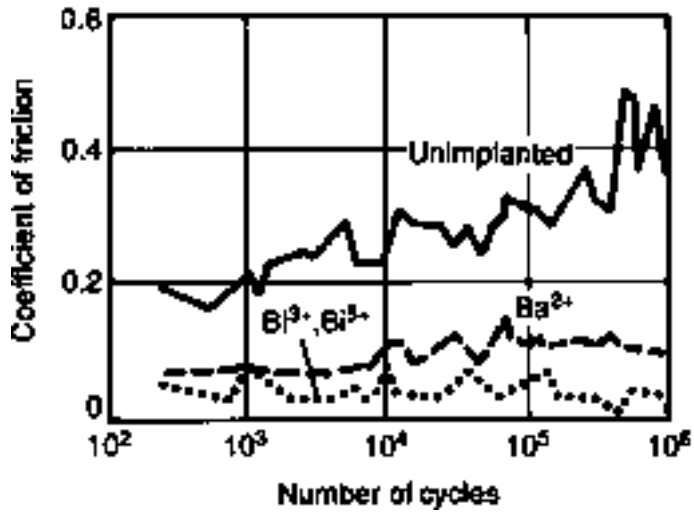


Fig. 28 Effect of ion implantation on the coefficient of friction in fretting of IMI 550 titanium alloy at 500 °C (930 °F)

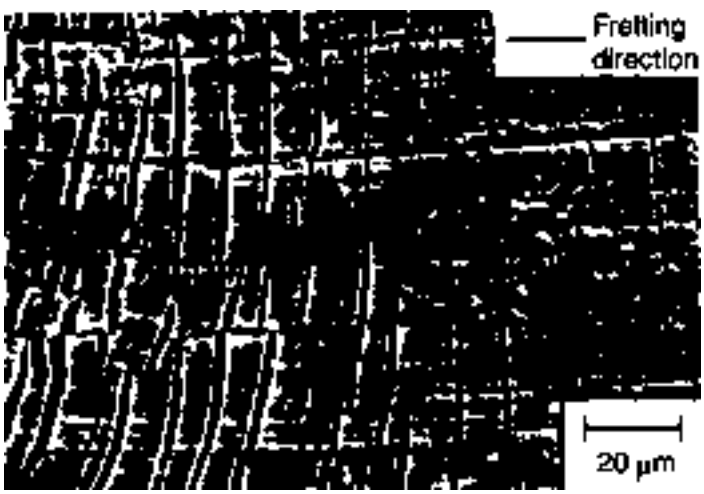


Fig. 29 Cracking of "glaze" oxide layer because of substrate creep in the fretting contact regions

Measurement of Fretting Wear

Profilometry. The very small amount of material that is removed in most fretting experiments renders the traditional method of measuring wear by weight loss impracticable. The method of surveying the surface by profilometry, where the instrument can plot out an isometric projection of the surface and the computer can produce a figure for the volume of the scar below die original surface and also the volume above the original surface, gives one of the best quantitative and descriptive assessments of the damage (Ref 92). When this method was checked with weight loss measurements, good agreement was found. Figure 30 shows the scar in a crossed-cylinder arrangement.

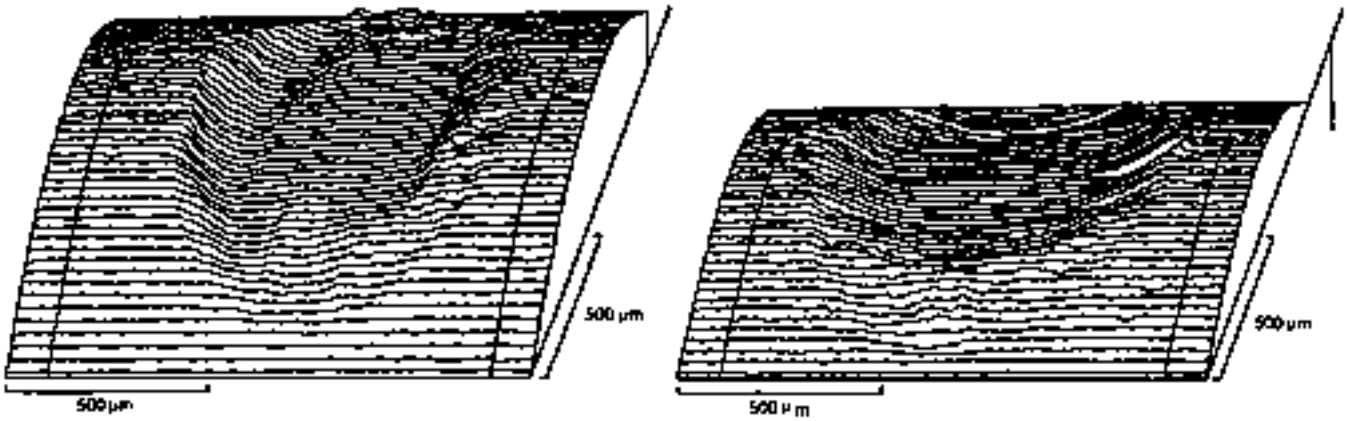


Fig. 30 Two views of an isometric projection of a scar in the crossed-cylinder arrangement

Holographic Interferometry. Another method that gives a visual impression of the damage is holographic interferometry. This practice enables a contour map of the surface to be generated prior to the wear process in order to compare it with a similar map generated after the wear has occurred. By using holographic image subtraction, very small changes in surface topography can be detected (Ref 93).

Thin-Layer Activation. A quantitative method applied to measuring fretting wear is known as *thin-layer activation* (TLA). A thin layer of radioactive atoms is produced on the surface by bombardment with a high-energy ion beam from an accelerator. After wear has taken place, debris is removed and the residual radioactivity is measured. A sensitivity of at least 0.2 µm (8 µin.) is claimed (Ref 94).

Measurement of Axial Distance. A less-sophisticated method for the case of crossed cylinders is the determination of wear from the decrease in distance between their axes (Ref 95). The advantage of this method is that it allows measurements to be taken without dismantling or disturbing the contact.

Mechanism of Fretting Wear

It generally is agreed that when fretting wear is taking place between two flat surfaces or two conforming cylindrical surfaces (that is, conditions in which escape of debris is not as easy), the progress with time (or number of cycles) can be recognized as occurring in three stages (Ref 96, 97):

1. Initial stage of a few thousand cycles when metal-to-metal contact is prevalent, resulting in local welding, roughening of the surface, high friction, and low contact resistance. Fatigue cracks are initiated in this stage if the movement is a result of cyclic stressing (Ref 98).
2. Formation of beds of compacted oxide with a fall in coefficient of friction and erratic behavior of contact resistance as it oscillates between high and low values
3. Onset of a steady state in which the friction is more or less constant and the contact resistance is generally high with occasional momentary falls to a low value

These changes are illustrated in Fig. 31, 32, 33 and can be recognized in the wear curve (Fig. 15). The length of the initial stage depends on the amplitude of slip and the normal load (Fig. 16).

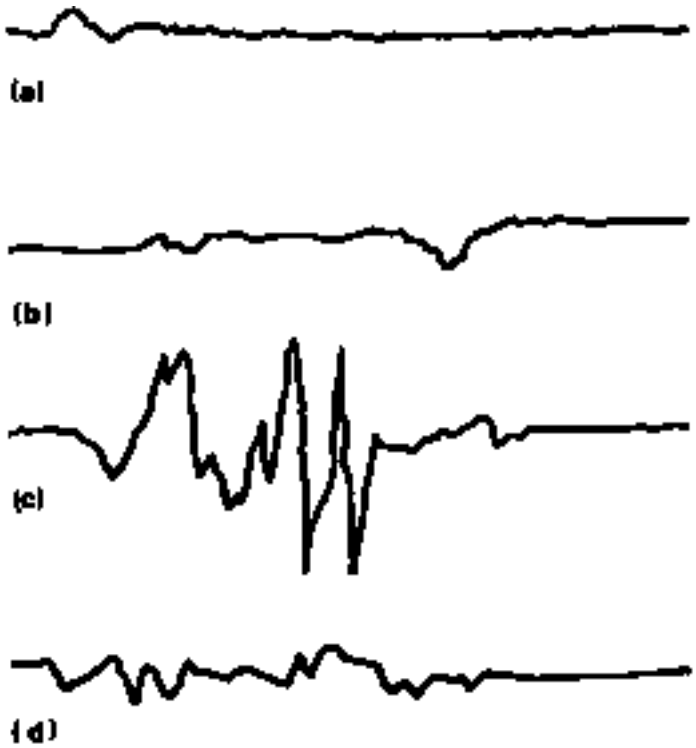


Fig. 31 Changes in surface roughness as a function of the number of fretting cycles for metal-to-metal contact. (a) As received. (b) 10^3 cycle. (c) 5×10^4 cycles. (d) 3×10^5 cycles

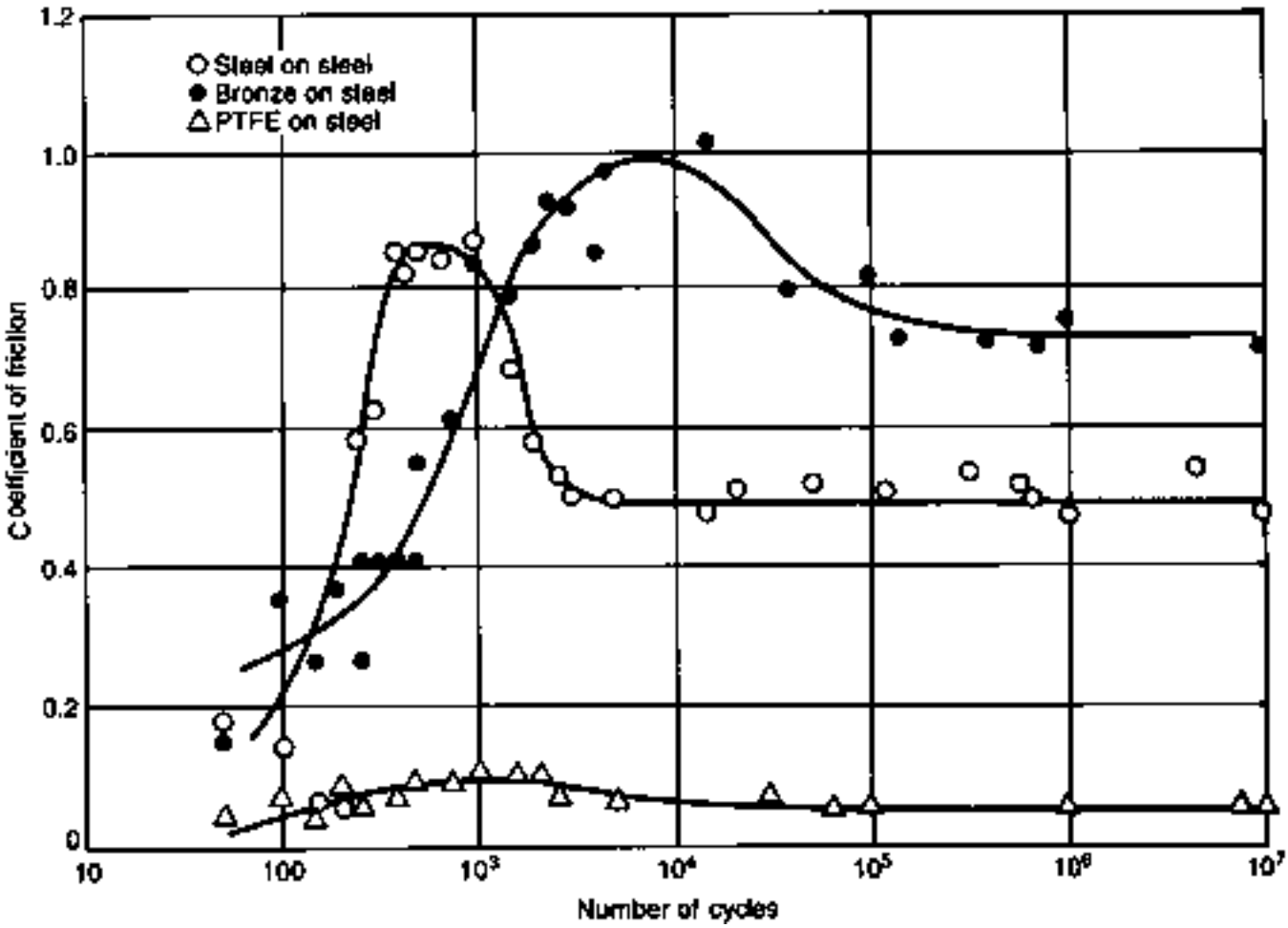


Fig. 32 Plot of the coefficient of friction versus the number of fretting cycles for three selected materials tested on steel. PTFE, polytetrafluoroethylene

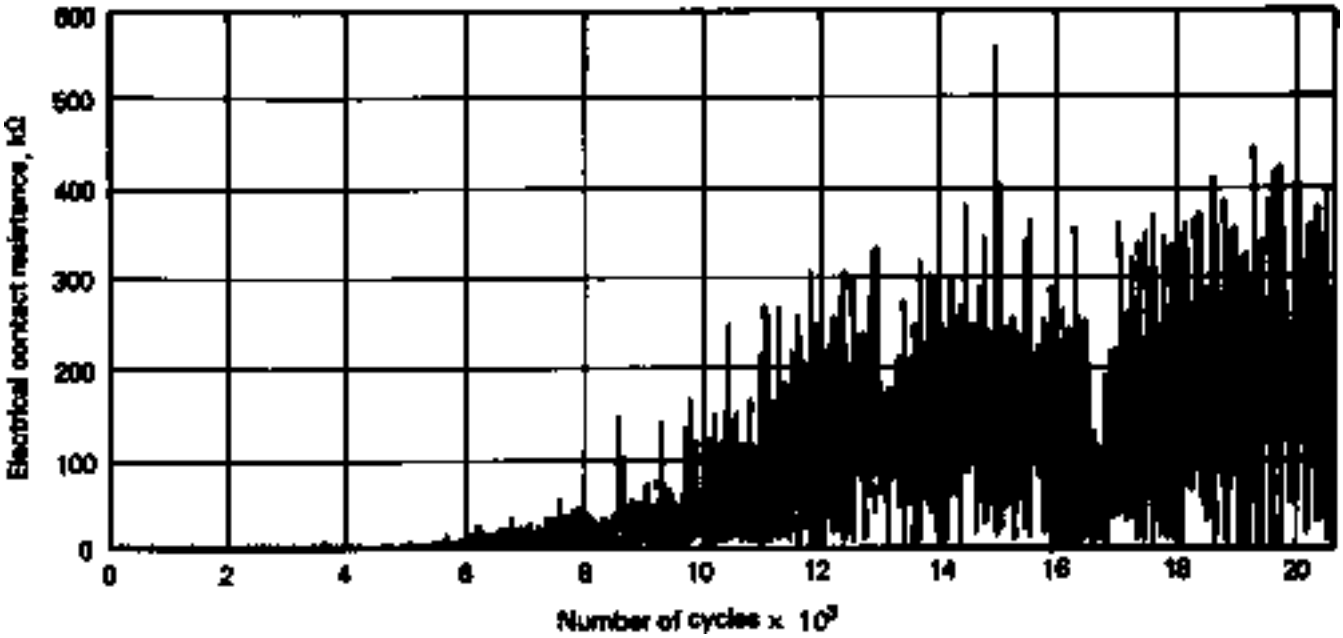


Fig. 33 Plot of contact resistance versus the number of fretting cycles

Observation of the debris on steel shows that it is platelike (Fig. 5). One of the early theories of fretting wear maintained that the debris arose by continual scraping and regrowth of oxide film (Ref 99). A modification of this idea suggests, from evidence of sections through a fretted surface on a carbon steel, that there are three well-defined zones (Ref 100):

- *Zone 1:* outer layer of compacted oxide
- *Zone 2:* severe deformation in which grains are comminuted and oriented in the fretting direction
- *Zone 3:* plastic deformation (from which it is concluded that in the steady-wear state, flakelike debris is formed by stripping oxide layers from the metal surface) (Ref 100)

The opposite point of view is that metal particles, which become progressively ground up and oxidized between the fretting surfaces, are removed. The evidence for this is that with reactive metals, such as aluminum and titanium, the debris is flakelike but contains metal (Ref 101). Furthermore, even after many millions of cycles, metal-to-metal contacts still occur in the fretting of steel flats (Ref 102). In experiments in which steel filings were introduced between two hard inert surfaces and subjected to fretting, the metal was soon converted to oxide (Ref 103). How the metal particles are generated is still a subject of controversy. However, in certain instances, there is evidence that the particles are generated in the later stages of a delamination process (Fig. 34).

Fig. 34 Typical delamination in the fretted region produced by metal-to-metal contact

It is possible that both processes, the removal of oxide and the removal of metal, could be occurring simultaneously in different regions of the surface. In the initial stage, peaks and valleys are formed due to local welding, and these variations appear to persist (Ref 104). The in-between areas become filled with compacted debris that has a layerlike structure, from which oxide flakes can become detached. These areas of compacted debris are maintained by metal particles produced by the continuing metal-to-metal contacts.
Although the oxide debris usually is harder than the metal from which it arises, abrasion by the oxide is not a contributory factor on steels to the wear. In fact, the debris, by limiting metal-to-metal contact and possibly taking up some of the movement between the surfaces, has more of a protective effect (Ref 60).

Prevention of Fretting Damage

The steps that can be taken to reduce or eliminate the damage due to fretting are extremely diverse, and each case or prospective case needs to be analyzed individually in order to select the most promising method to be applied. In this article, it is possible to give only some general indications of how die problems can be tackled. Reference 105 points out that it is useful to know whether the movement is “force controlled” or whether it is “displacement controlled.” Force-controlled fretting is experienced in the case of partial slip; therefore, increasing the coefficient of friction or increasing the normal load will reduce the slip. In displacement-controlled fretting, where slip is occurring over the whole interface, the amplitude is constant, and wear will be reduced by reducing the normal load as well as the coefficient of friction.
Improved Design. Because fretting arises only where there is relative movement between two surfaces, the elimination of this movement is a prime objective. Fretting often arises as a result of a stress concentration. In any case, reducing the stress via improved design (for example, raised wheel seats or stress-relieving grooves in hub/axle assemblies) is demanded by fatigue considerations. Increasing the normal pressure by reducing the area of contact (when the normal load is to be kept constant) or increasing the normal load will reduce the area of slip, but such action may introduce fatigue problems. The ultimate goal is to make the component in one piece and to get rid of the junction, but this is not always possible.
Surface Finish. Rough surfaces are less prone to fretting damage than highly polished surfaces are. Rough surfaces can be obtained by shot peening with glass beads or steel shot. Shot peening has the advantage of work hardening the surface, although the residual compressive stress appears to have little effect on the wear process; It has a considerable effect, however, if fatigue crack initiation and propagation are involved. Shot peening often is applied as a surface preparation for the application of coatings.
Coatings. Surface treatments that radically change the chemical composition of the surface can be divided into three categories:

- Foreign atoms or ions are introduced into the existing surface by diffusion (for example, carburizing, nitriding, chromizing, sherardizing, and aluminizing) or by bombardment (for example, ion implantation)
- The surface reacts with a chosen environment to form a compound (for example, oxidation, anodizing, and phosphating)
- An entirely foreign material is applied to the surface (for example, electrodeposition, plasma spray, ion plating, physical vapor deposition (PVD) and chemical vapor deposition (CVD)

Most of these coatings are harder and, consequently, more brittle than the substrate, which, if it is not strong enough, will plastically deform in the fretting contact, where local stresses can be high. This plastic deformation will crack and break up the coating. If the coating is particularly hard (for example, titanium nitride, with a hardness of 300 HV), further damage can occur by abrasion (Ref 106).
Carburizing and nitriding of steels is a well-established treatment for reducing wear in gears and is also effective against fretting (Ref 107). Electrodeposited coatings are used extensively in electrical contacts; gold alloys are the preferred coating because they are not subject to oxide formation, which is disastrous in its effect on conductivity (Ref 108).
The growth of oxide films at high temperatures has been specified to provide wear-resistant self-repairing coatings. These have been improved by ion implantation, particularly in titanium alloys, to give coatings that have low friction and low wear (Ref 90).

Anodizing is applied to aluminum alloys, and the hard coating provides protection against fretting, but it may cause wear of the opposing surface (Ref 109).

Inserts. Separation of the two surfaces by an insert, either a shim of a soft metal or a shim of a polymer with a low elastic modulus, sometimes will be effective. The intention is to take up the movement by either plastic deformation or elastic deformation of the material. A combination of metal and polymer can be more effective because it combines the advantages of both materials—the low elasticity of the polymer and the good conductivity of the metal. Bronze-filled polytetrafluorethylene (PTFE), when fretted against steel, has a very low friction coefficient and undetectable wear (Ref 110).

Lubricants are usually difficult to apply in fretting contacts because of the difficulty of maintaining the lubricant in the contact, but lubricants are included in the construction of steel ropes (for example, locked coil ropes) (Ref 111). In a survey of the effect of additives to a base mineral on the fretting wear of steel surfaces, zinc di-n-octyldithiophosphate was found to be most effective (Ref 112). Solid lubricants applied to the surfaces (for example, MoS_2) reduce the coefficient of friction initially but eventually tend to wear away (Ref 113).

References cited in this section

1. E.M. Eden, W.N. Rose, and F.L. Cunningham, Endurance of Metals, *Proc. Inst. Mech. Eng.*, Vol 4, 1911, p 839–974

2. G.A. Tomlinson, The Rusting of Steel Surfaces in Contact, *Proc. R. Soc. (London) A*, Vol A115, 1927, p 472–483

3. E.J. Warlow-Davies, Fretting Corrosion and Fatigue Strength, *Proc. Inst. Mech. Eng.*, Vol 146, 1941, p 32–38

4. J.R. McDowell, Fretting Corrosion Tendencies of Several Combinations of Materials, *Symposium on Fretting Corrosion*, STP 144, ASTM, 1953, p 2439

6. K.T. Styles, The Influence of Fret Corrosion upon the Fatigue Life of Machine Components, *Met. Australas.*, March 1980, p 4–6

7. J.F. Warriner, “Thin Shell Bearings for Medium Speed Diesel Engines,” No. 364, Diesel Engineers and Users Association, Feb 1975

8. R. Heinz, Fretting Wear—Examples, Test Methods and Preventive Measures, *Mater. Wiss. Werkstofftech.*, Vol 20, 1989, p 14–20

9. R.G. Mitchell, An Airline View of the Corrosion Problem, *Ind. Corros.*, July 1983, p 11–17

10. P. Greenfield and R.W. Suhr, The Factors Affecting the High-Cycle Fatigue Strength of Low-Pressure Turbine and Generator Rotors, *Proc. of International Conference on Fatigue of Engineering Materials and Structures*, 15–19 Sept 1986 (Sheffield), Inst. Mech. Eng., Vol 1, 1986, p 165–186

11. H.W. Fricker, Fretting in Tube Supports of Heat Exchangers, *Proc. of Conference on Component Design in High Temperature Reactors Using Helium as a Coolant*, Nuclear Engineering Group of Institution of Mechanical Engineers, 3–4 May 1972, p 213–231

12. H.J. Connors, Flow-Induced Vibration and Wear of Steam Generator Tubes, *Nucl. Technol.*, Vol 55, Nov 1981, p 311—331

13. P.J.W. Bongers, P.L.M. Dubs, and T.J.M. Linssen, Fatigue Crack Initiation on the Outside Surface of High-Pressure Tubes Caused by Tube Support, *Proc. of Conference on Safety in High Pressure*

Polyethylene Process, Symposium III, March 1977 (Houston), American Institute of Chemical Engineers

14. M. Antler, Electrical Effects of Fretting Connector Contact Materials: A Review, *Wear,* Vol 106, 1985, p 5–33

15. P.J. Kennedy, M.B. Peterson, and L. Stallings, An Evaluation of Fretting at Small Slip Amplitudes, *Materials Evaluation under Fretting Conditions,* STP 780, ASTM, 1982, p 30–48

16. P.J. Kennedy, L. Stallings, and M.B. Peterson, A Study of Surface Damage at Low-Amplitude Slip, *ASLE Trans.,* Vol 27 (No. 4), 1984, p 305–312

17. K.L. Johnson, Surface Interaction between Elastically Loaded Bodies under Tangential Forces, *Proc. R. Soc. (London) A,* Vol 230, 1955, p 531–548

19. M. Kuno, "Fretting Damage of High Carbon Chromium Steel," Ph.D. thesis, University of Nottingham, 1988

20. S. Goto and R.B. Waterhouse, Fretting and Fretting-Fatigue of Titanium Alloys under Conditions of High Normal Load, *Proc. 4th International Conference on Titanium,* TMS/AIME, Vol 3, 1980, p 1337–1384

21. O. Vingsbo and S. Söderberg, On Fretting Maps, *Wear,* Vol 126, 1988, p 131–147

22. U. Bryggman and S. Söderberg, Contact Conditions in Fretting, *Wear,* Vol 110, 1986, p 1–17

23. F.P. Bowden and D. Tabor, *The Friction and Lubrication of Solids,* Oxford University Press, 1950, p 19

24. I.M. Feng and H.H. Uhlig, Fretting Corrosion of Mild Steel in Air and Nitrogen, *J. Appl. Mech.,* Vol 21, 1954, p 395–400

25. R.B. Waterhouse, Fretting, *Treatise on Materials Science and Technology,* D. Scott, Ed., Academic Press, 1978, p 259–286

27. H. Kongsjorden, J. Kulsetas, and J. Sletbak, Degradation of Electrical Contacts Caused by Oscillatory Micromotion between the Contact Members, *Proc. 9th International Conference on Electrical Contact Phenomena,* 11–15 Sept 1978 (Chicago)

28. M. Antler, Electrical Effects of Fretting Connector Contact Materials: A Review, *Wear,* Vol 106, 1985, p 5–33

29. M. Braunovic, M.S. McIntyre, W.J. Chaurion, and I. Aitchison, Surface Analysis of Fretting Damage in Electrical Contacts of Aluminium with Different Contacting Materials, *Electrical Contacts 1983, Proc. 29th Holm Conference on Electrical Contacts (Chicago),* Illinois Institute of Technology, p 232–242

30. J.J. Mottine and B.T. Reagor, Investigation of Fretting Corrosion at Dissimilar Metal Interfaces in Socketed IC Device Applications, *Electrical Contacts 1983, Proc. 29th Holm Conference on Electrical Contacts (Chicago),* Illinois Institute of Technology, p 67–71

31. M. Kuno and R.B. Waterhouse, The Effect of Oscillatory Direction on Fretting Wear under Crossed Cylinder Contact Conditions, *Eurotrib 89, Proc. 5th International Congress on Tribology,* 12–15 June 1989 (Helsinki), Vol 3, p 30–35

32. K.B. Elliott, H.H. Mabie, M.J. Furey, and L.D. Mitchell, A Vibration Analysis of a Bearing/Cartridge Interface for a Fretting Corrosion Study, *J. Lubr. Technol. (Trans. ASME),* Vol 105, 1983, p 518–525

33. N.L. Golego, M.N. Rozhkov, V.P. Onoprienko, A.S. Nakrakui, and N.V. Kozii, Wear of Materials in Fretting Corrosion in Connection with Conditions of Vibrocontact Interaction, *Fiz. Khim. Mekh. Mater.,* Vol 21, 1985, p 71–75

34. J.H. Cha, M.W. Wambsganss, and J.A. Jendrzejcyk, Experimental Study on Impact/Fretting Wear in Heat Exchanger Tubes, *J. Pressure Vessel Technol. (Trans. ASME),* Vol 109, 1987, p 265–274

35. A.W.J. de Gee, C.P.L. Commissaris, and J.H. Zaat, The Wear of Sintered Aluminium Powder (SAP) under Directions of Vibrational Contact, *Wear,* Vol 7, 1964, p 535–550

36. P.L. Ko, Experimental Studies of Tube Fretting in Steam Generators and Heat Exchangers, *J. Pressure Vessel Technol. (Trans. ASME),* Vol 101, 1979, p 125–133

37. H.J. Connors, Flow-Induced Vibration and Wear in Steam Generator Tubes, *Nucl. Technol.,* Vol 55, 1981, p 311–331

38. S.F. Calhoun, "Effects of Metal Hardness and Surface Finish upon Fretting Corrosion," Technical Report 63-1835, Rock Island Arsenal, 28 May 1963

39. R.B. Waterhouse, Residual Stresses and Fretting, *International Guidebook on Residual Stresses,* A. Niku-Lari, Ed., Pergamon, Oxford, 1987, p 511–525

40. C.A. Bergman, R.C. Cobb, and R.B. Waterhouse, The Effect of Shot-Peening on the Friction and Wear of a Carbon Steel in Fretting Conditions, *Wear of Materials,* American Society of Mechanical Engineers, 1987, p 33–37

41. J. Labedz, Effect of Residual Stresses on Wear of the Surface Layer of Components Operating under Fretting Conditions, *Metal Treatments against Wear, Corrosion, Fretting and Fatigue,* A. Niku-Lari and R.B. Waterhouse, Ed., Pergamon, Oxford, 1988, p 87–98

42. K. Endo, Practical Observations of Initiation and Propagation of Fretting Fatigue Cracks, *Fretting Fatigue,* R.B. Waterhouse, Ed., Applied Science Publishers, London, 1981, p 127–141

43. A. Yournadj, G. Maeder, and A. Moison, Geometric and Mechanical Aspects of Fretting Corrosion, *Proc. 4th European Tribology Congress,* 9–12 Sept 1985 (Euiby, France), Vol 2, p 1–9

44. G. Fair, B. Noble, and R.B. Waterhouse, The Stability of Compressive Stresses Induced by Shot-Peening under Conditions of Fatigue and Fretting-Fatigue, *Advances in Surface Treatments,* Vol I, A. Niku-Lari, Ed., Pergamon, Oxford, 1984, p 3–8

45. M. Kuno, R.B. Waterhouse, D. Nowell, and D.A. Hills, Initiation and Growth of Fretting Fatigue Cracks in the Partial Slip Regime, *Fatigue Fract. Eng. Mater. Struct.,* Vol 12, 1989, p 387–398

46. T. Kayaba and A. Iwabuchi, Effect of the Hardness of Hardened Steels and the Action of Oxides on Fretting Wear, *Wear,* Vol 66, 1981, p 27–41

47. O. Vingsbo, M. Odfalk, and Ning-E Shen, Fretting Maps and Fretting Behaviour of Some FCC Metal Alloys, *Wear,* Vol 13, 1990, p 153–167

48. J. Dobromirski and I.O. Smith, Metallographic Aspects of Surface Damage, Surface Temperature and Crack Initiation in Fretting Fatigue, *Wear,* Vol 117, 1987, p 347–357

49. J. Beard and P.F. Thomason, The Mechanics of Fatigue Crack Nucleation in a Low Alloy Steel under Fretting Conditions, *Fatigue 87,* Conference Proceedings, 28 June–3 July 1987 (Charlottesville, VA), Vol 1, 1987, p 53–62

50. R. Becker and G. Sepold, Fretting of Amorphous Alloys, *Rapidly Solidified Amorphous and Crystalline Alloys,* B.H. Kear, B.C. Giessen, and M. Cohen, Ed., Elsevier Science Publishing, 1982, p 295–299

51. J. Sato, M. Sato, and S. Yamamoto, Fretting Wear of Stainless Steel, *Wear,* Vol 69, 1981, p 167–177

52. P.A. Higham, F.H. Stott, and B. Bethune, Mechanisms of Wear of the Metal Surface during Fretting Corrosion of Steel on Polymers, *Corros. Sci.,* Vol 18, 1978, p 3–13

53. P.A. Higham, F.H. Stott, and B. Bethune, The Influence of Polymer Composition on the Wear of the Metal Surface during the Fretting of Steel on Polymer, *Wear,* Vol 47, 1978, p 71–80

54. O. Jacobs, K. Friedrich, G. Marom, K. Schulte, and H.D. Wagner, Fretting Wear Performance of Glass-Carbon and Aramid Fibre/Epoxy-and PEEK-Composites, *Wear of Materials,* 1989, American Society of Mechanical Engineers, 1989, p 495–500

55. M. Kuno, R.B. Waterhouse, and B.R. Pearson, Fretting Wear in Sintered Alumina and Tungsten Carbide Cermet-Metal Couples, *Wear of Materials,* 1987, American Society of Mechanical Engineers, 1987, p 371–380

56. D. Klaffke Fretting Wear of Ceramics, *Tribol. Int.,* Vol 22, 1989, p 89–101

57. D.R. Horn, R.B. Waterhouse, and B.R. Pearson, Fretting Wear of Sintered Alumina, *Wear,* Vol 113, 1986, p 225–232

58. D. Godfrey, "Investigation of Fretting by Microscopic Observation," Report 1009, NACA, 1951

59. J. Sato, M. Shima, and T. Sugiwara, A Fundamental Study of Fretting Damage to Glass Using an Improved Apparatus, *Wear,* Vol 106, 1985, p 53–61

60. M. Godet, The Third Body Approach: Mechanical View of Wear, *Wear,* Vol 100, 1984, p 437–452

61. J.F. Andrews, P.D. Donovan, and J. Stringer, Fretting Corrosion Products of Aluminium Alloys, *Br. Corros. J.,* Vol 3, 1968, p 85–87

62. C. Colombié, Y. Berthier, A. Floquet, L. Vincent, and M. Godet, Fretting: Load Carrying Capacity of Wear Debris, *J. Tribol. (Trans. ASME),* Vol 106, 1984, p 194–201

63. D. Aldham, J. Warburton, and R.E. Pendlebury, The Unlubricated Fretting Wear of Mild Steel in Air, *Wear,* Vol 106, 1985, p 177–201

64. J. Warburton, The Fretting of Mild Steel in Air, *Wear,* Vol 131, 1989, p 365–386

65. R.E. Pendlebury, Unlubricated Fretting of Mild Steel in Air at Room Temperature, Part II: Electrical Contact Resistance Measurements and the Effect of Wear on Intermittent Loading, *Wear,* Vol 118, 1987, p 341–364

66. A. Iwabuchi, T. Kayaba, and K. Kata, Effect of Atmospheric Pressure on Friction and Wear of 0.45% C Steel in Fretting, *Wear,* Vol 91, 1983, p 289–305

67. A. Iwabuchi, K. Kato, and T. Kayaba, Fretting Properties of SUS304 Stainless Steel in Vacuum Environment, *Proc. JSLE Tribology Conference,* 8–10 July 1985 (Tokyo), Vol 1, Japan Society of Lubrication Engineers, 1985, p 29–34

68. K. Kato, T. Suzuki, A. Iwabuchi, and K. Hokkirigawa, Fretting Damage of SUS304 in a Vacuum, *JSME Int. J.,* Vol 30, 1987, p 330–336

69. A. Iwabuchi, H. Matsuzaki, and K. Hori, The Effect of Ambient Pressure on the Tribological Properties of an Amorphous Alloy in Fretting, *Wear of Materials 1989,* American Society of Mechanical Engineers, 1989, p 595–600

70. H. Goto and D.H. Buckley, The Influence of Water Vapour in Air on the Friction Behaviour of Pure Metals during Fretting, *Tribol. Int.,* Vol 18, 1985, p 237–245

71. H. Goto, M. Ashida, and K. Endo, The Influence of Oxygen and Water Vapour on the Friction and Wear of an Aluminum Alloy under Fretting Conditions, *Wear,* Vol 116, 1987, p 141–155

72. M.P. Sherwin, D.E. Taylor, and R.B. Waterhouse, An Electrochemical Investigation of Fretting Corrosion in Stainless Steel, *Corros. Sci.,* Vol 11, 1971, p 419–429

73. S.D. Cook, G.J. Gianoli, A.J.T. Clemow, and R.J. Haddad, Fretting Corrosion in Orthopaedic Alloys, *Biomater. Med. Dev. Art. Org.,* Vol 11, 1983–84, p 281–292

74. S.A. Brown and K. Merritt, Fretting Corrosion in Saline and Serum, *J. Biomed. Mater. Res.,* Vol 15, 1981, p 479–488

75. S.A. Brown and J.P. Simpson, Crevice and Fretting Corrosion of Stainless Steel Plates and Screws, *J. Biomed. Mater. Res.,* Vol 15, 1981, p 867–878

76. S.A. Brown and K. Merritt, Fretting Corrosion of Plates and Screws: An in vitro Test Method, *Corrosion and Degradation of Implant Materials: Second Symposium,* STP 859, A.C. Fraker and C.D. Griffin, Ed., ASTM, 1985, p 105–116

77. K. Merritt and S.A. Brown, Effect of Proteins and PH on Fretting Corrosion and Metal Ion Release, *J. Biomed. Mater. Res.,* Vol 22, 1988, p 111–120

78. M.P. Overs and R.B. Waterhouse, Fretting Wear of Two Weldable Structural Steels (HY80 and 50D) in Seawater, *Wear of Materials 1983,* American Society of Mechanical Engineers, 1983, p 541–545

79. B.R. Pearson and R.B. Waterhouse, The Fretting Corrosion in Seawater of Materials Used in Offshore Structures, *Proc. 9th International Congress on Metallic Corrosion,* 5–9 June 1984 (Toronto), Vol 2, NRC Canada, 1984, p 334–341

80. B.R. Pearson, P.A. Brook, and R.B. Waterhouse, Fretting in Aqueous Media, Particularly of Roping Steels in Seawater, *Wear,* Vol 106, 1985, p 225–260

81. M.W.J. Lewis and C.S. Campbell, Fretting and Wear of Stainless and Ferritic Steels in Liquid-Metal Fast Breeder Reactor Steam Generators, *Nucl. Technol.,* Vol 55, 1981, p 460–469

82. P.L. Hurricks and K.S. Ashford, The Effect of Temperature on the Fretting Wear of Mild Steel, *Proc. Inst. Mech. Eng.,* Vol 184, Pt. 3L, 1969–70, p 165–175

83. P.L. Hurricks, The Fretting Wear of Mild Steel from Room Temperature to 200 °C, *Wear,* Vol 19, 1972, p 207–229

84. P.L. Hurricks, The Fretting Wear of Mild Steel from 200 to 500 °C, *Wear,* Vol 30, 1974, p 189–212

85. F.H. Stott, D.S. Lui, and G.C. Wood, The Structure and Mechanism of the "Glaze" Oxide Layers Produced on Nickel-Based Alloys during Wear at High Temperatures, *Corros. Sci.,* Vol 13, 1973, p 449–469

86. M.M. Hamdy and R.B. Waterhouse, The Fretting-Fatigue Behaviour of a Nickel-Based Alloy (Inconel 718) at Elevated Temperatures, *Wear of Materials 1979,* American Society of Mechanical Engineers, 1979, p 351–355

87. M.M. Hamdy and R.B. Waterhouse, The Fretting Wear of Ti-6Al-4V and Aged Inconel 718 at Elevated Temperatures, *Wear,* Vol 71, 1981, p 237–248

88. A. Iwabuchi, Fretting Wear of Inconel 625 at High Temperature and in High Vacuum, *Wear,* Vol 106, 1985, p 163–175

89. M.M. Hamdy and R.B. Waterhouse, The Stability of the Protective Glaze Oxide Formed during High Temperature Fretting of an Aged Nickel-Based Alloy, Inconel 718, *Wear of Materials 1983,* American Society of Mechanical Engineers, 1983, p 546–549

90. A. Iwabuchi and R.B. Waterhouse, The Effect of Ion Implantation on the Fretting Wear of Four Titanium Alloys at Temperatures up to 600 °C, *Wear of Materials 1985,* American Society of Mechanical Engineers, 1985, p 471–476

91. M.M. Hamdy, B. Noble, and R.B. Waterhouse, Fatigue Crack Initiation in IMI829 Caused by High Temperature Fretting, *J. Mater. Sci.,* Vol 18, 1983, p 493–502

92. D.G. Chetwynd, F.A. McKee, and C.H. Wride, The Measurement of Wear Damage, *Proc. Inst. Mech. Eng.,* Vol 201 (No. 4C), 1987, p 251–258

93. J.T. Atkinson and M.J. Lalor, Holographic Interferometry Applied to Minimal Wear Measurement, 1st European Congress, Appl. to Metrology, *SPIE J.,* Vol 136, 1977, p 107–113

94. T.W. Conlon, The Scope for TLA in Corrosion and Wear Studies, *Metall. and Mater. Technol.,* Vol 2, 1979, p 143–146

95. J. Warburton and R. Bradford, The Progressive Wear of Tubes: The Volumes of the Intersections of Cylinders with Each Other and with Flats, *Wear,* Vol 110, 1986, p 331–352

96. P.L. Hurricks, The Mechanism of Fretting—A Review, *Wear,* Vol 15, 1970, p 389–409

97. D. Aldham, J. Warburton, and R.E. Pendlebury, The Unlubricated Fretting of Mild Steel in Air, *Wear,* Vol 106, 1985, p 177–201

98. R.B. Waterhouse, The Role of Adhesion and Delamination in the Fretting Wear of Metallic Materials, *Wear,* Vol 45, 1977, p 355–364

99. H.H. Uhlig, Mechanism of Fretting Corrosion, *J. Appl. Mech.,* Vol 21, 1954, p 401–407

100. X. Zhang, C. Zhang, and C. Zhu, Slip Amplitude Effects and Microstructural Characteristics of Surface Layers in Fretting Wear of Carbon Steel, *Wear,* Vol 134, 1989, p 297–309

101. R.B. Waterhouse, Fretting Corrosion and Fretting Fatigue in the Transport Industry, *Proceedings of the 6th European Congress on Metallic Corrosion,* 19–23 Sept 1977 (London)

102. R.E. Pendlebury, Unlubricated Fretting and Sliding Wear of Steels in Air, *Tribology—Friction, Lubrication and Wear, Fifty Years On,* Conference Proceedings, 1–3 July 1987 (London), Institute of Mechanical Engineers, Vol 1, 1987, p 267–275

103. R.E. Pendlebury and P.A. Tempest, "Unlubricated Fretting Wear of Mild Steel Surfaces in Air at Room Temperature, Part III: The Conversion of (Artificial) Metallic Wear Particles to Oxide," Report TPRD/B/0985/R87, AGR/CIF/TWG/P(87)232, Central Electricity Generating Board, Oct 1987

104. R.E. Pendlebury, "Unlubricated Fretting Wear of Mild Steel Surfaces in Air at Room Temperature," Report TPRD/B/0674/N85, AGR/CIF/TWG/P(85)164, Central Electricity Generating Board, June 1985

105. J. Beard, The Rational Selection of Palliatives for Avoidance of Fretting, *Tribology—Friction, Lubrication and Wear, Fifty Years On,* Conference Proceedings, 1–3 July 1987 (London), Institute of Mechanical Engineers, Vol 1, 1987, p 311–319

106. R.C. Cobb and R.B. Waterhouse, The Fretting Wear of Certain Hard Coatings Including TiN Applied to a 0.4C Steel, *Tribology—Friction, Lubrication and Wear, Fifty Years On,* Conference Proceedings, 1–3 July 1987 (London), Institute of Mechanical Engineers, Vol 1, 1987, p 303–310

107. J.C. Gregory, A Salt Bath Treatment to Improve the Resistance of Ferrous Metals to Scuffing, Wear, Fretting, and Fatigue, *Wear,* Vol 9, 1966, p 249–281

108. J.W. Souter and W. Staunton, The Fretting Wear of Electrodeposited Contact Coatings, *Trans. Inst. Met. Finish.,* Vol 66, 1988, p 8–14

109. R.B. Waterhouse, The Effect of Surface Treatments on the Fretting Wear of an Aluminum Alloy (RR58)/Steel (BS970 080 M40) Couple, *Proc. of Conference on Surface Engineering, Current Trends and Future Prospects,* 25–27 June 1990 (Toronto), S.A. Meguid, Ed., Elsevier Applied Science, 1990, p 325–334

110. D.J. Varley and R.B. Waterhouse, Fretting of Bronze-Filled PTFE against Mild Steel, *Proc. Japanese International Tribology Conference,* 29 Oct–1 Nov 1990 (Nagoya) Japanese Society of Tribologists, Vol 2, 1990, p 821–825

111. D.E. Taylor and R.B. Waterhouse, Fretting Fatigue in Steel Ropes, *Lubr. Eng.,* Vol 27, 1971, p 123–127

112. J. Sato, M. Shima, T. Sugiwara, and A. Tahara, Effect of Lubricants on Fretting Wear of Steel, *Wear,* Vol 125, 1988, p 83–95

113. D. Godfrey and E.E. Bisson, "Bonding of Molybdenum Disulphide to Various Materials to Form a Solid Lubricating Film, Part II, Friction and Endurance Characteristics of Films Bonded by Practical Methods," Technical Note 2802, NACA, 1952

Examples of Fretting Failures

When steels parts are lubricated properly, the presence of oxide debris often is indicative of fretting. If a part is unlubricated or experiences a lack of lubrication, oxide debris may not necessarily signify fretting, but rather, other types of wear that can be prevented by proper lubrication. In contrast, fretting may not be completely

remedied by lubrication (Ref 114). This doesn't mean that fretting damage cannot be mitigated by lubrication. In wire ropes, for example, fretting between strands can be prevented by proper lubrication (Ref 114).

Fretting differs from ordinary wear further in that the bulk of the debris produced is retained at the site of fretting. In ferrous materials, the fretting process creates a mass of reddish oxide particles. Fretting also occurs in nonoxidizing materials, such as gold, platinum, and cupric oxide. Fretting debris formed by many nonferrous alloys is largely unoxidized and larger in particle size than that of ferrous materials. On the other hand, in hard materials such as a tool steel and chromium, the initial wear particles are very small, with much oxide present (114). Little debris is present in an inert or protective atmosphere, although surface damage can be extensive. Fretting differs from ordinary wear in that the bulk of the debris produced is retained at the site of fretting.

If fretting is present, macroscopic examination of the affected surface area can be done after the corrosion has been scrubbed off with a paste of alumina and hexane, for example. The examination will reveal very reflective areas, as well as depressions and pits containing black patches of Fe_3O_4, which is produced when the oxygen supply is limited. If only ordinary corrosion is present, microscopic examination will reveal small but well-defined corrosion pits that have produced rust rosettes consisting of the voluminous hydrated iron oxide (Ref 114). Figure 35 shows a screw hole with fretting and fretting corrosion of a type 316LR stainless steel bone plate for an orthopedic implant. The overview (Fig. 35a) of wear on the plate hole shows mechanical and pitting corrosion attack, shallow pitting corrosion attack from the periphery of the contact area (Fig. 35b), and the effects of mechanical material transport (material tongue in right upper corner), fretting and wear (burnished areas), and corrosion (shallow pitting) in Fig. 35(c).

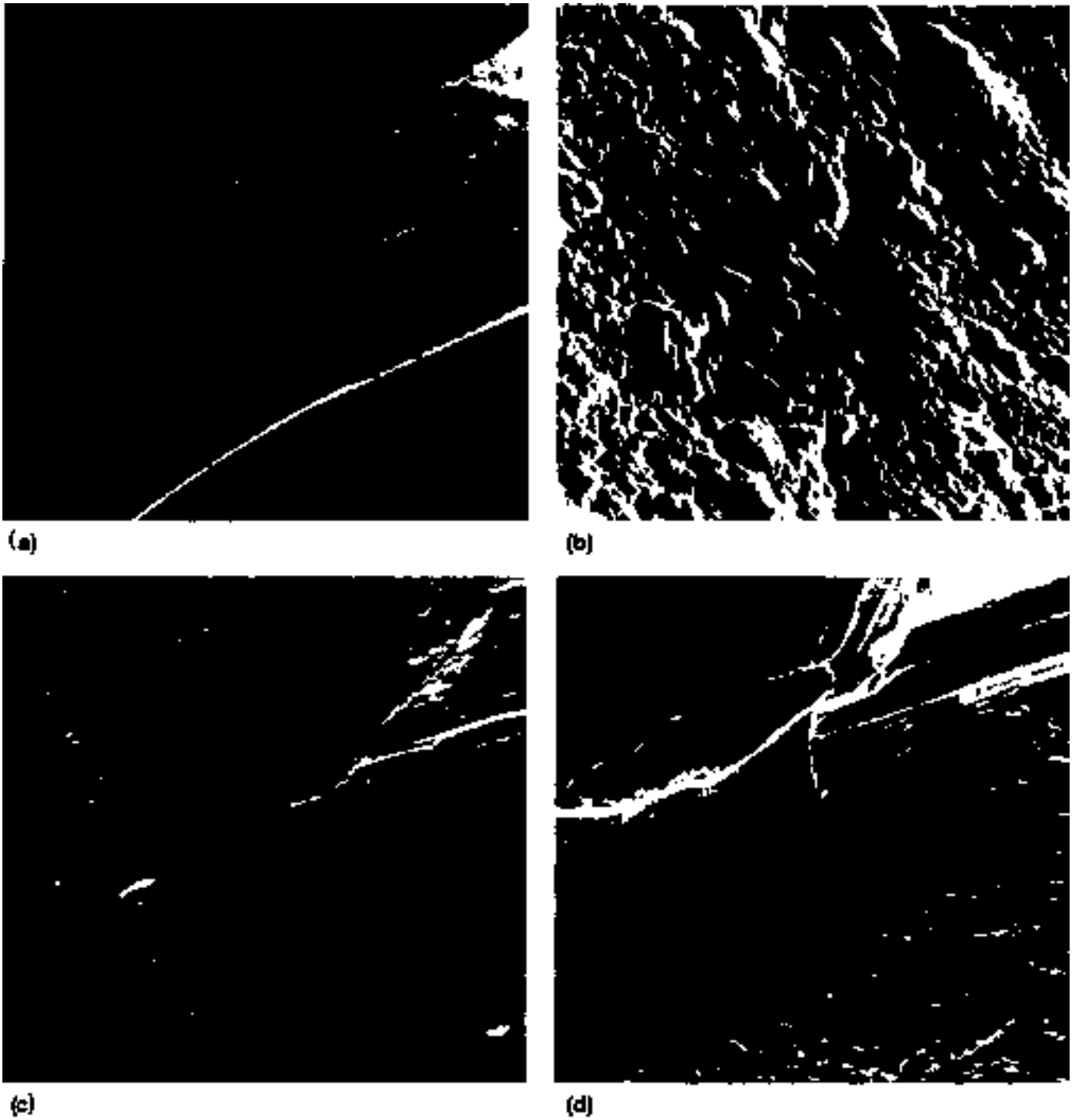


Fig. 35 Fretting and fretting corrosion at the contact area between the screw hole of a type 316LR stainless steel bone plate and the corresponding screw head. (a) Overview of wear on plate hole showing mechanical and pitting corrosion attack. 15×. (b) Higher-magnification view of shallow pitting corrosion attack from periphery of contact area. 355×. (c) Contact area showing the effects of mechanical material transport (material tongue in right upper corner), fretting and wear (burnished areas), and corrosion (shallow pitting). 355×. (d) Higher-magnification view of Fig. 30(c). 650×

Another example in Fig. 36(a) shows a portion of a titanium screw head implant with a lip of material that was transported by fretting at a plate-hole edge. A flat fretting zone is visible on the screw surface over the material lip, which is magnified in Fig. 36(b). A cellular wear structure containing wear debris is present. Figure 36(c) shows material destruction and the formation of thin flakes of wear debris. Mechanical deformation is superimposed on wear structures in Fig. 36(d). No morphological signs of corrosion have been observed in connection with fretting structures. This has been typically found on all contact areas of plate holes and screw heads of the titanium implants that were investigated.

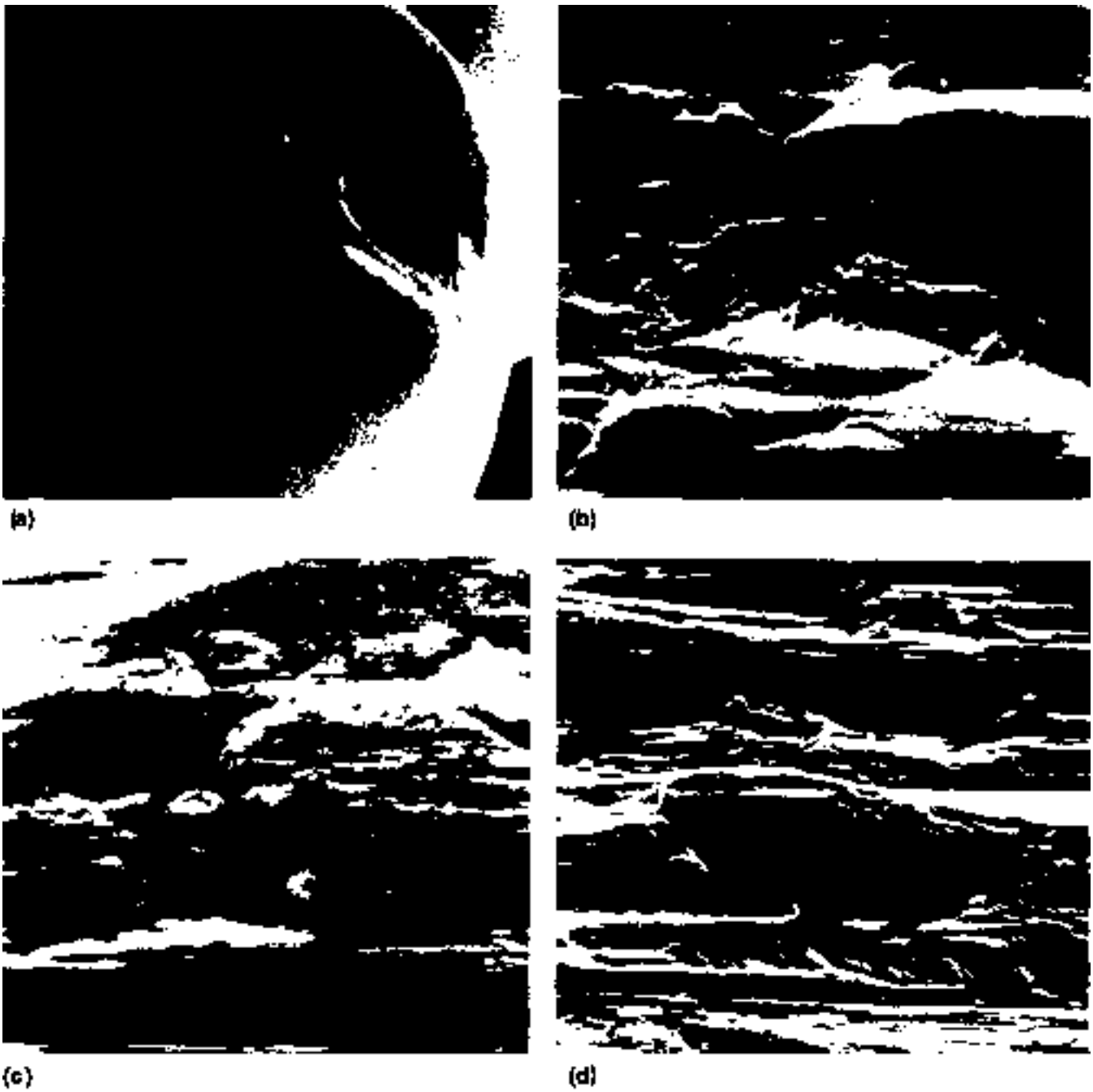


Fig. 36 Wear on head of titanium screw. (a) Material transport and fretting zone. (b) Close-up view of wear structures showing fine wear products. 120×. (c) Wear structures showing generation of small wear particles. 1200×. (d) Wear structures with additional fretting structures. 305×

Fretting in Rolling-Element Bearings

False Brinelling. The term *false brinelling* refers to a type of fretting-type damage that occurs in machine components such as ball bearings, where craters caused by the vibration of the ball against the race are circular and resemble Brinell impressions. Another type of fretting damage observed in machinery is known as *fitting rust,* where gages or shims are clamped together and experience vibration. Fitting rust is the contact corrosion that takes place between the bore of the bearing and the shaft or between the outside surface of the bearing and the bore of the housing.

False Brinelling is caused by vibrations or oscillations over a few degrees of arc between rolling elements and raceways in a nonrotating bearing. At the contact areas between the rolling elements and raceways, lubricant is squeezed out, resulting in metal-to-metal contact and localized wear. False Brinelling does not occur during normal running but is found in the bearings of machines subjected to vibration while at rest. False Brinelling also has been found to occur during transit. Transport by rail or truck is conducive to this form of damage. Even

transport by sea does not guarantee immunity, particularly where pronounced vibrations from diesel engines may be present. False Brinelling also has been found in the roller-bearing axle boxes of railway cars that have been standing for some time in sidings and that are subjected to vibrations from passing trains. When equipment is to be transported, the machine should be dismantled, and one of the following preventive measures should be adopted: remove the rolling-element bearings and fit temporary wooden packings in their place, or use clamps to lock rotors of electric motors rigidly to the frames and, at the same time, relieve the bearing of the dead-weight load. In other instances, the machinery could be packed so that the axis of the rotating portion is in the vertical plane, thereby preventing damage by stationary indentation.

There are several features that permit false Brinelling to be distinguished from true Brinelling. True Brinelling, or denting, is permanent deformation produced by excessive pressure or impact loading of a stationary bearing. Improper mounting procedures, such as forcing of a tight-fitting or cocked outer ring into a housing, are common causes of true Brinelling. Improper handling, such as dropping or pounding of bearings, will cause true Brinelling. True Brinelling also can be caused by vibrations in ultrasonic cleaning.

True Brinelling is characterized by hollows or dents produced by plastic flow of metal. Such indentations in raceways will be spaced in correspondence to ball or roller spacing, thus differentiating true Brinelling from dents made by debris, such as dirt or chips, which usually are spaced randomly. False Brinelling is found on those regions of a bearing that are subjected to the heaviest loads, for example, at the six o'clock position in a bearing supporting a horizontal shaft; the effect diminishes in a fairly regular manner on either side of this position. Figure 37 is a comparison of true Brinelling with false Brinelling. The final grinding marks made during the manufacture of raceways can be seen, virtually undisturbed, within the true Brinelled areas of Fig. 37.

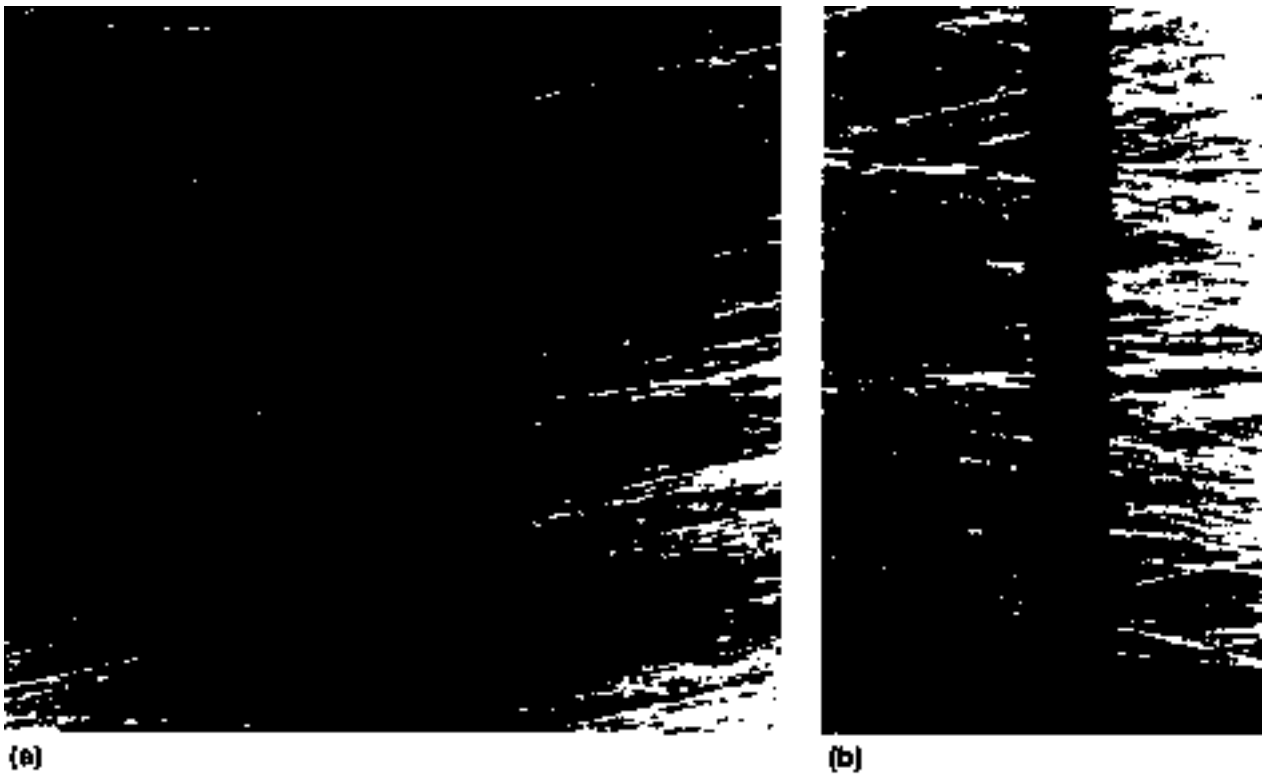


Fig. 37 Comparison of true Brinelling with false Brinelling. (a) In true Brinelling, the surface texture is essentially undisturbed. Note grinding marks in the impression. (b) In false Brinelling (fretting), the surface is worn away or material is transferred. No grinding marks are visible in the impression. Both 15×. Source: Ref 116

In many instances, false Brinelling remains undetected until the machine is started, at which time noisy operation is immediately evident. The phenomenon is thought to be a manifestation of fretting. The minute movements initiated by vibration cause wear of the contact surfaces, and the fine particles produced rapidly oxidize and result ultimately in production of characteristic grooves, with the oxide acting as an abrasive. To prevent false Brinelling in bearings of standby equipment, it may be necessary to arrange for continuous slow rotation of shafts while nearby machines are running. For a given angular rotation of a shaft resulting from

vibration, slightly greater movements of rolling elements relative to raceways will occur in a bearing with a greater number of elements. These movements may be greater than those at which false Brinelling develops. Therefore, where false Brinelling is not severe, changing to a bearing with a larger number of rolling elements or to one of the needle-roller types may be sufficient.

The corrugated surfaces produced by fretting or false Brinelling form stress raisers that, under subsequent conditions of rotation, may produce excessive noise and may cause premature spalling by rolling-contact fatigue. Figure 38 shows a portion of a shaft that served as the inner raceway for a drawn-cup needle-roller bearing. The rollers left deep and clearly defined impressions on the shaft. The damage was identified as fretting, or false Brinelling, because of the dull surface with little or no trace of the original surface finish remaining at the bottoms of the indentations, although the shaft also was Brinelled by a very heavy overload.

Fig. 38 Severe damage from fretting (false Brinelling) on the surface of a shaft that served as the inner raceway for a needle-roller bearing

Recommendations for reducing fretting in rolling-element bearings include:

- Keep radial play in the bearings at the lowest practical value.
- Increase the angle of oscillation (if possible) to secure roller or ball overlap in order to drag fresh lubricant into the area. If the surfaces can be separated by lubricant, fretting of the metal cannot occur.
- Relubricate frequently to purge the red iron oxide debris and reinstate the lubricating film.
- Use a larger bearing of higher capacity to reduce contact loads.
- Increase the hardness of the elements as much as possible. The commercial antifriction bearing already has fully hardened components. For best results, the shaft, if used as an inner raceway, should be hardened to 58 HRC minimum.
- Use a grease that has been specially formulated to provide maximum feeding of lubricant to areas susceptible to fretting damage. If the bearing is oil lubricated, flood it if possible.

Example 1: Fretting Failure of Raceways on 52100 Steel Rings of an Automotive Front- Wheel Bearing. The front-wheel outer angular-contact ball bearing shown in Fig. 39 generated considerable noise shortly after delivery of the vehicle. The entire bearing assembly was removed and submitted to the laboratory for failure analysis.

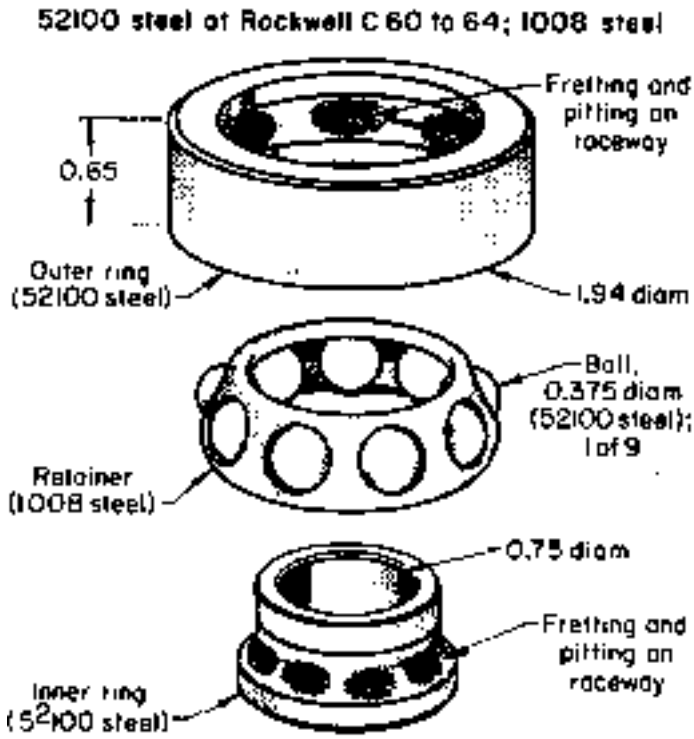


Fig. 39 Automotive front-wheel bearing that failed by fretting of raceways on inner and outer 52100 steel rings. Dimensions given in inches

The inner and outer rings were made of seamless cold-drawn 52100 steel tubing, the balls were forged from 52100 steel, and the retainer was stamped from 1008 steel strip. The inner ring, outer ring, and balls were austenitized at 845°C (about 1550°F), oil quenched, and tempered to a hardness of 60 to 64 HRC.

Investigation. Visual examination of the outer raceway revealed severe fretting and pitting in the ball-contact areas at spacings equivalent to those of the balls in the retainer. The areas were elongated, indicating that only a slight oscillation of the ring or wheel had occurred. Similarly spaced but less severely damaged areas were observed on the inner raceway.

Conclusions. Failure was caused by fretting due to vibration of the stationary vehicle position without bearing rotation. These vibrations were incurred during transportation of the vehicle. This bearing had experienced little, if any, service but showed conclusive evidence of fretting between the balls and raceway.

Recommendations. Methods of securing the vehicle during transportation should be improved to eliminate vibrations.

Example 2: Fretting of Freon-Compressor Shaft because of a Loose Bearing (Ref 114). The shaft-and-bearing assembly (Fig. 40a) in a freon compressor was subjected to severe pounding and vibration after six years of service. After about one year of service, the compressor had been shut down to replace a bearing seal. One month before the shaft failed, a second seal failure occurred, requiring the collar, spacer sleeve, seal, roller bearing, and lockwasher to be replaced.

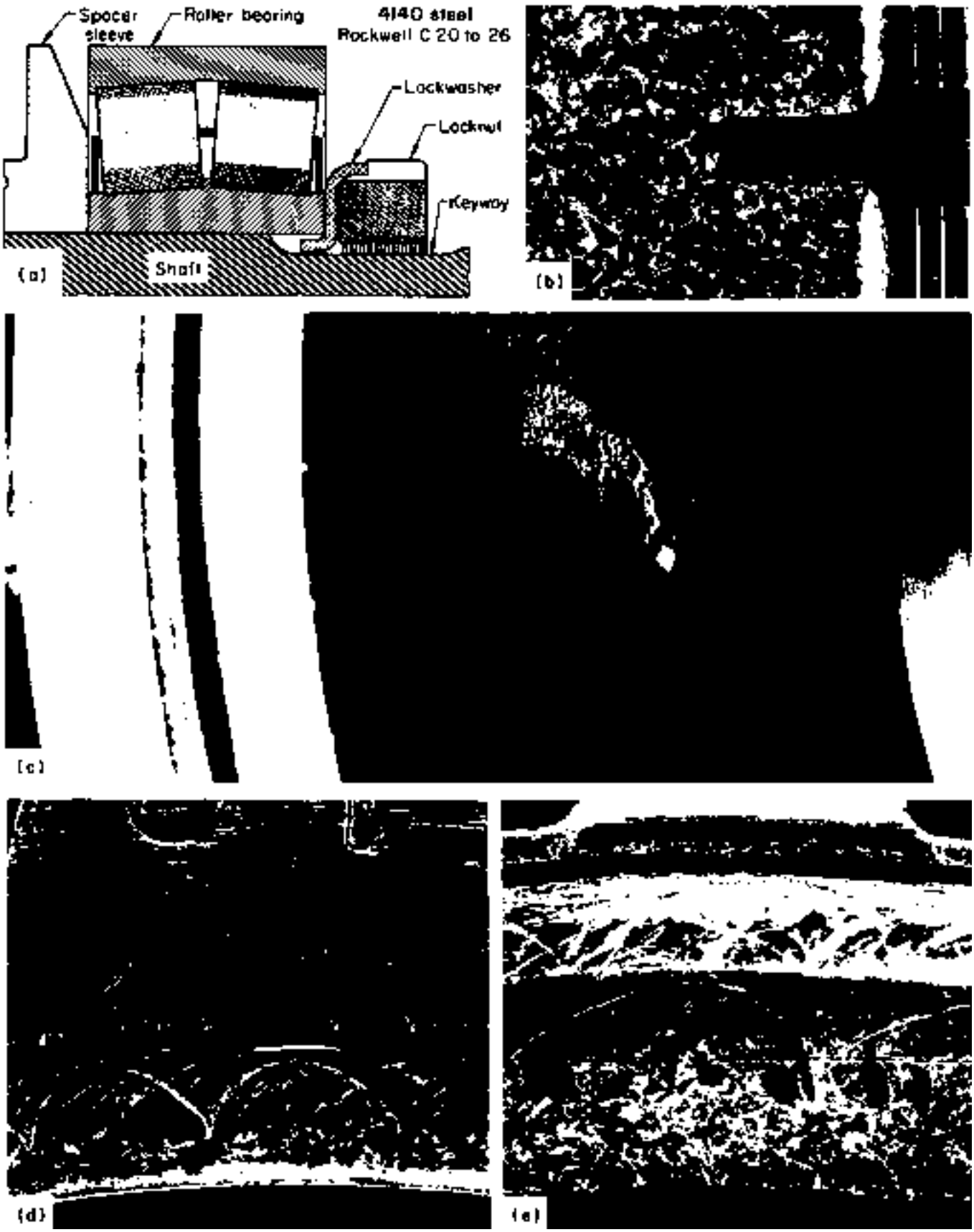


Fig. 40 Freon-compressor shaft of 4140 steel that failed by fretting corrosion in the bearing area (Example 2). (a) Shaft and bearing assembly. (b) Failed region of shaft, at 2×, showing disturbed metal and partly closed keyway. (c) Shaft side of bearing inner ring, at 3×, showing pitting similar to fretting damage. (d) Bearing side of spacer sleeve, at 4.5×, showing disturbed metal. (e) Bearing side of lockwasher, at 4.5×, showing disturbed metal

The shaft was made of 4140 steel, heat treated to a hardness of 20 to 26 HRC. The seal, bearing, and lockwasher were commercial components. Although the specifications did not call for hard facing of the bearing surface on the shaft, maintenance department personnel thought the shaft had been hard faced.

The shaft, spacer sleeve, roller bearing, and lockwasher were sent to the laboratory for examination.

Visual examination of the compressor shaft disclosed quantities of disturbed metal and partial closing of the keyway at the surface where it was contacted by a roller bearing as shown in Fig. 40(b). The failed surface contained a groove approximately 1 mm (0.040 in.) deep and was coated with a black corrosion deposit. The spacer sleeve, roller bearing, and lockwasher also were coated with the black deposit.

Deposits collected from the failed area and from the spacer sleeve, roller bearing, and lockwasher were identified by x-ray diffraction as ferric oxide (α-Fe_2O_3) with minor constituents of metallic iron and copper. Discoloration of the deposit was caused by the grease lubricant and mishandling during disassembly.

Spectrochemical analysis of scrapings taken from an affected area and an unaffected area of the bearing surface on the shaft and from areas adjacent to and away from the bearing surface of the shaft indicated that all surfaces were fabricated from 4140 steel. These findings discounted the theory that the bearing surface had been hard faced.

Readings taken along the shaft showed that the hardness of the area beneath the bearing was 24 HRC, while the area beneath the sleeve had a hardness of 90 HRB (approximately 10 HRC). This suggests that the area of the shaft beneath the bearing had been heat treated. The hardness of the inner bearing race was 54 HRC.

Microscopic examination of the shaft side of the inner bearing race showed pitting similar to that associated with fretting corrosion (see Fig. 40c). Examination of the surfaces of the spacer sleeve and of the lockwasher that contacted the roller bearing disclosed disturbed metal, indicating there had been movement between the sleeve and the bearing and between the bearing and the lockwasher (see Fig. 40d and e). The direction of the disturbed metal on the sleeve and the lockwasher showed that movement was both radial and circumferential.

The peened appearance of the metal that partly covered the keyway indicated movement between the shaft and the inner bearing race. The inner bearing race could not move radially, because it was backed by the rollers, outer race, and compressor shell, which were in good condition after the failure. Therefore, the shaft must have moved radially during failure. Because the lockwasher and spacer sleeve were attached to the shaft and moved with it, the circular scratches on the contacting surfaces of the bearing and the sleeve, and those of the bearing and the lockwasher, indicate radial motion of the shaft. The scratches on the sleeve and lockwasher were approximately 2 mm (0.080 in.) long, which was equal to the loss in diameter of the shaft at the failed area.

Conclusions. Shaft failure was initiated by fretting between the bearing race and the bearing surface on the shaft because of improper bearing installation. Once clearance was established between the bearing and the shaft, the shaft began pounding on the inner bearing race, causing final failure of the shaft surface.

Recommendations. Proper fitting of the shaft and bearing race is essential in preventing movement of the bearing on the shaft. Also, the lockwasher and locknut must be installed properly.

References cited in this section

114. F. Eric Krueger, Fretting Failures, *Failure Analysis and Prevention*, Vol 10, *Metals Handbook*, 8th ed., ASM International, 1975, p 155

116. R.L. Widner and W.E. Littmann, “Bearing Damage Analysis,” NBS 423, National Bureau of Standards, Washington, April 1976

Acknowledgments

The section “Fretting Wear” has been adapted from R.B. Waterhouse, “Fretting Wear,” *Friction, Lubrication, and Wear Technology*, Volume 18, *ASM Handbook*, p 242–246. The section “Examples of Fretting Failures” has been adapted from Ref 114 and 115.

References cited in this section

114. F. Eric Krueger, Fretting Failures, *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., ASM International, 1975, p 155

115. O.E.M. Pohler, Failures of Metallic Orthopedic Implants, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., ASM International, 1986, p 670–694

References

1. E.M. Eden, W.N. Rose, and F.L. Cunningham, Endurance of Metals, *Proc. Inst. Mech. Eng.,* Vol 4, 1911, p 839–974

2. G.A. Tomlinson, The Rusting of Steel Surfaces in Contact, *Proc. R. Soc. (London) A,* Vol A115, 1927, p 472–483

3. E.J. Warlow-Davies, Fretting Corrosion and Fatigue Strength, *Proc. Inst. Mech. Eng.,* Vol 146, 1941, p 32–38

4. J.R. McDowell, Fretting Corrosion Tendencies of Several Combinations of Materials, *Symposium on Fretting Corrosion,* STP 144, ASTM, 1953, p 2439

5. A.J. Fenner and J.E. Field, La Fatigue dans les Conditions de Frottement, *Rev. Métall.,* Vol 55, 1958, p 475–485

6. K.T. Styles, The Influence of Fret Corrosion upon the Fatigue Life of Machine Components, *Met. Australas.,* March 1980, p 4–6

7. J.F. Warriner, “Thin Shell Bearings for Medium Speed Diesel Engines,” No. 364, Diesel Engineers and Users Association, Feb 1975

8. R. Heinz, Fretting Wear—Examples, Test Methods and Preventive Measures, *Mater. Wiss. Werkstofftech.,* Vol 20, 1989, p 14–20

9. R.G. Mitchell, An Airline View of the Corrosion Problem, *Ind. Corros.,* July 1983, p 11–17

10. P. Greenfield and R.W. Suhr, The Factors Affecting the High-Cycle Fatigue Strength of Low-Pressure Turbine and Generator Rotors, *Proc. of International Conference on Fatigue of Engineering Materials and Structures,* 15–19 Sept 1986 (Sheffield), Inst. Mech. Eng., Vol 1, 1986, p 165–186

11. H.W. Fricker, Fretting in Tube Supports of Heat Exchangers, *Proc. of Conference on Component Design in High Temperature Reactors Using Helium as a Coolant,* Nuclear Engineering Group of Institution of Mechanical Engineers, 3–4 May 1972, p 213–231

12. H.J. Connors, Flow-Induced Vibration and Wear of Steam Generator Tubes, *Nucl. Technol.,* Vol 55, Nov 1981, p 311—331

13. P.J.W. Bongers, P.L.M. Dubs, and T.J.M. Linssen, Fatigue Crack Initiation on the Outside Surface of High-Pressure Tubes Caused by Tube Support, *Proc. of Conference on Safety in High Pressure*

Polyethylene Process, Symposium III, March 1977 (Houston), American Institute of Chemical Engineers

14. M. Antler, Electrical Effects of Fretting Connector Contact Materials: A Review, *Wear,* Vol 106, 1985, p 5–33

15. P.J. Kennedy, M.B. Peterson, and L. Stallings, An Evaluation of Fretting at Small Slip Amplitudes, *Materials Evaluation under Fretting Conditions,* STP 780, ASTM, 1982, p 30–48

16. P.J. Kennedy, L. Stallings, and M.B. Peterson, A Study of Surface Damage at Low-Amplitude Slip, *ASLE Trans.,* Vol 27 (No. 4), 1984, p 305–312

17. K.L. Johnson, Surface Interaction between Elastically Loaded Bodies under Tangential Forces, *Proc. R. Soc. (London) A,* Vol 230, 1955, p 531–548

18. D. Nowell and D.A. Hills, Crack Initiation Criteria in Fretting Fatigue, *Wear,* Vol 136, 1990, p 329–343

19. M. Kuno, "Fretting Damage of High Carbon Chromium Steel," Ph.D. thesis, University of Nottingham, 1988

20. S. Goto and R.B. Waterhouse, Fretting and Fretting-Fatigue of Titanium Alloys under Conditions of High Normal Load, *Proc. 4th International Conference on Titanium,* TMS/AIME, Vol 3, 1980, p 1337–1384

21. O. Vingsbo and S. Söderberg, On Fretting Maps, *Wear,* Vol 126, 1988, p 131–147

22. U. Bryggman and S. Söderberg, Contact Conditions in Fretting, *Wear,* Vol 110, 1986, p 1–17

23. F.P. Bowden and D. Tabor, *The Friction and Lubrication of Solids,* Oxford University Press, 1950, p 19

24. I.M. Feng and H.H. Uhlig, Fretting Corrosion of Mild Steel in Air and Nitrogen, *J. Appl. Mech.,* Vol 21, 1954, p 395–400

25. R.B. Waterhouse, Fretting, *Treatise on Materials Science and Technology,* D. Scott, Ed., Academic Press, 1978, p 259–286

26. S. Söderberg, U. Bryggman, and T. McCullough, Frequency Effects in Fretting Wear, *Wear,* Vol 110, 1986, p 19–34

27. H. Kongsjorden, J. Kulsetas, and J. Sletbak, Degradation of Electrical Contacts Caused by Oscillatory Micromotion between the Contact Members, *Proc. 9th International Conference on Electrical Contact Phenomena,* 11–15 Sept 1978 (Chicago)

28. M. Antler, Electrical Effects of Fretting Connector Contact Materials: A Review, *Wear,* Vol 106, 1985, p 5–33

29. M. Braunovic, M.S. McIntyre, W.J. Chaurion, and I. Aitchison, Surface Analysis of Fretting Damage in Electrical Contacts of Aluminium with Different Contacting Materials, *Electrical Contacts 1983, Proc. 29th Holm Conference on Electrical Contacts (Chicago),* Illinois Institute of Technology, p 232–242

30. J.J. Mottine and B.T. Reagor, Investigation of Fretting Corrosion at Dissimilar Metal Interfaces in Socketed IC Device Applications, *Electrical Contacts 1983, Proc. 29th Holm Conference on Electrical Contacts (Chicago),* Illinois Institute of Technology, p 67–71

31. M. Kuno and R.B. Waterhouse, The Effect of Oscillatory Direction on Fretting Wear under Crossed Cylinder Contact Conditions, *Eurotrib 89, Proc. 5th International Congress on Tribology,* 12–15 June 1989 (Helsinki), Vol 3, p 30–35

32. K.B. Elliott, H.H. Mabie, M.J. Furey, and L.D. Mitchell, A Vibration Analysis of a Bearing/Cartridge Interface for a Fretting Corrosion Study, *J. Lubr. Technol. (Trans. ASME),* Vol 105, 1983, p 518–525

33. N.L. Golego, M.N. Rozhkov, V.P. Onoprienko, A.S. Nakrakui, and N.V. Kozii, Wear of Materials in Fretting Corrosion in Connection with Conditions of Vibrocontact Interaction, *Fiz. Khim. Mekh. Mater.,* Vol 21, 1985, p 71–75

34. J.H. Cha, M.W. Wambsganss, and J.A. Jendrzejcyk, Experimental Study on Impact/Fretting Wear in Heat Exchanger Tubes, *J. Pressure Vessel Technol. (Trans. ASME),* Vol 109, 1987, p 265–274

35. A.W.J. de Gee, C.P.L. Commissaris, and J.H. Zaat, The Wear of Sintered Aluminium Powder (SAP) under Directions of Vibrational Contact, *Wear,* Vol 7, 1964, p 535–550

36. P.L. Ko, Experimental Studies of Tube Fretting in Steam Generators and Heat Exchangers, *J. Pressure Vessel Technol. (Trans. ASME),* Vol 101, 1979, p 125–133

37. H.J. Connors, Flow-Induced Vibration and Wear in Steam Generator Tubes, *Nucl. Technol.,* Vol 55, 1981, p 311–331

38. S.F. Calhoun, "Effects of Metal Hardness and Surface Finish upon Fretting Corrosion," Technical Report 63-1835, Rock Island Arsenal, 28 May 1963

39. R.B. Waterhouse, Residual Stresses and Fretting, *International Guidebook on Residual Stresses,* A. Niku-Lari, Ed., Pergamon, Oxford, 1987, p 511–525

40. C.A. Bergman, R.C. Cobb, and R.B. Waterhouse, The Effect of Shot-Peening on the Friction and Wear of a Carbon Steel in Fretting Conditions, *Wear of Materials,* American Society of Mechanical Engineers, 1987, p 33–37

41. J. Labedz, Effect of Residual Stresses on Wear of the Surface Layer of Components Operating under Fretting Conditions, *Metal Treatments against Wear, Corrosion, Fretting and Fatigue,* A. Niku-Lari and R.B. Waterhouse, Ed., Pergamon, Oxford, 1988, p 87–98

42. K. Endo, Practical Observations of Initiation and Propagation of Fretting Fatigue Cracks, *Fretting Fatigue,* R.B. Waterhouse, Ed., Applied Science Publishers, London, 1981, p 127–141

43. A. Yournadj, G. Maeder, and A. Moison, Geometric and Mechanical Aspects of Fretting Corrosion, *Proc. 4th European Tribology Congress,* 9–12 Sept 1985 (Euiby, France), Vol 2, p 1–9

44. G. Fair, B. Noble, and R.B. Waterhouse, The Stability of Compressive Stresses Induced by Shot-Peening under Conditions of Fatigue and Fretting-Fatigue, *Advances in Surface Treatments,* Vol I, A. Niku-Lari, Ed., Pergamon, Oxford, 1984, p 3–8

45. M. Kuno, R.B. Waterhouse, D. Nowell, and D.A. Hills, Initiation and Growth of Fretting Fatigue Cracks in the Partial Slip Regime, *Fatigue Fract. Eng. Mater. Struct.,* Vol 12, 1989, p 387–398

46. T. Kayaba and A. Iwabuchi, Effect of the Hardness of Hardened Steels and the Action of Oxides on Fretting Wear, *Wear,* Vol 66, 1981, p 27–41

47. O. Vingsbo, M. Odfalk, and Ning-E Shen, Fretting Maps and Fretting Behaviour of Some FCC Metal Alloys, *Wear,* Vol 13, 1990, p 153–167

48. J. Dobromirski and I.O. Smith, Metallographic Aspects of Surface Damage, Surface Temperature and Crack Initiation in Fretting Fatigue, *Wear,* Vol 117, 1987, p 347–357

49. J. Beard and P.F. Thomason, The Mechanics of Fatigue Crack Nucleation in a Low Alloy Steel under Fretting Conditions, *Fatigue 87,* Conference Proceedings, 28 June–3 July 1987 (Charlottesville, VA), Vol 1, 1987, p 53–62

50. R. Becker and G. Sepold, Fretting of Amorphous Alloys, *Rapidly Solidified Amorphous and Crystalline Alloys,* B.H. Kear, B.C. Giessen, and M. Cohen, Ed., Elsevier Science Publishing, 1982, p 295–299

51. J. Sato, M. Sato, and S. Yamamoto, Fretting Wear of Stainless Steel, *Wear,* Vol 69, 1981, p 167–177

52. P.A. Higham, F.H. Stott, and B. Bethune, Mechanisms of Wear of the Metal Surface during Fretting Corrosion of Steel on Polymers, *Corros. Sci.,* Vol 18, 1978, p 3–13

53. P.A. Higham, F.H. Stott, and B. Bethune, The Influence of Polymer Composition on the Wear of the Metal Surface during the Fretting of Steel on Polymer, *Wear,* Vol 47, 1978, p 71–80

54. O. Jacobs, K. Friedrich, G. Marom, K. Schulte, and H.D. Wagner, Fretting Wear Performance of Glass-Carbon and Aramid Fibre/Epoxy-and PEEK-Composites, *Wear of Materials,* 1989, American Society of Mechanical Engineers, 1989, p 495–500

55. M. Kuno, R.B. Waterhouse, and B.R. Pearson, Fretting Wear in Sintered Alumina and Tungsten Carbide Cermet-Metal Couples, *Wear of Materials,* 1987, American Society of Mechanical Engineers, 1987, p 371–380

56. D. Klaffke Fretting Wear of Ceramics, *Tribol. Int.,* Vol 22, 1989, p 89–101

57. D.R. Horn, R.B. Waterhouse, and B.R. Pearson, Fretting Wear of Sintered Alumina, *Wear,* Vol 113, 1986, p 225–232

58. D. Godfrey, “Investigation of Fretting by Microscopic Observation,” Report 1009, NACA, 1951

59. J. Sato, M. Shima, and T. Sugiwara, A Fundamental Study of Fretting Damage to Glass Using an Improved Apparatus, *Wear,* Vol 106, 1985, p 53–61

60. M. Godet, The Third Body Approach: Mechanical View of Wear, *Wear,* Vol 100, 1984, p 437–452

61. J.F. Andrews, P.D. Donovan, and J. Stringer, Fretting Corrosion Products of Aluminium Alloys, *Br. Corros. J.,* Vol 3, 1968, p 85–87

62. C. Colombié, Y. Berthier, A. Floquet, L. Vincent, and M. Godet, Fretting: Load Carrying Capacity of Wear Debris, *J. Tribol. (Trans. ASME),* Vol 106, 1984, p 194–201

63. D. Aldham, J. Warburton, and R.E. Pendlebury, The Unlubricated Fretting Wear of Mild Steel in Air, *Wear,* Vol 106, 1985, p 177–201

64. J. Warburton, The Fretting of Mild Steel in Air, *Wear,* Vol 131, 1989, p 365–386

65. R.E. Pendlebury, Unlubricated Fretting of Mild Steel in Air at Room Temperature, Part II: Electrical Contact Resistance Measurements and the Effect of Wear on Intermittent Loading, *Wear,* Vol 118, 1987, p 341–364

66. A. Iwabuchi, T. Kayaba, and K. Kata, Effect of Atmospheric Pressure on Friction and Wear of 0.45% C Steel in Fretting, *Wear,* Vol 91, 1983, p 289–305

67. A. Iwabuchi, K. Kato, and T. Kayaba, Fretting Properties of SUS304 Stainless Steel in Vacuum Environment, *Proc. JSLE Tribology Conference,* 8–10 July 1985 (Tokyo), Vol 1, Japan Society of Lubrication Engineers, 1985, p 29–34

68. K. Kato, T. Suzuki, A. Iwabuchi, and K. Hokkirigawa, Fretting Damage of SUS304 in a Vacuum, *JSME Int. J.,* Vol 30, 1987, p 330–336

69. A. Iwabuchi, H. Matsuzaki, and K. Hori, The Effect of Ambient Pressure on the Tribological Properties of an Amorphous Alloy in Fretting, *Wear of Materials 1989,* American Society of Mechanical Engineers, 1989, p 595–600

70. H. Goto and D.H. Buckley, The Influence of Water Vapour in Air on the Friction Behaviour of Pure Metals during Fretting, *Tribol. Int.,* Vol 18, 1985, p 237–245

71. H. Goto, M. Ashida, and K. Endo, The Influence of Oxygen and Water Vapour on the Friction and Wear of an Aluminum Alloy under Fretting Conditions, *Wear,* Vol 116, 1987, p 141–155

72. M.P. Sherwin, D.E. Taylor, and R.B. Waterhouse, An Electrochemical Investigation of Fretting Corrosion in Stainless Steel, *Corros. Sci.,* Vol 11, 1971, p 419–429

73. S.D. Cook, G.J. Gianoli, A.J.T. Clemow, and R.J. Haddad, Fretting Corrosion in Orthopaedic Alloys, *Biomater. Med. Dev. Art. Org.,* Vol 11, 1983–84, p 281–292

74. S.A. Brown and K. Merritt, Fretting Corrosion in Saline and Serum, *J. Biomed. Mater. Res.,* Vol 15, 1981, p 479–488

75. S.A. Brown and J.P. Simpson, Crevice and Fretting Corrosion of Stainless Steel Plates and Screws, *J. Biomed. Mater. Res.,* Vol 15, 1981, p 867–878

76. S.A. Brown and K. Merritt, Fretting Corrosion of Plates and Screws: An in vitro Test Method, *Corrosion and Degradation of Implant Materials: Second Symposium,* STP 859, A.C. Fraker and C.D. Griffin, Ed., ASTM, 1985, p 105–116

77. K. Merritt and S.A. Brown, Effect of Proteins and PH on Fretting Corrosion and Metal Ion Release, *J. Biomed. Mater. Res.,* Vol 22, 1988, p 111–120

78. M.P. Overs and R.B. Waterhouse, Fretting Wear of Two Weldable Structural Steels (HY80 and 50D) in Seawater, *Wear of Materials 1983,* American Society of Mechanical Engineers, 1983, p 541–545

79. B.R. Pearson and R.B. Waterhouse, The Fretting Corrosion in Seawater of Materials Used in Offshore Structures, *Proc. 9th International Congress on Metallic Corrosion,* 5–9 June 1984 (Toronto), Vol 2, NRC Canada, 1984, p 334–341

80. B.R. Pearson, P.A. Brook, and R.B. Waterhouse, Fretting in Aqueous Media, Particularly of Roping Steels in Seawater, *Wear,* Vol 106, 1985, p 225–260

81. M.W.J. Lewis and C.S. Campbell, Fretting and Wear of Stainless and Ferritic Steels in Liquid-Metal Fast Breeder Reactor Steam Generators, *Nucl. Technol.,* Vol 55, 1981, p 460–469

82. P.L. Hurricks and K.S. Ashford, The Effect of Temperature on the Fretting Wear of Mild Steel, *Proc. Inst. Mech. Eng.,* Vol 184, Pt. 3L, 1969–70, p 165–175

83. P.L. Hurricks, The Fretting Wear of Mild Steel from Room Temperature to 200 °C, *Wear,* Vol 19, 1972, p 207–229

84. P.L. Hurricks, The Fretting Wear of Mild Steel from 200 to 500 °C, *Wear,* Vol 30, 1974, p 189–212

85. F.H. Stott, D.S. Lui, and G.C. Wood, The Structure and Mechanism of the "Glaze" Oxide Layers Produced on Nickel-Based Alloys during Wear at High Temperatures, *Corros. Sci.,* Vol 13, 1973, p 449–469

86. M.M. Hamdy and R.B. Waterhouse, The Fretting-Fatigue Behaviour of a Nickel-Based Alloy (Inconel 718) at Elevated Temperatures, *Wear of Materials 1979,* American Society of Mechanical Engineers, 1979, p 351–355

87. M.M. Hamdy and R.B. Waterhouse, The Fretting Wear of Ti-6Al-4V and Aged Inconel 718 at Elevated Temperatures, *Wear,* Vol 71, 1981, p 237–248

88. A. Iwabuchi, Fretting Wear of Inconel 625 at High Temperature and in High Vacuum, *Wear,* Vol 106, 1985, p 163–175

89. M.M. Hamdy and R.B. Waterhouse, The Stability of the Protective Glaze Oxide Formed during High Temperature Fretting of an Aged Nickel-Based Alloy, Inconel 718, *Wear of Materials 1983,* American Society of Mechanical Engineers, 1983, p 546–549

90. A. Iwabuchi and R.B. Waterhouse, The Effect of Ion Implantation on the Fretting Wear of Four Titanium Alloys at Temperatures up to 600 °C, *Wear of Materials 1985,* American Society of Mechanical Engineers, 1985, p 471–476

91. M.M. Hamdy, B. Noble, and R.B. Waterhouse, Fatigue Crack Initiation in IMI829 Caused by High Temperature Fretting, *J. Mater. Sci.,* Vol 18, 1983, p 493–502

92. D.G. Chetwynd, F.A. McKee, and C.H. Wride, The Measurement of Wear Damage, *Proc. Inst. Mech. Eng.,* Vol 201 (No. 4C), 1987, p 251–258

93. J.T. Atkinson and M.J. Lalor, Holographic Interferometry Applied to Minimal Wear Measurement, 1st European Congress, Appl. to Metrology, *SPIE J.,* Vol 136, 1977, p 107–113

94. T.W. Conlon, The Scope for TLA in Corrosion and Wear Studies, *Metall. and Mater. Technol.,* Vol 2, 1979, p 143–146

95. J. Warburton and R. Bradford, The Progressive Wear of Tubes: The Volumes of the Intersections of Cylinders with Each Other and with Flats, *Wear,* Vol 110, 1986, p 331–352

96. P.L. Hurricks, The Mechanism of Fretting—A Review, *Wear,* Vol 15, 1970, p 389–409

97. D. Aldham, J. Warburton, and R.E. Pendlebury, The Unlubricated Fretting of Mild Steel in Air, *Wear,* Vol 106, 1985, p 177–201

98. R.B. Waterhouse, The Role of Adhesion and Delamination in the Fretting Wear of Metallic Materials, *Wear,* Vol 45, 1977, p 355–364

99. H.H. Uhlig, Mechanism of Fretting Corrosion, *J. Appl. Mech.,* Vol 21, 1954, p 401–407

100. X. Zhang, C. Zhang, and C. Zhu, Slip Amplitude Effects and Microstructural Characteristics of Surface Layers in Fretting Wear of Carbon Steel, *Wear,* Vol 134, 1989, p 297–309

101. R.B. Waterhouse, Fretting Corrosion and Fretting Fatigue in the Transport Industry, *Proceedings of the 6th European Congress on Metallic Corrosion,* 19–23 Sept 1977 (London)

102. R.E. Pendlebury, Unlubricated Fretting and Sliding Wear of Steels in Air, *Tribology—Friction, Lubrication and Wear, Fifty Years On,* Conference Proceedings, 1–3 July 1987 (London), Institute of Mechanical Engineers, Vol 1, 1987, p 267–275

103. R.E. Pendlebury and P.A. Tempest, "Unlubricated Fretting Wear of Mild Steel Surfaces in Air at Room Temperature, Part III: The Conversion of (Artificial) Metallic Wear Particles to Oxide," Report TPRD/B/0985/R87, AGR/CIF/TWG/P(87)232, Central Electricity Generating Board, Oct 1987

104. R.E. Pendlebury, "Unlubricated Fretting Wear of Mild Steel Surfaces in Air at Room Temperature," Report TPRD/B/0674/N85, AGR/CIF/TWG/P(85)164, Central Electricity Generating Board, June 1985

105. J. Beard, The Rational Selection of Palliatives for Avoidance of Fretting, *Tribology—Friction, Lubrication and Wear, Fifty Years On,* Conference Proceedings, 1–3 July 1987 (London), Institute of Mechanical Engineers, Vol 1, 1987, p 311–319

106. R.C. Cobb and R.B. Waterhouse, The Fretting Wear of Certain Hard Coatings Including TiN Applied to a 0.4C Steel, *Tribology—Friction, Lubrication and Wear, Fifty Years On,* Conference Proceedings, 1–3 July 1987 (London), Institute of Mechanical Engineers, Vol 1, 1987, p 303–310

107. J.C. Gregory, A Salt Bath Treatment to Improve the Resistance of Ferrous Metals to Scuffing, Wear, Fretting, and Fatigue, *Wear,* Vol 9, 1966, p 249–281

108. J.W. Souter and W. Staunton, The Fretting Wear of Electrodeposited Contact Coatings, *Trans. Inst. Met. Finish.,* Vol 66, 1988, p 8–14

109. R.B. Waterhouse, The Effect of Surface Treatments on the Fretting Wear of an Aluminum Alloy (RR58)/Steel (BS970 080 M40) Couple, *Proc. of Conference on Surface Engineering, Current Trends and Future Prospects,* 25–27 June 1990 (Toronto), S.A. Meguid, Ed., Elsevier Applied Science, 1990, p 325–334

110. D.J. Varley and R.B. Waterhouse, Fretting of Bronze-Filled PTFE against Mild Steel, *Proc. Japanese International Tribology Conference,* 29 Oct–1 Nov 1990 (Nagoya) Japanese Society of Tribologists, Vol 2, 1990, p 821–825

111. D.E. Taylor and R.B. Waterhouse, Fretting Fatigue in Steel Ropes, *Lubr. Eng.,* Vol 27, 1971, p 123–127

112. J. Sato, M. Shima, T. Sugiwara, and A. Tahara, Effect of Lubricants on Fretting Wear of Steel, *Wear,* Vol 125, 1988, p 83–95

113. D. Godfrey and E.E. Bisson, “Bonding of Molybdenum Disulphide to Various Materials to Form a Solid Lubricating Film, Part II, Friction and Endurance Characteristics of Films Bonded by Practical Methods,” Technical Note 2802, NACA, 1952

114. F. Eric Krueger, Fretting Failures, *Failure Analysis and Prevention,* Vol 10, *Metals Handbook,* 8th ed., ASM International, 1975, p 155

115. O.E.M. Pohler, Failures of Metallic Orthopedic Implants, *Failure Analysis and Prevention,* Vol 11, *Metals Handbook,* 9th ed., ASM International, 1986, p 670–694

116. R.L. Widner and W.E. Littmann, “Bearing Damage Analysis,” NBS 423, National Bureau of Standards, Washington, April 1976

Rolling Contact Fatigue

R. Ahmed, Heriot-Watt University (United Kingdom)

Introduction

A MAJOR CAUSE of failure in components subjected to rolling or rolling/sliding contacts (e.g., rolling-contact bearings, gears, and cam/tappet arrangements) is contact fatigue. Rolling-contact fatigue (RCF) can be defined as the mechanism of crack propagation caused by the near-surface alternating stress field within the rolling-contact bodies, which eventually leads to material removal. The mechanism of RCF thus differs from the delamination theory of wear (Ref 1, 2), which also relies on cyclic loading but in sliding conditions and at asperity level. The alternating stress field in RCF is either in pure rolling (e.g., rolling-element bearings), or in rolling/sliding conditions (e.g., gear-tooth loading), depending on the absence or presence of gross sliding within the rolling-contact region. Microslip (e.g., Heathcote or Reynolds slip [Ref 3]) within the contact region is, however, inevitable in both pure rolling and rolling/sliding conditions. The term “RCF” in this article, however, refers to pure rolling configurations, except where specific references to rolling/sliding conditions are made.

The material removal in a RCF failure varies from micropitting, macropitting and spalling (Ref 4, 5) in conventional bearing steels (Fig. 1) to delamination in hybrid ceramics (Ref 6) and overlay coatings (Ref 7), as discussed later. Contact geometry of the bodies subjected to RCF can be conforming (e.g., contact between the outer race and roller in a rolling-element roller bearing) or nonconforming (e.g., contact between the inner race and roller in a rolling-element roller bearing). The alternating stress field responsible for RCF failure can generally be idealized from Hertzian contact conditions in conventional metallic and ceramic materials (e.g., bearing steel and Si_3N_4 ceramics), but the interpretation of stress fields needs to be cautiously approached when dealing with layered surfaces (e.g., overlay coatings). The preliminary focus in this article is on RCF of coated surfaces, although the stress field conditions are briefly reviewed for the RCF of a homogeneous material surface.

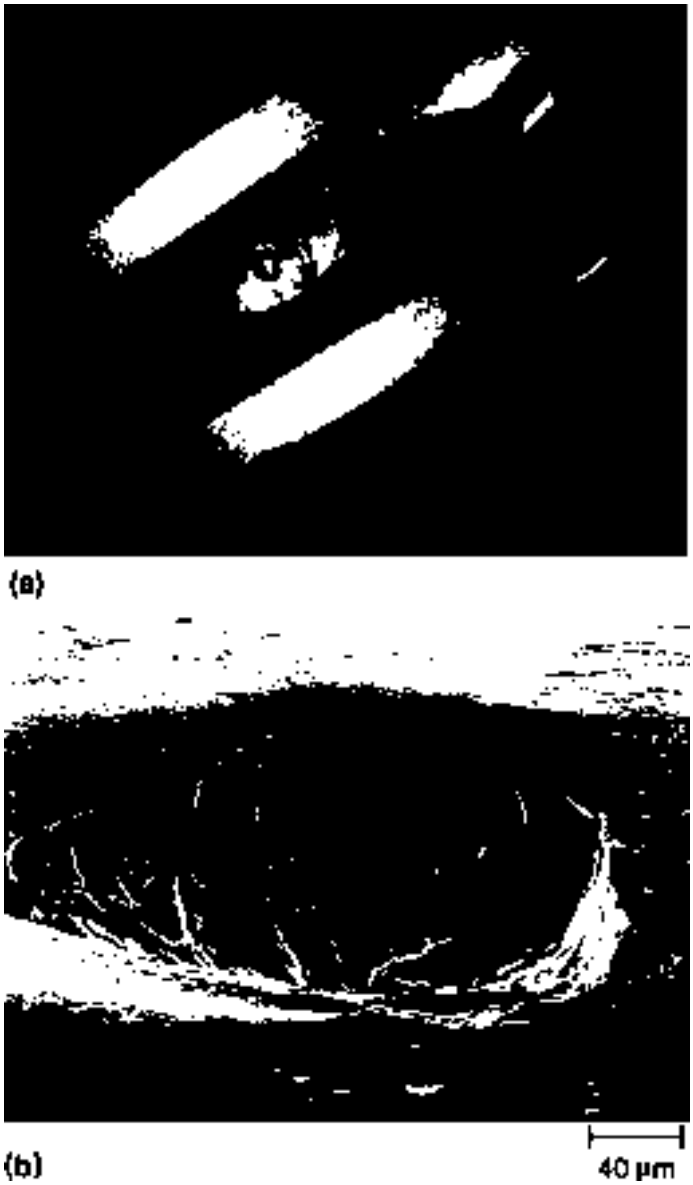


Fig. 1 Typical morphology of fatigue spall in rolling-element bearings. (a) Fatigue spall centered on a ball bearing raceway. (b) Fatigue spall on 12.7 mm (0.5 in.) diameter steel ball obtained using rolling four-ball machine

Prediction of statistical fatigue life (Ref 8, 9, 10) and failure modes (Ref 4, 11, 12, 13, 14, 15, 16, 17) during RCF in conventional bearing steels also has been the focus of hundreds of papers and numerous books (Ref 18, 19, 20) in published literature. The scope of this article focuses principally on the RCF performance and failure modes of overlay coatings, such as those deposited by physical vapor deposition, chemical vapor deposition, and thermal spraying. Some background description of RCF in bearing steels is, however, useful and necessary, because RCF of steels is important. General background on RCF in bearing steels also helps develop an understanding of failure modes in overlay coatings discussed in this article. More detailed description of RCF in conventional bearing steels can be seen elsewhere (e.g., Ref 11, 12, 13, 14, 15, 16, 17), along with a general discussion of contact fatigue in the article “Fatigue Failures” in this Volume.

References cited in this section

1. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear*, Vol 44, 1977, p 1–16

2. N.P. Suh, The Delamination Theory of Wear, *Wear*, Vol 25, 1973, p 111–124

3. D. Tabor, The Mechanism of Rolling Friction, Part II: The Elastic Range, *Proc. R. Soc. A,* Vol 229, 1955, p 198–220

4. S. Way, Pitting Due to Rolling Contact, *J. Appl. Mech.,* 1935, p A49–A58

5. T.E. Tallian, On Competing Failure Modes in Rolling Contact, *ASLE Trans.,* Vol 10, 1967, p 418–439

6. M. Hadfield, T.A. Stolarski, and R.T. Cundill, Failure Modes of Ceramics in Rolling Contact, *Proc. R. Soc. A,* Vol 443, 1993, p 607–621

7. R. Ahmed and M. Hadfield, Failure Modes of Plasma Sprayed WC-Co Coated Rolling Elements, *Wear,* Vol 230, 1999, p 39–55

8. J.V. Poplawski, S.M. Peters, and E.V. Zaretsky, Effect of Roller Profile on Cylindrical Roller Bearing Life, Part I: Comparison of Bearing Life Theories, *Tribol. Trans.,* Vol 44 (No. 3), 2001, p 339–350

9. T.E. Tallian, Data Fitted Bearing Life Prediction Model for Variable Operating Conditions, *Tribol. Trans.,* Vol 42 (No. 1), 1999, p 241–249

10. Y. Li, T.R. Kurfess, and S.Y. Liang, Stochastic Prognostics for Rolling Element Bearings, *Mech. Syst. Signal Process.,* Vol 14 (No. 5), 2000, p 747–762

11. R.S. Hyde, Contact Fatigue of Hardened Steels, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 680–690

12. W.A. Glaeser and S.J. Shaffer, Contact Fatigue, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 331–336

13. P.J. Blau, Rolling Contact Wear, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* 1992, p 257–262

14. W.D. Syniuta and C.J. Corrow, A Scanning Electron Microscope Fractographic Study of Rolling Contact Fatigue, *Wear,* Vol 15, 1970, p 187–199

15. J.A. Martin and A.D. Eberhardt, Identification of Potential Failure Nuclei in Rolling Contact Fatigue, *J. Basic Eng. (Trans. ASME),* 1967, p 932–942

16. W.E. Littmann and R.L. Widner, Propagation of Contact Fatigue from Surface and Subsurface Origins, *J. Basic Eng. (Trans. ASME),* 1966, p 624–636

17. W.E. Littmann, The Mechanism of Contact Fatigue, Conf. Proc. *An Interdisciplinary Approach to the Lubrication of Concentrated Contacts,* National Aeronautics and Space Administration, 1969, p 309–377

18. E.V. Zaretsky, "Life Factors for Rolling Bearings," Society of Tribology and Lubrication Engineers (STLE) publication, 1999

19. R.A. Guyer, *Rolling Bearings Handbook and Trouble Shooting Guide,* CRC Press, 1996

20. T. Harris, *Rolling Bearing Analysis,* 3rd ed., John Wiley & Sons, 1991

R. Ahmed, Heriot-Watt University (United Kingdom)

General Principles of RCF

General Background. The origin of RCF failure is understood to be stress concentrations, which initiate and propagate fatigue crack under cyclic loading. These stress concentrations occur due to surface or subsurface stress risers or to the geometry and kinematics of the contacting pair. Figure 2 summarizes a list of these stress risers, which have been the subject of numerous scientific investigations that have resulted in the improved life of rolling-element bearings (Ref 11). With the introduction of cleaner steels and greater precision in the manufacture of bearings, most of the surface and subsurface stress risers listed in Fig. 2 have been addressed. Nevertheless, the demand to operate rolling-element bearings in harsh tribological environments of lubrication, load, contamination, and temperature push for higher fatigue limits, and thus call for improved understanding of the RCF failure modes.

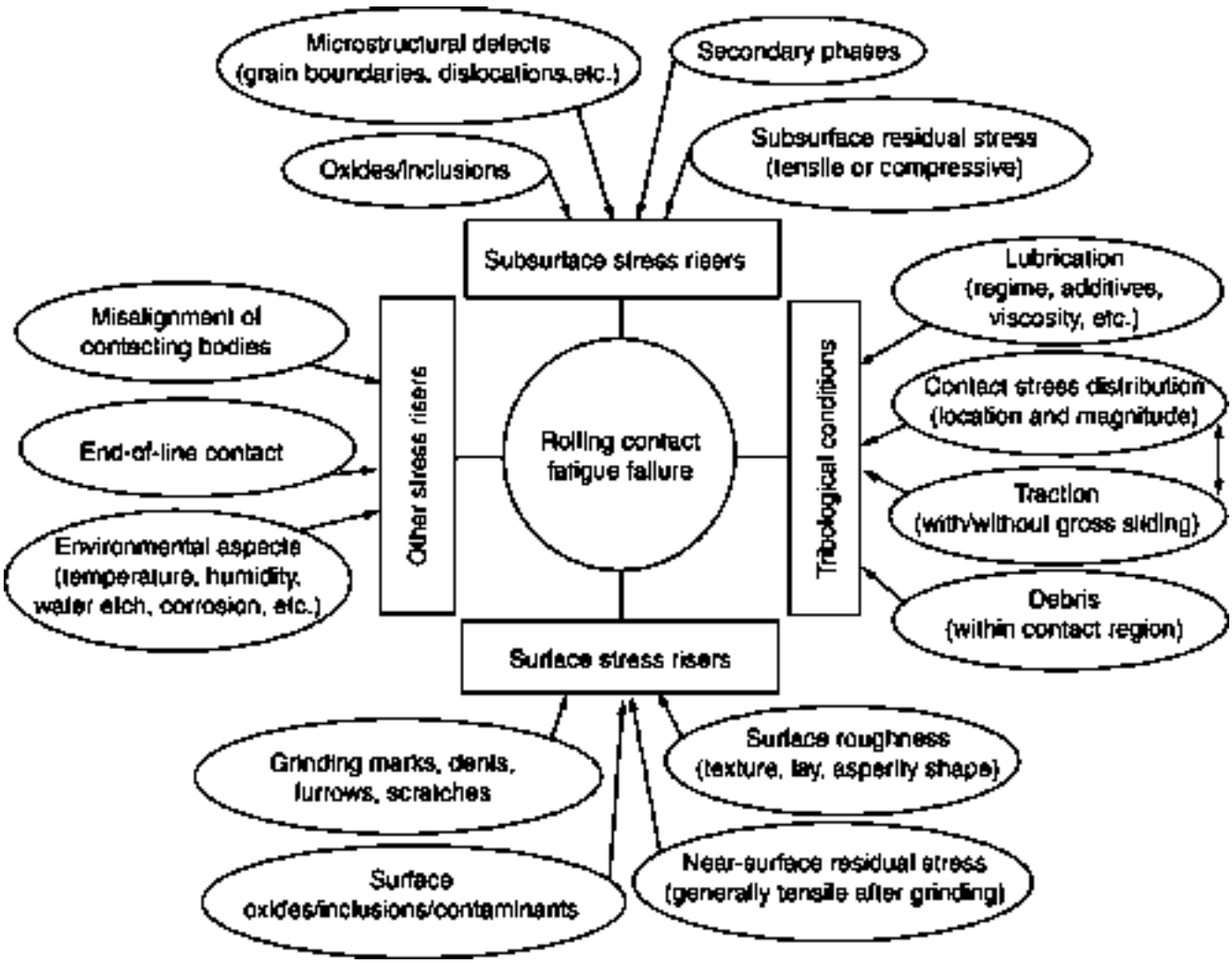


Fig. 2 Stress risers initiating rolling-contact fatigue failure

Four distinct failure modes have been established in rolling-contact bearings (Ref 5). These classifications include wear-type failures, plastic flow, contact fatigue, and bulk failures. Although the aim of this article is to comprehend RCF failures, wear-type failures that include surface removal and material transfer do not form a part of this background; Blau (Ref 13) has given a detailed account of rolling-contact wear (RCW), and he differentiates between RCF and RCW in the sense that RCF is a damage accumulation process under cyclic

loading, whereas, RCW can be thought of as nucleation sites for initiating fatigue damage. Rolling-contact wear is thus critical in components operating in rolling-sliding contact (e.g., gears where the lubrication regime is either boundary or mixed). However, the full film lubrication generally seen in rolling-element bearings should prevent such damage, except at the start/stop of rolling motion. Similarly, plastic flow and bulk failure depend on the bulk thermal and mechanical properties of the bearing materials and can lead to permanent dimensional changes.

The most classical failure mode in rolling-contact components is RCF. Rolling-contact fatigue failure modes have previously been classified as pitting, case crushing, peeling, frosting, glazing, surface distress, and so on (Ref 15, 16, 17, 21, 22). Some of these terminologies are similar; for example, pitting and spalling are often used interchangeably in the European literature (Fig. 1). Littmann (Ref 17) has given an account of the contact fatigue damage classification system, and he identified six modes of fatigue failure—inclusion origin, subcase fatigue, surface origin, geometric stress concentrations, peeling, and section failure. Later studies classified these failure modes as surface originated and subsurface originated (23, 24). A more comprehensive classification of these failure modes can be found in Ref 25.

Regardless of the classification system used to categorize failure modes, the mechanism of RCF failure generally involves the following characteristics:

- Stress concentrations either at the surface or subsurface (Fig. 2)
- Crack initiation either at the surface or subsurface due to stress concentrations
- Crack propagation due to cyclic loading
- Spalling or pitting leading to debris
- Postfailure damage due to misalignment or debris within the contact region

The exact mode of fatigue failure can substantially vary from failure to failure, and a number of failure modes compete to reduce the overall life of rolling bearings. This is mainly due to the synergetic effect of competing failure modes in which both surface and subsurface origins of crack initiation and propagation can be significantly affected by tribological conditions. Additional factors, such as the role of residual stress; hydraulic pressure propagation, coupled with the shape, size, and location of individual surface or subsurface defects; and variations in tribological conditions within the contact region, thus present a complex interdependency.

Characteristics of RCF in Steel.* The stress condition associated with RCF is illustrated schematically in Fig. 3. With any condition of rolling, the maximum stress being applied at or very near the contact area is the shear stress parallel to the rolling surface at some point below the surface. For normally loaded gear teeth, this distance is from 0.18 to 0.30 mm (0.007 to 0.012 in.) below the surface just ahead of the rolling point of contact. If sliding is occurring in the same direction, the shear stress increases at the same point. If the shear plane is close to the surface, then light pitting can occur. If the shear plane is deep due to a heavy rolling-load contact, then the tendency is for the crack propagation to turn inward (Fig. 4). The cracks continue to propagate under repeated stress until heavy pitting or spalling takes place (Fig. 5).

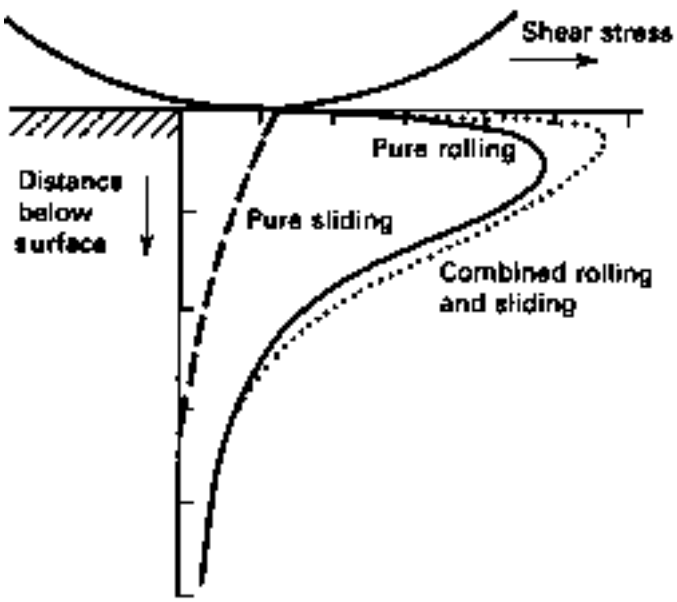


Fig. 3 Stress distribution in contacting surfaces due to rolling, sliding, and combined effect

Fig. 4 Gear-tooth section. Rolling-contact fatigue. Crack origin subsurface. Progression was parallel to surface and inward away from surface. Not etched. 60×

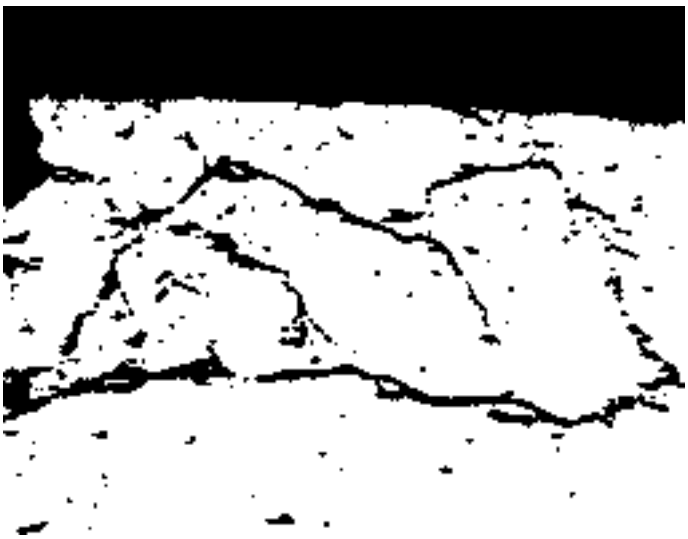


Fig. 5 Gear-tooth section. Rolling-contact fatigue. Crack origin subsurface. Progression was parallel with surface, inward, and finally to the surface to form a large pit or spall. Not etched. 60×

There is always one characteristic (and often two) of RCF that will distinguishes it from other modes of surface-contact fatigue. Both characteristics can be observed only by an examination of the microstructure. In rolling contact, the surface does not show a catastrophic movement; it remains as the original structure. For example, an unetched, polished sample taken near the origin of a subsurface fatigue crack (Fig. 6) very clearly shows undisturbed black oxides at the surface. The subsurface cracking could not have been caused by either abrasive or adhesive contact; it had to be by rolling. This is always the first evidence to look for when determining the type of applied stress.

Fig. 6 Gear-tooth section. Rolling-contact fatigue distinguished by subsurface shear parallel to surface. Note the undisturbed black oxides at the surface, indicating no surface-material movement. Not etched. 125×

The second characteristic is common only in a martensitic steel that contains very little or no austenite and is found only at, along, or in line with the shear plane. This is a microstructural feature that has been termed "butterfly wings." If the sample in Fig. 6 is etched properly with 3% nital, the result is the microstructure shown in Fig. 7(a). Increasing the magnification to 310× shows more detail (Fig. 7b). This type of microstructural alteration is typically associated with an inclusion present, but not always. Extensive and very detailed studies of RCF refer to the gray substructure as white bands of altered martensite. It is believed that these substructures are caused when, under an extreme shearing stress, movement is called for but is restrained and contained to such an extent that the energy absorbed institutes a change in the microstructure ahead of a progressing crack. They are never observed when significant amounts of austenite are present; austenite quickly absorbs the energy and is converted to untempered martensite. They are also in an area that has not been deformed but has definitely been transformed. Each area has distinct boundaries, and the oncoming cracks appear to follow these boundaries. It has been noted that some academic studies refer to this same structure as being a transformed shear band product formed by adiabatic shear. A more detailed review of microstructural change in contact fatigue of steel is provided in the article "Contact Fatigue of Hardened Steel" in *Fatigue and Fracture*, Volume 19, *ASM Handbook*, (Ref 11).

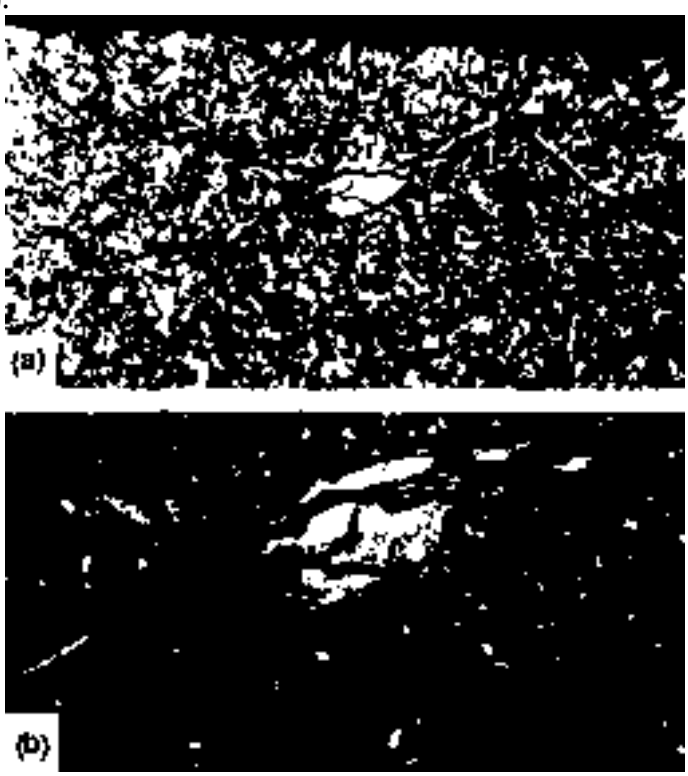


Fig. 7 Same sample as in Fig. 6, showing details of submicrostructure called butterfly wings. 3% nital etch. (a) 125×. (b) 310×

Rolling-Contact Fatigue Testing. The aforementioned complexity in underpinning the exact failure mechanism has led to simplified experimental contact model configurations (RCF tribometers), which have been extensively used to investigate the influence of changes in bearing materials, residual stress, and tribological conditions. Although the correlation between the fatigue life of these model contact configurations and the actual life of bearings in service has not been satisfactorily achieved, these RCF tribometers serve three important functions:

- They provide a method to benchmark the performance of existing and new generation of bearing materials prior to full-scale testing.
- They provide an insight to the mechanisms of individual failure modes by allowing the flexibility to vary individual parameters, such as lubrication, material cleanliness, contact load, surface roughness, and so on.
- Progressive tests allow the possibility to catalog the history of specific failures, thereby indicating the boundaries of failure initiation and propagation either in terms of progressive failure morphology or vibration levels, which can be useful for condition monitoring and wear mapping.

Although the failure mode depends on the tribological conditions selected for individual investigation, some test conditions adopted in RCF tribometers can accelerate the RCF failure, for example, by either increased contact loading or increased rolling velocity. This alters the elastic/plastic (shakedown) response, lubrication regime, and kinematics within the contact region when compared with those in the actual rolling-contact bearings. These variations can thus influence the mechanism of fatigue failure, and results of such investigations need to be interpreted in accordance with the adapted test conditions. Various researchers have compared the underpinning failure mechanisms observed during accelerated testing versus field performance. The influence of higher contact pressure on the residual stress profile and changes in white bands are considered in Ref 26. The elastic plastic shakedown can be thought responsible for changes that influences both the microstructure and also residual stress profile. Results have indicated that martensitic decay is possible for long duration tests and high toughness of bearing material of acceptable hardness leads to high RCF life in both accelerated and field tests.

The correct choice of tribometer is also critical for a given application. For example, a four-ball machine can simulate the kinematics of a deep-groove rolling-element ball bearing; however, the model contact in this tribometer considers a nonconforming contact between the inner race and rolling-element ball, which is not the case in an actual bearing. The capabilities of various RCF testing methods have been tabulated (Table 1) to compare the features of 13 different RCF tribometers (Ref 13). These tribometers (Ref 26, 27) have the capability to vary tribological conditions, for example, lubrication regime, contact configuration, and roll-to-slip ratio to mimic pure rolling or rolling/sliding motion for various rolling-contact applications. Even within a given category of RCF testers, there are generally numerous combinations of test configurations to allow the flexibility in experimental design; for example, in a rolling four-ball tester, there are numerous configurations (type I, II, III) possible to vary ball kinematics (Ref 28). Another modification to RCF testing has been the investigations of artificially induced surface defects (Ref 29, 30), which are thought to act as crack initiation sites during the RCF failure. Such investigations are useful in understanding the influence of surface defects on RCF failure and may also reduce the RCF testing time; the relevance of such investigations to the tribological conditions and failure modes in real bearing applications, however, requires a careful consideration of the differences in tribological conditions between them.

Table 1 Summary of RCF testing methods

Method	Description	Ref
NASA five-ball testing apparatus	Four lower balls, freely rotation 90° apart in a separator; simulates the kinematics of a thrust-loaded bearing; the contact angle can be varied; vibration sensor detects failure in unattended tests; low- (cryo) and high-temperature testing (to 1000 °C, or 1830 °F)	26 (a)
Flat-washer	16 retained balls rolling in a circle on a flat washer with a 75 mm (3 in.) outside	26

testing apparatus	diameter, 50 mm (2 in.) bore, and 6.4 mm ($\frac{1}{4}$ in.) thickness; 4.17 GPa (605 ksi) contact stress; 1500 rpm; filtered lubricant delivery system; piezo sensor detects vibration	(b)
Unisteel testing apparatus	Flat washer on retained balls; hanging dead-weight load; contact stress approximately 4.5 GPa (650 ksi); 1500 rpm; drip feed of lubricant; vibration detection system; thermocouples monitor temperature (typically 50 to 60 °C, or 120 to 140 °F)	26 (c)
Rolling-contact testing apparatus	Two hemispherically ground, toroidal rollers loaded against a round bar; 40:1 ratio of roller diameter to bar diameter; 2.7 to 5.5 GPa (390 to 800 ksi) contact stress; 12,500 rpm; drip-feed lubrication; velocity-vibration sensor	26 (d)
Ball-rod testing apparatus (Federal-Mogul)	Three 12.5 mm ($\frac{1}{2}$ in.) balls loaded against a rotating 9.5 mm ($\frac{3}{8}$ in.) outside diameter center rod; 3600 rpm; spring load on opposing tapered retaining rings; accelerometer coupled with a shutdown device; drip-feed lubrication; stress per ball typically 6 GPa (870 ksi)	26 (e)
Cylinder-to-ball testing apparatus	Symmetrical arrangement of two 19 mm ($\frac{3}{4}$ in.) balls rolling on a 12.5 mm ($\frac{1}{2}$ in.) outside diameter captive cylinder; coiled-spring load through a multiplying lever; small cylinder rpm = 22,677; splash lubrication; maximum contact stress, 5.8 GPa (840 ksi)	26 (f)
Cylinder-to-cylinder testing apparatus	Symmetrical arrangement of two 12.5 mm ($\frac{1}{2}$ in.) cylinders on two 20 mm (0.8 in.) outside diameter captive cylinders; coiled-spring load through a multiplying lever; small cylinder, cpm = 20,400; splash lubrication; maximum contact stress less than 4.4 GPa (640 ksi); vibration sensor terminates test	26 (g)
Ring-on-ring testing apparatus	Crowned rings rolling on their peripheries; ring diameters of 50 and 53 mm (2 and 2.1 in.) provide “no-slip” condition, but various degrees of slip are possible by changing ring diameters; typically 2000 rpm; contact ratio measured by electrical resistance; contact stress range typically 0.98 to 3.9 GPa (140 to 570 ksi)	26 (h)
Various types	Method of testing for rolling contact fatigue of bearing steels	26 (i)
Multiple bearing testing apparatus	Deep-groove ball bearing design; typically 3000 rpm; four bearings on a single center shaft; maximum contact stress, 2.9 GPa (420 ksi); accelerometers on the outer housing monitor failure	26 (j)
Rolling four-ball testing apparatus	Top ball drives three lower balls in a tetragonal arrangement; lower balls free to rotate in the cup; all balls 12.5 ($\frac{1}{2}$ in.) diam; upper balls spindle speed, 1500 rpm; 5.9 kN (1325 lbf) load applied vertically	26 (k)
High-speed four-ball testing apparatus	Same arrangement as above, but speeds of 15,000 to 20,000 rpm; operating temperatures often exceed 100 °C (210 °F) (Plint machine)	26 (k)
“AOL” vertical testing apparatus	11 retained balls clamped between two flat washers; thrust load; recirculating lubricant system	26 (k)
Inclined ball-on-disk testing apparatus	Spindle-held 20.5 mm (0.8 in.) ball rolling on a disk; up to 800 °C (1470 °F); ball speed up to 7200 rpm; disk speed up to 3600 rpm; variable slide/roll ratios; traction measurements; designed for ceramics	27

Note: NASA, National Aeronautics and Space Administration. Further information can be found on the following pages of Ref 26:
(a) p 5–45,
(b) p 46–66,
(c) p 67–84,
(d) p 85–106,

(e) p 107–124,
(f) p 125–135,
(g) p 136–149,
(h) p 150–165,
(i) p 169–189,
(j) p 206–218,
(k) p 219–236.
Source Ref 13

Theories of RCF Failure. The exact mechanism of RCF failure in engineering components needs to be appreciated in view of tribological conditions and various crack initiation and propagation sites; nevertheless, various theories (discussed subsequently) linking the location and magnitude of cyclic tensile or shear stress components to RCF failure are useful in understanding the failure mechanism. Although such theories based on cyclic stress components of the Hertzian stress field can sometimes oversimplify the RCF failure mechanism, the microscopic investigations comparing the cyclic stress components to RCF failure are compelling. For example, the existence of subsurface etchings and butterflies have been related to shear stress components under the contact surface (Ref 5, 11, 14, 15, 17).

Figure 8 illustrates a two-dimensional schematic of the location and magnitude of maximum shear (τ_{max}), orthogonal shear (q_{orth}) and maximum tensile (T_{max}) stress for a circular, dry-rolling frictionless contact. The contact diameter is assumed to be $2a$, and τ_{max} has a maximum value vertically below the center of contact region in a plane inclined at 45° to the coordinate axis (Ref 11). Its magnitude at any point can be calculated using the equation:

$$\tau_{max} = (\sigma_1 - \sigma_3)/2 = 0.35P_o \qquad \text{(Eq 1)}$$

where σ_1 and σ_3 are the values of maximum and minimum principal stress, respectively, and P_o is the peak compressive stress (Hertzian stress). Similarly, the orthogonal shear stress (q_{orth}) acts on planes parallel and perpendicular to the surface and is the vector sum of q_{yz} and q_{zy} for a circular contact, that is:

$$\left|q_{orth}\right| = \left(q_{yz}^2 + q_{zx}^2\right) = 0.21P_o \qquad \text{(Eq 2)}$$

where x and y are the two axes of the contact circle. For elliptical contacts, x any y indicate the major and minor axes of the contact ellipse. Although the magnitude of orthogonal shear stress is the same at the leading and trailing edge of the contact region for a dry, frictionless contact, the direction is reversed to give rise to maximum $q_{orth(max)}$ and minimum $q_{orth(min)}$ values of orthogonal shear stress (Fig. 8). Hence, the maximum amplitude of orthogonal shear stress reversal, Δq_{orth}, is given as:

$$\Delta q_{orth} = (q_{orth(max)} - q_{orth(min)}) = 0.42P_o \qquad \text{(Eq 3)}$$

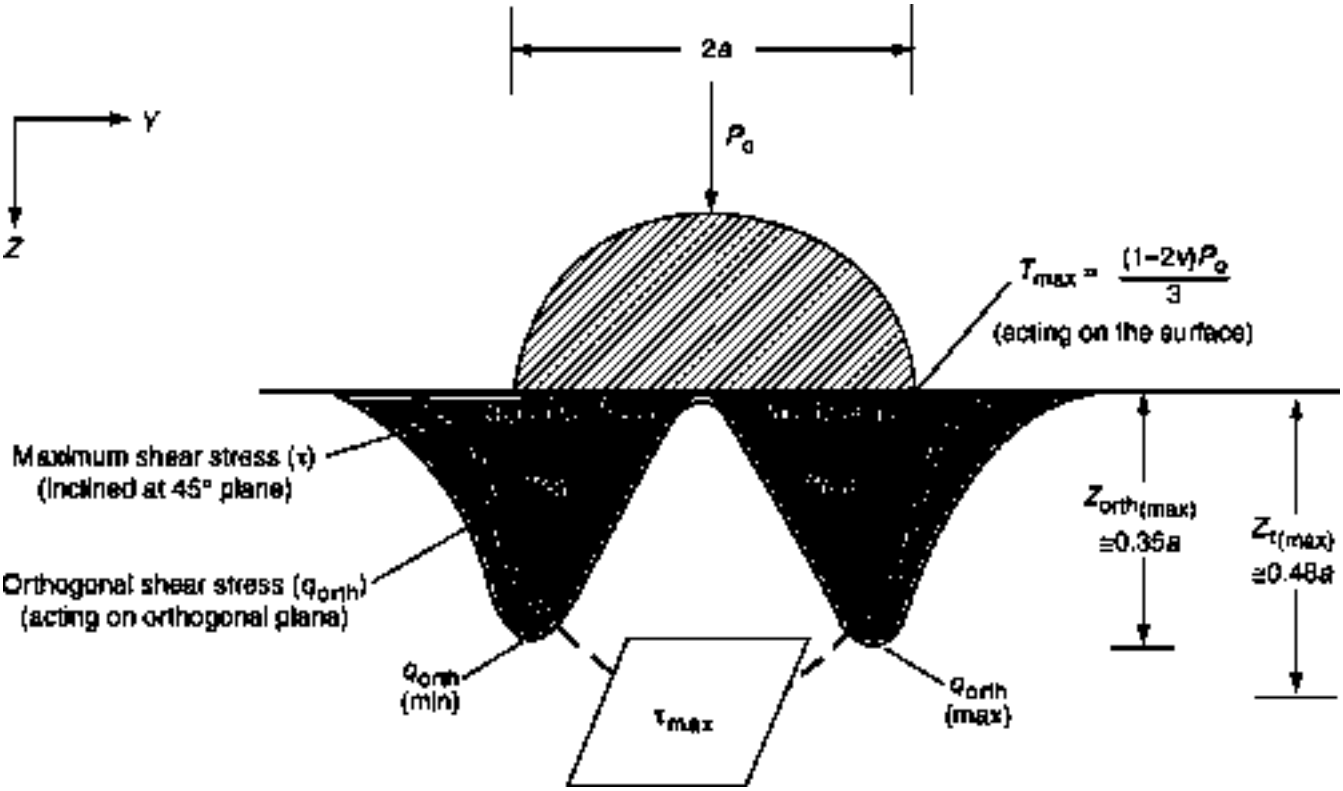


Fig. 8 Two-dimensional schematic of variation in critical stress components with depth for a circular rolling contact of diameter 2*a*

Values relative to P_o indicated in Eq 1–3 will be different for elliptical contacts. Also, the location and magnitude of these stresses strongly depend on the value of the friction coefficient within the contact region. As the friction coefficient increases, the tangential loading shifts the location of maximum shear stress toward the surface. Similarly, the lubrication regime, surface roughness, and residual stress all affect the stress fields. Engineering Science Data Units (Ref 31, 32) tabulate the variation in these stress fields with friction coefficient and contact ellipse ratio.

Both maximum and orthogonal shear stress theories have been proposed to act as the subsurface failure modes during RCF. Fatigue life prediction theories, for example, the Lundberg-Palmgren theory (Ref 33), is also associated with the location of orthogonal shear stress. More recently, equivalent or effective shear stress (q_e) based upon the von Mises criteria (Ref 32, 33, 34) has been shown to support experimental investigations, that is:

$$q_e = \left[q_{xy}^2 + q_{yz}^2 + q_{zx}^2 + \tfrac{1}{6}\left\{ (\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2 \right\} \right]^{1/2} \qquad \text{(Eq 4)}$$

where σ_x, σ_y, and σ_z are the direct stress, and q_{xy}, q_{yz}, and q_{zx} are the shear stress values, respectively. This failure criterion for homogeneous, isotropic, and ductile materials can be related to the critical value of effective shear stress ($q_{e(crit)}$) using the relation:

$$q_{e(crit)} = \sigma_{yield}^2 / 3 = \tau_y^2 \qquad \text{(Eq 5)}$$

where σ_{yield} is the yield strength of material in uniaxial tension or compression, and τ_y is the yield stress in simple shear (Ref 32). Apart from the subsurface shear stress and von Mises criteria mentioned previously, maximum tensile (T_{max}) stress criteria is a useful indicator for appreciating RCF failure mechanism in brittle materials. It is postulated that cracks initiate and propagate from the tensile stress at the edge of the contact region (Fig. 8). The magnitude of T_{max} increases with the increase in friction coefficient; however, its location remains very near the surface. For a frictionless, dry contact, the value of T_{max} can be approximated using the relation:

$$T_{max} = (1 - 2\nu)P_o/3 \qquad \text{(Eq 6)}$$

where ν is the Poisson's ratio.

Rolling Bearing Life. Life prediction of a roller or ball bearing is based on the statistical treatment of full-scale bearing tests conducted under controlled environments, for example, full-film lubrication regime, dust-free environment, and so on. Although these controlled environments are useful indicators of the RCF performance of rolling-element bearings, the actual life of a given bearing can significantly vary from the predicted life. It was reported that only 10% of all bearing replacements in the field can be attributed to classic RCF failure (Ref 35), whereas the remaining 90% are made for reasons and conditions not even closely related to RCF (Ref 26). Although this highlights the improvements achieved in designing and manufacturing better-quality bearings, it also indicates that other factors in addition to RCF need to be considered to predict the life of a bearing. Even under the controlled conditions of load, lubrication, and alignment in a dust-, corrosion-, and moisture-free environment, there is generally a large scatter in the RCF performance for a given bearing. Bearing manufacturers thus provide a statistical probability of the life of a bearing on the basis of experimental results conducted at a given load, speed, and lubrication regime. Weibull analysis (Ref 17) can then be used to estimate the life expectancy of the bearing. Bearing life is generally referred to as L_{10}, L_{50} or L_{90} and indicate the probability of failure (e.g., L_{10} indicates that 10% of the bearings in a given population will fail before a fixed number of stress cycles are reached). Lundberg and Palmgren developed a theory that indicates the life of a roller or ball bearing at a given load (P) can be approximated using the relation:

$$L = (C/P)^n \qquad \text{(Eq 7)}$$

where L is the fatigue life in a million revolutions, C is the load that gives an L_{10} life of 1 million revolutions, and n is a constant for a bearing type (e.g., $n = 3$ for ball bearings and 10/3 for roller bearings). A detailed list of various values of n for different bearing types can be seen elsewhere (Ref 36). The use of a specific value of n is critical for a given bearing type, and the product law of probability is in effect (e.g., as the design changes from single-row to double-row bearings). Hence, the life of a double-row bearing under similar conditions to that of a single-row bearing of identical design will be less than that of a single-row bearing. A detailed account of the statistical treatment of life predictions for a variety of bearing types can be found in Ref 36, which also indicates the importance of lubrication regime for such analysis. In addition to lubrication regime, material shakedown effects (e.g., Ref 37, 38) and the role of residual stress (e.g., Ref 39, 40, 41, 42) on the RCF life can be significant. Apart from Weibull analysis and the Lundberg-Palmgren theory, other theories of fatigue-life prediction (Ref 18) can also be used. These theories were compared recently using a model roller-race contact and the study indicated that the accuracy of individual theories depends not only on the life equation used but also on the assumed Weibull slope (Ref 8). The final choice of these parameters should therefore be consistent with the experimental investigations.

Footnote

* *This section was adapted from L.E. Alban, Failures of Gears, *Failure Analysis and Prevention,* Vol 11 in *ASM Handbook,* 1986, p 593–594.

References cited in this section

5. T.E. Tallian, On Competing Failure Modes in Rolling Contact, *ASLE Trans.,* Vol 10, 1967, p 418–439

8. J.V. Poplawski, S.M. Peters, and E.V. Zaretsky, Effect of Roller Profile on Cylindrical Roller Bearing Life, Part I: Comparison of Bearing Life Theories, *Tribol. Trans.,* Vol 44 (No. 3), 2001, p 339–350

11. R.S. Hyde, Contact Fatigue of Hardened Steels, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 680–690

13. P.J. Blau, Rolling Contact Wear, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* 1992, p 257–262

14. W.D. Syniuta and C.J. Corrow, A Scanning Electron Microscope Fractographic Study of Rolling Contact Fatigue, *Wear,* Vol 15, 1970, p 187–199

15. J.A. Martin and A.D. Eberhardt, Identification of Potential Failure Nuclei in Rolling Contact Fatigue, *J. Basic Eng. (Trans. ASME),* 1967, p 932–942

16. W.E. Littmann and R.L. Widner, Propagation of Contact Fatigue from Surface and Subsurface Origins, *J. Basic Eng. (Trans. ASME),* 1966, p 624–636

17. W.E. Littmann, The Mechanism of Contact Fatigue, Conf. Proc. *An Interdisciplinary Approach to the Lubrication of Concentrated Contacts,* National Aeronautics and Space Administration, 1969, p 309–377

18. E.V. Zaretsky, "Life Factors for Rolling Bearings," Society of Tribology and Lubrication Engineers (STLE) publication, 1999

21. Anon, Friction, Wear, and Lubrication Terms and Definitions, 1st revision, Organization for Economic Co-Operation and Development, Paris, 1968

22. F.J. Wren and C.A. Moyer, Modes of Fatigue Failures in Rolling Element Bearings, *Proc. Inst. Mech. Eng.,* Vol 179 (Part 3D), 1964-1965, p 236–247

23. Y.P. Chiu, T.E. Tallian, and J.I. McCool, An Engineering Model of Spalling Fatigue in Rolling Contact—The Subsurface Model, *Wear,* Vol 17, 1971, p 433–446

24. T.E. Tallian and J.I. McCool, An Engineering Model of Spalling Fatigue in Rolling Contact—The Surface Model and Engineering: Discussion and Illustrative Examples, *Wear,* Vol 17, 1971, p 447–480

25. T.E. Tallian, *Failure Atlas for Hertz Contact Machine Elements,* 2nd ed. ASME Press, 1999

26. J.J.C. Hoo, "Rolling Contact Fatigue Testing of Bearing Steels," STP 771, ASTM, 1982

27. L.D. Wedeven, R.A. Pallini, and C.G. Hingley, "Selection and Use of Wear Tests for Ceramics," STP 1010, ASTM, 1988, p 58–73

28. R. Tourret and E.P. Wright, *Rolling Contact Fatigue: Performance Testing of Lubricants,* Int. Symp., Inst. of Petroleum, Heyden and Sons, London, 1977

29. W. Cheng, W.S. Cheng, and L.M. Keer, Experimental Investigation on Rolling Sliding Contact Fatigue Crack Initiation with Artificial Defects, *Tribol. Trans.,* Vol. 37, 1994

30. G. Xu, F. Sadeghi, and M.R. Hoeprich, Dent Initiated Spall Formation in EHL Rolling/Sliding Contact, *J. Tribology (Trans. ASME),* Vol 120, 1998, p 453–462

31. Engineering Science Data Unit, "Contact Phenomena, Part I: Stresses, Deflections and Contact Dimensions for Normally-Loaded Unlubricated Elastic Components," ESDU 78035, Inst. of Mech. Eng., 1984

32. Engineering Science Data Unit, "Contact Phenomena, Part II: Stress Fields and Failure Criteria in Concentrated Elastic Contacts under Combined Normal and Tangential Loading," ESDU 84017, Inst. of Mech. Eng., 1984

33. G. Lundberg and A. Palmgren, Dynamic Capacity of Rolling Bearings," *Acta Polytech. Scand.,* Vol 1 (No. 3), 1947

34. N.G. Popinceanu, E. Diaconescu, and S. Cretu, Critical Stresses in Rolling Contact Fatigue, *Wear,* Vol 71, 1981, p 265–282, 1981

35. W.J. Anderson and E.V. Zaretsky, Rolling Element Bearings—A Review of the State of the Art, NASA TMX-71441, National Aeronautics and Space Administration, Cleveland, OH, 1973

36. T.A. Harris, Friction and Wear of Rolling-Element Bearings, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* 1992, p 499–514

37. A. Kapoor and J.A. Williams, Shakedown Limits in Sliding Contacts on a Surface-Hardened Half-Space, *Wear,* Vol 172, 1994, p 197–206

38. A. Kapoor and K.L. Johnson, Effects on Changes in Contact Geometry on Shakedown of Surfaces in Rolling Sliding Contact, *Int. J. Mech. Sci.,* Vol 34 (No. 3), 1992, p 223–239

39. O. Orringer, W.R. Paxton, D.E. Gray, and P.K. Raj, Residual Stress and Its Consequences on Both Sides of the Wheel-Rail Interface, *Wear,* Vol 191, 1996, p 25–34

40. Q. Chen, G.T. Hahn, C.A. Rubin, and V. Bhargava, The Influence of Residual Stresses on Rolling Contact Mode II Driving Force in Bearing Raceways, *Wear,* Vol 126, 1988, p 17–30

41. H. Muro, N. Tsushima, and K. Nunome, Failure Analysis of Rolling Bearings by X-Ray Measurement of Residual Stress, *Wear,* Vol 25, 1973, p 345–356

42. R.J. Pomeroy, Measurement of Residual Stress in Rolling Contact, *Wear,* Vol 16, 1970, p 393–412

Rolling Contact Fatigue

R. Ahmed, Heriot-Watt University (United Kingdom)

Rolling-Contact Fatigue of Vapor-Deposited Coatings

Advancements in rolling-bearing technology, such as the introduction of hybrid ceramic bearings that use the superior mechanical and thermal properties of ceramic balls, indicated that their full potential could only be realized by improving the performance of bearing races. Conventional materials and manufacturing processes for the fabrication of steel races were, however, at the limit of established technology, and there was a technological gap demanding an innovative approach to further improve rolling-bearing life in hostile environments. Overlay coatings such as those deposited by the physical vapor deposition (PVD) processes, which had already shown remarkable improvements in the cutting-tool technology, provided this innovation and improved the fatigue resistance of coated steel races to match that of ceramic balls. This combination of coated races and ceramic balls resulted in rolling bearings of improved performance, for example, higher rotational speeds in roller bearings for machine tool spindles (Ref 43) and pump bearings for the space shuttle main engines (Ref 44). In addition to steel races, the uses of PVD coatings have now been commercially extended to coated rollers and also coated rolling-element balls, thereby extending the performance of conventional steel bearings. The versatility of materials available for PVD coating processes provides an exotic combination of single- and multilayer functionally graded materials, for example, TiN, CrN, diamond-like carbon (DLC), and anti-friction layers of solid lubricants for space technology, for example, MoS_2 coatings. The scope of this article is, however, limited to the RCF performance of hard overlay coatings in general, with specific details of underpinning failure mechanisms of TiN and DLC coatings.

Physical vapor deposition is a gaseous-state process in which coating material is atomized or vaporized to deposit a coating. The success of PVD coatings in rolling-bearing technology (Ref 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, and 60) owes its existence to the advancements in PVD coating processes and, in particular, their ability to deposit these overlay coatings at a lower temperature, that is, with minimum residual stress in the coating layer(s) to avoid premature delamination. The commercialization of competing overlay coating technologies for RCF applications, for example, chemical vapor deposition (CVD) and thermal

spraying (TS), is in its infancy in comparison to PVD. However, despite the enormous success of PVD coatings, these competing technologies offer additional technological and economical incentives. For example, the coating thickness of CVD and TS coatings can be orders of magnitude greater than that of PVD deposits (which is seldom over 5 μm). Rolling-contact fatigue properties of substrate become less important as the coating thickness increases beyond the depth of maximum shear stress, providing the ability to combat both surface- and subsurface-initiated RCF. In addition, the high deposition rates and low cost of TS in comparison to PVD/CVD coatings provide economical incentives. Because the nature of the TS process and the resulting coating microstructure are very different from that of PVD/CVD coatings, the investigations of TS coatings for RCF applications are covered in a separate section of this article. This section deals with the RCF investigations of PVD/CVD coatings. For the description of various PVD, CVD, and TS coating processes considered in this article, readers are referred to Ref 48 and 49.

Despite the enormous success of overlay coatings in improving the RCF performance of rolling bearings, theories relating to the life prediction of coated rolling bearings are in their infancy. Most recently, research attempts in this field were made for particular bearing types, for example, NU1008 (Ref 44, 45, 46). The researchers concluded that RCF life of TiAlN coatings, in general, is consistent with the loading recommendations of bearing catalogs. Such investigations are crucial for generic solutions leading to RCF life prediction in coated components. Future studies in this area to include the effects of surface roughness, lubrication regime, and residual stress (Ref 61) are therefore inevitable before such models can be generally applied to estimate the RCF performance of coated rolling bearings. In one such attempt to include surface roughness effects (Ref 47), it was concluded that significant improvements in the RCF life of coated bearings can be made by applying relatively thicker (coating thickness ≥ 3 μm), well-adhered coatings with fine microstructure to avoid cohesive failure in thick coatings. However, due to the inherent complexity of interaction of various tribological factors that influence the RCF life prediction of coated surfaces, most of the investigations in this field are based on experimental evaluations under accelerated RCF test conditions using various RCF tribometers referred to in Table 1.

Rolling-Contact Fatigue Performance of PVD Coatings. Table 2catalogs the results of various RCF investigations of PVD coatings (Ref 43, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, and 60). It can be appreciated from Table 2 that even at relatively high contact stress levels (in excess of 4.0 GPa, or 0.6×10^6 psi), the RCF performance of PVD coatings can be well over several hundred-million cycles without failure. These investigations indicate that the RCF performance of nitride (especially TiN and HfN) and DLC coatings can easily outperform conventional bearing materials, whereas other coating materials such as CrN show a promise for the future. Sputtering (either reactive, ionic, or magnetron) and ionic beam deposition have been the preferred PVD coating process for nitride and DLC coatings, respectively.

Table 2 Published findings for RCF performance of PVD coatings

Coating process	Coating material	Substrate	RCF tester	Contact stress, GPa	Coating thickness, μm	Coating roughness R_a, μm	Coating hardness	Substrate hardness	RCF life, $\times 10^6$ cycles	Ref (year)
Ionic sputtering	CrN Mo TiAlN TiAlCN TiCN TiN + C	100Cr6	Ball on rod, full scale bearing tests (lubricated)	Contact load of 755 N for ball on rod	2.5 2.5 2.0 2.0 2.0 1.0	N/A	1300 650 2900 1800 2200 900	N/A	<100 (average) <100 (average) >250 (average) >250 (average) >250 (average) >1000 (average)	43 (1996)
Ion plating	TiN	Tool steel	Ball on cylinder	3.5, 4.6, 5.1	2 → 5	0.03 ± 0.01 0.15 ± 0.01	50 → 60 HRC	2300 HV	<1.3 (average)	50 (1998)
Reactive sputtering	TiN	M50 bearing steel	Three-ball on rod (lubricated)	5.5	0.25, 0.5, 0.75, 1, 2, 3, 5	0.05 → 0.1	30 GPa	59–60 HRC	>100 (for coating thickness <1 μm)(a)	51 (1998)
Ion plating, magnetron sputtering	TiN Ti(CN) CrN	High speed steel	Amsler wear tester (dry)	50 N (load)	1.8–3.5 1.8–3.4 3.5	0.9 ± 0.1	N/A	N/A	1 1 5	52 (1998)
Reactive magnetron sputtering	TiN ZrN	440C stainless steel (RCF life of 6.3 and 1.7 × 10^6 cycles at 4.0	Three-ball on rod (lubricated)	4.0 and 5.4	0.25, 0.5, 0.75, 1.00	N/A	55 → 59 HRC	59 HRC	36 and 8.7 (L_{10})(b) 36 and 5.5	53 (1993)

	HfN CrN Mo TiAlN TiZrN (TiAlV)N	and 5.4 GPa)							(L_{10})(b) 81 and 9.7 (L_{10})(b) 21.6 and 2.0 (L_{10})(b) 39.9 and 5.0 (L_{10})(b) 2.6 and 1.6 (L_{10})(b) 21.5 and 10.8 (L_{10})(b) 47.9 and 10.3 (L_{10})(b)	
Reactive sputtering	TiN	AISI 4118 steel	Two-disc (lubricated)	2.3	0.25–0.5 1.0 2.5 5.0	N/A	N/A	N/A	>60 10–37 1.0–4.1 <0.1	54 (1990)
Reactive sputtering	TiN	AISI 4118 steel	Two-disc (lubricated) roll/slip ≈ 0.25	2.3	1.0	0.45	2300 kgf/mm^2	62–64 HRC	>33 (average)	55 (1991)
Reactive sputtering	TiN TiAlN CrN	Si_3N_4 and M50	Four-ball (lubricated)	5.8 and 6.8	0.3–1.0	N/A	N/A	N/A	<100 for 0.75 μm TiN at 5.84 GPa(c)	56 (1994)
Magnetron sputtering	TiN/NbN	M50	Three-ball on rod	3.4	0.5	0.075–0.1	5200 kgf/mm^2	561–63 HRC	191(d)	57 (1998)

			(lubricated)	5.2					15(d) (L_{10} in hours)	
Ion beam	DLHC	M50 AISI 52100 AISI 4118 AISI 440C	Three-ball on rod (lubricated)	5.5	0.5 → 1	0.06 → 0.07	N/A	N/A	76.4 (L_{50} life), 10(e) 232 (L_{50} life), 14(e) 327 (L_{50} life), 10(e) 22.7 (L_{50} life), 3.7(e)	58 (1993)
Ion beam	DLC	M50	Three ball on rod (lubricated)	5.5	0.5 → 1	N/A	11.3 → 14.7 GPa at 25 g load	10.8 GPa at 25 g load	1 to 300(f)	59 (1992)
Ion beam	DLC	M50	Three ball on rod (lubricated)	4.8	0.33	N/A	N/A	N/A	91.5 at 23 °C (73 °F) (L_{50})(g) 45.9 at 177 °C (73 °F) (L_{50})(g)	60 (1997)

Note: AISI, American Iron and Steel Institute; DLHC, diamond-like hydro-carbon. (a) Best performance at L_{50} of 25×10^6 and $>100 \times 10^6$ with rough and smooth balls, respectively, for coating thickness <1 μm. (b) Best L_{10} values quoted, which relate to coating thickness range of 0.5–0.75 μm at two stress levels of 4.0 and 5.4 GPa, respectively. (c) All coatings in hybrid ceramic combinations showed improvement in average RCF life, with best average improvement by a factor of 2.5 for 0.75 μm TiN coating over uncoated case. (d) L_{10} values quoted in hours for superlattice period of 3–6 nm at respective stress level. Uncoated M50 substrate had L_{10} of 39 and 9 h at stress levels of 3.4 and 5.2 GPa, respectively. (e) L_{50} improvement factor over uncoated case. (f) L_{50} and L_{10} of 13×10^6 and 0.6×10^6, respectively, with no significant influence of coating thickness. (g) L_{10} of 36.7×10^6 and 38.4×10^6 at 23 and 177 °C (73 and 351 °F), respectively.

Although PVD coating deposition parameters used to deposit the coatings for these RCF investigations have not been indicated in Table 2, they can have significant influence on coating microstructure and thus its RCF performance; for example, for reactive sputtering of TiN coatings, intercolumnar porosity of TiN coatings can be decreased by increasing the substrate temperature (Ref 62). However, the influence of PVD deposition conditions (parameters) on the RCF performance has not yet been clearly addressed in published literature. For now, it can be assumed that the coatings referred to in Table 2 were deposited under the best-known deposition conditions available for specific coating material and PVD process. Wherever possible, the RCF performance levels in Table 2 are quoted in terms of L_{10}, L_{50}, or average life for statistical comparison. However, these results are generally qualitative and should be used as the basis of rating the RCF performance of various coatings rather than to predict the statistical RCF life.

Table 2 also indicates that the RCF results are sensitive to parameters such as coating thickness, substrate hardness, and lubrication regime. Even for the best possible combination of coating thickness, substrate hardness, and lubrication (e.g., in TiN coatings of less than 1 μm coating thickness), potential RCF improvements have been shown to vary, depending on the surface roughness of the contacting pair (Ref 51). Such investigations indicate that for very smooth surfaces of rolling-contact bodies, the improvements in RCF performance by the application of PVD coatings can be marginal. However, rolling bearings generally operate under harsh tribological environments, and potential RCF improvements by surface coatings are generally realized in most industrial applications. Table 2 also indicates the significant scatter in RCF performance while comparing the results from various accelerated RCF test methods. Even for a given test method and coating, variations in stress levels have shown significant variations in potential RCF improvements. Nevertheless, the performance indicators referred to in Table 2 have also been shown to match that of full-scale bearing tests (Ref 63). It is worth appreciating that this dependency of RCF performance on various tribological design parameters and RCF test methods is not unique to overlay coatings. Conventional steel bearings often display such dependency, with the exception that this interdependency in the case of surface coatings is much more complex due to additional influential factors such as coating microstructure, thickness, and interfacial bond strength.

The dependence of RCF performance on the tribological design parameters, such as coating thickness, substrate hardness, tribological condition of contact stress, lubrication regime, and surface topography, can actually be understood by a generic understanding of underlying failure modes. What dictates the influence of the previously mentioned design factors is their sensitivity to initiate and propagate various RCF failure modes, the understanding of which cannot only provide better life prediction models but also improved RCF performance of coated rolling bearings. As indicated earlier, the lack of data on all coating materials has made it almost impossible to ascertain their failure modes, although their RCF performance is represented in Table 2. The scope of the remaining section is hence focused to ascertain the underlying failure mechanisms of the two most widely used PVD coating materials, that is, TiN and DLC coatings. Before embarking on the discussion of the failure mechanism of these coating materials, it is important to appreciate the generics of the tribological design approach adapted to improve RCF performance by the application of PVD coatings.

Tribological Concept of Depositing Thin Coatings to Improve RCF Performance. Physical vapor deposition coatings, which are strongly adhered to the substrate, are very thin and seldom over a few microns in thickness. These thin coatings generally follow the topography of the substrate material, so that there is no appreciable improvement in the surface roughness of the coated substrate. So, how does such a thin coating improve the RCF performance? Figures 2 and 8 indicated various crack initiation sites during a contact loading. Obviously, such a thin coating in contact with a smooth counterbody marginally affects the depth of orthogonal, maximum shear or von Mises stress. Also, the influence of a thin coating on the subsurface stress distribution is marginal for low values of coating-to-substrate modulus ratio (Ref 64, 65). Hence, these thin coatings do not improve the

resistance to subsurface fatigue crack initiation and propagation. However, the near-surface stress field and near-surface crack initiation and propagation can be seriously affected by the application of such hard coatings. This is achieved by the interaction of the hard coating layer on the counterbody, which initiates its micropolishing either in two-body abrasion (e.g., for TiN coatings) or in three-body abrasion (e.g., for DLC coatings) to improve the overall RCF performance of the contacting pair by minimizing the stress protrusions at individual asperity contact. In essence, these coatings tend to delay near-surface crack initiation, which is thought to constitute 90% of the total life to RCF failure (Ref 60). In addition, improvement in frictional properties of the contacting pair, for example, by DLC and molybdenum coatings, provides additional improvements in RCF performance.

Rolling-Contact Fatigue Failure Modes of PVD Coatings. Despite numerous studies to comprehend the RCF performance of PVD coatings (Table 2), only a few have attempted to outline a fundamental understanding of failure mechanisms. Some studies (Ref 51, 53, 54, 55, 58, 59, and 60) only begin to understand the underpinning failure mechanisms of PVD coatings, as discussed subsequently for TiN and DLC coatings.

Failure Modes of PVD-TiN Coatings. The structure of TiN coating deposited by various PVD processes is generally columnar with some evidence of equiaxial grains near the interface (Ref 50, 51). These coatings have shown good bonding with various substrate steels, for example, M50, 440C, and so on, and also, TiN coatings on Si_3N_4 ceramics (Ref 56) have been successfully applied. Apart from good bonding of TiN coating either directly onto the substrate or by functional grading, the most important substrate property, which dictates the RCF performance of coated rolling elements, is substrate hardness. The choice of substrate material and coating deposition conditions are therefore critical to allow the retention of substrate hardness after deposition. Failure to retain substrate hardness after deposition adversely affects the RCF performance (Ref 50, 53). The underlying failure mechanism caused by a reduction in substrate hardness is principally due not only to a compromise in the ability of substrate to support the coating, but also to microstructural changes associated with substrate softening that can result in stress concentrations and create fatigue initiation sites. Hence, tribological design of these coatings must not only aim for a well-bonded coating but also for the retention of substrate hardness.

In addition to substrate bonding and its hardness, the most critical parameter that has been shown to significantly influence the RCF performance of TiN coatings is coating thickness. TiN-coated cutting tools are generally coated up to a coating thickness of 3 μm. However, investigations have shown that such coating thickness can adversely affect the RCF performance. In fact, the optimal thickness for improved performance is generally reported as approximately 0.5–0.75 μm (Ref 51, 53, 54, and 56). So, why does any increase in coating thickness beyond this level adversely affect the RCF performance? The answer to this question lies in the TiN coating microstructure, and there are two schools of thought that explain this dramatic influence of coating thickness: adhesive failure of thicker coatings (Ref 50, 52, 54, 55) and cohesive failure (Ref 51).

Polonsky et al. (Ref 51) argued that even in a good-quality PVD (TiN) coating with negligible porosity, the columnar coating microstructure becomes coarser away from the interface. Hence, an increase in coating thickness causes competitive growth during film deposition. (Although it was assumed that the best available deposition conditions were used to deposit the PVD coatings listed in Table 2, it is worth appreciating that these are not entirely defect-free coatings, and further improvements in RCF performance of PVD coatings can be realized by improved coating microstructure.) The interfaces between these columns thus represent the crack initiation sites within the coating microstructure, the probability of which (i.e., defects at column boundaries) increases with the coating thickness. Hence, the coating fails cohesively with the increase in coating thickness—and the failed coating area is very rough due to shear within the columnar coating microstructure—rather than adhesively at the interface. Contrary to this, some studies (Ref 50, 52, 54, 55) suggest that the weakest section in the coated rolling element is the coating-substrate interface, and hence, the failure is adhesive. The substrate hardness can further influence the failure mechanism in TiN coatings, (Ref 50) as shown in Fig. 9—that is, cracks propagate perpendicular to the coating substrate interface for softer substrate coatings (Fig. 9a), and parallel to the coating-substrate interface for harder substrates (Fig. 9b, c). From Fig. 9, it is striking to note that for the scanning electron microscope representation of failed areas for harder substrates (Fig. 9b, c), which are more closely related to the tribological conditions reported in Ref 50, the failure is just above the coating-substrate interface (i.e., similar to a cohesive failure) rather than at the coating-substrate interface. Hence, there seems to be consistent experimental evidence pointing to cohesive, rather than adhesive failure of thicker TiN coatings. The theory of adhesive failure presented in Ref 50, 52, 54, and 55 thus relies on weakness of the coating-substrate interface to explain the influence of increase in coating thickness on reduced

RCF performance. Reference 50 also suggests the theory put forward in Ref 66, which relates the cyclic shear stress to crack initiation; this seems unlikely, because the depth of maximum and orthogonal shear stress for the smooth, lubricated contacts considered in the experimental study can be an order of magnitude deeper than the coating thickness.

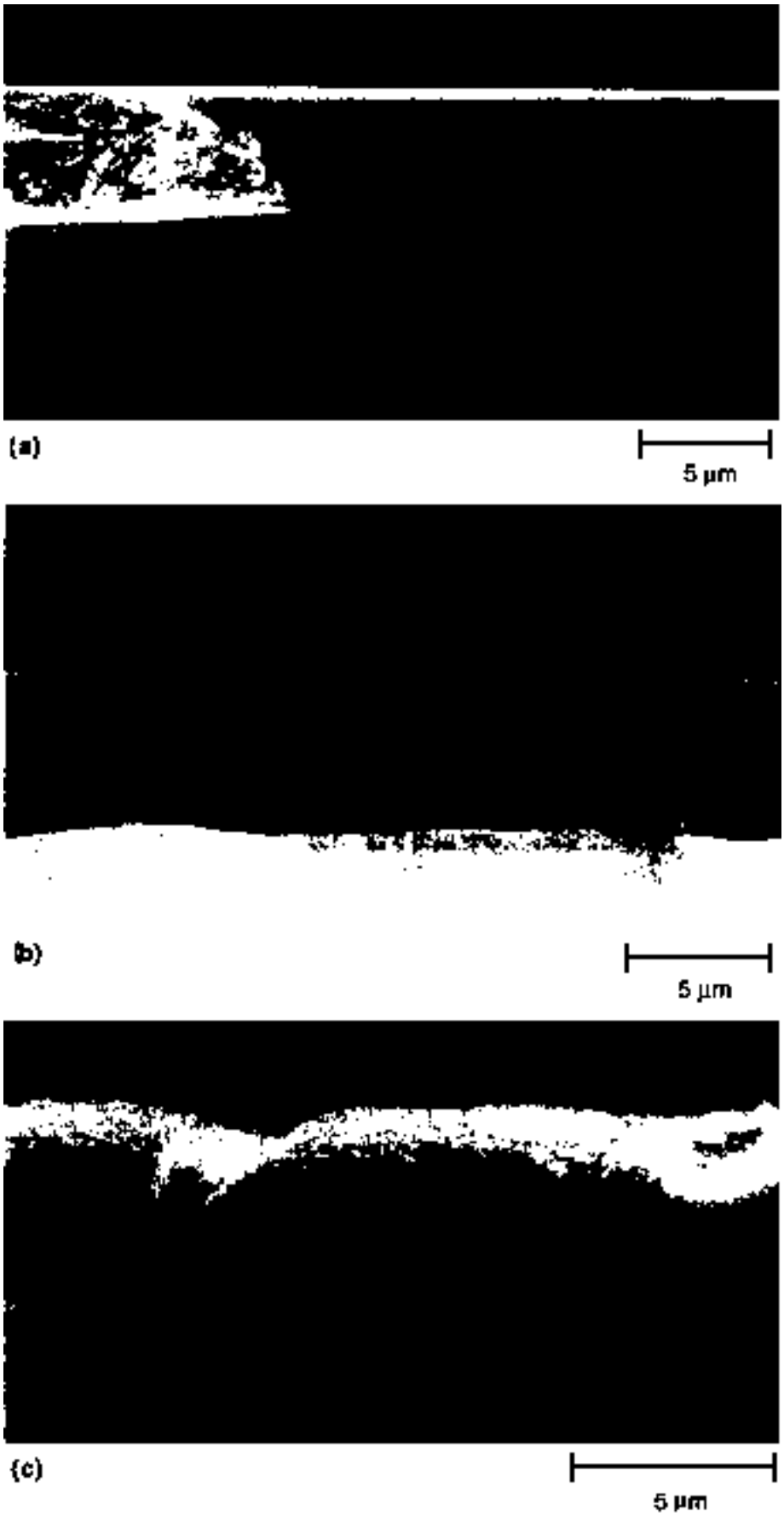


Fig. 9 Morphology of cracks leading to rolling-contact fatigue failure of PVD (TiN) coatings. (a) Crack parallel to the interface leading to spalled area for hard substrate (60 HRC) TiN coating. (b) Cracks parallel to the coating-substrate interface for hard substrate (60 HRC) TiN coating. (c) Cracks perpendicular to the coating-substrate interface for soft substrate (50 HRC) TiN coating. Source: Ref 50

Another reason for the adhesive failure of thicker coatings can be understood by comparing the Young's modulus of the coating to that of the substrate material. It is postulated that as the difference in Young's modulus increases, thicker coatings resist following the elastically deformed profile of the substrate (under the contact region in response to cyclic loading), whereas thinner coatings tend to follow a similar profile to the elastically deformed substrate. This could lead to premature failure in thicker TiN coatings. More recently, however, work reported on TiAlN coatings (Ref 44) has indicated that as the difference in the modulus between the coating and substrate material increases, even for smooth contact surfaces, the coating-substrate interface can be subjected to a high degree of shear stress in thin (2–3 μm thick) PVD coatings. Similar results were reported in Ref 64. Functional grading of the coating can thus be useful to improve the RCF performance of thicker coatings.

Even if some of the reported failures (Ref 50, 52, 54, 55) were considered truly adhesive, the test methodologies in these studies were very much alike (i.e., two-disc machine or its modification) and very different to the ball-on-rod test method (Ref 51), which shows cohesive failure. Hence, the model contact geometry can be responsible for the change in failure mode, which is also known to influence the failure mode in conventional ceramics and steels (Ref 67). Also, there were considerable differences in substrate shakedown behavior. Some studies (Ref 51) indicated substrate yielding in the first few cycles, due to high contact stress that was above the yield strength of the substrate M50 steel. This shakedown effect (Ref 37, 38) cannot only improve the RCF performance by substrate work hardening and improved coating-substrate conformity near their interface, but also influence the RCF failure mode. Contrary to this, studies reported in Ref 50, 52, 54, and 55 did not indicate any such shakedown effects. Hence, the difference between the two test conditions can also be thought responsible for the reported differences in cohesive and adhesive failures.

The tribological explanation of how thinner (0.25–0.75 μm) PVD coatings avoid cohesive and adhesive failure and thus improve the RCF life lies in the microcontacts within the Hertzian contact area. Thinner TiN coatings, which do not fail during the RCF tests, do not show any change in their surface roughness, apart from surface glazing. The surface of the counterbody, generally steel, however, undergoes micropolishing during the RCF test. This micropolishing reduces the stress protrusions at the asperity level, which is generally thought to be responsible for near-surface initiation of RCF failure. Indeed, it has been confirmed that for very smooth counterbodies—for example, grade 24 steel balls in their study (R_a = 0.01 μm)—the contacting surfaces did not have any further scope of micropolishing, and hence, no substantial increase in RCF performance was obtained (Ref 51). However, it is worth appreciating that it is not only the surface roughness but also the hardness difference of the contacting pair and their friction properties, that also dictate the RCF performance of coated rolling-element bearings. Rolling bearings generally operate in harsh tribological environments, and the scope of improvement in RCF performance by coating application is always realized, for example for self-mated TiN couples in pure rolling or rolling/sliding contacts (Ref 52), and by Igartua et al. (Ref 43) for hybrid ceramic bearings where the counterbody is a ceramic of similar hardness to a TiN coating.

The theory of interaction of surface asperities to provide micropolishing and subsequent improvements in RCF life also provides some clues as to why nitride coatings (Ref 53), thinner than 0.25 μm provide only marginal RCF improvements. According to the aforementioned theories of cohesive and adhesive failure in TiN coatings, a thinner coating should theoretically provide a better resistance to both failure mechanisms, that is, thinner coatings have less probability of intercolumnar coarsening and hence low probability of cohesive failure. Also, thinner coatings should adequately follow the deformed substrate profile, so low stress concentrations at the interface should resist adhesive failure. Marginal RCF improvements for coatings thinner than 0.25 μm can, however, be explained in terms of the influence of the substrate surface roughness. Even for a smooth substrate, the peak-to-valley height of the surface profile can be significantly greater than the average (R_a) surface roughness. A thinner coating may, therefore, only provide partial protection to surface asperities, especially in the valleys of the surface profile, for example, due to shadowing effects. This can minimize the rate of micropolishing of the counterbody, and hence, the delay in crack initiation is not maximized, which is essential for optimizing the RCF performance.

Failure Modes of PVD-Diamond-Like Hydrocarbon (DLHC)/DLC Coatings. Superior tribological properties of DLC coatings (Ref 48) have shown considerable promise in improving the RCF life of coated rolling elements. References 58 and 59 and, more recently Ref 60, reported most of the work in this field. As with the TiN coatings, these results relate to investigations in mixed lubrication regimes. Except for the room-temperature RCF studies reported in Ref 60 (Table 2), where coatings were much thinner (~33 nm) than those considered in Ref 58 and 59 (0.5–1 μm), both DLC and DLHC coatings have been shown to provide considerable improvements in the RCF life of coated rolling elements. The physical mechanism underpinning such improvement is different from that of TiN coatings (mentioned previously); however, the underlying analogy to micropolishing is the same.

In the studies reported in Ref 58 and 59, two tribological mechanisms are thought to be responsible for the improvement of RCF life. The first mechanism is three-body abrasion, leading to the micropolishing of the counterbody and thus lowering stress protrusions within the contact region at the asperity level. This three-body abrasion is thought to be caused by the delaminated DLC or DLHC coating particles during RCF testing, which, when mixed with test lubricant, polish the softer surface within the contact region. Wei et al. (Ref 58, 59) postulated that this mechanism was due to the absence of diamond film at the time of spall or at the end of the terminated tests. This essentially refers to the same mechanism of micropolishing as was reported for TiN coatings, except for the fact that the micropolishing is in three-body abrasion. The second mechanism thought responsible for improved RCF performance is the gradual graphitization of the DLHC film during the RCF test, leading to reduced friction at asperity level, and hence a delay in the surface-initiated RCF. This also explains the less dramatic influence of the DLC/DLHC coating thickness in the range 0.5–1.0 μm, and further supports the theory that coating detachment is responsible for RCF improvement, namely, three-body abrasion. Although for much thinner coatings (~33 nm), in which there was no RCF improvement seen at room-temperature tests (Ref 60), it is worth appreciating that there was no evidence of delamination, which is thought to be essential for micropolishing of asperities and hence, RCF improvement. Tests conducted at a higher temperature (175 °C, or 350 °F) with these thinner (~33 nm) coatings, however, showed some improvement to the RCF life. The exact mechanism leading to this improvement is not clear at this stage.

Rolling-Contact Fatigue of CVD Coatings. By using smoother contact bodies to suppress the mechanism of micropolishing within the coating region (Ref 51), very thin coatings could not provide significant improvement in RCF performance. In such cases, coating needs to be relatively thick (coating thickness ≥3 μm) and should have very high cohesive strength to provide RCF improvements. Chemical vapor deposition coatings, which have the versatility to deposit much thicker layers, can thus provide useful improvements in RCF life. Coatings in the thickness range of 3–190 μm have therefore been the focus of some RCF studies. However, the establishment of tribological data that could lead to generic design and understanding of failure modes in CVD coatings is in its infancy. Rolling-contact fatigue performance data of specific coating types is, however, available, as listed in Table 3 (Ref 68, 69, 70, 71, 72, 73). Among these coatings, titanium-base nitride coatings have shown RCF performance up to a maximum of 50 million stress cycles. This does not match the extraordinary performance indicated earlier for PVD coatings; nevertheless, it is worth appreciating that some of these investigations (Ref 68, 69) were carried out to measure the bond strength rather than the RCF performance of these coatings, and tests were suspended after a specific number of stress cycles. The best performance for CVD coatings was reported in Ref 73 for TiC coatings. They showed that for CVD-coated TiC rolling-element balls, bearings can last in excess of 900 million revolutions in full-scale testing, and also that CVD coatings delayed the onset of lubricant degradation by reducing the effect of cold welding within the microcontact region.

Table 3 Published findings for RCF performance of CVD coatings

Coating process	Coating	Substrate	RCF tester	Contact stress, GPa	Coating thickness, µm	Surface roughness, µm	Coating hardness	Substrate hardness, HRC	Fatigue life, $\times 10^6$ cycles	Ref (year)
PACVD	TiN (TiSi)N	High-speed steel	N/A	Equivalent bond stress of 800 MPa	3 → 4	0.6	2000 → 2500 HV	63–65	5 (suspended tests)	68 (1995)
PCVD	TiN TiCN	High-speed steel	Spherical	0.8 1.7 (estimated)	2.7	0.04	N/A	63 45–63	>5 >50 <0.1 at higher stress for both coatings	69 (1997)
PECVD	TiN	High speed steel	Cylindrical spherical	Critical stress of 0.2–0.7	N/A	0.04	N/A	65	50 (at lower stress level)	70 (1998)
CVD	CrC-TiC	M50 bearing steel	Ball on rod, disc	2.76 → 4.83	5.4	0.02	64 HRC	65	5–10 × 10^6 (L_{50} life)	71 (1985)
CVD	SiC	SiC-TiC graphite	Three ball on rod	5.5	192	0.08	N/A	N/A	>50	72 (1995)
CVD	TiC	440C	Full scale	40 N load	3	0.005	35,000 MPa	N/A	>900 × 10^6 at 1400 rpm	73 (1993)

Note: PACVD, plasma-assisted chemical vapor deposition; PCVD, plasma chemical vapor deposition; PECVD, plasma-enhanced chemical vapor deposition

The real advantage of CVD coatings is, however, in their ability to deposit layers that are an order of magnitude thicker than PVD coatings. Chao et al. (Ref 72) have recently reported that not only the coating thickness but also the beneficial compressive stress tailored in these thicker CVD coatings can be used to combat RCF failure. Although the best performance achieved in their setup was in excess of 50 million stress cycles for a 190 μm thick SiC coating with 680 MPa (100 ksi) compressive residual stress, these tests were suspended due to the failure of the steel balls rather than the coating. For those coatings that failed, the failure was cohesive, and the authors argued that the failure was due to discontinuous growth of SiC film. Such discontinuous growth may, however, be combated by appropriate selection of deposition conditions. In general, limited studies indicate that CVD-coated bearings outperform conventional bearing steels (Ref 72, 73), and future studies are inevitable to comprehend the underpinning failure mechanisms of these coatings.

References cited in this section

37. A. Kapoor and J.A. Williams, Shakedown Limits in Sliding Contacts on a Surface-Hardened Half-Space, *Wear,* Vol 172, 1994, p 197–206

38. A. Kapoor and K.L. Johnson, Effects on Changes in Contact Geometry on Shakedown of Surfaces in Rolling Sliding Contact, *Int. J. Mech. Sci.,* Vol 34 (No. 3), 1992, p 223–239

43. A. Igartua, J. Laucircia, A. Aranzabe, T. Leyendecker, O. Lemmer, G. Erkens, M. Weck, and G. Hanrath, Application of Low Temperature PVD Coatings in Roller Bearings: Tribological Tests and Experiences with Spindle Bearing Systems, *Surf. Coat. Technol.,* Vol 86–87, 1996, p 460–466

44. A. Erdemir, Rolling Contact Fatigue and Wear Resistance of Hard Coatings on Bearing Steel Substrates, *Surf. Coat. Technol.,* Vol 54–55, 1992, p 482–489

45. K.D. Bouzakis, N. Vidakis, A. Lontos, S. Mitsi, and K. David, Implementation of Low Temperature-Deposited Coating Fatigue Parameters in Commercial Roller Bearings Catalogues, *Surf. Coat. Technol.,* Vol 133–134, 2000, p 489–496

46. K.D. Bouzakis, N. Vidakis, and S. Mitsi, Fatigue Prediction of Thin Hard Coatings on the Steel Races of Hybrid Bearings Used in High Speed Machine Tool Spindles, *J. Tribology, (Trans. ASME),* Vol 120, 1998, p 835–842

47. I.A. Polonsky, T.P. Chang, L.M. Keer, and W.D. Sproul, An Analysis of the Effect of Hard Coatings on Near-Surface Rolling Contact Fatigue Initiation Induced by Surface Roughness, *Wear,* Vol 208, 1997, p 204–219

48. K. Holmberg and A. Matthews, *Coatings Tribology: Properties, Techniques, and Applications in Surface Engineering, Tribology Series,* Vol 28, Elsevier Science, 1994

49. L. Pawlowski, The Science and Engineering of Thermal Spray Coatings, John Wiley & Sons, 1995

50. N.J.M. Carvalho, A.J. Huis in't Veld, and J.T. De Hossen, Interfacial Fatigue Stress in PVD TiN Coated Tool Steels Under Rolling Contact Fatigue Conditions, *Surf. Coat. Technol.,* Vol 105, 1998, p 109–116

51. I.A. Polonsky, T.P. Chang, L.M. Keer, and W.D. Sproul, A Study of Rolling-Contact Fatigue of Bearing Steel Coated with Physical Vapor Deposition TiN films: Coating Response to Cyclic Contact Stress and Physical Mechanisms Underlying Coating Effect on the Fatigue Life, *Wear,* Vol 215, 1998, p 191–204

52. G. Meier zu Kocker, E. Santner, and B. Loffelbein, Functional Behavior of PVD Coatings under Sliding and Rolling Stress Conditions, *Tribotest J.,* Vol 4 (No. 3), 1998, p 241–260

53. R. Thom, L. Moore, W.D. Sproul, and T.P. Chang, Rolling Contact Fatigue Tests of Reactively Sputtered Nitride Coatings of Ti, Zr, Hf, Mo, Ti-Al, Ti-Zr, Ti-Al-V on 440C Stainless Steel Substrates, *Surf. Coat. Technol.,* Vol 62, 1993, p 423–427

54. T.P. Chang, P. Taung-Sea, H.S. Cheng, and S. Herbert, Influence of Coating Thickness on Lubricated Rolling Contact Fatigue Life, *Surf. Coat. Technol.,* Vol 44 (No. 1–3), 1990, p 699–708

55. T.P. Chang, H.S. Cheng, H.S. Chiou, and W.D. Sproul, Comparison of Fatigue Life Morphology between TiN Coated and Uncoated Rollers, *Tribol. Trans.,* Vol 34 (No. 3), 1991, p 408–416

56. P.J. Rudnik and W.D. Sproul, "Ceramic Coated Bearing Elements for Improved Durability and Reliability," final report, Project F33615-92-C-5295, Advanced Research Projects Agency (APRA), 1994

57. M.A. Liston, Rolling Contact Fatigue Properties of TiN/NbN Superlattice Coatings on M-50 Steel, *ASTM Spec. Tech. Publ.,* Vol 1327, 1998, p 499–510

58. R. Wei, P.J. Wilbur, M. Liston, and G. Lux, Rolling Contact Fatigue Wear Characteristics of Diamond-Like Carbon Coatings on Steels, *Wear,* Vol 162–164, 1993, p 558–568

59. R. Wei, P.J. Wilbur, and F.M. Kustas, A Rolling Contact Fatigue Study of Hard Carbon Coated M-50 Steel, *Trans ASME,* Vol 114, 1992, p 298–302

60. L. Rosado, V.K. Jain, and H.K. Trivedi, The Effect of Diamond-Like Carbon Coatings on the Rolling Fatigue and Wear of M-50 Steel, *Wear,* Vol 212, 1997, p 1–6

61. A.J. Perry, M. Jagner, W.D. Sproul, and P.J. Rudnik, Residual Stress and Strain Distribution Anomalies in TiN Films Deposited by Physical Vapour Deposition, *Surf. Coat. Technol.,* Vol 42 (No. 1), 1990, p 49–68

62. I. Petrov, L. Hultman, U. Helmersson, and J.E. Sundgren, Microstructure Modification of TiN by Ion Bombardment during Reactive Sputter Deposition, *Thin Solid Films,* Vol 169, 1989, p 299–314

63. K.D. Bouzakis, N. Vidakis, A. Igartua, T. Leyendecker, G. Erkens, M. Weck, G. Hanrath, and A. Schumacher, "Application of Low Temperature PVD Coatings in Roller Bearings-Fatigue Prediction of Thin Hard Coatings—Tribological Tests and Experiments with Spindle Bearing Systems," final report of the BRITE project Eurobearings, European Commission technical report, BRE-CT92-0333, 1997

64. H. Djabella and R.D. Arnell, Finite Element Analysis of Contact Stress in Elastic Coating on an Elastic Substrate, *Thin Solid Films,* Vol 213, 1992, p 205–209

65. P.K. Gupta and J.A. Walowit, Contact Stress between an Elastic Cylinder and a Layered Elastic Solid, *Trans. ASME,* 1974, p 250–257

66. A. Otsuka, H. Sugawara, and M. Shomura, A Test Method for Mode-II Fatigue Crack Growth Relating to a Model for Rolling Contact Fatigue, *Fatigue Fract. Eng. Mater. Struct.,* Vol 19, 1996, p 1265–1275

67. M. Hadfield and T.A. Stolarski, The Effect of the Test Machine on the Failure Mode in Lubricated Rolling Contact of Silicon Nitride, *Tribol. Int.,* Vol 28, 1995, p 377–382

68. J. He, C. Bai, K. Xu, and N. Hu, Improving the Anticorrosion and Mechanical Behaviour of PACVD TiN, *Surf. Coat. Technol.,* Vol 74–75, 1995, p 387–393

69. J. He, K. Xu, and N. Hu, Evaluation of Bonding Strength of Thin Hard Films by Spherical Rolling Test, *Surf. Coat. Technol.,* Vol 97, 1997, p 295–298

70. K. Xu, N. Hu, and J. He, Evaluation of the Bond Strength of Hard Coatings by the Contact Fatigue Test, *J. Adhes. Sci. Technol.,* Vol 12, 1998, p 1055–1069

71. J.F. Dill, M.N. Gardos, H.E. Hintermann, and H.J. Boving, Rolling Contact Fatigue Evaluation of Hardcoated Bearing Steels, *ASCE Publ.,* 1984, p 230–240

72. L. Chao, R. Lakshminarayanan, D.K. Shetty, and R. Cutler, Rolling Contact Fatigue and Wear of CVD SiC with Residual Surface Compression, *J. Am. Ceram. Soc.,* Vol 78 (No. 9) 1995, p 2307–2313

73. A. Savan, H.J. Boving, F. Fluehmann, and H.E. Hintermann, Increased Performance of Bearings Using TiC-Coated Balls, *J. Phys. (France) IV,* Colloque C7, Supplement au *J. Phys. (France) III,* Vol 3, Nov 1993, p 943–947

Rolling Contact Fatigue

R. Ahmed, Heriot-Watt University (United Kingdom)

Rolling-Contact Fatigue of TS Coatings

Thermal spraying is a molten- or semimolten-state process in which coating material in the form of a wire, rod, or powder is heated by combustion, gas plasma, or electric arc and accelerated toward the substrate to deposit a coating. The most commonly used classification of TS processes is based on the method of heat generation, and the most commonly used processes considered for RCF investigations are high-velocity oxyfuel (HVOF), detonation-gun (D-gun), and plasma spraying (PS). Within each category of these processes, there are sub-classifications, for example, HVOF can be liquid fueled or gas fueled, and PS can be air plasma sprayed (APS) or vacuum plasma sprayed (VPS). For details of various TS processes, readers are referred to Ref 49.
The coating materials available for TS processes range from high-abrasion-resistant ceramics (e.g., Al_2O_3) and cermets (e.g., WC-Co) to abradabale coatings (e.g., polymer-based coatings for aeroengine applications). The coating microstructure of TS coatings is lamellar, and their industrial applications are found in aerospace, biomedical, automotive, and manufacturing industries. The advancements in TS technology, for example, liquid-fuel HVOF systems and VPS, made it possible to attain TS coatings with negligible porosity and minimum powder degradation, thereby improving the fracture toughness of TS coatings. Despite recent innovations in TS technology and its ever-expanding market share, RCF investigations of these coatings are in their infancy (Ref 74, 75, 76, 77, 78, 79, 80, 81, 82, 83). The aim of this section is to comprehend the RCF performance and failure modes of TS coatings, with a view to defining the tribological design considerations. The rationale behind the RCF applications of TS coatings lies in their competitive advantages, which include:

- Very high deposition rates (typically 0.0025 kg/s, or 20 lb/h, for HVOF coatings)
- Ability to deposit thick coatings (up to a few mm) to resist both surface- and subsurface- initiated RCF
- Ability to produce coatings on large components
- Ability to restore worn or undersized components, making them environmentally friendly
- Wide range of coating and substrate materials
- Low deposition cost

Rolling-Contact Fatigue Performance of TS Coatings. Table 4summarizes the RCF performance results for various TS coatings (Ref 74, 75, 76, 77, 78, 79, 80, 81, 82, 83). As with the RCF performance of vapor deposition coating processes (Tables 2 and 3), these results are qualitative and should be considered to rank the RCF performance rather than for statistical fatigue-life prediction. These results indicate that the most commonly used TS coating material for RCF investigations is WC-Co cermet and Al_2O_3 ceramic. Among these

coating materials, the RCF performance of cermets is superior to that of ceramics. The RCF performance of these coatings is, however, not as high as that associated with PVD (Table 2) coatings at stress levels greater than 3 GPa (0.4×10^6 psi) in point-contact loading. At moderate-to-low stress levels (up to 3 GPa, or 0.4×10^6 psi), these coatings, however, exhibit a RCF life in excess of 70 million stress cycles without failure. The scatter in RCF performance data shown in Table 4 is due to the influence of variations in tribological conditions as well as design parameters such as coating thickness, substrate hardness, and complexities in the coating microstructure. These design parameters significantly influence the RCF failure modes, as discussed subsequently.

Table 4 Published findings for RCF performance of thermally sprayed coatings

Coating process	Coating material	Substrate material	RCF tester	Contact stress, MPa	Coating thickness, μm	Coating roughness (R_a), μm	Average coating hardness	Substrate hardness	RCF life, $\times 10^6$ cycles	Ref (year)
APS, HVOF	WC-12%Co WC-17%Co WC-10%Co-4%Cr	DIN 17200 steel	Two-disc (dry)	420–600	100	0.12–0.51	758–1208 HV	32–37 HRC	>10 [(a)]	74 (1997)
HVOF	WC-Cr-Ni	JIS G4105 steel	Two-disc (lubricated)	2400–3000	20–90	0.2	940	790 HV	>20 (average)[(b)]	75 (1995)
D-gun	WC-15%Co Al_2O_3	M50 steel	Four-ball (lubricated)	3000–3400	70	0.04	1200 1050	658 HV	0.16–0.6 <0.1	76 (1997)
D-gun	WC-15%Co	M50 steel	Four-ball (lubricated)	3400 5200	50	0.02	1200	600 HV	2.9 0.4	77 (1996)
APS	Ni-B-Si-Cr-Fe-C	Steel	Two-disc (lubricated)	822, 845, 960, 1086	200 450	0.24	N/A	60 HRC	<5.5 <3	78 (2000)
HVOF	WC-Cr-Ni	S45C steel	Two-disc (lubricated)	1400	40–90	0.1	920	308–670 HV	>20[(c)]	79 (2000)
HVOF, D-gun	WC-12%Co	Steel	Three-disc Two-disc	410	100	0.35	1050, 1150	710 HV	…	80 (1994)
HVOF	WC-12%Co	Mild steel	Four-ball (lubricated)	1700–1900	20 50 150	0.05	1318	218 HV	>70[(d)] 2–20 15–30	81 (1997)
HVOF	WC-12%Co	440C steel	Four-ball (lubricated)	2700–3100	50	0.06	1296	850	<1.5	82 (2001)

					150 250				>70[d] >70[d]	
APS	Al_2O_3-TiO_2 Cr_2O_3-SiO_2-TiO_2 Mo	A2017 aluminium alloy	Two-disc (lubricated)	1500	300–350	N/A	805, 1089, 491	N/A	<10[e]	83 (1990)

(a) Best performance for HVOF WC-Co coating.
(b) For coatings thicker than 40 μm.
(c) For positive (≈12%) and zero slip ratio.
(d) Test suspended after 70×10^6 stress cycles without failure.
(e) Best performance for molybdenum coatings

Rolling-Contact Fatigue Failure Modes in TS Cermet (WC-Co) and Ceramic (Al_2O_3) Coatings. Investigations relating to the RCF failure modes of TS coatings are reported in Ref 7, 74, 79, 84, and 85. These investigations have classified the fatigue failure modes on the basis of surface and subsurface observations in pre- and post-RCF conditions. Rolling-contact fatigue failure of TS coatings are generally categorized in four main modes and named as abrasion, delamination, bulk failure, and spalling (M1–M4), as indicated in Fig. 10. A summary of the underpinning failure mechanisms leading to these failure modes is given subsequently, details of which can be seen in the aforementioned references.

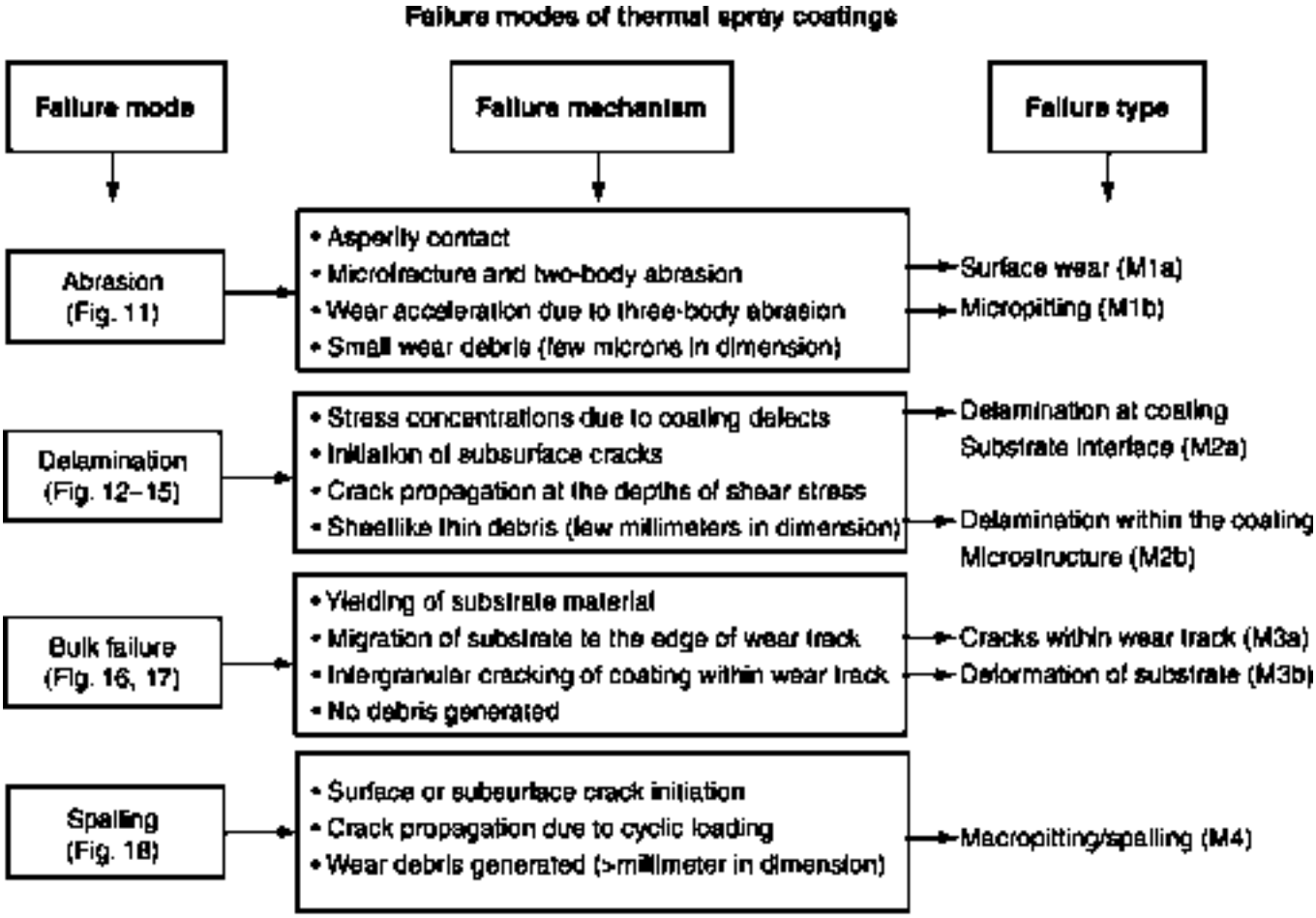


Fig. 10 Rolling-contact fatigue failure modes of thermal spray cermet and ceramic coatings. Source: Ref 84

Abrasive Failure of Coated Rolling Elements. The combined effect of micropitting and surface wear on the wear track of components in rolling contact is collectively termed as abrasive failure in TS coatings. This failure mode is seen with both ceramic (Al_2O_3) and cermet (WC-Co) coatings and with all TS coating techniques. A typical example of this failure is shown in Fig. 11. Noncontacting three-dimensional interferometry of failed wear track indicates that the micropits leading to abrasive failure are, on average, 50 μm wide and a maximum of 5 μm deep. Abrasive failure mode in TS coatings is thus not significantly different to those associated with rolling contact wear rolling contact wear (RCW) in conventional steel bearings; for example, Blau (Ref 13) has considered RCW as nucleation sites for initiating RCF. Similar failure was characterized as “peeling” during a study of fatigue failure modes of conventional steel ball bearings (Ref 16), whereas in Ref 5, this type of failure was characterized as surface distress. In spite of the various terminologies used to distinguish similar failure in steel bearings, the underpinning failure mechanism is associated with

asperity contact in the presence of microslip within the contact region. Gross sliding, though not necessary for this type of failure, is thought to promote micropitting.

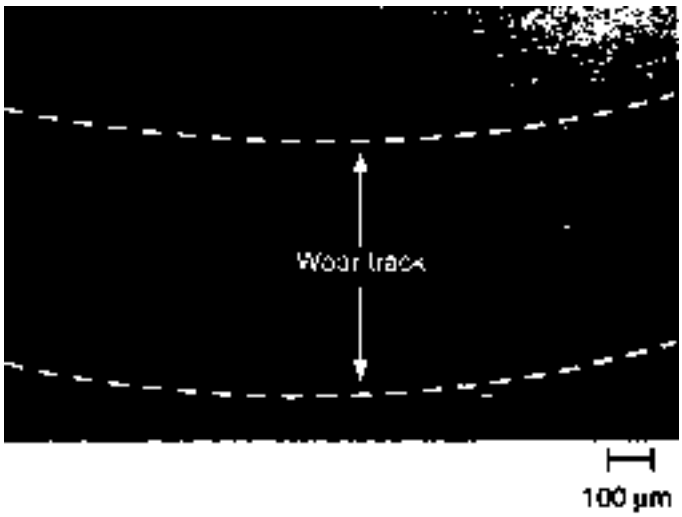


Fig. 11 Abrasive failure of the rolling wear track in thermally sprayed WC-Co coating

Mechanism of Coating Abrasion in Rolling/ Sliding Contacts. In the case of TS coatings, a similar mechanism of asperity contact in the presence of microslip and sliding is responsible for the micropitting and surface wear. During partial elastohydrodynamic lubrication (EHL) conditions, that is, when lubrication regime is in the mixed region ($1 < \lambda < 3$, where λ is nondimensional film thickness), surface asperities come into contact. Under these conditions, it was shown that the asperity contact and the EHL film share the load (Ref 86). According to Ref 87, these asperity contacts produce high stress concentrations very close to the surface but do not change the subsurface Hertzian stress pattern. These stress concentrations due to the interaction of asperities in the presence of microslip within the contact region result in shear stress beneath the asperity. Hard coatings such as WC-Co and Al_2O_3 respond to this stress concentration by microcracking and eventual micropitting, which also results in the attenuation of near-surface compressive residual stress within the coating material (Ref 88).

Apart from the mechanism of asperity interactions and subsequent microcracking leading to abrasive failure in TS coatings, the criterion of maximum tensile stress at the edge of the contact area for brittle materials is also thought to be responsible for the abrasive wear of TS cermet and ceramic coatings. The tensile stress is very sharply localized and decays very rapidly at small depths below the surface. For elliptical contacts, this tensile stress (T_{max}) for a given value of peak compressive stress (P_o) can be evaluated from the following relation, reproduced here for clarity (Ref 32, 33).

$$\frac{T_{max}}{P_o} = \frac{0.33 \times \beta}{n^3}\left[0.5 \ln\left(\frac{1+n}{1-n}\right) - n\right] \qquad \text{(Eq 8)}$$

where β is the ratio of minor to major axes of contact ellipse, and $n = (1 - \beta^2)^{0.5}$. For most of the RCF test conditions listed in Table 4, for example, in a rolling four-ball machine, T_{max} can be approximated as 324 MPa (47 ksi), for example, at a compressive contact stress of 2.7 GPa (0.39×10^6 psi). The fracture stress of WC-Co coatings (using the tensile test technique) has been shown to be in the range of 380–690 MPa (55–100 ksi) for high-velocity plasma sprayed and D-gun coatings (Ref 89). These values of fracture stress are similar to the tensile stresses associated with the Hertzian stress distribution at the edge of the contact region. This indicates that even under a fully developed EHL ($\lambda > 3$) regime, the microcracks in the coating material, either due to coating defects or asperity contact, can propagate due to tensile stress at the edge of the contact area. Hence, coating fracture toughness, which is related to coating microstructure via various TS processes and process conditions, is critical in controlling abrasive failure; for example, Ahmed and Hadfield (Ref 84) have related the poor fracture toughness of APS coatings in comparison to the liquid-fuel HVOF system, with the higher rate of abrasive wear in APS coatings.

Delamination Failure in TS Coatings. Suh initially proposed a delamination theory of sliding wear in 1973 (Ref 2). Flemming and Suh (Ref 90), Suh and Saka (Ref 91), and Suh (Ref 92) have since performed experimental and theoretical analysis supporting the delamination theory. The mechanism of delamination wear includes the

propagation of cracks parallel to the surface at a depth governed by material properties and the friction coefficient. Although rolling friction prevails in RCF tests and the delamination theory of wear is based on sliding friction, nevertheless the similarities of the failure mechanism in both cases are compelling. Typical observation of coating delamination at the coating-substrate interface can be seen from Fig. 12 for Al_2O_3 coating, whereas the delamination failure within the coating material can be appreciated from Fig. 13 for WC-Co coating. Coatings of Al_2O_3, however, have not shown delamination within the coating material but only at the coating-substrate interface (Table 4). Sheetlike debris that reached a few millimeters in major dimension is produced during this process.

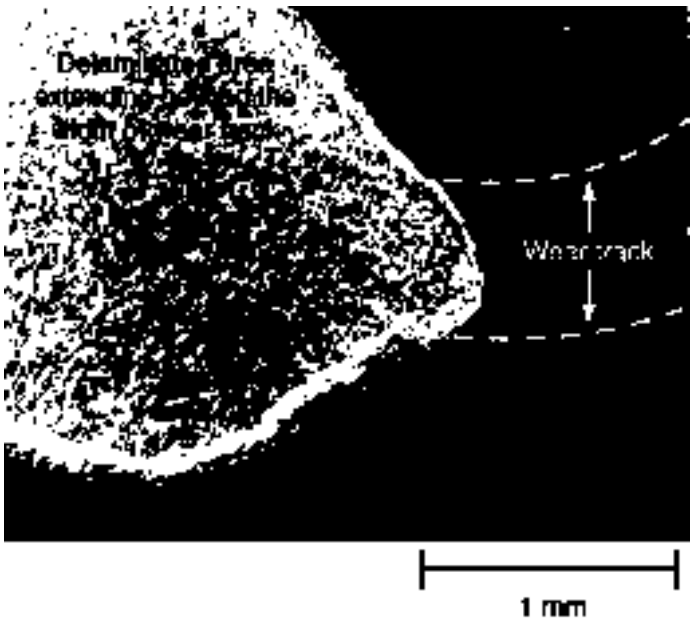


Fig. 12 Adhesive delamination in thermally sprayed Al_2O_3 coating

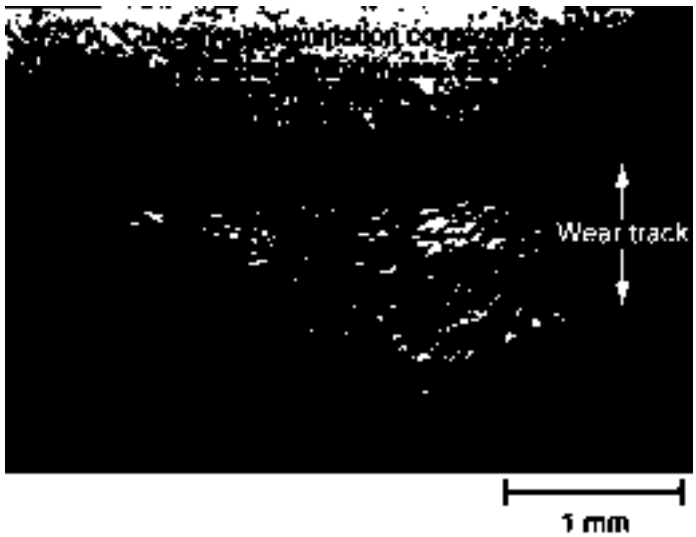


Fig. 13 Cohesive delamination in thermally sprayed WC-Co coating (backscattered electron image)

Mechanism of Coating Delamination in Cermet and Ceramic Coatings. The damage theory of materials begins with the premise that material contains a multitude of defects in the form of microvoids (Ref 93) that undergo extension due to loading and unloading. A similar approach is adapted to explain the mechanism of coating delamination. A TS coating microstructure contains varying levels of micropores, microcracks, and secondary-phase particles, which act as stress concentration during cyclic loading. The extent of these microdefects varies for different coating techniques and also depends on the starting powder, coating process, and its parameters. Such is the complexity of microstructure that stress concentrations are inevitable during cyclic loading. During rolling contact, these microdefects have a higher tendency for crack propagation at the depths of maximum shear stress (τ_{max}) and orthogonal shear stress reversal (τ_{orth}) (Fig. 8). The location of these stresses in the

vicinity of the coating-substrate interface can thus significantly influence crack propagation due to the mismatch of coating and substrate properties and a high level of quenching stress at this interface (Ref 94). Circumferential cracks similar to those shown in Fig. 14 appear beneath the surface at the depths of maximum shear and orthogonal shear stress for WC-Co coatings, whereas in the case of Al_2O_3 coatings, cracks generally appear at the coating-substrate interface. Hence, cohesive and adhesive delamination is observed in cermets whereas adhesive delamination is the preferred failure mode in ceramic coatings.

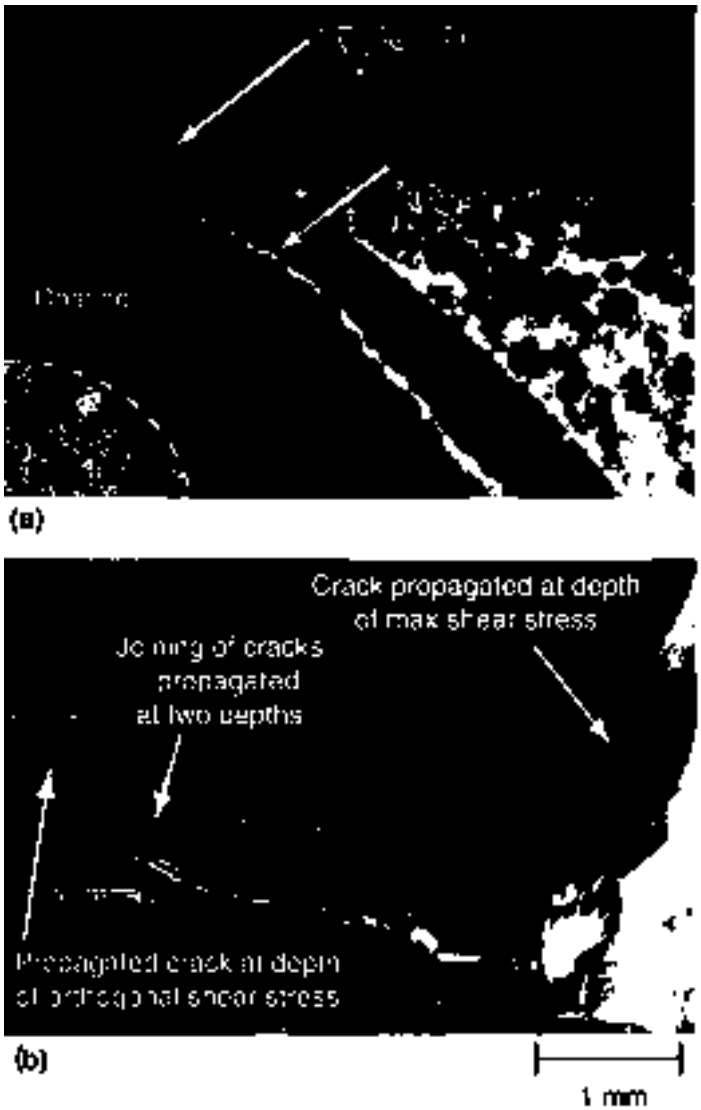


Fig. 14 Subsurface crack observations during delamination failure of thermal spray coatings. (a) Subsurface cracks in WC-Co-coated rolling cone at the depths of maximum and orthogonal shear stress. (b) Propagated subsurface cracks leading to coating delamination during RCF failure of WC-Co coating

Figure 15 gives a schematic of crack propagation behavior in ceramic and cermet coatings. The process can be summarized as follows:

- Crack initiation: For cermet (WC-Co) coatings, cracks initiate subsurface at different depths under the contact region but propagate at the depths of maximum shear stress and orthogonal shear stress (Fig. 14a). These cracks initiate due to stress concentrations in the presence of microstructural defects. For WC-Co coatings thinner than the depth of maximum shear stress, an additional factor of stress concentrations is the mismatch of coating-substrate properties at the interface. For WC-Co coatings thicker than the depth of maximum shear stress, these cracks initiate within the coating microstructure; for coatings thinner than the depth of maximum shear stress, these cracks initiate within the coating microstructure as well as at the coating-substrate interface. For ceramic (Al_2O_3) coatings, cracks generally initiate at the coating-substrate interface.

- Crack propagation: For WC-Co coatings thicker than the depth of maximum shear stress, initiated cracks propagate slowly in response to cyclic loading at their respective depths within the coating microstructure (Fig. 14b). For WC-Co or ceramic (Al_2O_3) coatings thinner than the depth of maximum shear stress, the cracks at the coating-substrate interface propagate much faster than the cracks at the depth of orthogonal shear stress (i.e., cracks within the coating microstructure) and result in poor RCF performance of thinner coatings.
- Crack extension: For WC-Co coatings thicker than the depth of maximum shear stress, the cracks at maximum shear stress extend to greater lengths and ultimately join the cracks at the depth of orthogonal shear stress (Fig. 14b). For coatings thinner than the depth of maximum shear stress or for ceramic (Al_2O_3) coatings, the cracks at interface had the greater tendency of reaching the surface independently, and, in some cases, they combine with the cracks at the depth of orthogonal shear stress and eventually reach the surface.
- Delamination: For WC-Co coatings thicker than the depth of maximum shear stress, the cracks from the depth of orthogonal shear stress reach the surface much more quickly than the cracks at the depth of orthogonal shear stress, thereby leading to coating delamination at the depth of orthogonal shear stress. Contrary to this, for thinner WC-Co or ceramic (Al_2O_3) coatings, interfacial delamination takes place.

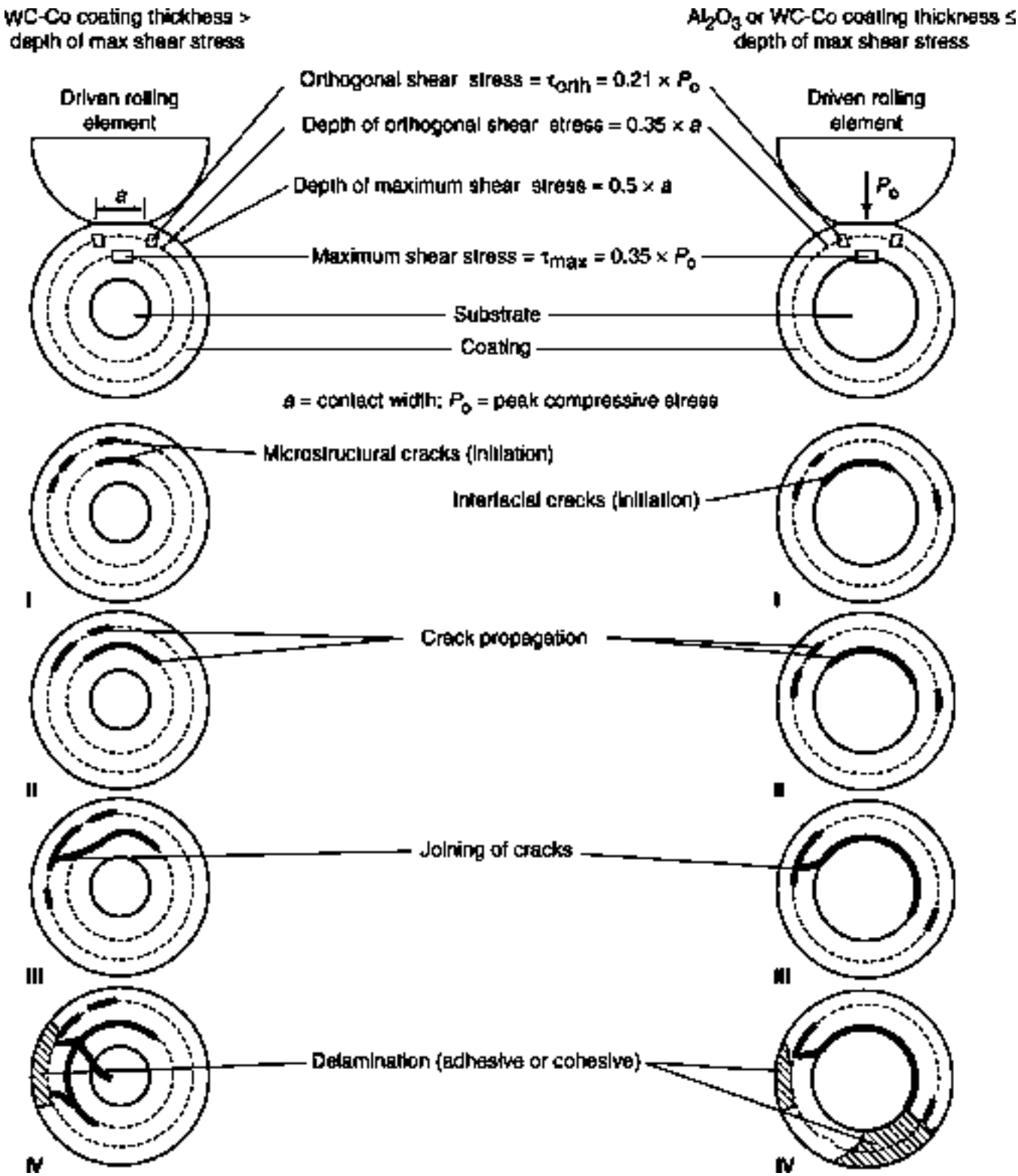


Fig. 15 Schematic of coating delamination process for cermet and ceramic coatings

Coating Failure Due to Bulk Deformation. Bulk deformation of substrate material is of primary importance for the cases of hard coatings (WC-Co and Al_2O_3) on a soft substrate. This is because the contact stress can be in the elastic range of the coating material and in the plastic range of the substrate. The primary effect of this is the plastic flow of substrate, leading to conformity of the contact region and a hump at the edge of the wear track. Figure 16 shows a typical example of such a failure in which the substrate could no longer support the coating, leading to bending and cracking of coating material in the initial stages of RCF failure. As cyclic loading continues, the coating cracks in the middle of the wear track due to its inability to plastically deform under the tensile stress caused by the plastic flow of substrate material. During this failure, the plastic flow of the substrate continues, and the substrate is pushed up at the edges of the wear track as it conforms to the geometry

of the counterbody, leading to subsequent cracking at the edge of the wear track. Further cyclic loading leads to crack propagation in tension, and the substrate finally emerges at the edge of wear track. Because the crack propagation is progressive due to cyclic loading, this failure mode is categorized as a RCF failure mode. Once the conformity of contact is such that the stresses in the substrate are no longer in the plastic region, the substrate migration terminates, and a steady state is reached. Figure 17 shows a schematic of this failure mode and indicates that substrate hardness should be carefully selected (based on contact loading) to combat this failure. The mechanism of bulk deformation is thus strongly dependent on the ability of the substrate material to support the coating in relation to the contact stress and is marginally affected by the changes in either coating material or process.

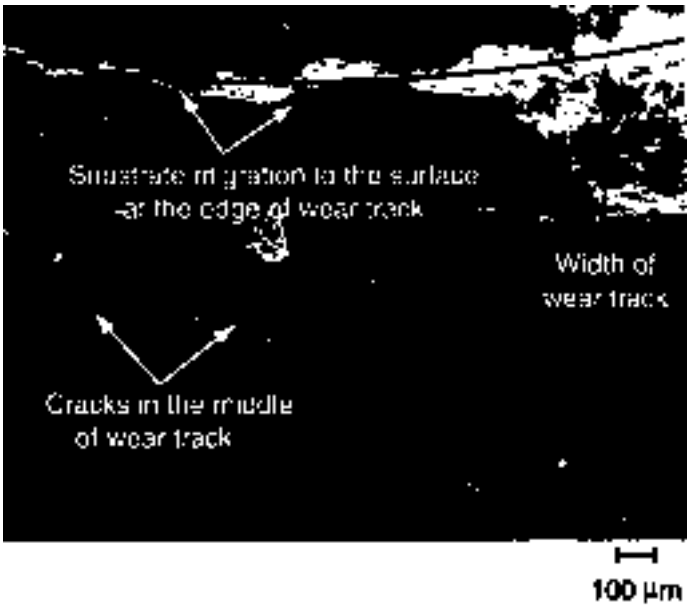


Fig. 16 Bulk deformation failure of WC-Co coating, indicating cracks in the middle of the wear track and substrate migration at the edge of the wear track

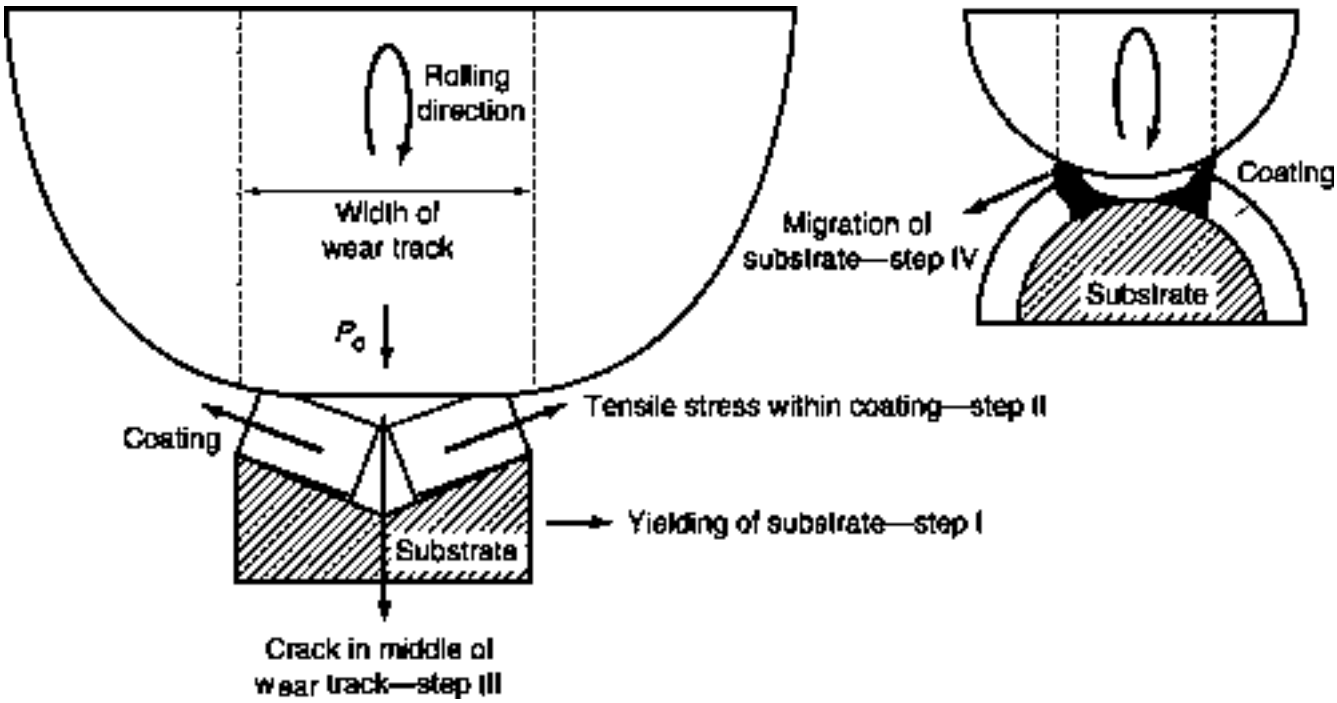


Fig. 17 Mechanism of bulk deformation

Coating Failure Due to Spalling. Spalling is the most commonly seen RCF failure in steel rolling-element bearings. Spalling fatigue is, however, the most rare mode of fatigue failure in TS coatings. Spall in TS coatings resembles in appearance the spalls in conventional bearings, as shown in Fig. 18. It differs from delamination

failure (discussed previously) in the sense that spall is contained within the wear track and is circular or elliptical in appearance, with its surface area (or width-to-depth ratio) much smaller than that of delaminated coating. A comparison of Fig. 1, 12 and 18 can distinguish the appearance between the two failures. As with conventional steel bearings, spalls in TS coatings can initiate from micropits, furrows, grinding marks, or dents on the surface of a wear track. Also, subsurface inclusions and defects are known to lead to spalling of rolling elements. Examination of the wear track of a spalled specimen in TS coatings indicates that substantial micropitting of the wear track occurs before fatigue spall is produced. This highlights the possibility that the fatigue spall in TS coatings initiates from the micropits, and subsequent crack propagation takes place due to cyclic loading. It is also possible that once initiated, spalling can be assisted by hydraulic pressure crack propagation (Ref 4). However, the exact mechanism of fatigue spall, that is, surface or subsurface initiation and propagation, in TS coatings is not completely understood at this stage.

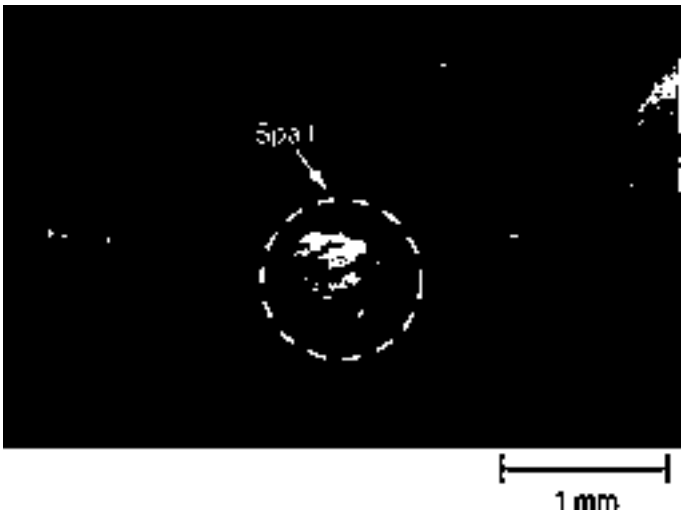


Fig. 18 Spalling of thermally sprayed WC-Co coating

Design Considerations for RCF Applications of TS Coatings. The design of surface overlay coatings for tribological applications involving Hertzian contact loading not only requires a thorough understanding of the tribological conditions, for example, contact stress, lubrication, and friction, but also the influence of coating processes, material, thickness, and the role of substrate material properties. Tribologists often have little choice about the former but can influence the coating performance by appropriate selection of the latter. The following thus provides some insight into the design considerations of coating-substrate material properties, coating thickness, and coating processes to combat RCF failure in TS coatings.

Substrate Material. The mechanism of bulk and interfacial delamination failure mode (discussed above) indicated that the four most important design considerations while selecting the substrate material are:

- Ability to support the coating
- Higher coefficient of thermal expansion than the coating material, to induce a certain degree of compressive residual stress within the coating material
- Ability to withstand preheat temperature prior to thermal spraying
- Ability to plastically deform during shot peening prior to the coating process, to promote mechanical interlock at the coating-substrate interface

The ratio ($\delta = P_o/\sigma$, where δ is normalized pressure) of the Hertzian contact stress (P_o) to the substrate yield strength (σ) is thus a useful indicator in determining the ability of the substrate to support the coating. Figure 19 shows this effect and indicates that normalized contact stress (P_o/σ) can be used to benchmark candidate substrate materials.

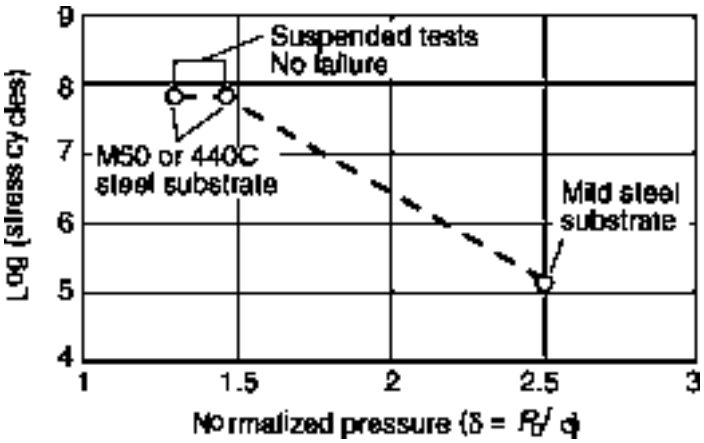


Fig. 19 Influence of normalized pressure on stress cycles ($\Delta \geq 1.5$, P_o = 2.7 GPa, or 0.39 × 10^6 psi)

Coating Thickness. Investigations of the delamination failure mode (both cohesive and adhesive) indicated that while selecting the coating thickness for a given coating-substrate system, it is critical to shift the maximum and orthogonal shear stresses away from the coating substrate interface. Using this technique can also simplify the coating design by not having to functionally grade the material to avoid the sharp stress gradient at the interface during the Hertzian loading. Hence, the ratio ($\Delta = \xi/\Psi$, where Δ is the normalized coating thickness, ξ is the average coating thickness, and Ψ is the depth of maximum shear stress) of the coating thickness to the depth of shear stress ($\Psi = 0.65b$, where b is the length of the minor axis of the contact ellipse) is found to be a good indicator of the coating substrate system to combat adhesive delamination. Figure 20 shows this effect and indicates that Δ value can have a significant influence on the fatigue performance of coatings.

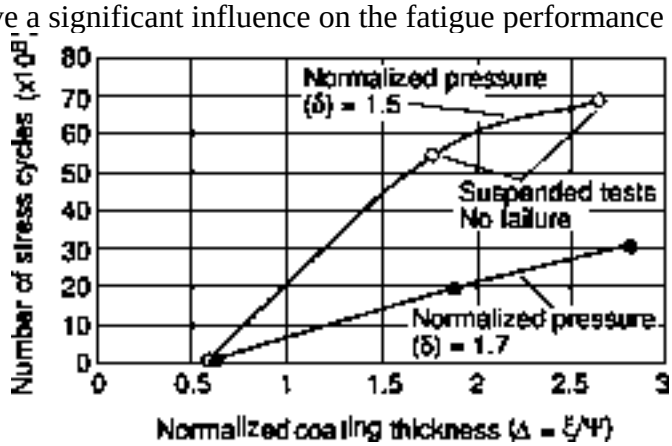


Fig. 20 Influence of normalized coating thickness on the performance of HVOF coating on 440C steel substrate (σ = 1840 MPa, or 267 ksi)

Coating Material. Investigation of failure modes in TS coatings indicated that the selection of a hard, wear-resistant coating material requires a consideration of its coefficient of thermal expansion and the ability of the coating process to produce a dense coating microstructure. Within the boundaries of materials considered in published research (WC-Co and Al_2O_3), test results indicate that the cermet coatings performance was superior to that of the ceramic coatings. This was mainly because of the following three reasons:

- The lower melting point of cermet material allowed the melting of lamellas during the HVOF processes, whereas ceramic was partially melted, resulting in a lower cohesive and adhesive strength.
- Although higher temperatures during the APS process melted the ceramic particles, the velocities achieved are generally much lower than the HVOF system, resulting in porous microstructure and thus lower strength.
- A higher lamella temperature during the APS process would have a high rate of cooling (quenching) on impact with the underlying surface, which can lead to internal cracking.

These factors contributed toward lower adhesive and cohesive strength of ceramic coatings and indicate the need for an interfacial layer. With the advancements of high-velocity plasma systems, however, it may be possible to improve RCF of TS ceramics in the future.

Coating Process(es). The most important consideration while selecting the coating process for RCF applications is its ability to produce a dense coating microstructure, resulting in a higher mechanical strength. There are a number of ways to analyze this, for example, porosity measurements, x-ray diffraction pattern, microhardness and bond strength. Experimental research has indicated that a useful measure for RCF applications is to compare the indentation fracture toughness (K_{Ic}) of the coating material. So far, HVOF-deposited WC-Co coatings with $K_{Ic} \geq 1.7$ MPa$\sqrt{m}$ have shown a fatigue life in excess of 70 million stress cycles for coating design parameters of $\Delta \geq 1.5$ and $\delta \leq 1.5$ (e.g., Ref 84).

References cited in this section

2. N.P. Suh, The Delamination Theory of Wear, *Wear,* Vol 25, 1973, p 111–124

4. S. Way, Pitting Due to Rolling Contact, *J. Appl. Mech.,* 1935, p A49–A58

5. T.E. Tallian, On Competing Failure Modes in Rolling Contact, *ASLE Trans.,* Vol 10, 1967, p 418–439

7. R. Ahmed and M. Hadfield, Failure Modes of Plasma Sprayed WC-Co Coated Rolling Elements, *Wear,* Vol 230, 1999, p 39–55

13. P.J. Blau, Rolling Contact Wear, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* 1992, p 257–262

16. W.E. Littmann and R.L. Widner, Propagation of Contact Fatigue from Surface and Subsurface Origins, *J. Basic Eng. (Trans. ASME),* 1966, p 624–636

32. Engineering Science Data Unit, “Contact Phenomena, Part II: Stress Fields and Failure Criteria in Concentrated Elastic Contacts under Combined Normal and Tangential Loading,” ESDU 84017, Inst. of Mech. Eng., 1984

33. G. Lundberg and A. Palmgren, Dynamic Capacity of Rolling Bearings,” *Acta Polytech. Scand.,* Vol 1 (No. 3), 1947

49. L. Pawlowski, The Science and Engineering of Thermal Spray Coatings, John Wiley & Sons, 1995

74. R. Nieminen, P. Vouristo, K. Niemi, T. Mantyla, and G. Barbezat, Rolling Contact Fatigue Failure Mechanisms in Plasma and HVOF Sprayed WC-Co Coatings, *Wear,* Vol 212, 1997, p 66–77

75. M. Yoshida, K. Tani, A. Nakahira, A. Nakajima, and T. Mawatari, Durability and Tribological Properties of Thermally Sprayed WC Cermet Coatings in Rolling/Sliding Contact, *Proc. ITSC,* May 1995 (Kobe), p 663–668

76. R. Ahmed and M. Hadfield, Rolling Contact Fatigue Performance of Detonation Gun Coated Elements, *Tribol. Int.,* Vol 30, 1997, p 129–137

77. R.Ahmed, M. Hadfield, Rolling Contact Fatigue Behaviour of Thermally Sprayed Rolling Elements, *Surf. Coat. Technol.,* Vol 82, 1996, p 176–186

78. B.Y. Sarma and M.M. Mayuram, Some Studies on Life Prediction of Thermal Sprayed Coatings under Rolling Contact Conditions, *J. Tribology (Trans ASME),* Vol 122, 2000, p 503–510

79. A. Nakajima, T. Mawatari, M. Yoshida, K. Tani, and A. Nakahira, Effects of Coating Thickness and Slip Ratio on Durability of Thermally Sprayed WC Cermet Coatings in Rolling/Sliding Contact, *Wear,* Vol 241, 2000, p 166–173

80. A. Makela, P. Vouristo, M. Lahdensuo, K. Niemi, and T. Mantyla, Rolling Contact Fatigue Testing of Thermally Sprayed Coatings, *Proc. Seventh Int. Thermal Spray Conf.,* 20–24 June 1994 (Boston, MA), ASM International, 1994, p 759–763

81. R. Ahmed and M. Hadfield, Wear of High-Velocity Oxy-Fuel (HVOF)-Coated Cones in Rolling Contact, *Wear,* Vol 203–204, 1997, p 98–106

82. R. Ahmed and M. Hadfield, Influence of Coating Thickness and Contact Stress on the Fatigue Failure of HVOF Coatings, *Int. Thermal Spray Conf.,* (Singapore), ASM International, 2001, p 1009–1015

83. S. Tobe, S. Kodama, and H. Misawa, Rolling Fatigue Behaviour of Plasma Sprayed Coatings on Aluminum Alloy, *Proc. Nat. Thermal Spray Conf.* 1990 (Tokyo, Japan), ASM International, 1990, p 171–178

84. R. Ahmed and M. Hadfield, Mechanisms of Fatigue Failure in Thermal Spray Coatings, *J. Therm. Spray Technol.,* in press 2002

85. R. Ahmed, Contact Fatigue Failure Modes of HVOF Coatings, *Wear,* submitted for publication 2001

86. K.L. Johnson, J.A. Greenwood, and S.Y. Poon, A Simple Theory of Asperity Contact in Elastohydrodynamic Lubrication, *Wear,* Vol 19, 1972, p 91–108

87. D. Berthe, L. Flamand, and M. Godet, Micropitting in Hertzian Contacts, *J. Lubr. Technol.,* (*Trans. ASME*), Vol 102, 1980, p 479–485

88. R. Ahmed and M. Hadfield, Experimental Measurement of Residual Stress Field within Thermally Sprayed Rolling Elements, *Wear,* Vol 209, 1997, p 84–95

89. R.C. Tucker, Jr., Structure Property Relationship in Deposits Produced by Plasma Spray and Detonation Gun Techniques, *J. Vac. Sci. Technol.,* Vol 11, 1994, p 725–734

90. J.W. Flemming and N.P. Suh, Mechanics of Crack Propagation in Delamination Wear, *Wear,* Vol 44, 1977, p 39–56

91. N.P. Suh and N. Saka, Fundamentals of Tribology, *Int. Conf. Fundamentals of Tribology,* Massachusetts Institute of Technology Press, 1980

92. N.P. Suh, *Tribophysics,* Prentice Hall, 1986

93. P.J. RabIer, Some Remarks on Damage Theory, *Int. J. Eng. Sci.,* Vol 27, 1989, p 29–54

94. S.C. Gill, “Residual Stress in Plasma Sprayed Deposits,” Ph.D. dissertation, Gonville and Caius College, Cambridge, U.K., 1993

R. Ahmed, Heriot-Watt University (United Kingdom)

References

1. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear,* Vol 44, 1977, p 1–16

2. N.P. Suh, The Delamination Theory of Wear, *Wear,* Vol 25, 1973, p 111–124

3. D. Tabor, The Mechanism of Rolling Friction, Part II: The Elastic Range, *Proc. R. Soc. A,* Vol 229, 1955, p 198–220

4. S. Way, Pitting Due to Rolling Contact, *J. Appl. Mech.,* 1935, p A49–A58

5. T.E. Tallian, On Competing Failure Modes in Rolling Contact, *ASLE Trans.,* Vol 10, 1967, p 418–439

6. M. Hadfield, T.A. Stolarski, and R.T. Cundill, Failure Modes of Ceramics in Rolling Contact, *Proc. R. Soc. A,* Vol 443, 1993, p 607–621

7. R. Ahmed and M. Hadfield, Failure Modes of Plasma Sprayed WC-Co Coated Rolling Elements, *Wear,* Vol 230, 1999, p 39–55

8. J.V. Poplawski, S.M. Peters, and E.V. Zaretsky, Effect of Roller Profile on Cylindrical Roller Bearing Life, Part I: Comparison of Bearing Life Theories, *Tribol. Trans.,* Vol 44 (No. 3), 2001, p 339–350

9. T.E. Tallian, Data Fitted Bearing Life Prediction Model for Variable Operating Conditions, *Tribol. Trans.,* Vol 42 (No. 1), 1999, p 241–249

10. Y. Li, T.R. Kurfess, and S.Y. Liang, Stochastic Prognostics for Rolling Element Bearings, *Mech. Syst. Signal Process.,* Vol 14 (No. 5), 2000, p 747–762

11. R.S. Hyde, Contact Fatigue of Hardened Steels, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 680–690

12. W.A. Glaeser and S.J. Shaffer, Contact Fatigue, *Fatigue and Fracture,* Vol 19, *ASM Handbook,* 1996, p 331–336

13. P.J. Blau, Rolling Contact Wear, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* 1992, p 257–262

14. W.D. Syniuta and C.J. Corrow, A Scanning Electron Microscope Fractographic Study of Rolling Contact Fatigue, *Wear,* Vol 15, 1970, p 187–199

15. J.A. Martin and A.D. Eberhardt, Identification of Potential Failure Nuclei in Rolling Contact Fatigue, *J. Basic Eng. (Trans. ASME),* 1967, p 932–942

16. W.E. Littmann and R.L. Widner, Propagation of Contact Fatigue from Surface and Subsurface Origins, *J. Basic Eng. (Trans. ASME),* 1966, p 624–636

17. W.E. Littmann, The Mechanism of Contact Fatigue, Conf. Proc. *An Interdisciplinary Approach to the Lubrication of Concentrated Contacts,* National Aeronautics and Space Administration, 1969, p 309–377

18. E.V. Zaretsky, "Life Factors for Rolling Bearings," Society of Tribology and Lubrication Engineers (STLE) publication, 1999

19. R.A. Guyer, *Rolling Bearings Handbook and Trouble Shooting Guide,* CRC Press, 1996

20. T. Harris, *Rolling Bearing Analysis,* 3rd ed., John Wiley & Sons, 1991

21. Anon, Friction, Wear, and Lubrication Terms and Definitions, 1st revision, Organization for Economic Co-Operation and Development, Paris, 1968

22. F.J. Wren and C.A. Moyer, Modes of Fatigue Failures in Rolling Element Bearings, *Proc. Inst. Mech. Eng.,* Vol 179 (Part 3D), 1964-1965, p 236–247

23. Y.P. Chiu, T.E. Tallian, and J.I. McCool, An Engineering Model of Spalling Fatigue in Rolling Contact—The Subsurface Model, *Wear,* Vol 17, 1971, p 433–446

24. T.E. Tallian and J.I. McCool, An Engineering Model of Spalling Fatigue in Rolling Contact—The Surface Model and Engineering: Discussion and Illustrative Examples, *Wear,* Vol 17, 1971, p 447–480

25. T.E. Tallian, *Failure Atlas for Hertz Contact Machine Elements,* 2nd ed. ASME Press, 1999

26. J.J.C. Hoo, "Rolling Contact Fatigue Testing of Bearing Steels," STP 771, ASTM, 1982

27. L.D. Wedeven, R.A. Pallini, and C.G. Hingley, "Selection and Use of Wear Tests for Ceramics," STP 1010, ASTM, 1988, p 58–73

28. R. Tourret and E.P. Wright, *Rolling Contact Fatigue: Performance Testing of Lubricants,* Int. Symp., Inst. of Petroleum, Heyden and Sons, London, 1977

29. W. Cheng, W.S. Cheng, and L.M. Keer, Experimental Investigation on Rolling Sliding Contact Fatigue Crack Initiation with Artificial Defects, *Tribol. Trans.,* Vol. 37, 1994

30. G. Xu, F. Sadeghi, and M.R. Hoeprich, Dent Initiated Spall Formation in EHL Rolling/Sliding Contact, *J. Tribology (Trans. ASME),* Vol 120, 1998, p 453–462

31. Engineering Science Data Unit, "Contact Phenomena, Part I: Stresses, Deflections and Contact Dimensions for Normally-Loaded Unlubricated Elastic Components," ESDU 78035, Inst. of Mech. Eng., 1984

32. Engineering Science Data Unit, "Contact Phenomena, Part II: Stress Fields and Failure Criteria in Concentrated Elastic Contacts under Combined Normal and Tangential Loading," ESDU 84017, Inst. of Mech. Eng., 1984

33. G. Lundberg and A. Palmgren, Dynamic Capacity of Rolling Bearings," *Acta Polytech. Scand.,* Vol 1 (No. 3), 1947

34. N.G. Popinceanu, E. Diaconescu, and S. Cretu, Critical Stresses in Rolling Contact Fatigue, *Wear,* Vol 71, 1981, p 265–282, 1981

35. W.J. Anderson and E.V. Zaretsky, Rolling Element Bearings—A Review of the State of the Art, NASA TMX-71441, National Aeronautics and Space Administration, Cleveland, OH, 1973

36. T.A. Harris, Friction and Wear of Rolling-Element Bearings, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* 1992, p 499–514

37. A. Kapoor and J.A. Williams, Shakedown Limits in Sliding Contacts on a Surface-Hardened Half-Space, *Wear,* Vol 172, 1994, p 197–206

38. A. Kapoor and K.L. Johnson, Effects on Changes in Contact Geometry on Shakedown of Surfaces in Rolling Sliding Contact, *Int. J. Mech. Sci.,* Vol 34 (No. 3), 1992, p 223–239

39. O. Orringer, W.R. Paxton, D.E. Gray, and P.K. Raj, Residual Stress and Its Consequences on Both Sides of the Wheel-Rail Interface, *Wear,* Vol 191, 1996, p 25–34

40. Q. Chen, G.T. Hahn, C.A. Rubin, and V. Bhargava, The Influence of Residual Stresses on Rolling Contact Mode II Driving Force in Bearing Raceways, *Wear,* Vol 126, 1988, p 17–30

41. H. Muro, N. Tsushima, and K. Nunome, Failure Analysis of Rolling Bearings by X-Ray Measurement of Residual Stress, *Wear,* Vol 25, 1973, p 345–356

42. R.J. Pomeroy, Measurement of Residual Stress in Rolling Contact, *Wear,* Vol 16, 1970, p 393–412

43. A. Igartua, J. Laucircia, A. Aranzabe, T. Leyendecker, O. Lemmer, G. Erkens, M. Weck, and G. Hanrath, Application of Low Temperature PVD Coatings in Roller Bearings: Tribological Tests and Experiences with Spindle Bearing Systems, *Surf. Coat. Technol.,* Vol 86–87, 1996, p 460–466

44. A. Erdemir, Rolling Contact Fatigue and Wear Resistance of Hard Coatings on Bearing Steel Substrates, *Surf. Coat. Technol.,* Vol 54–55, 1992, p 482–489

45. K.D. Bouzakis, N. Vidakis, A. Lontos, S. Mitsi, and K. David, Implementation of Low Temperature-Deposited Coating Fatigue Parameters in Commercial Roller Bearings Catalogues, *Surf. Coat. Technol.,* Vol 133–134, 2000, p 489–496

46. K.D. Bouzakis, N. Vidakis, and S. Mitsi, Fatigue Prediction of Thin Hard Coatings on the Steel Races of Hybrid Bearings Used in High Speed Machine Tool Spindles, *J. Tribology, (Trans. ASME),* Vol 120, 1998, p 835–842

47. I.A. Polonsky, T.P. Chang, L.M. Keer, and W.D. Sproul, An Analysis of the Effect of Hard Coatings on Near-Surface Rolling Contact Fatigue Initiation Induced by Surface Roughness, *Wear,* Vol 208, 1997, p 204–219

48. K. Holmberg and A. Matthews, *Coatings Tribology: Properties, Techniques, and Applications in Surface Engineering, Tribology Series,* Vol 28, Elsevier Science, 1994

49. L. Pawlowski, The Science and Engineering of Thermal Spray Coatings, John Wiley & Sons, 1995

50. N.J.M. Carvalho, A.J. Huis in't Veld, and J.T. De Hossen, Interfacial Fatigue Stress in PVD TiN Coated Tool Steels Under Rolling Contact Fatigue Conditions, *Surf. Coat. Technol.,* Vol 105, 1998, p 109–116

51. I.A. Polonsky, T.P. Chang, L.M. Keer, and W.D. Sproul, A Study of Rolling-Contact Fatigue of Bearing Steel Coated with Physical Vapor Deposition TiN films: Coating Response to Cyclic Contact Stress and Physical Mechanisms Underlying Coating Effect on the Fatigue Life, *Wear,* Vol 215, 1998, p 191–204

52. G. Meier zu Kocker, E. Santner, and B. Loffelbein, Functional Behavior of PVD Coatings under Sliding and Rolling Stress Conditions, *Tribotest J.,* Vol 4 (No. 3), 1998, p 241–260

53. R. Thom, L. Moore, W.D. Sproul, and T.P. Chang, Rolling Contact Fatigue Tests of Reactively Sputtered Nitride Coatings of Ti, Zr, Hf, Mo, Ti-Al, Ti-Zr, Ti-Al-V on 440C Stainless Steel Substrates, *Surf. Coat. Technol.,* Vol 62, 1993, p 423–427

54. T.P. Chang, P. Taung-Sea, H.S. Cheng, and S. Herbert, Influence of Coating Thickness on Lubricated Rolling Contact Fatigue Life, *Surf. Coat. Technol.,* Vol 44 (No. 1–3), 1990, p 699–708

55. T.P. Chang, H.S. Cheng, H.S. Chiou, and W.D. Sproul, Comparison of Fatigue Life Morphology between TiN Coated and Uncoated Rollers, *Tribol. Trans.,* Vol 34 (No. 3), 1991, p 408–416

56. P.J. Rudnik and W.D. Sproul, "Ceramic Coated Bearing Elements for Improved Durability and Reliability," final report, Project F33615-92-C-5295, Advanced Research Projects Agency (APRA), 1994

57. M.A. Liston, Rolling Contact Fatigue Properties of TiN/NbN Superlattice Coatings on M-50 Steel, *ASTM Spec. Tech. Publ.,* Vol 1327, 1998, p 499–510

58. R. Wei, P.J. Wilbur, M. Liston, and G. Lux, Rolling Contact Fatigue Wear Characteristics of Diamond-Like Carbon Coatings on Steels, *Wear,* Vol 162–164, 1993, p 558–568

59. R. Wei, P.J. Wilbur, and F.M. Kustas, A Rolling Contact Fatigue Study of Hard Carbon Coated M-50 Steel, *Trans ASME,* Vol 114, 1992, p 298–302

60. L. Rosado, V.K. Jain, and H.K. Trivedi, The Effect of Diamond-Like Carbon Coatings on the Rolling Fatigue and Wear of M-50 Steel, *Wear,* Vol 212, 1997, p 1–6

61. A.J. Perry, M. Jagner, W.D. Sproul, and P.J. Rudnik, Residual Stress and Strain Distribution Anomalies in TiN Films Deposited by Physical Vapour Deposition, *Surf. Coat. Technol.,* Vol 42 (No. 1), 1990, p 49–68

62. I. Petrov, L. Hultman, U. Helmersson, and J.E. Sundgren, Microstructure Modification of TiN by Ion Bombardment during Reactive Sputter Deposition, *Thin Solid Films,* Vol 169, 1989, p 299–314

63. K.D. Bouzakis, N. Vidakis, A. Igartua, T. Leyendecker, G. Erkens, M. Weck, G. Hanrath, and A. Schumacher, "Application of Low Temperature PVD Coatings in Roller Bearings-Fatigue Prediction of Thin Hard Coatings—Tribological Tests and Experiments with Spindle Bearing Systems," final report of the BRITE project Eurobearings, European Commission technical report, BRE-CT92-0333, 1997

64. H. Djabella and R.D. Arnell, Finite Element Analysis of Contact Stress in Elastic Coating on an Elastic Substrate, *Thin Solid Films,* Vol 213, 1992, p 205–209

65. P.K. Gupta and J.A. Walowit, Contact Stress between an Elastic Cylinder and a Layered Elastic Solid, *Trans. ASME,* 1974, p 250–257

66. A. Otsuka, H. Sugawara, and M. Shomura, A Test Method for Mode-II Fatigue Crack Growth Relating to a Model for Rolling Contact Fatigue, *Fatigue Fract. Eng. Mater. Struct.,* Vol 19, 1996, p 1265–1275

67. M. Hadfield and T.A. Stolarski, The Effect of the Test Machine on the Failure Mode in Lubricated Rolling Contact of Silicon Nitride, *Tribol. Int.,* Vol 28, 1995, p 377–382

68. J. He, C. Bai, K. Xu, and N. Hu, Improving the Anticorrosion and Mechanical Behaviour of PACVD TiN, *Surf. Coat. Technol.,* Vol 74–75, 1995, p 387–393

69. J. He, K. Xu, and N. Hu, Evaluation of Bonding Strength of Thin Hard Films by Spherical Rolling Test, *Surf. Coat. Technol.,* Vol 97, 1997, p 295–298

70. K. Xu, N. Hu, and J. He, Evaluation of the Bond Strength of Hard Coatings by the Contact Fatigue Test, *J. Adhes. Sci. Technol.,* Vol 12, 1998, p 1055–1069

71. J.F. Dill, M.N. Gardos, H.E. Hintermann, and H.J. Boving, Rolling Contact Fatigue Evaluation of Hardcoated Bearing Steels, *ASCE Publ.,* 1984, p 230–240

72. L. Chao, R. Lakshminarayanan, D.K. Shetty, and R. Cutler, Rolling Contact Fatigue and Wear of CVD SiC with Residual Surface Compression, *J. Am. Ceram. Soc.,* Vol 78 (No. 9) 1995, p 2307–2313

73. A. Savan, H.J. Boving, F. Fluehmann, and H.E. Hintermann, Increased Performance of Bearings Using TiC-Coated Balls, *J. Phys. (France) IV,* Colloque C7, Supplement au *J. Phys. (France) III,* Vol 3, Nov 1993, p 943–947

74. R. Nieminen, P. Vouristo, K. Niemi, T. Mantyla, and G. Barbezat, Rolling Contact Fatigue Failure Mechanisms in Plasma and HVOF Sprayed WC-Co Coatings, *Wear,* Vol 212, 1997, p 66–77

75. M. Yoshida, K. Tani, A. Nakahira, A. Nakajima, and T. Mawatari, Durability and Tribological Properties of Thermally Sprayed WC Cermet Coatings in Rolling/Sliding Contact, *Proc. ITSC,* May 1995 (Kobe), p 663–668

76. R. Ahmed and M. Hadfield, Rolling Contact Fatigue Performance of Detonation Gun Coated Elements, *Tribol. Int.,* Vol 30, 1997, p 129–137

77. R.Ahmed, M. Hadfield, Rolling Contact Fatigue Behaviour of Thermally Sprayed Rolling Elements, *Surf. Coat. Technol.,* Vol 82, 1996, p 176–186

78. B.Y. Sarma and M.M. Mayuram, Some Studies on Life Prediction of Thermal Sprayed Coatings under Rolling Contact Conditions, *J. Tribology (Trans ASME),* Vol 122, 2000, p 503–510

79. A. Nakajima, T. Mawatari, M. Yoshida, K. Tani, and A. Nakahira, Effects of Coating Thickness and Slip Ratio on Durability of Thermally Sprayed WC Cermet Coatings in Rolling/Sliding Contact, *Wear,* Vol 241, 2000, p 166–173

80. A. Makela, P. Vouristo, M. Lahdensuo, K. Niemi, and T. Mantyla, Rolling Contact Fatigue Testing of Thermally Sprayed Coatings, *Proc. Seventh Int. Thermal Spray Conf.,* 20–24 June 1994 (Boston, MA), ASM International, 1994, p 759–763

81. R. Ahmed and M. Hadfield, Wear of High-Velocity Oxy-Fuel (HVOF)-Coated Cones in Rolling Contact, *Wear,* Vol 203–204, 1997, p 98–106

82. R. Ahmed and M. Hadfield, Influence of Coating Thickness and Contact Stress on the Fatigue Failure of HVOF Coatings, *Int. Thermal Spray Conf.,* (Singapore), ASM International, 2001, p 1009–1015

83. S. Tobe, S. Kodama, and H. Misawa, Rolling Fatigue Behaviour of Plasma Sprayed Coatings on Aluminum Alloy, *Proc. Nat. Thermal Spray Conf.* 1990 (Tokyo, Japan), ASM International, 1990, p 171–178

84. R. Ahmed and M. Hadfield, Mechanisms of Fatigue Failure in Thermal Spray Coatings, *J. Therm. Spray Technol.,* in press 2002

85. R. Ahmed, Contact Fatigue Failure Modes of HVOF Coatings, *Wear,* submitted for publication 2001

86. K.L. Johnson, J.A. Greenwood, and S.Y. Poon, A Simple Theory of Asperity Contact in Elastohydrodynamic Lubrication, *Wear,* Vol 19, 1972, p 91–108

87. D. Berthe, L. Flamand, and M. Godet, Micropitting in Hertzian Contacts, *J. Lubr. Technol.,* (*Trans. ASME*), Vol 102, 1980, p 479–485

88. R. Ahmed and M. Hadfield, Experimental Measurement of Residual Stress Field within Thermally Sprayed Rolling Elements, *Wear,* Vol 209, 1997, p 84–95

89. R.C. Tucker, Jr., Structure Property Relationship in Deposits Produced by Plasma Spray and Detonation Gun Techniques, *J. Vac. Sci. Technol.,* Vol 11, 1994, p 725–734

90. J.W. Flemming and N.P. Suh, Mechanics of Crack Propagation in Delamination Wear, *Wear,* Vol 44, 1977, p 39–56

91. N.P. Suh and N. Saka, Fundamentals of Tribology, *Int. Conf. Fundamentals of Tribology,* Massachusetts Institute of Technology Press, 1980

92. N.P. Suh, *Tribophysics,* Prentice Hall, 1986

93. P.J. RabIer, Some Remarks on Damage Theory, *Int. J. Eng. Sci.,* Vol 27, 1989, p 29–54

94. S.C. Gill, “Residual Stress in Plasma Sprayed Deposits,” Ph.D. dissertation, Gonville and Caius College, Cambridge, U.K., 1993

Rolling-Contact Fatigue of Ceramics

Y. Wang and M. Hadfield, Bournemouth University (United Kingdom)

Introduction

TECHNICAL CERAMICS used as materials for rolling-contact bearing components show some practical advantages over traditional bearing steels. The properties of ceramics, specifically low density and high stiffness, are of most interest to gas turbine (Ref 1) and machine tool manufacturers (Ref 2). High hardness, low coefficient of thermal expansion, and high-temperature capability are properties also suited to rolling-element materials. Silicon nitride has been found to have a good combination of properties suitable for these applications. Research over the past three decades on structure, quality control, and manufacturing techniques has produced ceramic materials that are a suitable alternative to steel for rolling-contact bearing design. This is especially true for hybrid ball bearings; that is, precision angular-contact ball bearings using ceramic rolling elements are now offered as standard components within the ball bearing manufacturing industry.

Rolling-contact fatigue (RCF) is the surface damage process due to the repeated application of stresses when the surfaces of two bodies roll on each other. Rolling-contact fatigue is encountered most often in rolling-element bearings and gears. The failure process of RCF involves fatigue crack initiation and propagation, which is related to the characteristics of surface quality, stress distributions, and lubrication condition. Surface crack defects are considered as a main factor that causes RCF failure, at the present. It is difficult to detect these surface cracks during high-volume production processes, and hence, it is crucially important to understand their

influence and the fundamental mechanism of the failures they cause. Additional information about RCF is provided in the article “Rolling-Contact Fatigue” in this Volume.

References cited in this section

1. G. Hamburg, P. Cowley, and R. Valori, Operation of an All-Ceramic Mainshaft Roller Bearing in J-402 Gas-Turbine Engine, *Lubr. Eng.* (*J. ASME*) Vol 37, 1981, p 407–415

2. H. Aramaki, Y. Shoda, Y. Morishita, and T. Sawamoto, The Performance of Ball Bearings with Silicon-Nitride Ceramic Balls in High Speed Spindles for Machine Tools, *J. Tribology* (*Trans. ASME*), Vol 110, 1988, p 693–698

Rolling-Contact Fatigue of Ceramics

Y. Wang and M. Hadfield, Bournemouth University (United Kingdom)

Surface Crack Defects

Surface cracks caused by manufacturing pressing faults or blunt impact loads are the most common types found on ceramic ball surfaces. Figure 1 illustrates typical images of the crack defects found on the ceramic ball surface. Most of the defects are ring cracks. Lateral and radial surface cracks are less common and occur due to poor sintering mixture or a sharp indention load. Surface ring cracks are circular, extending for approximately $\frac{1}{4}$ to $\frac{1}{3}$ of the circumference of a circle. These circular cracks have been studied extensively. The fundamental basis for an explanation of the ring and cone fracture is provided by the Hertz theory of elastic contact. Traditionally, Hertzian cone cracks have been most widely studied in flat, silicate glass plate, using spheres of hard steel or tungsten carbide. Extension to ceramics is becoming more prevalent, and the theory of Hertzian fracture has been significantly developed during recent years (Ref 3). A typical experiment involves pressing an indent into the surface of the solid with a progressively increasing load and observing the pattern of fracture that develops under the contact. Under a blunt indent, a well-defined cone-shaped crack pops in when the normal load reaches a critical magnitude (Ref 4, 5). The surface tensile stress for visible ring cracking is much greater than that required for fracture in bulk tensile test specimens. This has been attributed to the small volume acted on by the contact-induced tensile stress.

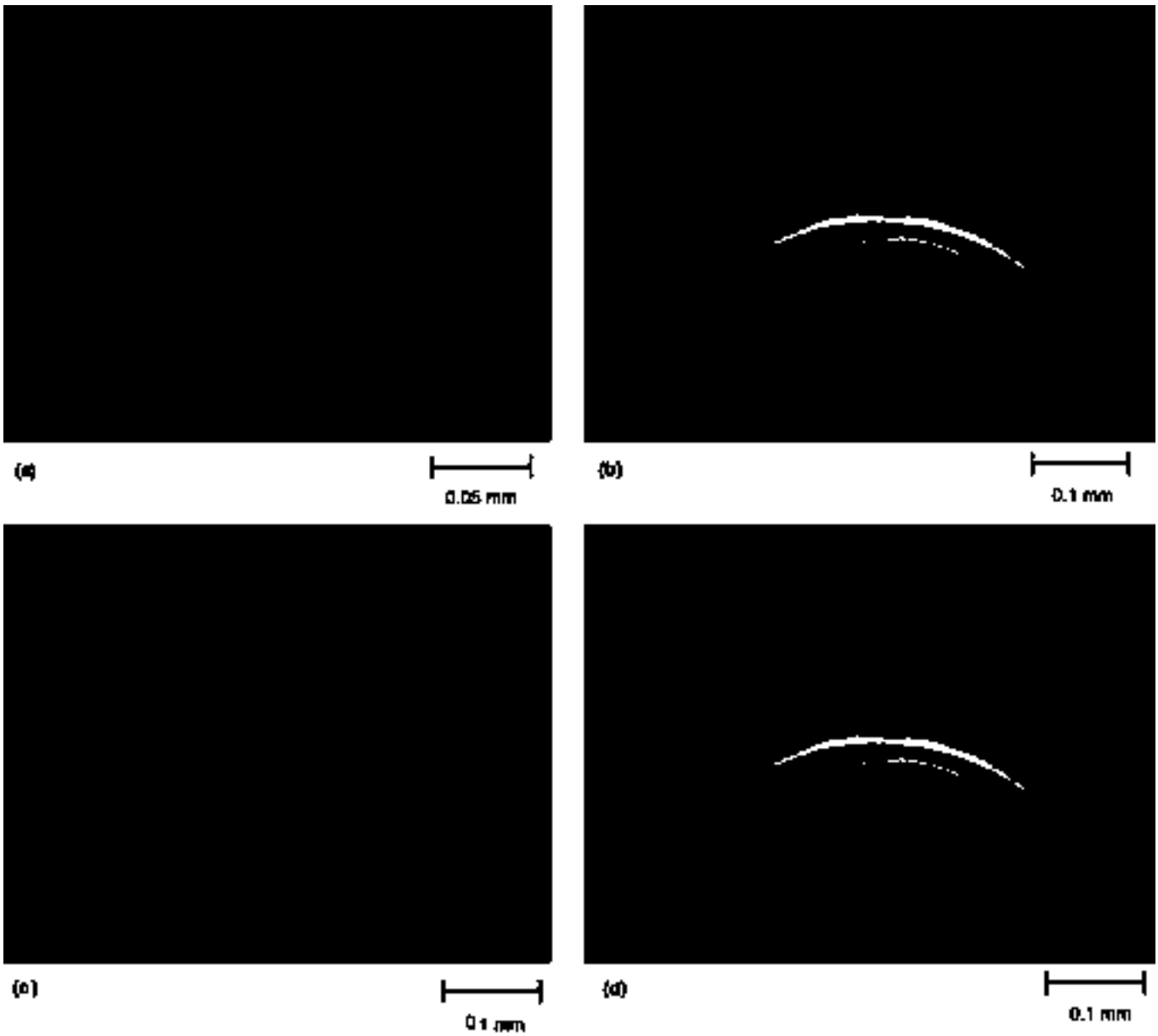


Fig. 1 Typical crack defects found on ceramic ball surfaces under ultraviolet light. (a) Star defect. (b) Pressing defect. (c) Single ring crack. (d) Double ring cracks

Ring crack initiation load of hot isostatically pressed (HIP) silicon nitride bearing balls was experimentally studied by some researchers (Ref 6, 7, 8). Initiation of a ring crack was detected using acoustic emission. The ring crack initiation load showed considerable scatter and ranged from 3 to 6 kN (corresponding to the maximum Hertz contact pressure of 14 and 18 GPa, or 2.0 and 2.6 × 10^6 psi). The ring crack initiation load is much smaller than the crushing load and approximately follows two-parameter Weibull distribution, with a shape parameter ranging from 9 to 11. Impact cracking of silicon nitride balls during ball finishing or by subsequent handling is a cause for concern, because a surface crack is the type of defect most likely to cause failure in a hybrid rolling-element bearing. Ring cracks formed prior to or during the final lapping or polishing stage are particularly difficult to detect. Such cracks are not visible with optical techniques, even at high magnifications, except after processing with high-sensitivity fluorescent penetrants. Impact cracking of silicon nitride balls was investigated, leading to a parameter for assessing impact resistance (Ref 9). Some material types were found to be more resistant to impact cracks than others. The types of silicon nitride densified by sintering and followed by HIPing and by gas pressure sintering have the best impact resistance. For ball sizes of 6.35 mm (0.25 in.), the maximum contact pressure of 20 to 25 GPa (2.9 to 3.6 × 10^6 psi) was required to form surface ring cracks; for ball sizes of 12.7 mm (0.50 in.), 10 to 15 GPa (1.5 to 2.2 × 10^6 psi) was required; for ball sizes of more than 20 mm (0.79 in.), 10 to 12 GPa (1.5 to 1.7 × 10^6 psi) was required.

A number of theoretical models have been developed that can predict both the fracture load and the subsequent length of the cracks (Ref 10, 11, 12). The development of the cone crack from the ring crack in the strongly inhomogeneous Hertzian stress field was investigated (Ref 10). Mode I stress-intensity factor at the cone crack front was calculated by modeling it as a two-dimensional plane-strain crack in an infinite medium loaded by the stresses acting on the location of the cone crack in the corresponding uncracked body. It was assumed that the inclined sides of the cone crack followed trajectories of the minimum principal stress, so that it was perpendicular to the maximum principal stress. This model predicts that the crack propagation arrests at a depth that is related to the applied load. The partial cone cracks can be generated if the indent slides over the surface under the action of a combined normal and tangential load. The cracks are initiated at the trailing edge of the contact and propagate almost perpendicular to the surface (Ref 13, 14).

A similar two-dimensional approximation was used to consider the effect of ring crack initiation some distance outside of the circle of contact (Ref 15). The model was combined with experiments to measure the fracture toughness of carbides based on the initiation of the ring crack. A two-dimensional approximation was used to analyze cracks generated by both spheres and flat punches (Ref 12). Circular crack initiation and propagation was fully analyzed using a numerical calculation method (Ref 16). A finite-element method was used. It was found that the mode II stress-intensity factor was about one-tenth of the mode I stress-intensity factor. A review paper discussed ring crack initiation and propagation using analytical techniques (Ref 17). The first two stages of surface defect growth show that surface flaws develop into complete rings. In the final phase, the crack develops into a cone.

References cited in this section

3. B.R. Lawn, Indentation of Ceramics with Spheres: A Century after Hertz, *J. Am. Ceram. Soc.*, Vol 81, 1998, p 1977–1994

4. F.C. Roesler, Brittle Fracture near Equilibrium, *Proc. Phys. Soc. (London)*, Vol B69, 1956, p 981–1012

5. M. Chaudhri and M.A. Phillips, Quasi-Static Indentation Cracking of Thermally Tempered Soda-Lime Glass with Spherical and Vickers Indenters, *Philos. Mag. A*, Vol 62, 1990, p 1–27

6. M. Ichikawa, T. Takamatsu, T. Matsuo, N. Okabe, and Y. Abe, Intra-Ball and Inter-Ball Variability of Ring Crack Initiation Load of Silicon-Nitride Bearing Balls, *JSME Int. J. A: Mech. Mater. Eng.*, Vol 38, 1995, p 231–235

7. M. Ichikawa, T. Takamatsu, N. Shindou, N. Okabe, and Y. Abe, Ring Crack Initiation Load of Silicon Nitride Bearing Balls, *JSME Int. J. A: Mech. Mater. Eng.*, Vol 38, 1995, p 226–230

8. K. Ohgushi and M. Ichikawa, Fracture Mechanics Study of Ring Crack Initiation in Ceramics by Sphere Indentation, *JSME Int. J. A: Mech. Mater. Eng.*, Vol 39, 1996, p 489–495

9. R.T. Cundill, “Impact Resistance of Silicon Nitride Balls,” Proceedings of the Sixth International Symposium on *Ceramic Materials and Components for Engines* (Arita, Japan), Japan Fine Ceramics Association, 1997 p 556–561

10. F.C. Frank and B.R. Lawn, On the Theory of Hertzian Fracture, *Proc. R. Soc.* (*London*) *A*, Vol 299, 1967, p 291–306

11. T.R. Wilshaw, The Hertzian Fracture Test, *J. Phys. D, Appl. Phys.*, Vol 4, 1971, p 1567–1581

12. R. Mouginot and D. Maugis, Fracture Indentation beneath Flat and Spherical Punches, *J. Mater. Sci.*, Vol 20, 1985, p 4354–4376

13. B.R. Lawn, Partical Cone Crack Formation in a Brittle Material Loaded with a Sliding Spherical Indenter, *Proc. R. Soc.* (*London*) *A*, Vol 299, 1967, p 307–316

14. B.J. Bethune, The Surface Cracking of Glassy Polymers under a Sliding Spherical Indenter, *J. Mater. Sci.*, Vol 11, 1971, p 199–205

15. R. Warren, Measurement of the Fracture Properties of Brittle Solids by Hertzian Indenters, *Acta Metall.*, Vol 46, 1978, p 1759–1769

16. Y.Z. Li and D.A. Hills, The Hertzian Cone Crack, *J. Appl. Mech.* (*Trans. ASME*), Vol 58, 1991, p 120–127

17. P.D. Warren, D.A. Hills, and D.N. Dai, Mechanics of Hertzian Cracking, *Tribol. Int.*, Vol 28, 1995, p 357–362

Rolling-Contact Fatigue of Ceramics

Y. Wang and M. Hadfield, Bournemouth University (United Kingdom)

Fatigue Crack Propagation in Rolling Contact

Fracture Mechanics Approach. Fatigue crack propagation is always involved in RCF failure. Many efforts have been completed to relate crack propagation behavior with RCF life performance since the publication of a delamination wear theory arising from rolling and sliding contact (Ref 18). The bulk of that research was focused on the experimental aspects of wear. Only a few attempts have been made to develop analytical models for predicting the wear process. The importance of determining analytical models for wear mechanisms has been discussed (Ref 19). Wear modes must be described mechanistically to better understand how they occur and to develop formulae for wear-material removal. Fracture mechanics has been applied to the analysis of wear modes in a number of studies. Numerous researchers (Ref 20, 21, 22) have presented a fracture mechanics approach to delamination wear of metals. They analyzed a subsurface crack parallel to the surface of a half-plane subjected to a Hertzian surface load. The stress-intensity factors at the crack tips were calculated from an approximate stress field induced by the contact load in a crack-free half-plane. Several researchers analyzed surface cracks at an angle to a surface acted on by a Hertzian contact load (Ref 23, 24). In one study (Ref 23), lubrication pressure was considered as a crack opening force; in another (Ref 24), sufficient lubrication was taken to exist in the crack, such that the crack-face friction was zero. The analyses were further extended to include the effects of subsurface inclusions and asperity perturbations on the pitting crack (Ref 25). Finally, a thorough analysis of the surface pitting crack summarized the effects of crack-face friction and lubrication pressures on crack propagation and included loading history dependence when regimes of slip and stick were considered for the crack (Ref 26). In all of these analyses, mode II stresses were the main cause of crack propagation. This caused crack propagation to be highly dependent on the coefficient of friction between the crack faces.

Several attempts have been made to explain the wear mechanisms in brittle materials (Ref 19). In the case of brittle materials, cracks may occur on the surface, due to tension or subsurface cracks parallel to the contact path due to shear. Friction in rolling and sliding contact can arise from asperity interlock, adhesion, or abrasion by debris trapped between the contacting surface and viscous drag of the lubricant. A qualitative model (Ref 27) was presented to describe the microchipping wear mode in ceramic rollers. It is stated that microchipping cracks can form as a result of the stress distribution due to a Hertzian contact. Because the direction of principal stresses varies rapidly at an indentation contact, cracks propagating at the surface link or branch to turn parallel to the surface, thus providing a mode for wear-material removal. A strong dependence was seen relating the wear particle size with the coefficient of friction. Researchers modeled near-surface cracks in zirconia ceramics that are subjected to contact loading (Ref 28). It is concluded that frictional contact may have a significant influence, especially on the sliding mode (mode II) deformation of the closed crack. In the analysis for a phase-transforming material, the friction contact between the crack faces may be of even greater importance, because the transformation zones at the crack tip(s) showed a marked closing effect on the tips, thus exerting significant

compressive stresses there. These compressive stresses would, in turn, induce frictional stresses opposing the sliding deformation of the crack tips.

Crack Propagation Testing. Fatigue crack propagation in ceramics has been studied in two modes: long-crack and short-crack studies. The long-crack studies (typically with crack lengths of about 3 mm, or 0.12 in.) were done with through cracks in precracked specimens, such as single-edge-notched specimens in three- and four-point bending or tapered double-cantilever beam specimens, or compact tension specimens. Crack lengths were monitored optically or by measuring the resistance of a conducting coating. Linear elastic stress-intensity solutions for the various specimen geometries are given in standard handbooks (Ref 29). Short-crack experiments (typically with cracks less than 250 μm, or 10 mil, in length) were done by bending beams using surface cracks introduced from a notch or by indentation. Stress-intensity factors for such surface cracks can be obtained from linear elastic solutions for three-dimensional semielliptical surface cracks in bending (and/or tension) (Ref 30). The long-crack propagation studies gave results that could be described by the Paris law, but the exponent m varied between 12 and 40, which is much larger than the values of 2 to 4 typical of metals. The short-crack studies gave much greater crack growth rates than the long-crack studies at equivalent applied stress-intensity values. Also, the crack growth takes place in the short-crack studies at stress intensities below the threshold value required in long-crack studies (Ref 31).

Studies of cyclic fatigue in ceramics using long-crack techniques are clearly important in understanding basic fatigue phenomena in ceramics, and data obtained in such experiments may apply to design considerations in some circumstances. However, the same problem arises as with other uses of long-crack data; that is, ceramics having only the naturally occurring short cracks are generally used for load-bearing applications. Thus, fatigue data for short cracks are needed. Results on small fatigue cracks in ceramics, however, are very limited. In contrast to long-crack results, the small cracks grow at progressively decreasing growth rates with increase in size, until finally linking together as the density of cracks across the specimen surface increases; the specimen then fails. Small-crack propagation rates display a negative dependency on stress intensity and occur at applied stress-intensity levels well below the fatigue crack propagation threshold.

Mechanisms of cyclic fatigue in ceramics have not yet been definitely established. The possible mechanisms have been classified into two categories: intrinsic and extrinsic (Ref 31, 32). The intrinsic mechanism would involve the creation of a fatigue-damaged microstructure ahead of the crack tip that would produce a crack-advance mechanism operating uniquely in cyclic fatigue. Possible mechanisms are alternating crack tip blunting and resharpening or, for whisker-reinforced composites, whisker breakage. Extrinsic mechanisms may not involve a change in the crack propagation mechanism under cyclic fatigue conditions. Instead, the unloading may change the value of the crack tip shielding. Thus, cyclic fatigue might reduce the crack tip shielding effect of transformation toughening or crack tip bridging.

References cited in this section

18. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear,* Vol 44, 1977, p 1–16

19. J.F. Braza, H.S. Cheng, M.E. Fine, A.K. Gangopadhyay, L.M. Keer, and R.E. Worden, Mechanical Failure Mechanisms in Ceramic Sliding and Rolling Contacts, *Tribol. Trans.,* Vol 32, 1989

20. J.R. Fleming and N.P. Suh, Mechanics of Crack Propagation in Delamination Wear, *Wear,* Vol 44, 1977, p 39–56

21. D.A. Hills and D.W. Ashelby, A Fracture Mechanics Approach to Rolling Contact Fatigue, *Tribol. Int.,* 1979, p 115–119

22. A.R. Rosenfield, A Fracture Mechanics Approach to Wear, *Wear,* Vol 61, 1980, p 125–132

23. L.M. Keer and M.D. Bryant, A Pitting Model for Rolling-Contact Fatigue, *J. Lubr. Technol.* (*Trans. ASME*), Vol 105, 1983, p 198–205

24. M.D. Bryant, G.R. Miller, and L.M. Keer, Line Contact between a Rigid Indenter and a Damaged Elastic Body, *Q. J. Mech. Appl. Math.,* Vol 37, 1984, p 468–478

25. G.R. Miller, L.M. Keer, and H.S. Cheng, On the Mechanics of Fatigue Crack-Growth due to Contact Loading, *Proc. R. Soc.* (*London*) *A: Math. Phys. Sci.,* Vol 397, 1985, p 197–209

26. A.F. Bower, The Influence of Crack Face Friction and Trapped Fluid on Surface Initiated Rolling-Contact Fatigue Cracks, *J. Tribology* (*Trans. ASME*), Vol 110, 1988, p 704–711

27. L.M. Keer and R.E. Worden, A Qualitative Model to Describe the Microchipping Wear Mode in Ceramic Rollers, *Tribol. Trans.,* Vol 33, 1990, p 411–417

28. N.B. Thomsen and B.L. Karihaloo, Subsurface and Surface Cracks under Contact Loading in Transformation-Toughened Ceramics, *J. Mech. Phys. Solids,* Vol 44, 1996, p 207–231

29. H. Tada, P.C. Paris, and G.R. Irwin, *Stress Analysis of Cracks Handbook,* Paris Publications Inc., St. Louis, MO, 1985

30. J.C. Newman and I.S. Raju, *Computational Methods in the Mechanics of Fracture,* S.N. Atluri, Ed., Elsevier Science Publishers B.V., 1986, p 312–334

31. R.O. Ritchie and R.H. Dauskardt, Cyclic Fatigue of Ceramics—A Fracture-Mechanics Approach to Subcritical Crack-Growth and Life Prediction, *J. Ceram. Soc. Jpn.,* Vol 99, 1991, p 1047–1062

32. R.O. Ritchie, C.J. Gilbert, and J.M. McNaanry, Mechanics and Mechanisms of Fatigue Damage and Crack Growth in Advanced Materials, *Int. J. Solids Struct.* Vol 37, 2000, p 311–329

Rolling-Contact Fatigue of Ceramics

Y. Wang and M. Hadfield, Bournemouth University (United Kingdom)

Rolling Contact Fatigue Test Machine

Many test-rig types are used to study RCF performance of ceramics. Typically, in such tests, a detector coupled with a shutdown device is used to monitor the vibration of the assembly. When a preset vibration level is exceeded, indicating the formation of a spall or excessive uneven wear, the test is automatically stopped, and the lifetime of the test is recorded. Alternatively, the tests are suspended at different durations to study the wear rate. The characteristics of various types of RCF test methods are summarized as follows.

Modified Four-Ball Machine. A model contact consisting of three lower balls driven by a fourth contacting upper ball simulates conditions within a standard deep-groove ball bearing. The upper ball models the bearing race, while the cup simulates the bearing outer race, and the three planetary balls represent the balls within a ball bearing. The loading geometry is shown in Fig. 2. The machine has a proven history as being useful to investigate the RCF resistance of materials under various tribological conditions. Previously, this machine has been mainly used by various research staff as an accelerated method to measure the RCF of bearing steels. The influence of lubrication type on steel ball RCF was also evaluated using the machine (Ref 33, 34, 35). The Institute of Petroleum gathered numerous papers (Ref 36) that describe various test results, ball dynamics, and kinematics using the modified four-ball machine.

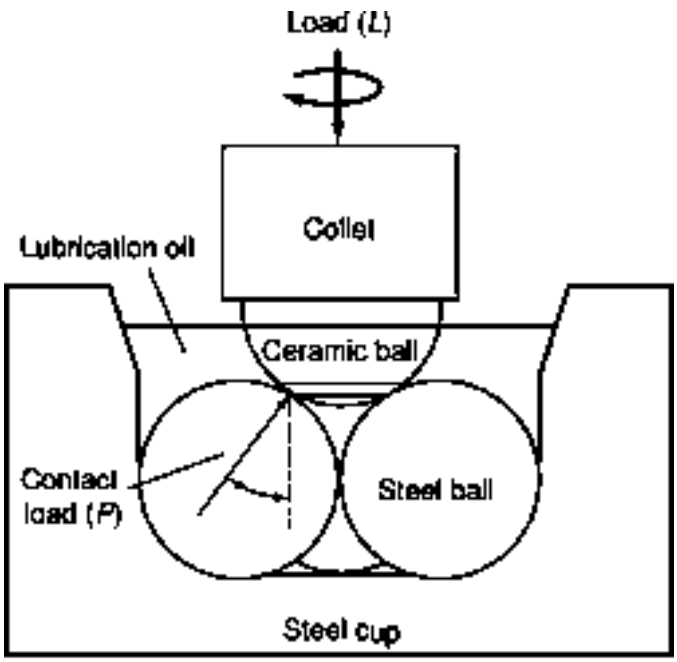


Fig. 2 Loading configuration of four-ball machine

This machine has been used to study the RCF performance of hot pressing silicon nitride bearing materials (Ref 37, 38). Recently, the modified four-ball machine has been used to simulate the rolling contact in hybrid ball bearings (Ref 39, 40, 41).

Five-Ball Machine. Figure 3 shows the contact geometry and the rotating mechanisms of the test rig. The test assembly consists of a driven test ball on top of four lower balls, positioned by a separator in the form of a pyramid. The four lower balls rotate in a race driven by the upper ball, thus simulating the rolling and sliding produced in angular-contact ball bearings. This five-ball machine has been used to investigate the fatigue life of high-speed ball bearings with hot pressed silicon nitride balls (Ref 42, 43).

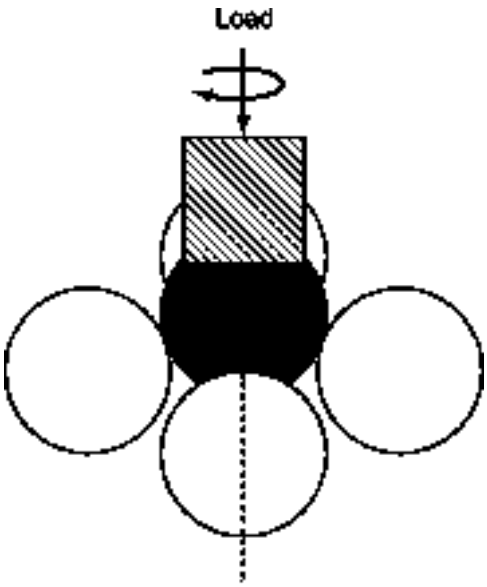


Fig. 3 Loading configuration of five-ball machine

Ball-on-Plate Machine. Figure 4 shows a cross-sectional view of the test rig. The test unit has a unidirectional thrust bearing configuration, which consists of three balls or rollers equally spaced at 120° by a retainer and loaded between a stationary flat washer and a rotating grooved washer. The rotating washer produces ball motion and serves to transmit load to the balls (or rollers) and the flat washer. The ball-on-plate machine was used to investigate the RCF performance of silicon nitride (Ref 44, 45).

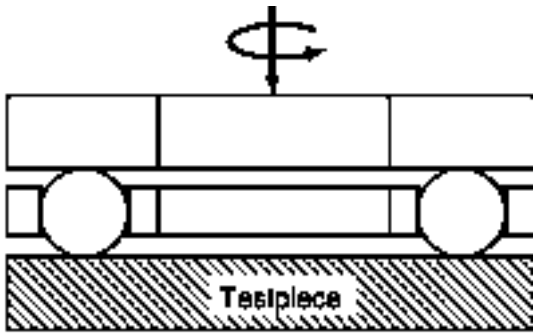


Fig. 4 Loading configuration of ball-on-plate machine

Ball-on-Rod Machine. Figure 5 shows the loading geometry of the ball-on-rod-type testing machine. Three steel balls, 12.7 mm (0.5 in.) in diameter, orbit a rotating 9.53 mm (0.375 in.) diameter cylinder test specimen. The thrust load on the bearing cups is applied mechanically by forcing the upper cup toward the lower cup with three precalibrated coil springs in the assembly. The rod specimen is rotated by a direct-drive electric motor mounted in line with the specimen below the table. The three balls are, in turn, driven by the rod and rotate around the rod. Lubricant is supplied by drip feeding onto the top of the rod. The machine has been employed to study rolling-contact wear performance (Ref 46, 47, 48, 49).

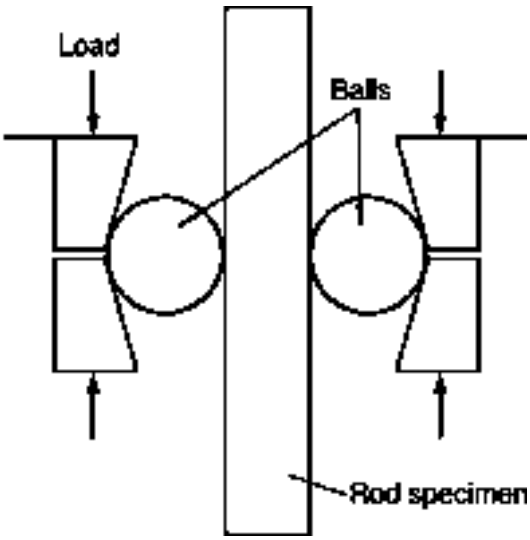


Fig. 5 Loading configuration of ball-on-rod machine

Disc-on-Rod Machine. A disc-on-rod machine is shown in Fig. 6. Two steel disks, 177.7 mm (7.0 in.) in diameter and 12.7 mm (0.5 in.) thick, are pressed against the test specimen, which is held in a spindle and driven at 10,000 rpm. The test specimen is a 76 mm (3 in.) long, straight cylinder with a 9.52 mm (0.375 in.) diameter. With the known geometry, the contact stresses as a function of load can be calculated. The machine provides a means of rapid testing in nearly pure rolling contact. The specimen receives 1,200,000 stress cycles/h when driven at 10,000 rpm. Some researchers performed earlier work using this machine (Ref 50, 51).

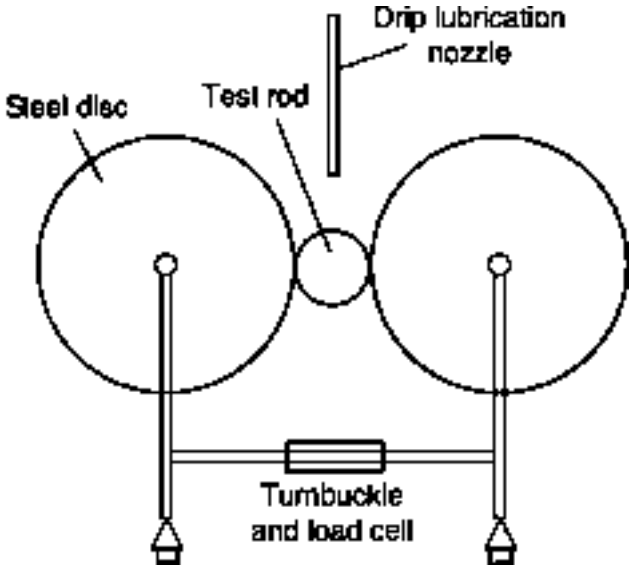


Fig. 6 Loading configuration of disc-on-rod machine

Contacting Ring Machine. Figure 7 shows a diagram of the test rig, which uses two disk-shaped specimens rotating against each other on their outer surfaces. Geometry of the test specimen can be disks or rings that can be mounted on shafts. The contacting outer surface of the specimens can both be flat, or one specimen may have a flat surface while the other has a toroidal surface. The rotational speeds of the two disks (ring) can be controlled individually with separate motors or by using the gear mechanism; thus, rolling contact with a differing amount of sliding (slippage) at the interface can be generated. Typically, the load is applied using a compressed coil spring or dead-weight loading systems. Tests can be conducted in either dry or lubricated conditions. Profiles of the wear scars on the contact surfaces can be used to estimate the wear volume. The contact ring machine has been used to investigate wear performance of ceramics (Ref 52, 53, 54).

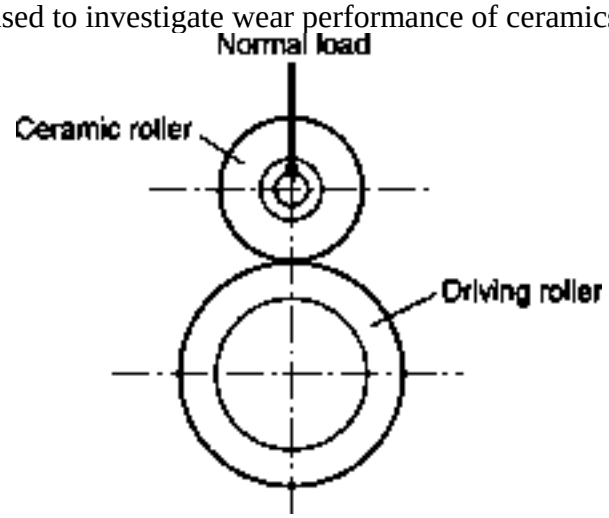


Fig. 7 Loading configuration of contacting ring machine

References cited in this section

33. F.T. Barwell and D. Scott, Effect of Lubricant on Pitting Failure of Ball Bearings, *Eng.*, 6 July 1956, p 9–12

34. G.S. Krivoshein, On Evaluating the Effect of Lubricants on the Fatigue Pitting of Steel, *Ind. Lab.*, Vol 26, 1960, p 405–407

35. D. Scott, Study of the Effect of Lubricant on Pitting Failure of Balls, *Lubrication and Wear Convention*, I Mech E, 1963, p 463–468

36. R. Tourret and E.P. Wright, *Rolling Contact Fatigue: Performance Testing of Lubricants,* Heyden & Son Ltd., London, 1977

37. A. Scott, J. Blackwell, and P.J. McCullagh, Silicon Nitride as a Rolling Bearing Material—A Preliminary Assessment, *Wear,* Vol 17, 1971, p 73–82

38. D. Scott and J. Blackwell, Hot Pressed Silicon Nitride as a Rolling Bearing Material—A Preliminary Assessment, *Wear,* Vol 24, 1973, p 61–67

39. M. Hadfield, T.A. Stolarski, and R.T. Cundill, Failure Modes of Ceramics in Rolling Contact, *Proc. R. Soc.* (*London*) *A,* Vol 443, 1993, p 607–621

40. M. Hadfield, T.A. Stolarski, R.T. Cundill, and S. Horton, Failure Modes of Ceramic Elements with Ring-Crack Defects, *Tribol. Int.,* Vol 26, 1993, p 157–164

41. M. Hadfield, Failure of Silicon Nitride Rolling Elements with Ring Crack Defects, *Ceram. Int.,* Vol 24, 1998, p 379–386

42. T.L. Carter and E.V. Zaretsky, "Rolling Contact Fatigue Life of a Crystallized Glass Ceramic," NASA TN D-259, National Aeronautics and Space Administration, 1960

43. R. Parker and E.V. Zaretsky, Fatigue Life of High-Speed Ball Bearings with Silicon Nitride Balls, *J. Lubr. Technol.,* (*Trans. ASME*), 1975, p 350–357

44. T. Fujiwara, T. Yoshioka, T. Kitahara, S. Koizumi, H. Takebayashi, and T. Tada, Study on Load Rating Property of Silicon Nitride for Rolling Bearing Material, *J. JSLE, Int. Ed.,* 1989, p 81–86

45. K. Kikuchi, T. Yoshioka, T. Kitahara, K. Ozaki, K. Nakayama, and T. Fujiwara, Rolling Contact Fatigue Life of Ceramics for Rolling Bearing Materials, *J. JSLE, Int. Ed.,* 1984, p 137–142

46. J.W. Lucek, "Rolling Wear of Silicon Nitride Bearing Materials," Paper Vol 90-GT-165, American Society of Mechanical Engineers, 1990

47. D.L. Allen, Effect of Composition and Physical Properties of Silicon Nitride on Rolling Wear and Fatigue Perfomance, *Tribol. Trans.,* Vol 37, 1994, p 410–414

48. H.I. Burrier, Optimizing the Structure and Properties of Silicon Nitride for Rolling Contact Bearing Performance, *Tribol. Trans.,* Vol 39, 1996, p 276–285

49. L.-Y. Chao, R. Lakshminarayanna, N. Iyer, G.-Y. Lin, and D.K. Shetty, Transient Wear of Silicon Nitride in Lubricated Rolling Contact, *Wear,* Vol 223, 1998, p 58–65

50. H.R. Baumgartner, "Evaluation of Roller Bearings Containing Hot-Pressed Silicon Nitride Rolling Elements," *Ceramics for High Performance Applications,* J.J. Burke, A.E. Gorum, and R.N. Katz, Ed., Brook Hill Publishing, 1974, p 713–727

51. J.W. Lucek and P.E. Cowley, "Investigation of the Use of Ceramic Material in Aircraft Engine Bearings," Code AIR-52032A, U.S. Dept. of Navy, 1978

52. M. Akazawa and K. Kato, Wear Properties of Silicon Nitride in Rolling-Sliding Contact, *Wear,* Vol 124, 1988, p 123–132

53. J.F. Braza, H.S. Cheng, and M.E. Fine, Silicon Nitride Wear Mechanisms: Rolling and Sliding Contact, *Tribol. Trans.,* Vol 32, 1989, p 439–446

54. M. Akazawa, K. Kato, and K. Umeya, Wear Properties of Silicon Nitride in Rolling Contact, *Wear*, Vol 110, 1986, p 285–293

Rolling-Contact Fatigue of Ceramics

Y. Wang and M. Hadfield, Bournemouth University (United Kingdom)

Failure Modes of RCF

As previously discussed, many test-rig types are used to investigate RCF of silicon nitride bearing materials. The RCF performance and failure modes were influenced by the test-rig type (Ref 55), and failure modes could be changed according to contact geometry. The experimental evaluation of silicon nitride should be conducted on a number of model tests before predicting its performance for rolling-element applications. Several failure modes of silicon nitride in rolling contact have been reported, that is, spalling fatigue failure, delamination failure, and rolling-contact wear.

Spalling Fatigue Failure. Spalling describes a specific form of surface damage, and it is the most common mode of failure in RCF tests. Silicon nitride spalls in a manner similar to bearing steels, which is, in fact, one of the reasons for silicon nitride being a good candidate for replacing bearing steels. Other structural ceramics, for example, SiC, typically fail by catastrophic fracture due to their low fracture toughness or severe wear by microfracture. Catastrophic failure can cause damage to the entire bearing assembly. Typical properties of engineering ceramics and bearing steel are shown in Table 1.

Table 1 Typical properties of engineering ceramics and bearing steel

Material	Density, g/cm^3	Elastic modulus, GPa (10^6 psi)	Hardness, kg/mm^2	Toughness, MPa $\sqrt{m}$ KSi $\sqrt{in}$	Failure mode
Silicon nitride	3.2	315 (46)	1400–1800	4–7 (3.6–6.4)	Spalling
Silicon carbide	3.1	420 (61)	2000–2400	2–4 (1.8–3.6)	Fracture
Alumina	3.9	390 (57)	1800–2000	3–5 (2.7–4.6)	Fracture
Zirconia	5.8	210 (30)	1100–1400	8–12 (7.3–10.9)	Spalling
Steel	7.8	200 (29)	1000	>16 (>14.6)	Spalling

The spalling fatigue mode has been found in various test machines (Ref 43, 44, 46, 48, 51, 56). For example, the fatigue life of hot pressed silicon nitride balls was investigated using a five-ball machine (Ref 43). Lubricated contact was studied with 12.7 mm (0.5 in.) diameter balls, Hertz stresses ranging from 4.3 to 6.2 GPa (0.62 to 0.90 × 10^6 psi), a shaft speed of 9600 rpm, and a contact angle 30°. There was no evidence of a wear mode of failure, only fatigue spalls similar to those found on steel. Life predictions showed that ceramic-steel contact compared equally to steel-steel contact. Also, the failure mode of hot pressed ceramic was investigated, using a disc-on-rod machine (Ref 51). All failures were of a spalling non-catastrophic nature, and test loads varied from 4.1 to 5.5 GPa (0.59 to 0.80 × 10^6 psi). The results inferred that silicon nitride tested under various tribological conditions fails in a noncatastrophic spalling mode. Additionally, silicon nitride hybrid bearings were tested, using 12.7 mm (0.5 in.) diameter balls, M50 steel rings, and a brass cage (Ref 56). Contact pressure ranged between 1.95 and 2.44 GPa (0.28 and 0.35 × 10^6 psi), and the film parameter was approximately 3.5 for all tests. Spalling was identified as the only mode of failure, with no cases of the silicon nitride balls fracturing. Statistical analysis showed that bearing life varies in accordance with a power law model. Another experimental work (Ref 44) examined the dynamic and static load rating of silicon nitride as a bearing material. The ball-on-plate machine was employed. Steel balls revolving around an inner race were loaded against a flat testpiece. The maximum contact stress was 6.4 GPa (0.93 × 10^6 psi), the rotational test speed was 1400 rpm, and the lubricant was a turbine mineral oil. Spalling was identified as the most common long-term mode of failure.

The microstructural variation is a significant factor in the fatigue failure of rolling contact. The RCF performance was studied, using eleven silicon nitride bearing materials (Ref 48). Tests were conducted using a ball-on-rod machine at maximum contact stresses of 5.93 GPa (0.86×10^6 psi) and rod rotation speed of 3600 rpm. Research showed that the materials exhibited differences in life span of several orders of magnitude, and the failure mode was spalling fatigue. Silicon nitride materials can exhibit a wide range of RCF endurance. Under relatively high-stress conditions, fatigue durability increased dramatically as the microstructure tended toward finer, more equiaxed grains and a uniform, minimum distribution of second phases.

Surface defects found on ceramic rolling elements can lead to early or catastrophic failure. Many efforts have been made in the past decade to investigate their influence on RCF performance (Ref 39, 40, 57). The failure modes of ring, radial, and lateral precracked ceramic elements under rolling contacts have been extensively studied (Ref 57). The stages of the ring crack failure process in the form of fatigue spalling on the ceramic surface and subsurface were also investigated (Ref 40). The results revealed that the precracked ceramic ball failed in a noncatastrophic spalling mode. The typical images of surface failure from artificially induced ring/cone cracks are shown in Fig. 8.

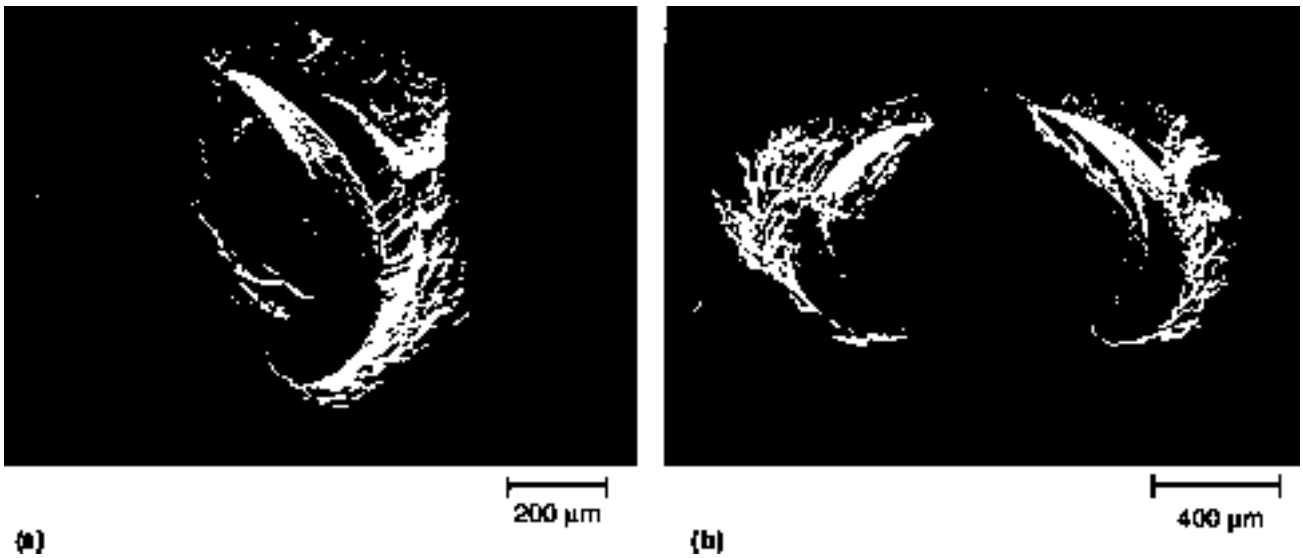


Fig. 8 Spalling fatigue failure resulting from artificially induced ring/cone cracks (SEM micrographs). (a) Single spall. (b) Double spall

All those results placed high reliance on random crack location and did not focus on the significance of specific crack locations. In a study (Ref 58), a new RCF test method was devised for positioning the ring crack, and hence, the surface ring crack was located precisely within the contact path. The results showed that the "natural" ring crack failure is dependent on the crack location within the contact path. Fatigue failure happens only in a few crack locations/orientations. The loading condition geometry of the test machine and the geometric orientations are shown in Fig. 9. Surface observations from different crack orientations are shown in Fig. 10. A fracture mechanics approach has been utilized to explain the influence of the crack location on RCF performance. The reason for differences in RCF performance is that different crack locations have different stress-intensity factor values that govern the crack propagation behavior (Ref 59). It should be noted that fatigue spalling processes are involved in the original ring crack propagation as well as the secondary surface cracks induced as the crack grows (Ref 60). Hence, the fatigue life cannot be simply determined in terms of the propagation life of the surface crack, and the subsequent surface damage also plays an important role in the whole life prediction. Figure 11 illustrates a typical section image of the spalling fatigue failure found on the ceramic balls (Ref 60).

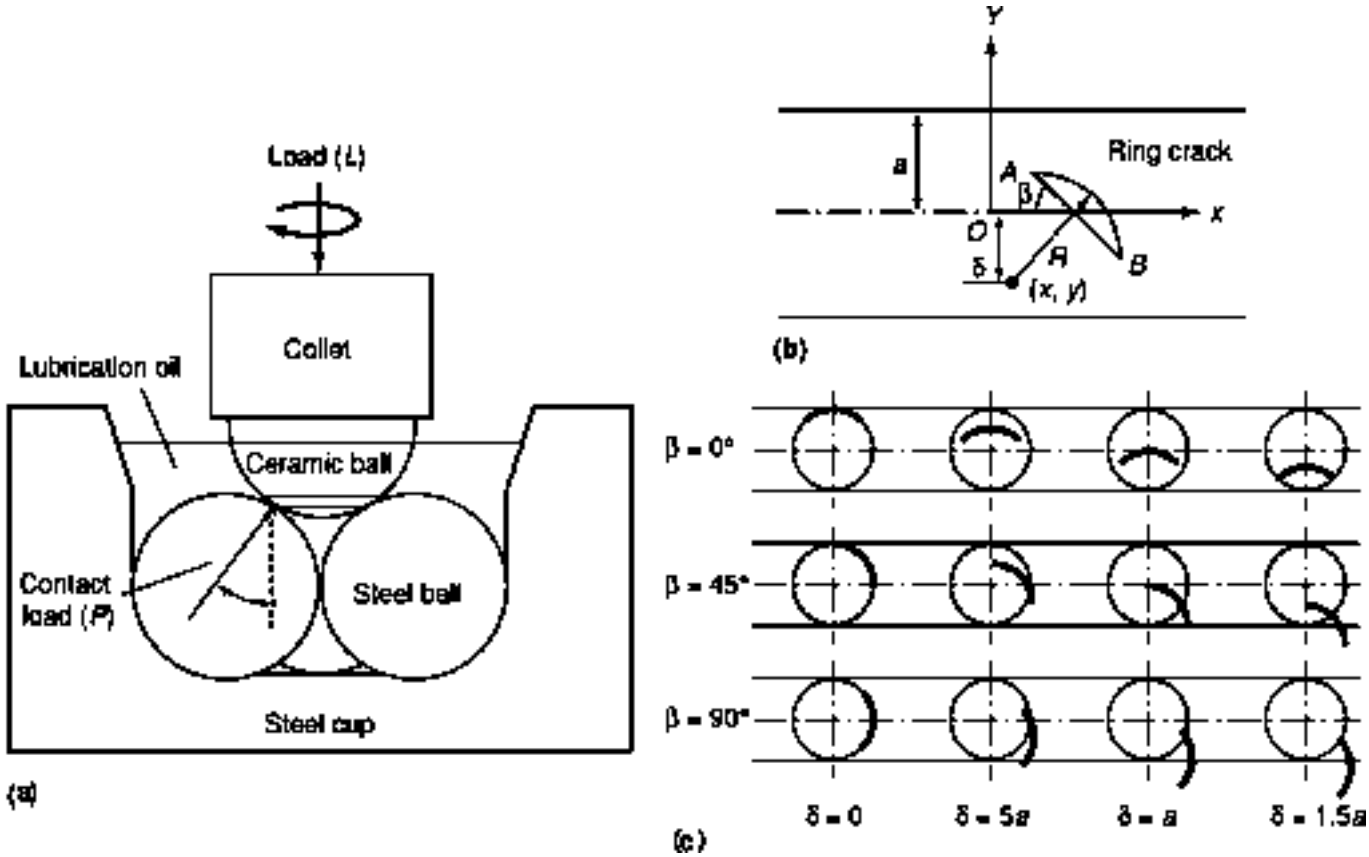


Fig. 9 Loading configuration and ring crack location on the contact track. (a) Loading system. (b) Geometric location on the contact track, where *a* is the radius of the contact circle. (c) Twelve typical locations within the contact path

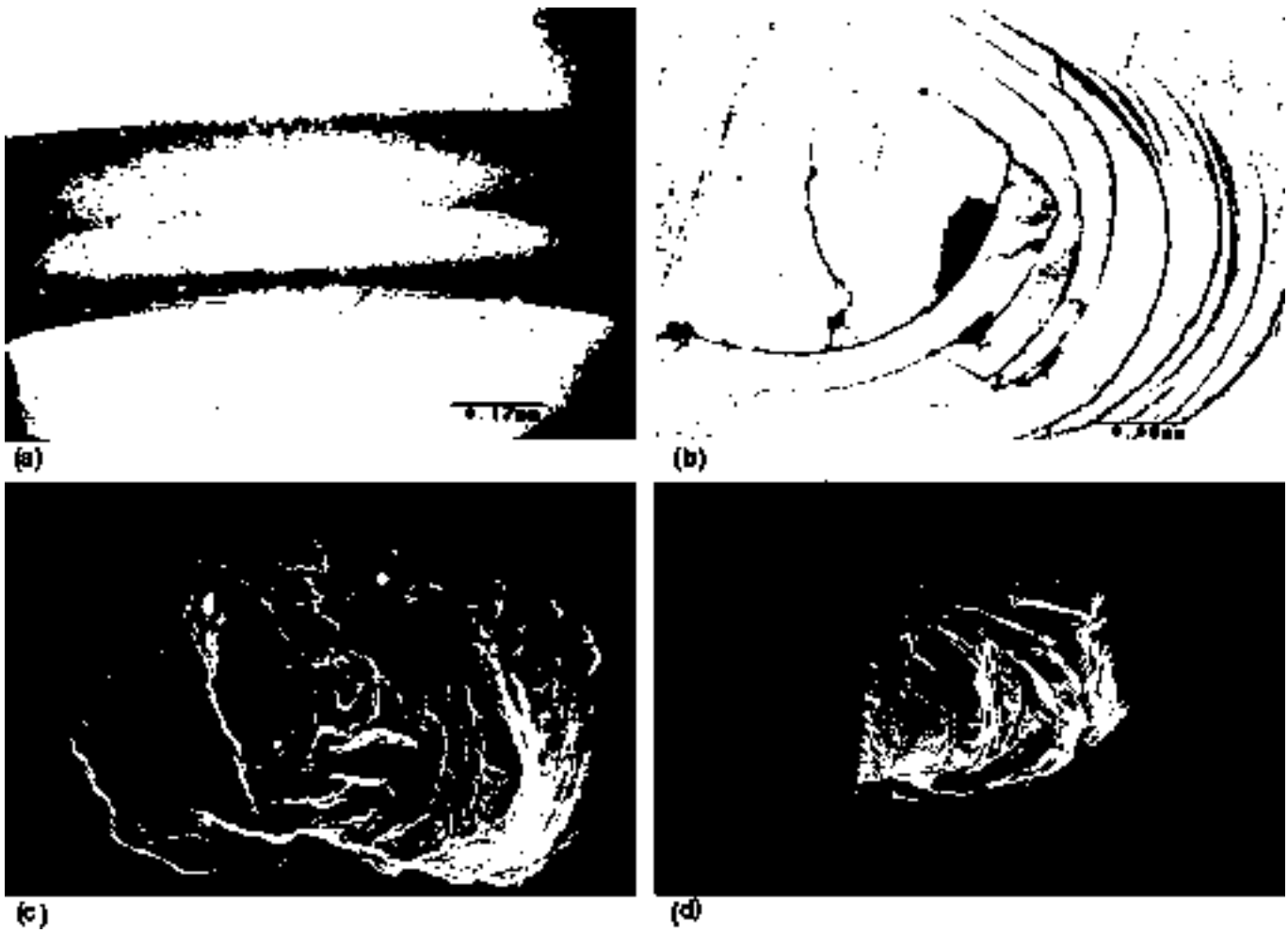


Fig. 10 (a)–(c) Surface fatigue damage resulting from "natural" ring cracks and (d) line defects. (a) Ring cracks and wear track after 113 million stress cycles at crack location β = 0° and δ = 0, where β is the angle of the chord of ring crack to the central line of the contact track, and δ is the distance of the center of the ring crack circle to the central line of the contact track (optical micrograph). (b) Incipient failure after 27 million stress cycles at the crack location β = 45° and δ = 0 (optical micrograph). (c) Spall SEM micrograph after 16 million stress cycles at crack location β = 90° and δ = 0. (d) Spall SEM micrograph after 1.4 million stress cycles at β = 90°

Fig. 11 Subsurface observation of spalling fatigue failure (optical micrograph)

Delamination Failure. Studies of delamination fatigue failure on perfect and artificially precracked silicon nitride balls were carried out using a modified four-ball machine (Ref 39, 57, 61, 62). The silicon nitride balls were in rolling contact with a steel ball to simulate the rolling condition of a hybrid bearing. The cracks were artificially induced, being in the forms of ring cracks or radial and lateral cracks. Testing under high compressive cyclic stress and various tribological conditions resulted in delamination failure. Scanning electron microscope (SEM) observations enabled the delamination failures to be classified by the location of crack initiation and subsequent propagation. For radial- and lateral-cracked balls, failure occurred in three ways: lateral-crack-induced spalling, radial-crack-induced surface bulging, and radial-crack-induced delamination. The failure mode was also sensitive to the lubricant used, because lubricant could penetrate into the subsurface cracks to produce significant hydrostatic pressure. Figure 12 shows typical images of surface damage.

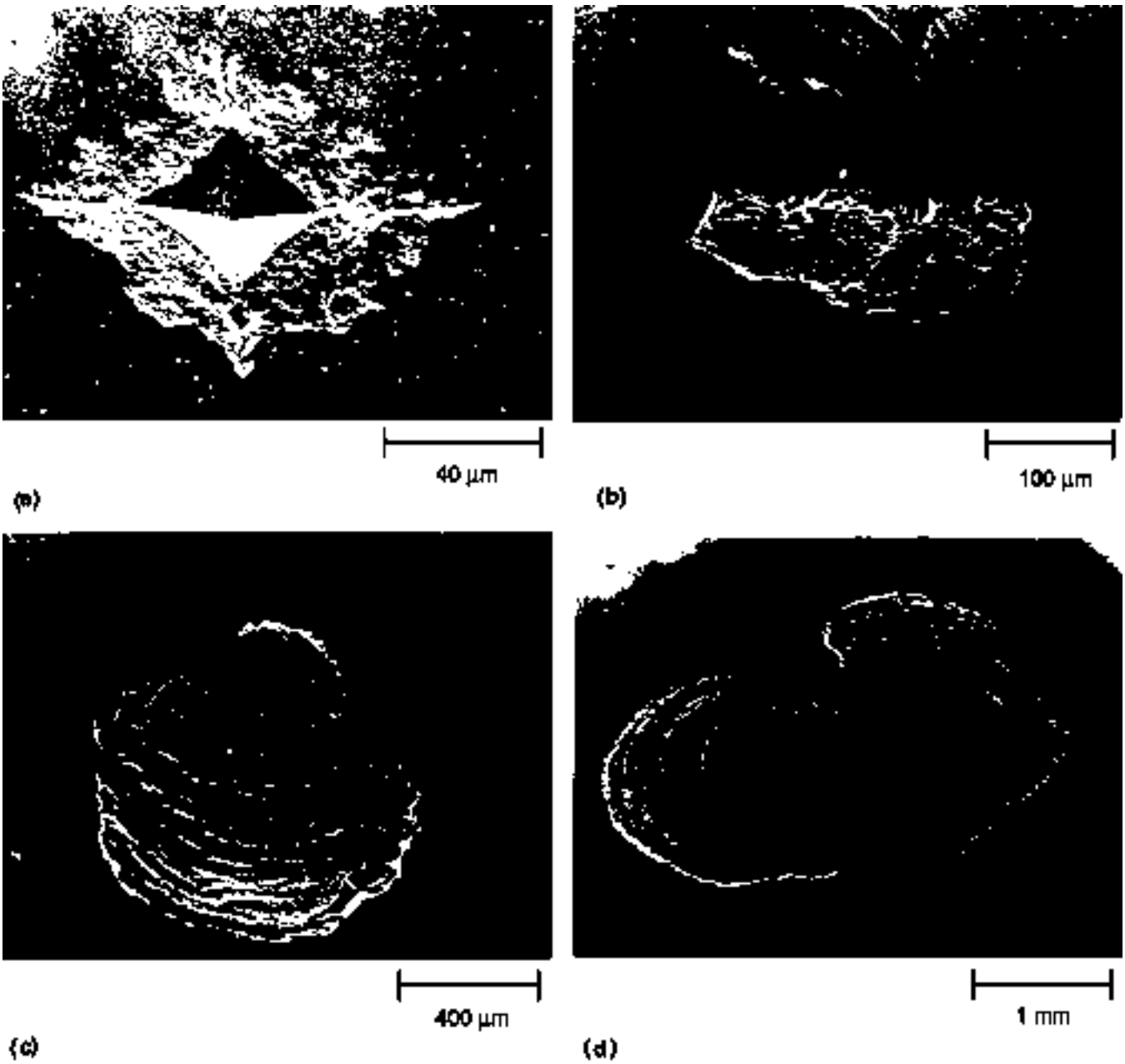


Fig. 12 Surface damage resulting from ceramic-steel contact (scanning electron microscope micrographs). (a) Lateral-crack spall. (b) Radial-crack propagation and delamination. (c) and (d) Ceramic-ceramic contact at high Hertz contact pressure

In the case of surface delamination, adjacent radial cracks were propagated and connected, and the delamination area showed typical fatigue striations. Residual stress measurement by x-ray diffraction indicated compressive stress in the delaminated area. In the case of surface bulging, ferrous debris originating from the upper steel ball was trapped under the ball surface. The steel debris, suspended in the pressurized lubricant, was forced up through the ball surface.

Rolling-contact wear is the most common type found on the contact path, and it has been reported by a number of researchers. Preliminary work used a modified four-ball machine to assess hot pressed silicon nitride as a rolling bearing material (Ref 38). Unlubricated contact was studied at different temperatures at 1500 rpm and a maximum Hertzian stress of 5.8 GPa (0.84×10^6 psi). Lubricated contact was also studied at 7.7 GPa (1.1×10^6 psi) and ambient temperature. Under unlubricated conditions, upper ball wear after 30 min was measured as 0.6 mg, and the wear track on the ball surface retained modest surface quality. For lubricated contact, failure occurred to the upper ball after 20 s, the failure mode being wear and fracture. A silicon nitride upper ball was tested in contact with an En31 steel lower ball, and this increased the time of failure to 70 s.

Wear properties of silicon nitride in dry rolling-sliding contact were studied using a ring-on-roller test machine (Ref 52). Under no lubrication and a maximum Hertzian pressure of 1.1 GPa (0.16×10^6 psi), results showed

thin, flakelike wear particles that were rich in oxygen and depended on the roughness of the contact surfaces. Rolling and sliding wear mechanism experiments were carried out (Ref 53). This work involved the use of a large-diameter cast iron roller contacting with a smaller-diameter silicon nitride roller, with an internal steel pin lubricated with a standard lubrication oil. Contact pressure was small, that is, 1.37 GPa (0.20×10^6 psi), and three types of silicon nitride were tested: hot pressed, pressureless sintered, and reaction bonded. The wear rate for the reaction-bonded type was higher, the wear mode being by "grain pullout." The other two types had similar wear rates, the mode of wear being submicron chipping. Rolling-contact friction coefficients were 0.0056 and all similar. Dry rolling tests (Ref 63) on several ceramics concluded that the silicon nitride wear rate was the smallest. A ring-on-disc testing machine was used, with the maximum Hertz stress ranging from 2.5 GPa to 5 GPa (0.36 to 0.72×10^6 psi). A relation between wear rate and mechanical properties was attempted without success, suggesting that a combination of material properties and roughness was needed. Brittle fracture was concluded to be the dominant wear mode.

The effect of composition and physical properties of silicon nitride on rolling wear and fatigue performance was studied using a ball-on-rod machine (Ref 47). Three commercial bearing materials were tested at maximum contact pressure of 6.4 GPa (0.93×10^6 psi). Wear and fatigue properties were compared to physical properties and material composition. It was found that high fracture toughness was not required for high fatigue life and wear resistance. Grain-boundary composition did not appear to affect the wear mechanism of these HIP materials. Transient wear of silicon nitride in lubricated rolling contact was investigated using a ball-on-rod test machine, with the maximum Hertz stress ranging from 5 to 6.5 GPa (0.72 to 0.94×10^6 psi). Wear rate and behavior were dependent on the initial surface roughness.

References cited in this section

38. D. Scott and J. Blackwell, Hot Pressed Silicon Nitride as a Rolling Bearing Material—A Preliminary Assessment, *Wear*, Vol 24, 1973, p 61–67

39. M. Hadfield, T.A. Stolarski, and R.T. Cundill, Failure Modes of Ceramics in Rolling Contact, *Proc. R. Soc.* (*London*) *A*, Vol 443, 1993, p 607–621

40. M. Hadfield, T.A. Stolarski, R.T. Cundill, and S. Horton, Failure Modes of Ceramic Elements with Ring-Crack Defects, *Tribol. Int.*, Vol 26, 1993, p 157–164

43. R. Parker and E.V. Zaretsky, Fatigue Life of High-Speed Ball Bearings with Silicon Nitride Balls, *J. Lubr. Technol.*, (*Trans. ASME*), 1975, p 350–357

44. T. Fujiwara, T. Yoshioka, T. Kitahara, S. Koizumi, H. Takebayashi, and T. Tada, Study on Load Rating Property of Silicon Nitride for Rolling Bearing Material, *J. JSLE, Int. Ed.*, 1989, p 81–86

46. J.W. Lucek, "Rolling Wear of Silicon Nitride Bearing Materials," Paper Vol 90-GT-165, American Society of Mechanical Engineers, 1990

47. D.L. Allen, Effect of Composition and Physical Properties of Silicon Nitride on Rolling Wear and Fatigue Perfomance, *Tribol. Trans.*, Vol 37, 1994, p 410–414

48. H.I. Burrier, Optimizing the Structure and Properties of Silicon Nitride for Rolling Contact Bearing Performance, *Tribol. Trans.*, Vol 39, 1996, p 276–285

51. J.W. Lucek and P.E. Cowley, "Investigation of the Use of Ceramic Material in Aircraft Engine Bearings," Code AIR-52032A, U.S. Dept. of Navy, 1978

52. M. Akazawa and K. Kato, Wear Properties of Silicon Nitride in Rolling-Sliding Contact, *Wear*, Vol 124, 1988, p 123–132

53. J.F. Braza, H.S. Cheng, and M.E. Fine, Silicon Nitride Wear Mechanisms: Rolling and Sliding Contact, *Tribol. Trans.,* Vol 32, 1989, p 439–446

55. M. Hadfield and T.A. Stolarski, The Effect of the Test Machine on the Failure Mode in Lubricated Rolling-Contact of Silicon-Nitride, *Tribol. Int.,* Vol 28, 1995, p 377–382

56. F.R. Morrison, J.I. McCool, and T.M. Yonushonis, The Load-Life Relationship for M50 Steel Bearings with Silicon Nitride Ceramic Balls, *Lubr. Eng. (J. ASME),* Vol 40, 1984, p 153–159

57. M. Hadfield, T.A. Stolarski, R.T. Cundill, and S. Horton, Failure Modes of Pre-Cracked Ceramic Elements under Rolling-Contact, *Wear,* Vol 169, 1993, p 69–75

58. Y. Wang and M. Hadfield, The Influence of Ring Crack Location on the Rolling Contact Fatigue Failure of Lubricated Silicon Nitride: Experimental Studies, *Wear,* Vol 243, 2000, p 167–174

59. Y. Wang and M. Hadfield, The Influence of Ring Crack Location on the Rolling Contact Fatigue Failure of Lubricated Silicon Nitride: Fracture Mechanics Analysis, *Wear,* Vol 243, 2000, p 157–166

60. Y. Wang and M. Hadfield, Ring Crack Propagation in Silicon Nitride under Rolling Contact, *Wear,* Vol 250, 2001, p 282–292

61. M. Hadfield, S. Tobe, and T.A. Stolarski, Subsurface Crack Investigation on Delaminated Ceramic Elements, *Tribol. Int.,* Vol 27, 1994, p 359–367

62. M. Hadfield and T.A. Stolarski, Observations of Delamination Fatigue on Precracked Ceramic Elements in Rolling-Contact, *Ceram. Int.,* Vol 21, 1995, p 125–130

63. S.S. Kim, K. Kato, K. Hokkirigawa, and H. Abe, Wear Mechanism of Ceramic Materials in Dry Rolling Friction, *Trans. ASME,* Vol 108, 1986, p 522–526

Rolling-Contact Fatigue of Ceramics

Y. Wang and M. Hadfield, Bournemouth University (United Kingdom)

References

1. G. Hamburg, P. Cowley, and R. Valori, Operation of an All-Ceramic Mainshaft Roller Bearing in J-402 Gas-Turbine Engine, *Lubr. Eng.* (*J. ASME*) Vol 37, 1981, p 407–415

2. H. Aramaki, Y. Shoda, Y. Morishita, and T. Sawamoto, The Performance of Ball Bearings with Silicon-Nitride Ceramic Balls in High Speed Spindles for Machine Tools, *J. Tribology* (*Trans. ASME*), Vol 110, 1988, p 693–698

3. B.R. Lawn, Indentation of Ceramics with Spheres: A Century after Hertz, *J. Am. Ceram. Soc.,* Vol 81, 1998, p 1977–1994

4. F.C. Roesler, Brittle Fracture near Equilibrium, *Proc. Phys. Soc. (London),* Vol B69, 1956, p 981–1012

5. M. Chaudhri and M.A. Phillips, Quasi-Static Indentation Cracking of Thermally Tempered Soda-Lime Glass with Spherical and Vickers Indenters, *Philos. Mag. A,* Vol 62, 1990, p 1–27

6. M. Ichikawa, T. Takamatsu, T. Matsuo, N. Okabe, and Y. Abe, Intra-Ball and Inter-Ball Variability of Ring Crack Initiation Load of Silicon-Nitride Bearing Balls, *JSME Int. J. A: Mech. Mater. Eng.*, Vol 38, 1995, p 231–235

7. M. Ichikawa, T. Takamatsu, N. Shindou, N. Okabe, and Y. Abe, Ring Crack Initiation Load of Silicon Nitride Bearing Balls, *JSME Int. J. A: Mech. Mater. Eng.*, Vol 38, 1995, p 226–230

8. K. Ohgushi and M. Ichikawa, Fracture Mechanics Study of Ring Crack Initiation in Ceramics by Sphere Indentation, *JSME Int. J. A: Mech. Mater. Eng.*, Vol 39, 1996, p 489–495

9. R.T. Cundill, "Impact Resistance of Silicon Nitride Balls," Proceedings of the Sixth International Symposium on *Ceramic Materials and Components for Engines* (Arita, Japan), Japan Fine Ceramics Association, 1997 p 556–561

10. F.C. Frank and B.R. Lawn, On the Theory of Hertzian Fracture, *Proc. R. Soc.* (*London*) *A*, Vol 299, 1967, p 291–306

11. T.R. Wilshaw, The Hertzian Fracture Test, *J. Phys. D, Appl. Phys.*, Vol 4, 1971, p 1567–1581

12. R. Mouginot and D. Maugis, Fracture Indentation beneath Flat and Spherical Punches, *J. Mater. Sci.*, Vol 20, 1985, p 4354–4376

13. B.R. Lawn, Partical Cone Crack Formation in a Brittle Material Loaded with a Sliding Spherical Indenter, *Proc. R. Soc.* (*London*) *A*, Vol 299, 1967, p 307–316

14. B.J. Bethune, The Surface Cracking of Glassy Polymers under a Sliding Spherical Indenter, *J. Mater. Sci.*, Vol 11, 1971, p 199–205

15. R. Warren, Measurement of the Fracture Properties of Brittle Solids by Hertzian Indenters, *Acta Metall.*, Vol 46, 1978, p 1759–1769

16. Y.Z. Li and D.A. Hills, The Hertzian Cone Crack, *J. Appl. Mech.* (*Trans. ASME*), Vol 58, 1991, p 120–127

17. P.D. Warren, D.A. Hills, and D.N. Dai, Mechanics of Hertzian Cracking, *Tribol. Int.*, Vol 28, 1995, p 357–362

18. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear*, Vol 44, 1977, p 1–16

19. J.F. Braza, H.S. Cheng, M.E. Fine, A.K. Gangopadhyay, L.M. Keer, and R.E. Worden, Mechanical Failure Mechanisms in Ceramic Sliding and Rolling Contacts, *Tribol. Trans.*, Vol 32, 1989

20. J.R. Fleming and N.P. Suh, Mechanics of Crack Propagation in Delamination Wear, *Wear*, Vol 44, 1977, p 39–56

21. D.A. Hills and D.W. Ashelby, A Fracture Mechanics Approach to Rolling Contact Fatigue, *Tribol. Int.*, 1979, p 115–119

22. A.R. Rosenfield, A Fracture Mechanics Approach to Wear, *Wear*, Vol 61, 1980, p 125–132

23. L.M. Keer and M.D. Bryant, A Pitting Model for Rolling-Contact Fatigue, *J. Lubr. Technol.* (*Trans. ASME*), Vol 105, 1983, p 198–205

24. M.D. Bryant, G.R. Miller, and L.M. Keer, Line Contact between a Rigid Indenter and a Damaged Elastic Body, *Q. J. Mech. Appl. Math.,* Vol 37, 1984, p 468–478

25. G.R. Miller, L.M. Keer, and H.S. Cheng, On the Mechanics of Fatigue Crack-Growth due to Contact Loading, *Proc. R. Soc.* (*London*) *A: Math. Phys. Sci.,* Vol 397, 1985, p 197–209

26. A.F. Bower, The Influence of Crack Face Friction and Trapped Fluid on Surface Initiated Rolling-Contact Fatigue Cracks, *J. Tribology* (*Trans. ASME*), Vol 110, 1988, p 704–711

27. L.M. Keer and R.E. Worden, A Qualitative Model to Describe the Microchipping Wear Mode in Ceramic Rollers, *Tribol. Trans.,* Vol 33, 1990, p 411–417

28. N.B. Thomsen and B.L. Karihaloo, Subsurface and Surface Cracks under Contact Loading in Transformation-Toughened Ceramics, *J. Mech. Phys. Solids,* Vol 44, 1996, p 207–231

29. H. Tada, P.C. Paris, and G.R. Irwin, *Stress Analysis of Cracks Handbook,* Paris Publications Inc., St. Louis, MO, 1985

30. J.C. Newman and I.S. Raju, *Computational Methods in the Mechanics of Fracture,* S.N. Atluri, Ed., Elsevier Science Publishers B.V., 1986, p 312–334

31. R.O. Ritchie and R.H. Dauskardt, Cyclic Fatigue of Ceramics—A Fracture-Mechanics Approach to Subcritical Crack-Growth and Life Prediction, *J. Ceram. Soc. Jpn.,* Vol 99, 1991, p 1047–1062

32. R.O. Ritchie, C.J. Gilbert, and J.M. McNaanry, Mechanics and Mechanisms of Fatigue Damage and Crack Growth in Advanced Materials, *Int. J. Solids Struct.* Vol 37, 2000, p 311–329

33. F.T. Barwell and D. Scott, Effect of Lubricant on Pitting Failure of Ball Bearings, *Eng.,* 6 July 1956, p 9–12

34. G.S. Krivoshein, On Evaluating the Effect of Lubricants on the Fatigue Pitting of Steel, *Ind. Lab.,* Vol 26, 1960, p 405–407

35. D. Scott, Study of the Effect of Lubricant on Pitting Failure of Balls, *Lubrication and Wear Convention,* I Mech E, 1963, p 463–468

36. R. Tourret and E.P. Wright, *Rolling Contact Fatigue: Performance Testing of Lubricants,* Heyden & Son Ltd., London, 1977

37. A. Scott, J. Blackwell, and P.J. McCullagh, Silicon Nitride as a Rolling Bearing Material—A Preliminary Assessment, *Wear,* Vol 17, 1971, p 73–82

38. D. Scott and J. Blackwell, Hot Pressed Silicon Nitride as a Rolling Bearing Material—A Preliminary Assessment, *Wear,* Vol 24, 1973, p 61–67

39. M. Hadfield, T.A. Stolarski, and R.T. Cundill, Failure Modes of Ceramics in Rolling Contact, *Proc. R. Soc.* (*London*) *A,* Vol 443, 1993, p 607–621

40. M. Hadfield, T.A. Stolarski, R.T. Cundill, and S. Horton, Failure Modes of Ceramic Elements with Ring-Crack Defects, *Tribol. Int.,* Vol 26, 1993, p 157–164

41. M. Hadfield, Failure of Silicon Nitride Rolling Elements with Ring Crack Defects, *Ceram. Int.,* Vol 24, 1998, p 379–386

42. T.L. Carter and E.V. Zaretsky, “Rolling Contact Fatigue Life of a Crystallized Glass Ceramic,” NASA TN D-259, National Aeronautics and Space Administration, 1960

43. R. Parker and E.V. Zaretsky, Fatigue Life of High-Speed Ball Bearings with Silicon Nitride Balls, *J. Lubr. Technol.,* (*Trans. ASME*), 1975, p 350–357

44. T. Fujiwara, T. Yoshioka, T. Kitahara, S. Koizumi, H. Takebayashi, and T. Tada, Study on Load Rating Property of Silicon Nitride for Rolling Bearing Material, *J. JSLE, Int. Ed.,* 1989, p 81–86

45. K. Kikuchi, T. Yoshioka, T. Kitahara, K. Ozaki, K. Nakayama, and T. Fujiwara, Rolling Contact Fatigue Life of Ceramics for Rolling Bearing Materials, *J. JSLE, Int. Ed.,* 1984, p 137–142

46. J.W. Lucek, “Rolling Wear of Silicon Nitride Bearing Materials,” Paper Vol 90-GT-165, American Society of Mechanical Engineers, 1990

47. D.L. Allen, Effect of Composition and Physical Properties of Silicon Nitride on Rolling Wear and Fatigue Perfomance, *Tribol. Trans.,* Vol 37, 1994, p 410–414

48. H.I. Burrier, Optimizing the Structure and Properties of Silicon Nitride for Rolling Contact Bearing Performance, *Tribol. Trans.,* Vol 39, 1996, p 276–285

49. L.-Y. Chao, R. Lakshminarayanna, N. Iyer, G.-Y. Lin, and D.K. Shetty, Transient Wear of Silicon Nitride in Lubricated Rolling Contact, *Wear,* Vol 223, 1998, p 58–65

50. H.R. Baumgartner, “Evaluation of Roller Bearings Containing Hot-Pressed Silicon Nitride Rolling Elements,” *Ceramics for High Performance Applications,* J.J. Burke, A.E. Gorum, and R.N. Katz, Ed., Brook Hill Publishing, 1974, p 713–727

51. J.W. Lucek and P.E. Cowley, “Investigation of the Use of Ceramic Material in Aircraft Engine Bearings,” Code AIR-52032A, U.S. Dept. of Navy, 1978

52. M. Akazawa and K. Kato, Wear Properties of Silicon Nitride in Rolling-Sliding Contact, *Wear,* Vol 124, 1988, p 123–132

53. J.F. Braza, H.S. Cheng, and M.E. Fine, Silicon Nitride Wear Mechanisms: Rolling and Sliding Contact, *Tribol. Trans.,* Vol 32, 1989, p 439–446

54. M. Akazawa, K. Kato, and K. Umeya, Wear Properties of Silicon Nitride in Rolling Contact, *Wear,* Vol 110, 1986, p 285–293

55. M. Hadfield and T.A. Stolarski, The Effect of the Test Machine on the Failure Mode in Lubricated Rolling-Contact of Silicon-Nitride, *Tribol. Int.,* Vol 28, 1995, p 377–382

56. F.R. Morrison, J.I. McCool, and T.M. Yonushonis, The Load-Life Relationship for M50 Steel Bearings with Silicon Nitride Ceramic Balls, *Lubr. Eng. (J. ASME),* Vol 40, 1984, p 153–159

57. M. Hadfield, T.A. Stolarski, R.T. Cundill, and S. Horton, Failure Modes of Pre-Cracked Ceramic Elements under Rolling-Contact, *Wear,* Vol 169, 1993, p 69–75

58. Y. Wang and M. Hadfield, The Influence of Ring Crack Location on the Rolling Contact Fatigue Failure of Lubricated Silicon Nitride: Experimental Studies, *Wear,* Vol 243, 2000, p 167–174

59. Y. Wang and M. Hadfield, The Influence of Ring Crack Location on the Rolling Contact Fatigue Failure of Lubricated Silicon Nitride: Fracture Mechanics Analysis, *Wear,* Vol 243, 2000, p 157–166

60. Y. Wang and M. Hadfield, Ring Crack Propagation in Silicon Nitride under Rolling Contact, *Wear,* Vol 250, 2001, p 282–292

61. M. Hadfield, S. Tobe, and T.A. Stolarski, Subsurface Crack Investigation on Delaminated Ceramic Elements, *Tribol. Int.,* Vol 27, 1994, p 359–367

62. M. Hadfield and T.A. Stolarski, Observations of Delamination Fatigue on Precracked Ceramic Elements in Rolling-Contact, *Ceram. Int.,* Vol 21, 1995, p 125–130

63. S.S. Kim, K. Kato, K. Hokkirigawa, and H. Abe, Wear Mechanism of Ceramic Materials in Dry Rolling Friction, *Trans. ASME,* Vol 108, 1986, p 522–526

Impact Wear Failures

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Introduction

IMPACT, OR PERCUSSIVE, WEAR has been defined as “the wear of a solid surface that is due to percussion, which is a repetitive exposure to dynamic contact by another body” (Ref 1). This is a restricted definition that distinguishes impact wear from erosive wear, which occurs by impact of streams of small particles and which is covered elsewhere in this Volume. Impact wear, however, is closely related to erosive wear. The main difference is that in impact wear, the bodies tend to be large and contact in a well defined location in a controlled way, whereas in erosion, the eroding particles are small and interact randomly with the target surface.

Many industrial situations involve impacting bodies in which failures due to wear can be costly and safety, reliability, and quality can be greatly reduced.

Excessive wear of impacting poppet valves in automotive engines can lead to loss of cylinder pressure and ultimately engine failure. Failure of tools used for drilling rock and other media raises cost concerns associated not only with the need for frequent replacement of parts, but also with the down time incurred. Impact wear of printer typefaces will lead to loss of definition and a subsequent reduction in print quality. Unintentional impacts, such as the chattering of tubes carrying, for example, waste from nuclear reactors, can give rise to critical safety and reliability concerns. Potential impact wear problems are also found in dental implants and heart valves, where health and well being are at stake.

In this article, some generic features of impact wear are described and one particular application is examined in detail.

Reference cited in this section

1. P.A. Engel, Impact Wear, *Friction, Lubrication and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 263–270

Impact Wear Failures

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Impact Wear of Metals

Deformation during Impact. In bodies subjected to repetitive impulse loading, characteristic subsurface zones have been observed for a large variety of materials under various test conditions (with both normal and compound impact) (Ref 4).

In general, three subsurface regions can be seen, as shown in Fig. 2. The first, furthest from the contact surface, consists of undisturbed base material. The *deformation zone,* the intermediate region, consists of plastically deformed base material. The deformation ranges from zero at the interface with zone 1, to a maximum at the interface with the third region, the *white layer*. The white layer, containing the contact surface, is usually homogeneous and finely structured.

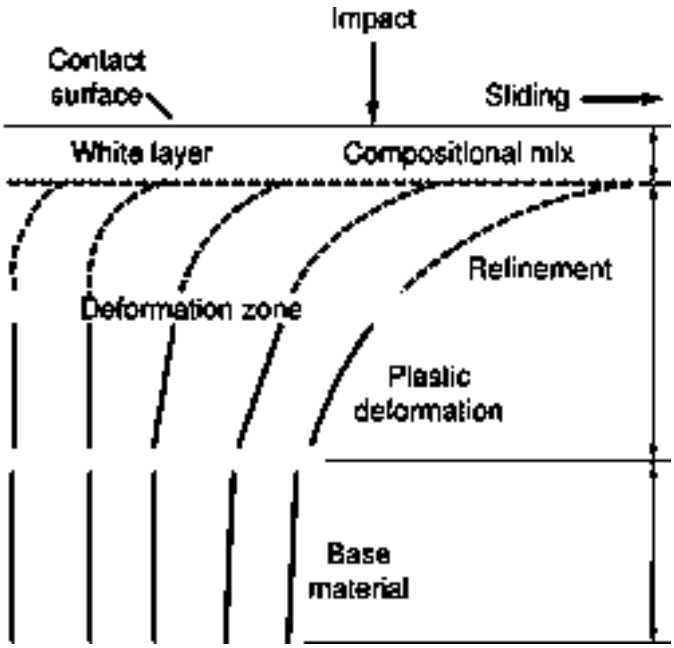


Fig. 2 Schematic representation of subsurface zones found beneath surfaces subjected to repetitive impact loading. Source: Ref 4

The characteristics of the subsurface regions vary with the impacting materials and the conditions of stress and sliding speed at the contact. In some cases, there is strong evidence of crack and void formation in subsurface regions (Ref 5), while in others, no cracks are visible (Ref 6). In other specimens, varying degrees of crack formation in the white layer or along the deformation zone/white layer interface (Ref 7) have been observed.

The composition of the white layer is often mixed, consisting of material from both impacting bodies as well as constituents of the operating environment. In one-body wear situations, the white layer in one body can be entirely made up of material transferred from the second body.

Subsurface regions form on both impacting bodies and grow very quickly, within a few hundred impacts (Ref 8). Once formed, the white layer and the deformation zone maintain a steady-state configuration as more impacts occur.

The formation of the white layer has been attributed to adhesion (Ref 9), although mechanical “mixing” and diffusion also take place. More probably, a combination of these three occurs to give the specific morphology and composition of the white layer. The deformation zone is clearly formed by the repetitive stress cycling.

Hardness has been noted to increase at the surface of impacting bodies (Ref 2, 5, 9). Within the deformation zone this is credited to work hardening, while in the white layer, it is a property of the near surface material developed in situ (Ref 9).

A different subsurface region system develops when a brittle or nondeforming matrix is subjected to repetitive impact (Ref 8). A white layer still forms, but the base material tends to remain undeformed (see Fig. 3).

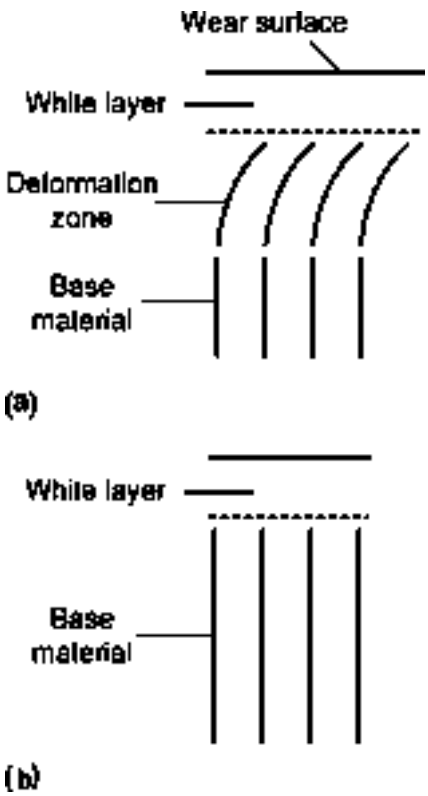


Fig. 3 Subsurface regions for (a) ductile and (b) brittle or non-deforming materials. Source: Ref 9

Material Removal Mechanisms. The mechanisms by which material can be removed from a surface by repeated impacts include the following (Ref 10):

- *Oxidative wear:* mild wear regime where contacting bodies are separated by oxide films reducing metal-to-metal contact
- *Adhesion:* a more severe wear regime characterized by material smearing and transfer
- *Abrasion:* equally severe, one-body or two-body modes lead to the formation of ploughing grooves
- *Surface fatigue:* more severe wear again, leading to subsurface crack nucleation and subsequent delamination or spalling of surface material
- *Plastic deformation:* while not leading to actual material removal, causes displacement of material from the contact area through “mushrooming”

Several of these mechanisms of impact wear are shown in Fig. 4.

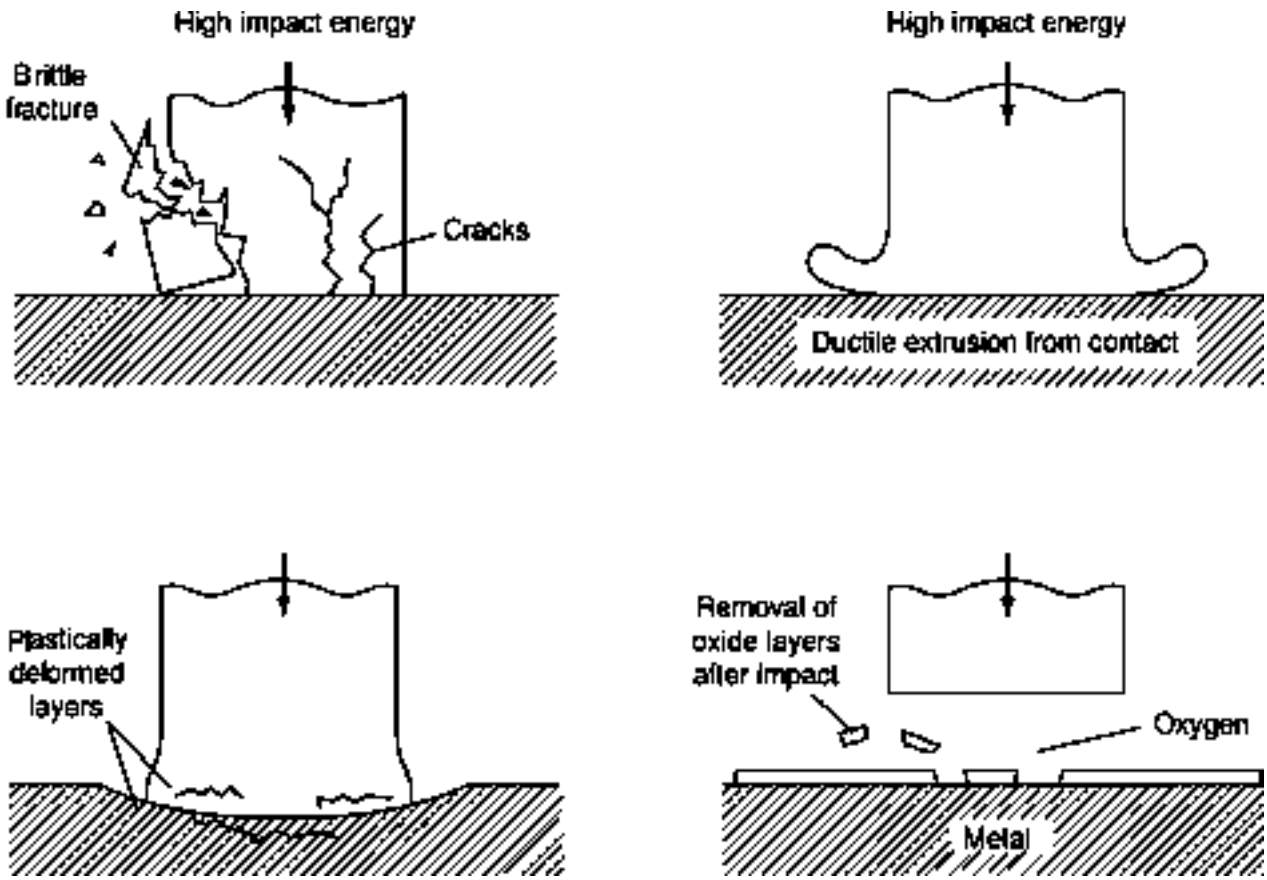


Fig. 4 Schematic diagrams of the different mechanisms of impact wear. Source: Ref 11

The mechanisms occurring in a given impact case are determined by the stress and sliding conditions within the contact. For example, in high-strength tungsten carbide bits of rock drilling equipment, surface fatigue leads to the formation of spalls (Ref 12). Repetitive impact of metallic components at low stress levels, however, is more likely to result in small, submicron wear particles and a highly oxidized surface state.

Under moderate conditions of normal or compound impact, it has been shown that that wear is caused by a process similar to that of adhesive wear during sliding (Ref 3, 13). More severe conditions at the contact lead to the occurrence of surface fatigue, subsurface crack growth and spalling (Ref 12, 14, 15, 16). Very large impact energies bring about failures due to surface fracture and produce rapid material removal (Ref 17, 18). This evolution of wear mechanisms, as the conditions in an impact contact become more severe, is illustrated in Fig. 5.

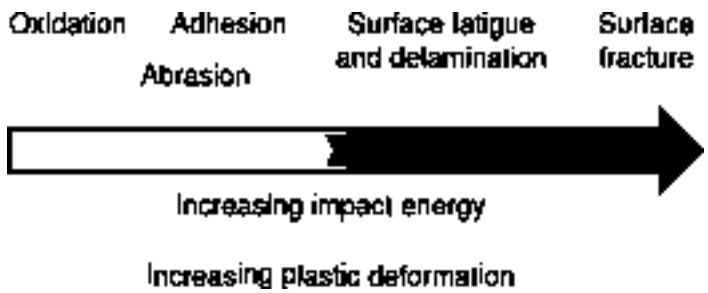


Fig. 5 Evolution of impact wear mechanisms as conditions in a contact become more severe

Oxidative wear tends to occur at low stress levels and can be found in both normal and compound impact contacts. Wear rates are low as the contacting surfaces are separated by oxide films, which tend to have better lubricating properties than bare metal surfaces. Wear debris is generally fine and predominantly made up of metal oxide. Oxidative wear is particularly prevalent in contacts with a small amount of sliding or those in which fretting occurs, such as the backprinter device of a hammer striking against a print band (Ref 19).

Adhesive wear occurs in both compound and normal impact contacts and is characterized by material smearing and transfer of material from one contacting body to the other (Ref 20). This occurs as a result of the formation

of welded junctions at asperity tips, as shown in Fig. 6. Cracks nucleate in the workhardened tips, which break away, remaining adhered to the lower surface. This can occur as one-body wear, when material is only transferred from one body, or as two-body wear, when transfer occurs from both.

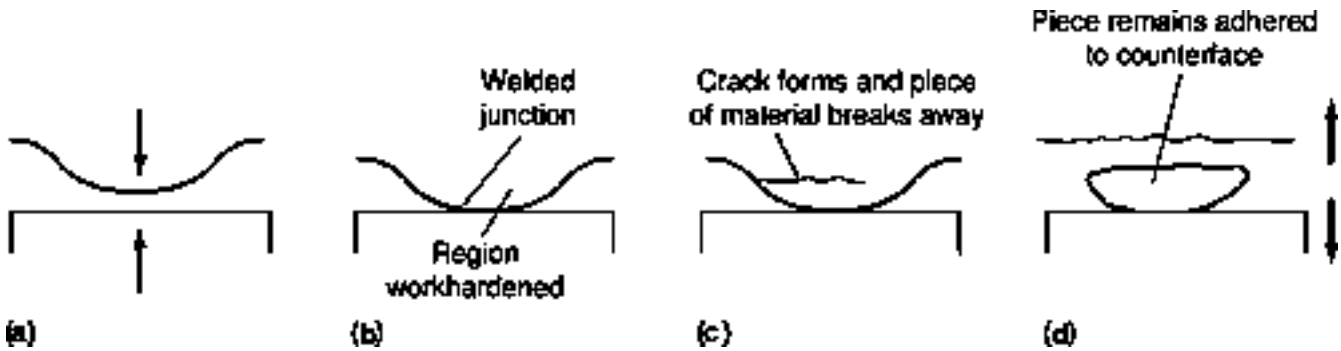


Fig. 6 Schematic diagram of the formation of an adhesive transfer particle

Abrasive wear, mainly found in compound impact contacts, manifests itself as plowing grooves and leads to a roughening of the contact surface (Ref 21). The grooves become more accentuated as impact stresses increase. However, under closer examination, it can be seen that at higher impact stresses, there is evidence of crack nucleation and the formation of flat platelike particles within the grooves (Ref 5).

Surface Fatigue and Delamination. In general, impact wear failure mechanisms are dependent on the formation of the subsurface layers described in the previous section. Fatigue failure in impact is mainly induced in the white layer—the subsurface zone including the wear surface (see Fig. 2) (Ref 8, 9, 22, 23).

Two mechanisms have been identified within the white layer. These explain the occurrence of delamination and spalling of impacting bodies respectively (Ref 22), as shown in Fig. 7.

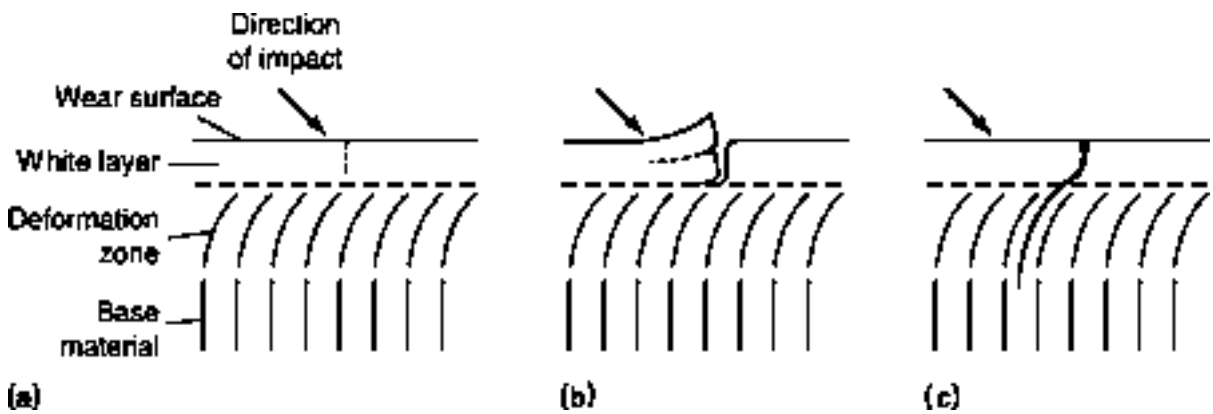


Fig. 7 Failure mechanisms induced by the white layer. Source: Ref 22

In the case of impacting bodies, the white layer builds up on the wear surface as impacts progress and becomes intensely work hardened. As a result, microcracks are easily nucleated within the white layer material (see Fig. 7a). The path of least resistance along which the cracks can propagate is the deformation zone/white layer interface and, hence, it is here that delamination occurs (see Fig. 7b).

This process is thought to be different to the conventional delamination theory (Ref 24, 25), in which increasing shear deformation, as a result of bodies sliding against one another under load, leads to the piling-up of subsurface dislocations. Under this theory, these cause the nucleation of microcracks, which subsequently lead to the occurrence of delamination.

In addition to the surface white layer, subsurface white layers have been observed at depths of tens to several hundred microns in bodies subjected to repetitive impacts (Ref 23). Propagation of cracks in a subsurface white layer leads to removal of material between the wear surface and the white layer.

Spalling occurs when microcracks within the white layer propagate in a direction perpendicular to the wear surface and parallel to the flow lines in the deformation zone (see Fig. 7c). The direction of the crack

propagation is affected by defects present in the deformation zone. Microcracks in the white layer tend to merge with cracks formed at defects in the deformation zone.

Plastic Deformation. While mass loss in impact wear mainly results from wear debris arising from the subsurface zones, wear also occurs as a result of plastic deformation and subsequent “mushrooming” of the impacting bodies, leading to the creation of leading and trailing edges (see Fig. 4) (Ref 5, 6, 11).

Influences on Wear. The largest influences on normal and compound impact wear failures are contact stress (Ref 9) and sliding velocity (Ref 8). The magnitude of both directly affects the formation of the subsurface zones in the impacting bodies. At low impact stresses, there is some evidence of local fracture and deformation. At significantly higher stresses, greater deformation occurs and a distinct deformation zone is formed. Wear rates increase with increasing contact stress, as shown in Fig. 8 (Ref 5). The data are plotted as change in specimen height rather than mass loss because impact stress was varied by altering the contact area and, therefore, the larger-diameter specimens would have lost more mass.

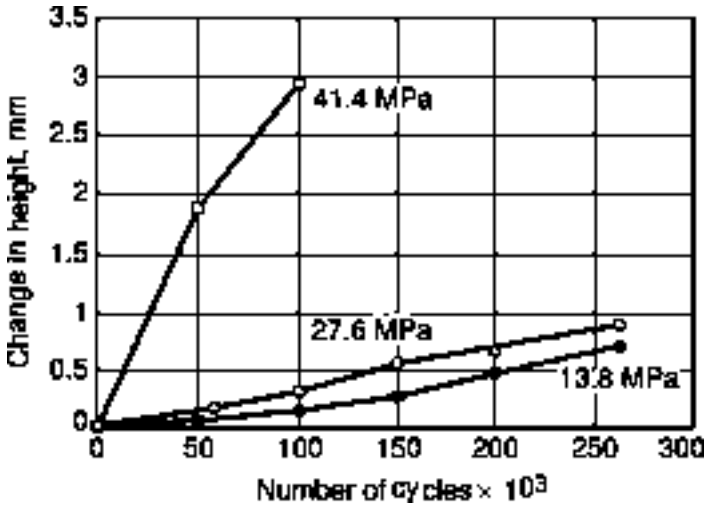


Fig. 8 Height change vs. number of compound impact cycles for aluminium 2011 T3 specimens tested against 17-4 PH stainless steel counterfaces with varying impact stresses (sliding velocity 5.33 m/s). Source: Ref 5

In bodies impacted with a low sliding velocity, crack formation occurs. At higher velocities, however, the white layer is usually quite homogeneous and integrally bonded to its substrate. These two mechanisms lead to observations that wear rates both increase and decrease with increasing sliding velocity (see Fig. 9, 10) (Ref 5, 26). This can be explained by studying data over a greater range of sliding velocities, as shown in Fig. 11 (Ref 20). As can be seen, the wear exhibits both a maximum and a minimum over the range of sliding velocities tested. This phenomenon has also been observed in high-speed sliding wear tests (Ref 27).

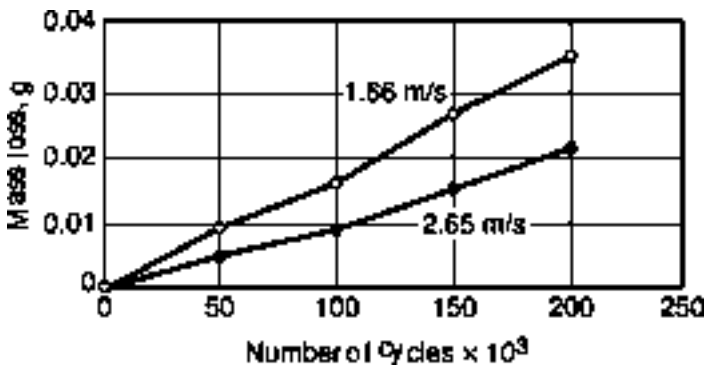


Fig. 9 Mass loss vs. number of compound impact cycles for aluminium 2011 T3 specimens tested against 17-4 PH stainless steel counterfaces with varying sliding velocities (impact stress 10.8 MPa). Source: Ref 5

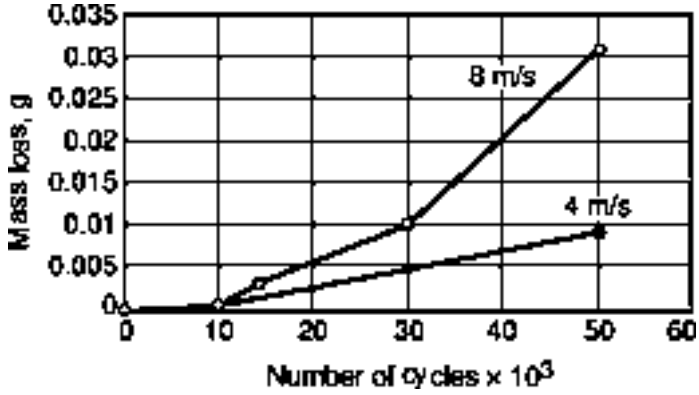


Fig. 10 Mass loss vs. number of compound impact cycles for 17-4 PH stainless steel counterfaces tested with CPM-10V steel specimens (impact stress 69 MPa). Source: Ref 26

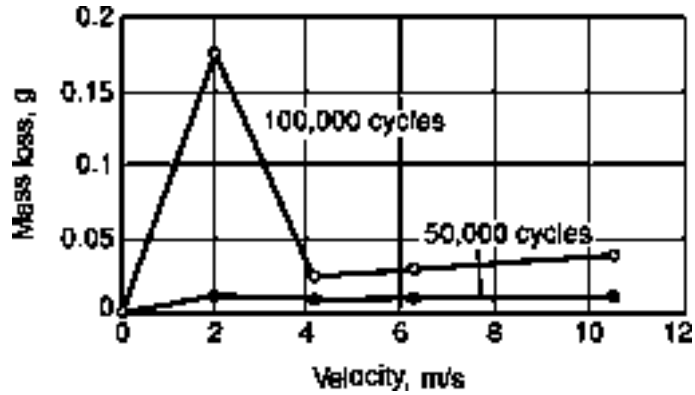


Fig. 11 Mass loss vs. sliding velocity for compound impact testing of titanium alloy RMI 5522S specimens against 17-4 PH stainless steel counterfaces (impact stress 18.6 MPa). Source: Ref 20

Specimen stiffness affects the wear rate of both impacting bodies (Ref 21). As shown in Fig. 12, the short, stiff specimen wore the counterface and yet hardly wore at all itself. In contrast, the longer, less stiff specimen wore considerably more, yet its counterface was worn much less than that with the stiffer specimen.

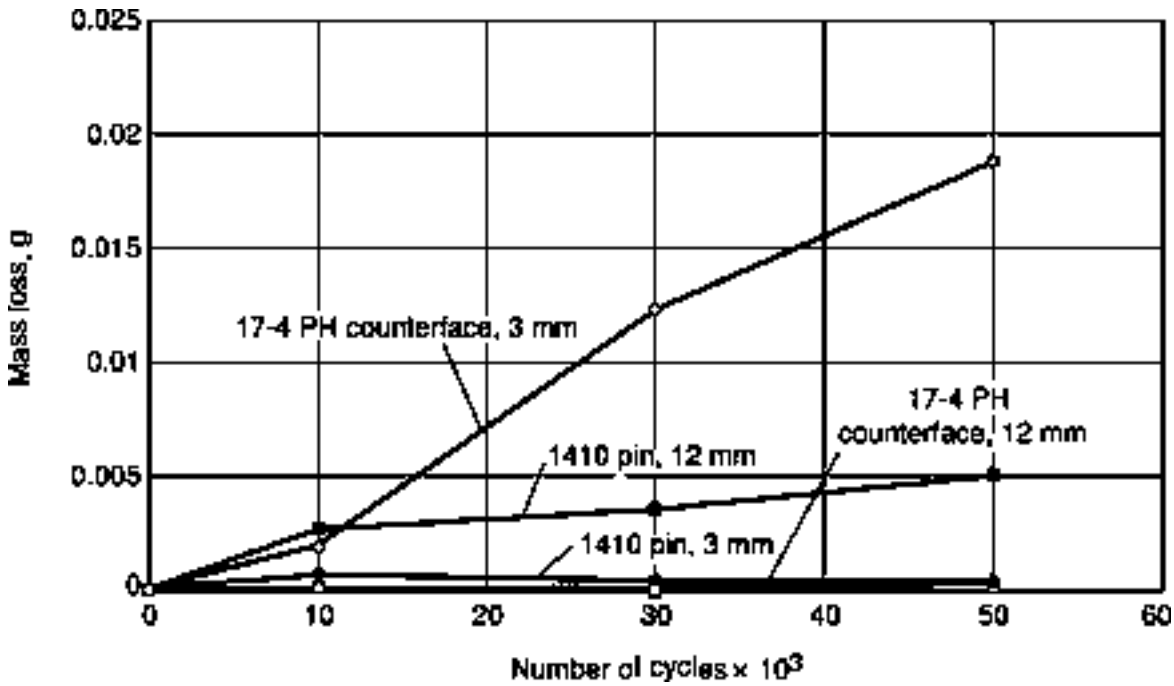


Fig. 12 Mass loss vs. number of compound impact cycles for 1410 steel specimens of varying length run against 17-4 PH stainless steel counterfaces (impact stress 69 MPa; sliding velocity 10 m/s). Source: Ref 21

Microstructure and bulk material hardness also play a role in impact wear (Ref 5, 28). These are critical in determining the development of the white layer and the subsequent crack nucleation and wear debris formation.

In repetitive normal impact applications with lubrication, it is probable that asperity contact will be frequent, except for combinations of low-contact pressure, high-viscosity oils, and low surface roughness. It is likely, therefore, that the contact will be boundary lubricated, in which case the surface shear stresses will not differ substantially from the dry case. Thus, a boundary-lubricated, normal-impact contact is essentially quasi-dry. Wear is, therefore, not likely to be reduced with lubrication of a normal contact.

In a lubricated compound-impact contact, due to the tendency for a full film to form, a superimposed sliding speed is conducive to less wear. This is in contrast to dry impact, where introducing sliding substantially increases wear. Lubrication tends to reduce sliding friction, so that even if a full film does not form, damage is still greatly reduced in comparison with the dry case.

In an experimental study, the sliding and impact speeds and the lubricant viscosity were varied to ascertain their effect on wear (Ref 29). The results are illustrated in Fig. 13. As can be seen, the highly viscous lubricant prevented all wear involving a sliding component. The less viscous lubricant allowed wear at high-impact velocities but prevented wear at low-impact viscosities. Results for the lighter lubricant indicate that normal impact is more damaging than compound impact.

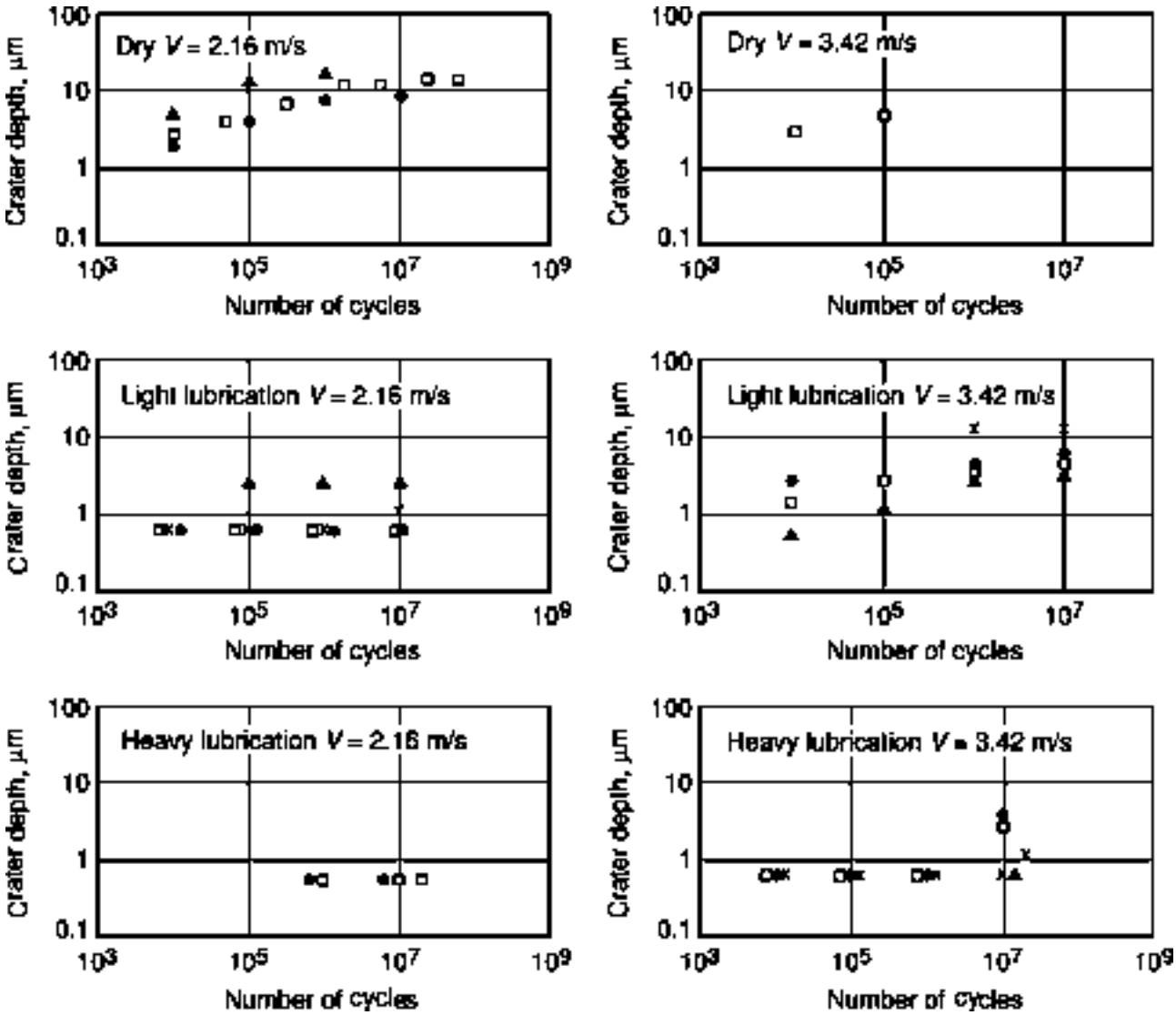


Fig. 13 Results of lubricated compound impact wear experiments (Ref 29). (*V* is impact velocity and *v* = sliding velocity; ×: *v* = 0 m/s; •: *v* = 0.25 m/s; □: *v* = 1.27 m/s; ▲: *v* = 3.81 m/s)

Wear Debris. The size and shape of wear debris particles arising from normal and compound wear varies, depending on the material pair and the nominal levels of stress and sliding velocity in the contact. Analysis of wear debris from a range of experiments carried out mainly with titanium alloy specimens impacting 17-4 PH stainless steel counterfaces at varying contact conditions (6.9 to 69 MPa impact stress and 0.01 to 10 m/s sliding velocity) (Ref 30) has shown that:

- Most observable wear particles fell in the range 0.3 to 0.7 mm.
- In general, the shape was platelike.
- A preferred crystallographic orientation exists, as would be expected from the sliding direction.

Observations were also made that indicate that:

- Debris stems from both specimen and counterface materials.
- Metallic particles consist of small and/or distorted primary crystals.
- Oxide formation occurs.
- Phase transformation of the metallic constituents may occur.

The occurrence of oxides in the wear debris indicates that oxygen plays a part in the wear process itself. It is probable that at least a portion of the oxides found in debris are formed while the particles are still attached to the substrate at the contact interface, although oxidation will also occur after the particles have been broken away from the wear surface.
Phase transformations can provide evidence of local temperatures and/or stress fields and/or interdiffusion of species from surrounding materials (Ref 30). In the tests outlined above it was found, for example, that the 17-4 PH stainless steel yielded debris containing γ-Fe (austenite). In order to transform the martensitic state of the 17-4 PH stainless steel, a temperature of at least 500 °C is required.

References cited in this section

2. A.W.L. De Gee, C.P.L. Commissaris, and J.H. Zaat, The Wear of Sintered Aluminium Powder (SAP) Under Conditions of Vibrational Contact, *Wear,* Vol 7, 1964, p 535–550

3. E. Rabinowicz and K. Hozaki, Impact Wear of Ductile Metals, *Proc. JSLE International Tribology Conf.,* Tokyo, 1985, p 263–268

4. S.L. Rice, A Review of Wear Mechanisms and Related Topics, *Proc. International Conf. on the Fundamentals of Tribology,* Massachusetts Institute of Technology Press, 1978, p 469–476

5. S.L. Rice, The Role of Microstructure in the Impact Wear of Two Aluminium Alloys, *Wear,* Vol 54 (No. 2), 1978, p 291–301

6. S.L. Rice and S.F. Wayne, Wear of Two Titanium Alloys Under Repetitive Compound Impact, *Wear,* Vol 61, 1980, p 69–76

7. S.L. Rice and R. Solecki, "Wear of Homogeneous and Composite Materials Under Conditions of Repeated Normal and Sliding Impact," AFOSR Tech Rep. 1241, 1979

8. S.L. Rice, H. Nowotny, and S.L. Wayne, Formation of Subsurface Zones in Impact Wear, *ASLE Trans.,* Vol 24 (No. 2), 1981, p 264–268

9. S.L. Rice, H. Nowotny, and S.L. Wayne, Characteristics of Metallic Subsurface Zones in Sliding and Impact Wear, Wear, Vol 74, 1981, p 131–142

10. P.A. Engel, *Impact Wear of Materials,* Elsevier, Amsterdam, 1976

11. G.W. Stachowiak and A.W. Batchelor, *Engineering Tribology,* Butterworth-Heinemann, 2001

12. R.S. Montgomery, The Mechanism of Percussive Wear of Tungsten Carbide Composites, *Wear,* Vol 12, 1968, p 309–329

13. E. Rabinowicz, Metal Transfer During Static Loading and Impacting, *Proc. Phys. Soc.,* Vol 65B, 1952, p 630–640

14. R.G. Bayer, P.A. Engel, and S.L. Sirico, Impact Wear Testing Machine, *Wear,* Vol 19, 1972, p 343–354

15. E.B. Iturbe, I.G. Greenfield, and T.W. Chow, Surface Layer Hardening of Polycrystalline Copper by Multiple Impact, *J. Mater. Sci.,* Vol 15, 1980, p 2331–2334

16. G. Laird, Repetitive- and Single-Blow Impact Testing of Wear Resistant Alloys, *J. Test. Eval.,* Vol 23 (No. 5), 1995, p 333–340

17. C.J. Studman and F.E. Field, A Repeated Impact Testing Machine, *Wear,* Vol 41, 1977, p 373–381

18. I.R. Sare, Repeated Impact-Abrasion of Ore-Crushing Hammers, *Wear,* Vol 87, 1983, p 202–225

19. P.A. Engel, H.C. Lee, and J.L. Zable, Dynamic Response of a Print Belt System, *IBM J. Res. Dev.,* Vol 23 (No. 4), 1979, p 403–410

20. S.L. Rice, S.F. Wayne, and H. Nowotny, Material Transport Phenomena in the Impact Wear of Titanium Alloys, *Wear,* Vol 61, 1980, p 215–226

21. S.L. Rice, H. Nowotny, and S.F. Wayne, The Role of Specimen Stiffness in Sliding and Impact Wear, *Wear,* Vol 77 (No. 1), 1982, p 13–28

22. Y. Yang, H. Fang, Y. Zheng, Z. Yang, and Z. Jiang, The Failure Models Induced by White Layers During Impact Wear, *Wear,* Vol 185, 1995, p 17–22

23. B. Zhang, Y. Liu, W. Shen, Y. Wang, X. Tang, and X. Wang, A Study on the Behaviour of Adiabatic Shear Bands in Impact Wear, *Wear,* Vol 198, 1996, p 287–292

24. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear,* Vol 44, 1977, p 1–16

25. N.P. Suh, The Delamination Theory of Wear, *Wear,* Vol 25, 1973, p 111–124

26. H. Nowotny, S.L. Rice, and S.F. Wayne, Impact Wear of a Tool Steel, *Wear,* Vol 81 (No. 1), 1982, p 175–181

27. N. Saka, A. Eleiche, and N.P. Suh, Wear of Metals at High Sliding Speeds, *Wear,* Vol 44, 1977, p 109–126

28. S.L. Rice, H. Nowotny, and S.F. Wayne, Specimen Counterface Bulk Hardness Effects in Impact Wear of 17-4 PH Steel Pairs, *Wear,* Vol 103 (No. 2), 1985, p 175–185

29. P.A. Engel and J.L. Siroco, Impact Wear Study of Lubricated Contacts, *Trans. ASLE,* Vol 18, 1975, p 279–289

30. H. Nowotny, S.L. Rice, and S.F. Wayne, Characteristics of Wear Debris in Impact Sliding, *Wear,* Vol 68, 1981, p 159–167

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Impact Wear of Ceramics

Material Removal Mechanisms. In both normal impact and compound impact with tangential contact involving ceramic bodies, mechanical energy is expended partly in frictional heating and in mechanical deformation that leads to wear (Ref 31). Brittle and plastic modes of material removal occur during impact. Brittle deformation is linked to the creation of pits (formed when material is adhesively *plucked* from the surface of one impacting body by the other), partial ring cracks, and delamination layers. The removed material is pulverized and compacted. At acute contact angles, where the largest slip occurs, the wear debris is also sheared, transferred, and redeposited.

Delamination is generally the dominant wear process. It causes extensive material removal and occurs at all contact angles. Figure 14 illustrates four stages of delamination (Ref 31). Stage 1 involves the fracturing of substrate material and crushing of wear debris. The depth of fracturing, which is less than the average grain dimension, is governed by the normal pressure applied. Eventually the layer of debris becomes thick enough to cushion and redistribute the normal stress in the substrate. The impact energy then acts to extrude the debris into interstices, and a thin, fine-grained film is created (stage 2). The interface between the film and the substrate is a zone of weakness in which cracks can nucleate (stage 3). Lateral propagation and intersection of the cracks results in the delamination of platelike debris.

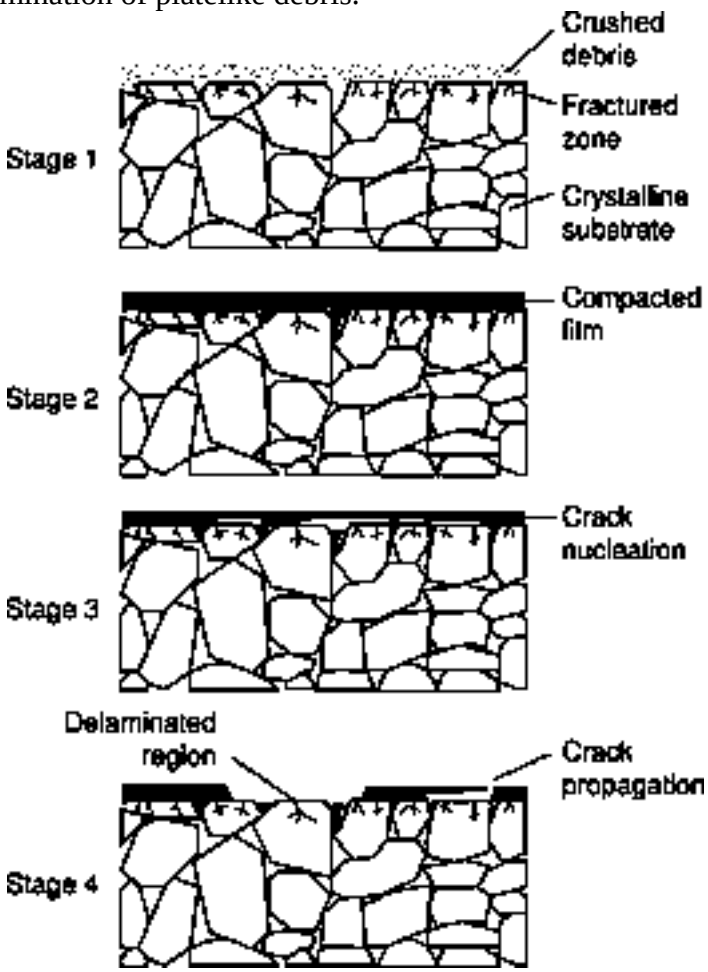


Fig. 14 Diagram of the stages of delamination caused by repeated impact on a ceramic surface. Stage 1 fracturing on the surface and crushing of debris; stage 2, extrusion of pulverized debris in interstices and

compaction of a fine grained film; stage 3, nucleation of cracks along the weak film/substrate interface; and stage 4, lateral propagation of interfacial cracks and delamination of the compacted film. Source: Ref 31

Influences on Wear. In repetitive impact of ceramics, two factors influence the magnitude of wear experienced. One is associated with the normal stress that is partly accommodated in the brittle fracture of the parent materials. The effect of normal stress on wear increases as the angle of contact increases. The second factor is related to the tangential stress and accelerates delamination of the protective surface film. The influence of shear decreases as the angle (from 45 to 90°) of contact increases. A model illustrating the dependency of impact wear to angle of contact is shown in Fig. 15 (Ref 31). The model shows the combined effect of the normal and tangential stresses are minimized between these angles.

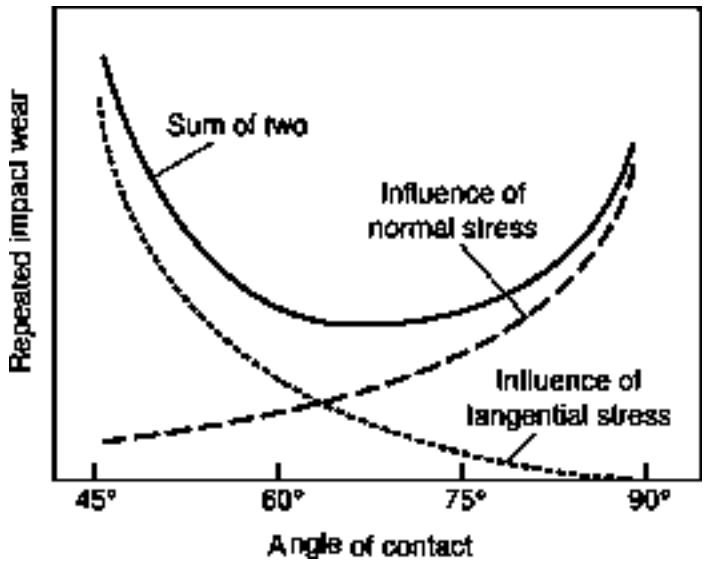


Fig. 15 Schematic illustration of the dependency of the repeated impact of a ceramic object on the angle of contact. Source: Ref 31

Relative humidity has a pronounced effect on the wear mechanism and material removal rate during the repetitive impact of ceramics (see Fig. 16) (Ref 32). In dry conditions, as described in the previous section, extensive fracture and growth of cracks is promoted along grain boundaries. In humid conditions, however, fracture only happens on a small scale due to the occurrence of crack-arrest mechanisms, and, consequently, wear rates are reduced.

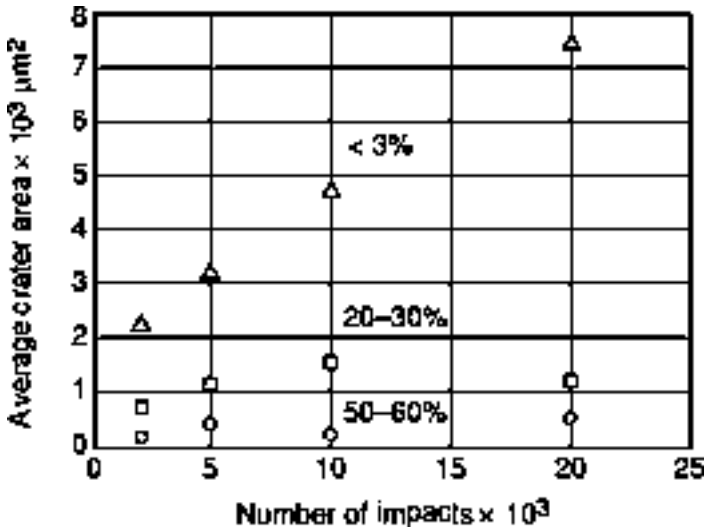


Fig. 16 Average crater area vs. number of impacts for SN220M silicon nitride counterfaces impacted with a SN220M silicon nitride ball at varying humidities. Source: Ref 32

The machined surface finish has an influence on wear during early cycles of impact, as shown in Fig. 17 (Ref 33). Wear rates of ground bodies are higher than polished bodies because the latter is protected by a film of oxidized wear debris. The wear behavior of a ground surface is characterized by the lateral extrusion of elongated ridges and then microfracture of extruded slivers. At higher numbers of impacts, the wear rates of polished and ground specimens are almost the same and are characterized by the mechanism outlined in Fig. 14.

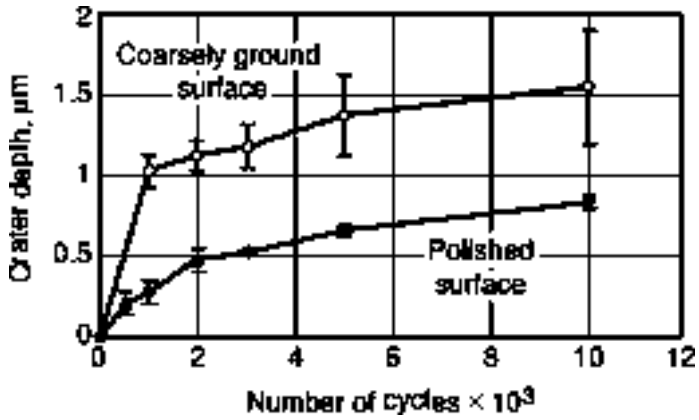


Fig. 17 Crater depth vs. number of impacts for GS-44 silicon nitride counterfaces of varying surface finish impacted with a NBD-200 silicon nitride ball. Source: Ref 33

Machining residual stresses also affect wear. The results shown in Fig. 18 (Ref 34) indicate that impact wear decreases with increasing compressive residual stress. It has been suggested that this reduction in wear at high residual stresses results from crack growth being impeded (Ref 33).

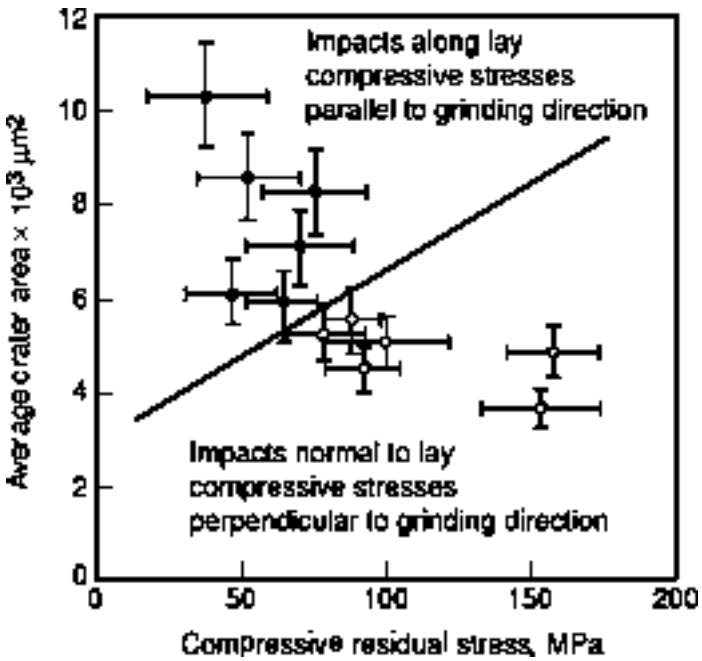


Fig. 18 Impact crater area vs. compressive residual stress for impacts normal to the lay and compressive stresses perpendicular to the grinding direction (○), and for impacts perpendicular to the lay and compressive stresses parallel to the grinding direction (●), using sintered reaction-bonded silicon nitride (SRBSN) counterfaces and NBD-200 silicon nitride balls. Source: Ref 34

It can also be seen that grinding results in compressive residual stresses that are typically larger in magnitude in the direction perpendicular to the grinding direction than in the direction of grinding.

Impact Failure of Ceramic Coatings. Hard coatings are becoming widely used to increase wear resistance of machine parts subjected to repetitive normal impacts or tools used in cutting or cold forming. The thin layers of coating can be deposited on any steel component using sputtering techniques, allowing economy in the use of expensive materials. Three different failure zones are apparent in ceramic coatings, such as TiN and TiC (Ref

35), as shown in Fig. 19, which illustrates a typical indentation formed in a ceramic coating subjected to repetitive impact by a hardened steel ball. A central zone of cohesive failure exists in which cracking and coating delamination occurs. An intermediate zone is found between the central and peripheral regions on the internal side of the rim formed around the indentation. The intermediate zone exhibits adhesive-cohesive failure of the coating under the buildup of shear stresses, which arise from plastic strain of the substrate during impact. In hard brittle coatings, the high tensile stress in the immediate vicinity of the impact causes the development of a large number of cracks in the peripheral region of the impact.

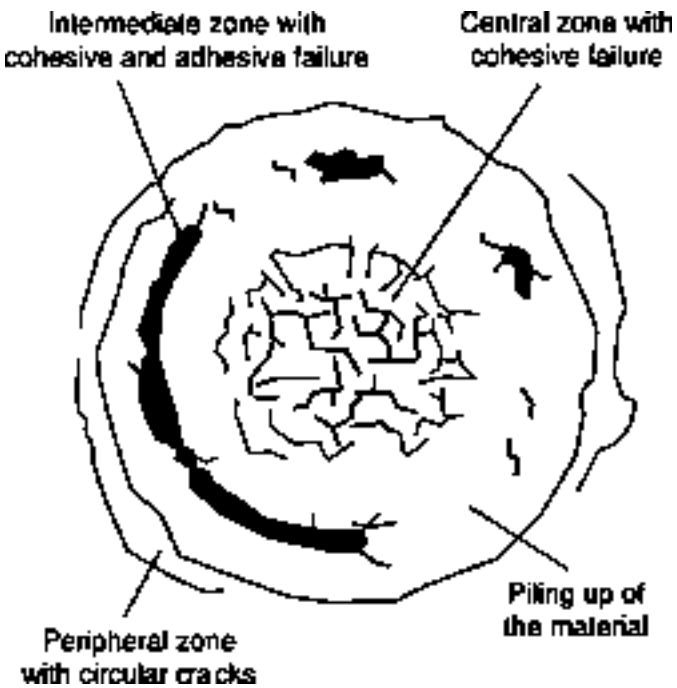


Fig. 19 Schematic diagram of a typical indentation formed in a ceramic coating under repetitive impact loading showing different failure zones. Source: Ref 35

A good correlation exists between increasing failure and the number of impacts. Increasing the number of impacts not only increases the number of cracks but also increases the wear at each crack before the coating eventually starts to delaminate or spall. Cracks arise when a hard coating is not tough or ductile enough to accommodate the stress of deformation under an impact.

References cited in this section

31. E.S. Zanoria and P.J. Blau, Effect of Incidence Angle on the Impact-Wear Behaviour of Silicon Nitride, *J. Am. Ceram. Soc.,* Vol 81 (No. 4), 1998, p 901–909

32. S.R. Srinivasan and P.J. Blau, Effect of Relative Humidity on Repetitive Impact Behaviour of Machined Silicon Nitride, *J. Am. Ceram. Soc.,* Vol 77 (No. 3), 1994, p 683–688

33. E.S. Zanoria and P.J. Blau, Effects of Machined Surface Condition on the Repeated Impact Behaviour of Silicon Nitride, *Wear,* Vol 218, 1998, p 66–77

34. S.R. Srinivasan, P.J. Blau, and J.L. Bjerke, Effect of Machining Residual Stresses on the Repetitive Impact Behaviour of Silicon Nitride, *J. Mater. Res.,* Vol 10 (No. 1), 1995, p 95–100

35. R. Bantle and A. Matthews, Investigation into the Impact Wear Behaviour of Ceramic Coatings, *Surf. Coat. Technol.,* Vol 74–75, 1995, p 857–868

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Impact Wear of Polymers

Two modes of impact wear or damage of polymers have been observed (Ref 36). The first mode relates primarily to the plastic flow or deformation of the material without material loss. A second mode appears as the wear rate increases, and this is associated with material loss. In the material loss mode, cracks appear in the surface.

For fluoropolymers and acetal resins tested against 17-4 PH stainless steel counterfaces, wear rates for normal impacts at low stress were very low (Ref 37). Compound impact wear rates were higher and increased with sliding speed. Wear occurred by a mechanism of delamination. Wear debris in the form of platelets is particularly apparent in cases of compound impact. Increasing sliding speed tends to accelerate the process of surface and subsurface crack initiation, and increasing numbers of impact cycles tend to refine the microcrack patterns.

An acetal resin is a relatively crystalline polymer and, therefore, the conventional delamination theory based on the accumulation of dislocations leading to the nucleation of subsurface cracks (Ref 24) may apply. Fluoropolymers, however, are relatively amorphous and therefore an alternative to the theory of delamination outlined previously is required. Part of the answer may be related to microvoids observed in polyimide films (Ref 38) as well as water clusters found to be introduced during processing of polymers (Ref 39). Both, if present, could act as sites for the initiation of microcracks.

For impacts at higher stresses, several critical phenomena have been identified, such as charring and chemical degradation and the existence of a stress threshold beyond which catastrophic wear occurs (Ref 40).

References cited in this section

24. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear,* Vol 44, 1977, p 1–16

36. R.G. Bayer, Impact Wear of Elastomers, *Wear,* Vol 112, 1986, p 105–120

37. S.L. Rice, Reciprocating Impact Wear Testing Apparatus, *Wear,* Vol 45, 1977, p 85–95

38. R.G. Bayer and E. Sacher, The Effect of Voids on the Sliding Friction of Polyimide Film, *Wear,* Vol 37, 1976, p 15–20

39. G.E. Johnson, H.E. Bair, E.W. Anderson, and J.H. Daane, The Nature of Water in Polyethylene and its Relationship to Dielectric Loss, *Proc. National Academy of Sciences National Research Council Conf. On Electrical Insulation and Dielectric Phenomena,* 1976

40. R.G. Bayer, P.A. Engel, and E. Sacher, Impact Wear Phenomena in Thin Polymer Films, *Wear,* Vol 32, 1975, p 181–194

Impact Wear Failures

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Impact Wear Testing

Two different types of test rigs have been developed for impact wear studies: ballistic impact testers and pivotal hammer testers (Ref 10). The following describes the features of each.

Ballistic Impact Wear Apparatus. In ballistic impacter wear testers, repetitive impacts are achieved by firing projectiles at a counterface. To ensure the projectiles impact at the same point each time, guides are provided. Some test rigs use one projectile repeatedly impacting against a counterface (Ref 2, 14), while others use a number of projectiles (usually balls) that are recycled (see Fig. 20) (Ref 16, 22). Those using one projectile can be used for one-body wear investigations on the projectile with a hard, negligibly wearing counterface. By proper synchronization of projectile firing with rotational speed of the counterface to ensure impacts occur at the same point, two-body wear can also be studied. In those test rigs using a number of projectiles, clearly only the wear of the counterface can be investigated. Ballistic test rigs have two main disadvantages. The first is that, because of cleaning problems with the projectile guides, lubricated tests are difficult to carry out. The second is that it is sometimes difficult to retune the rigs to take different weight projectiles.

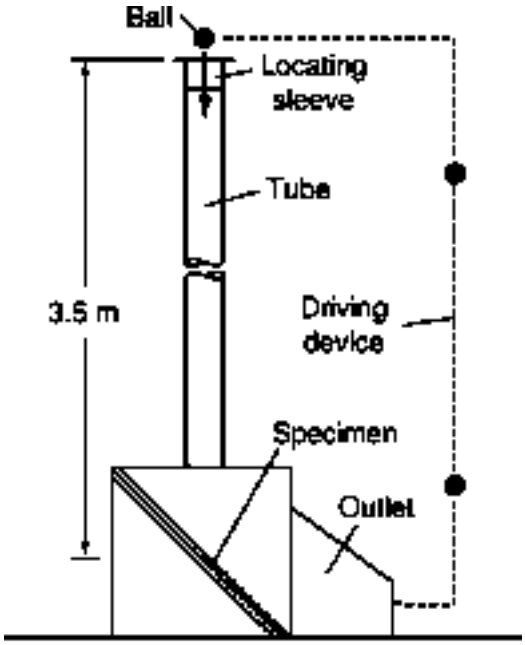


Fig. 20 Ballistic impact wear tester. Source: Ref 22

Pivotal Hammer Impact Wear Apparatus. In pivotal hammer impact wear test rigs, a hammer is made to impact a counterface. A stationary counterface is used for normal impact studies, while for compound studies, shear is imparted by using a rotating counterface. Altering the tilt of the counterface changes the angle of contact. Again, synchronization is required for rotating counterfaces to ensure impact occurs at the same point during each cycle. A number of pivotal hammer test rigs are in existence (Ref 40, 41, 42). A typical example is shown in Fig. 21. Various mechanisms are employed to move the hammer up and down, including electromagnets and cams. Pivotal hammer test rigs overcome the contact lubrication and retuning problems encountered with ballistic rigs. However, there are problems associated with the flexure arising in the hammers, and the elastic vibrational modes of the hammers may complicate analysis. Alignment problems are also more likely to arise with pivotal hammers than in ballistic rigs. This can mean that small tangential oscillations are induced between hammer and counterface, leading to surface damage by shear tractions.

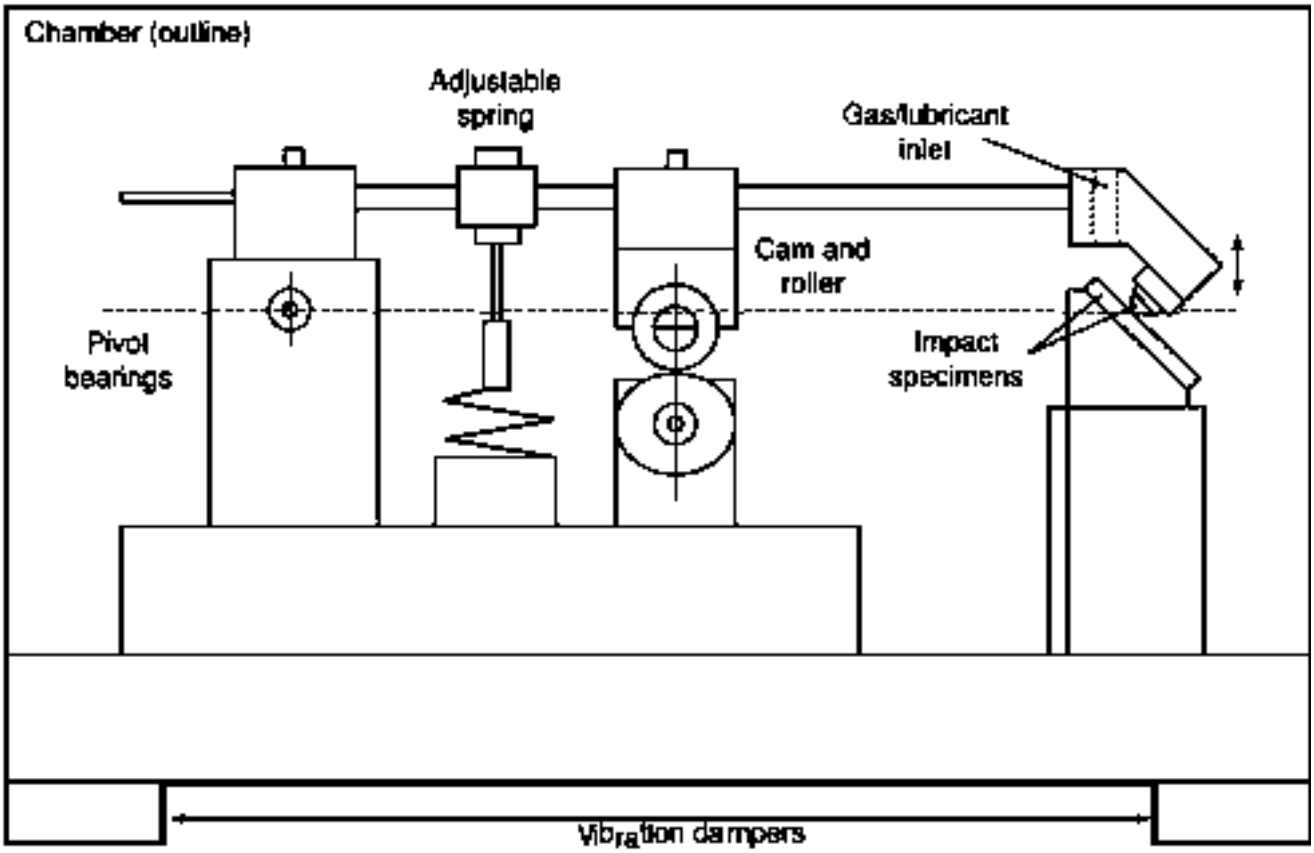


Fig. 21 Pivotal hammer impact tester. Source: Ref 41

References cited in this section

2. A.W.L. De Gee, C.P.L. Commissaris, and J.H. Zaat, The Wear of Sintered Aluminium Powder (SAP) Under Conditions of Vibrational Contact, *Wear,* Vol 7, 1964, p 535–550

10. P.A. Engel, *Impact Wear of Materials,* Elsevier, Amsterdam, 1976

14. R.G. Bayer, P.A. Engel, and S.L. Sirico, Impact Wear Testing Machine, *Wear,* Vol 19, 1972, p 343–354

16. G. Laird, Repetitive- and Single-Blow Impact Testing of Wear Resistant Alloys, *J. Test. Eval.,* Vol 23 (No. 5), 1995, p 333–340

22. Y. Yang, H. Fang, Y. Zheng, Z. Yang, and Z. Jiang, The Failure Models Induced by White Layers During Impact Wear, *Wear,* Vol 185, 1995, p 17–22

40. R.G. Bayer, P.A. Engel, and E. Sacher, Impact Wear Phenomena in Thin Polymer Films, *Wear,* Vol 32, 1975, p 181–194

41. P.J. Blau and T.A. Hanft, Quantitative Measurement of Repetitive Impact Damage on Ground Silicon Nitride Surfaces, *Tribol. Int.,* Vol 27 (No. 2), 1994, p 109–118

42. N.J. Mahoney, R.J. Grieve, and T. Ellis, A Simple Experimental Method for Studying the Impact Wear of Metals, *Wear,* Vol 98 (No. 1–3), 1984, p 79–87

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Impact Wear Modelling

An impact-wear modelling approach has been established based on two stages of wear progression (Ref 10). The first is an *induction period* (see Fig. 22), during which deformation occurs and a wear scar is formed, but there is no measurable material loss. The end of this stage is defined as the *zero wear limit.* This constitutes the initial point of the *measurable wear* region. As indicated in Fig. 22, the progression of wear in this region can take a number of forms. These depend on the mechanism of wear, which is related to the impacting material and the severity of the contact conditions.

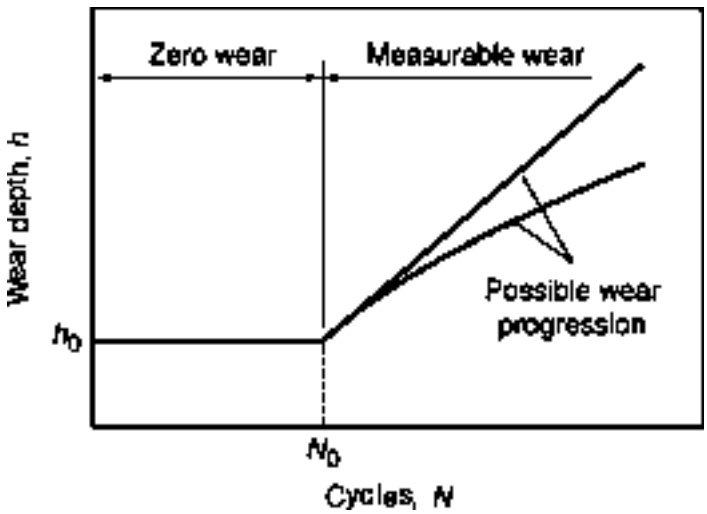


Fig. 22 Zero and measurable wear. Source: Ref 10

Most mechanical components continue to be functional beyond the zero wear limit, and their usefulness is normally connected with the loss of a specific depth of material. For that reason, the measurable wear process is probably of more interest to an analyst attempting to define a life prediction for a repetitively impacted component.

Zero Impact Wear Model. The following equation has been developed to establish a quantitative, semiempirical evaluation of the *zero wear limit* for specimens subjected to repetitive compound impacts in the elastic stress region (Ref 10). It considers wear damage to be contributed from the surface shear stress, τ_2, and the maximum subsurface shear stress, τ_1, during each impact.

$$N_0 = \frac{1}{1+\beta}\left(\frac{\gamma\sigma_y}{\sigma}\right)^9 2000 \qquad \text{(Eq 1)}$$

where N_0 is the number of cycles at which the zero limit is reached, σ_y is the uniaxial tensile yield stress, σ is the peak Hertzian pressure, and γ is a material wear factor (experiments have shown that γ is approximately 1.1 for carbon and tool steel specimens) (Ref 43). The quantity β is the ratio of the surface damage contribution, D_1, to the subsurface damage contribution, D_2, where:

$$D_1 = \int_0^{t^*} \tau_1^9(t)dt \qquad \text{(Eq 2)}$$

$$D_2 = \int_0^{\bar{t}} \tau_2^9(t)dt \qquad \text{(Eq 3)}$$

$$\beta = \frac{D_1}{D_2} \qquad \text{(Eq 4)}$$

where t^* is the duration of the impact force pulse. The integral in Eq 3 is evaluated between $t = 0$ and $t = \bar{t}$, where $\bar{t}$ is the *slipping time.* This signifies that in compound impact with sliding, wear is only occurring in the initial phase of contact, while slipping is taking place between the specimen and the counterface.

Measurable Impact Wear Models. Approaches for modelling *measurable* impact wear include a linear model based on an Archard-type wear law (Ref 3). Starting with the relationship for wear volume, *W*:

$$W = \frac{kPx}{H} \qquad \text{(Eq 5)}$$

where *P* is the load (N), *x* is the sliding distance (m), *H* is the hardness, and *k* is a nondimensional wear coefficient. Load can be replaced by friction force divided by the friction coefficient, μ. Friction force multiplied by the sliding distance represents the energy dissipated in sliding. If α is defined as the proportion of the total impact energy expended in interfacial slip and the term (α*k*/μ) is replaced by a parameter *K*, and if *I* is the energy of impact and *n* the number of impacts, wear volume can be defined as:

$$W = \frac{KIn}{H} \qquad \text{(Eq 6)}$$

where *K* is now the nondimensional impact wear coefficient.

As stated above, *K* is related to *k* by the following expression:

$$K = \frac{k\alpha}{\mu} \qquad \text{(Eq 7)}$$

The nonelastic part of an impact has been estimated as being 70% (based on rebound measurements of a striking hammer) (Ref 3), and only a small amount of this would be expended in shear at the interface, so α would be small, for example, 10%. The friction coefficient, μ, for an unlubricated contact generally falls in the range 0.5 to 1, it would be expected that *K* would be about an order of magnitude smaller than the sliding wear coefficient, *k*. For unlubricated metals, *k* is generally in the range 5×10^{-3} to 3×10^{-5}, while *K* values between 5×10^{-4} and 1×10^{-6} would be anticipated. The majority of the values for *K* given in Table 1, determined from a range of experiments carried out with different material and contact conditions, fall within this range.

Table 1 Impact wear coefficient values

Reference	Surface measured	Impact energy, kgf · mm	No. of impacts	Lubrication	Impact wear coefficient, *K*
Ref 13	Copper (hard)	2	1	N	53×10^{-6}
	Steel	20	1	N	67×10^{-6}
	Steel	20	1	Y	10×10^{-6}
Ref 12	Tungsten	138	2.4×10^{3}	N	130×10^{-6}
Ref 44	Copper	60	5×10^{5}	N	20×10^{-6}
	Soft steel	120	5×10^{5}	N	19×10^{-6}
	Aluminium alloy	60	5×10^{5}	N	42×10^{-6}
	Titanium alloy	120	5×10^{5}	N	15×10^{-6}
	Hard steel	160	5×10^{5}	N	28×10^{-6}
	Hard steel	700	4×10^{3}	N	43×10^{-6}
	Hard steel	42	2×10^{4}	N	104×10^{-6}
Ref 29	Hard steel	0.36	6.5×10^{4}	N	14×10^{-6}
	Hard steel	0.36	4.3×10^{5}	Y	1.2×10^{-6}
	Soft steel	0.36	1×10^{4}	N	247×10^{-6}
	Soft steel	0.36	3.2×10^{6}	Y	0.8×10^{-6}

Source: Ref 3

A relationship of the same form as that used in erosion studies to model wear mass, *W* (Ref 44), which gives a declining wear rate, has also been used successfully to predict wear in percussive impact of large bodies (Ref 45, 46):

$$W = KNe^n \qquad \text{(Eq 8)}$$

where *e* is the impact energy per cycle (J), *N* is the number of cycles, and *K* and *n* are empirically determined wear constants, and:

$$e = \frac{1}{2}mv^2 \qquad \text{(Eq 9)}$$

where *m* is mass in kg and *v* is the velocity at impact in m/s.

References cited in this section

3. E. Rabinowicz and K. Hozaki, Impact Wear of Ductile Metals, *Proc. JSLE International Tribology Conf.,* Tokyo, 1985, p 263–268

10. P.A. Engel, *Impact Wear of Materials,* Elsevier, Amsterdam, 1976

12. R.S. Montgomery, The Mechanism of Percussive Wear of Tungsten Carbide Composites, *Wear,* Vol 12, 1968, p 309–329

13. E. Rabinowicz, Metal Transfer During Static Loading and Impacting, *Proc. Phys. Soc.,* Vol 65B, 1952, p 630–640

29. P.A. Engel and J.L. Siroco, Impact Wear Study of Lubricated Contacts, *Trans. ASLE,* Vol 18, 1975, p 279–289

43. P.A. Engel, Percussive Impact Wear, *Tribol. Int.,* Vol 11, 1978, p 169–176

44. K. Wellinger and H. Breckel, Kenngrossen und Verscheiss Beim Stoss Metallischer Werkstoffe (Deformation and Wear in the Impact of Metallic Workpieces), *Wear,* Vol 13, 1969, p 257–281

45. R.W. Fricke and C. Allen, Repetitive Impact-Wear of Steels, *Wear,* Vol 163, 1993, p 837–847

46. R. Lewis and R.S. Dwyer-Joyce, "Design Tools for Predicting Inlet Valve Recession and Solving Valve Failure Problems," SAE Paper 2001-01-1987, presented at the SAE International Spring Fuel and Lubricants meeting (Orlando, FL), Society of Automotive Engineers, 7–9 May 2001, in SP-1624, 2001, p 105–114

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

Impact Wear Failure Case Study: Automotive Engine Inlet Valve and Seat Wear

The drive for reduced oil consumption and exhaust emissions has led to a reduction in the amount of lubricant present in the air stream in automotive engines. This, combined with the effort to lengthen service intervals and increase engine performance, has led to an increase in the wear of inlet valves and seat inserts.

Engines are typically designed to tolerate a certain amount of valve recession (defined in Fig. 23). After this has been exceeded, the gap between the valve tip and the follower must be adjusted to ensure that the valve

continues to seat correctly. If the valve is not able to seat, cylinder pressure will be lost and the hot combustion gases that leak will cause valve guttering or torching to occur, which will rapidly lead to valve failure.

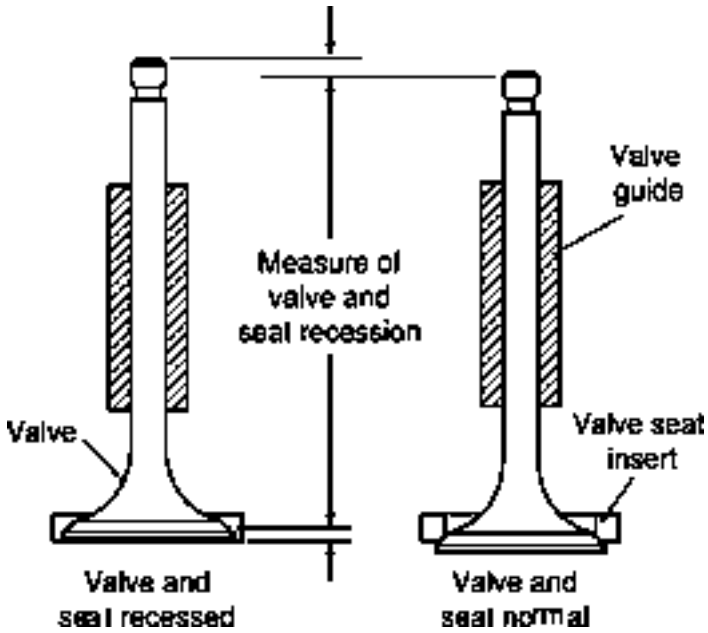


Fig. 23 Valve recession

Work has been carried out to isolate the fundamental mechanisms of automotive engine inlet valve and seat wear. This work used experimental apparatus designed to simulate the loading environment and contact conditions to which a valve and seat insert are subjected (Ref 47, 48, 49). The results have shown that recession originates from two processes; impact of the valve on the seat on valve closure and sliding of the valve against the seat as the valve head deflects and wedges into the seat as combustion occurs in the cylinder. These processes produce characteristic wear features on valves and seats. Impact on valve closure leads to the formation of a series of ridges and valleys on valve seating faces (see Fig. 24) and surface cracking on seat seating faces (see Fig. 25). While both mechanisms are significant, impact on valve closure is the main contributor to valve recession (Ref 49).

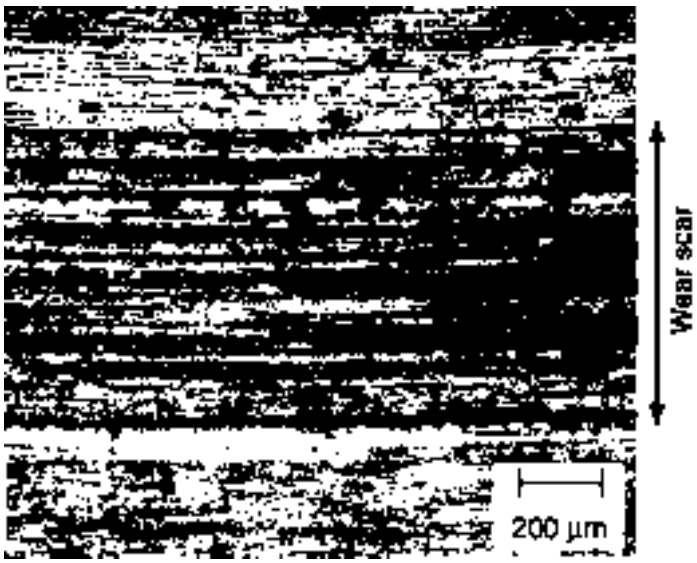


Fig. 24 Valve seating face showing a series of ridges and valleys formed circumferentially around the axis of the valve

Fig. 25 Seat insert seating face showing surface cracking and subsequent material spalling

Impact wear resistance has been related to material properties, as shown in Fig. 26, which plots the results of bench tests isolating impact of the valve on the seat, where the tougher cast seat insert material has worn less than the sintered seat insert material (both materials are based on a tool steel matrix). This is further exemplified in Fig. 27, where results of bench tests run on a range of materials are shown. The tougher materials gave lower valve recession.

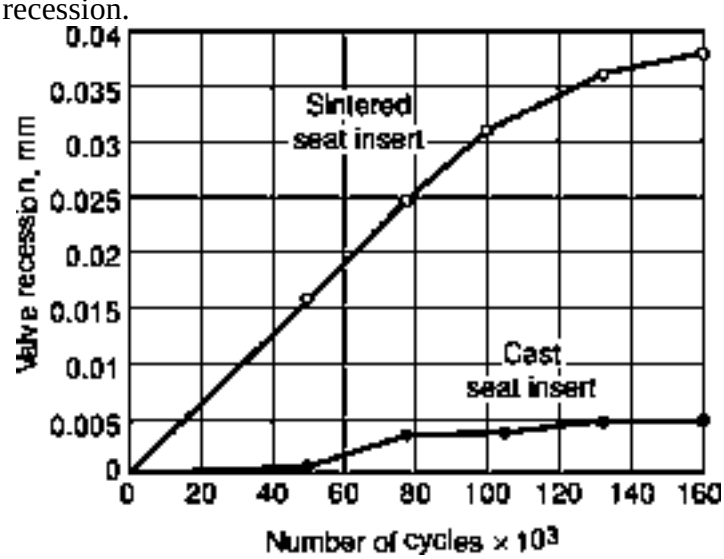


Fig. 26 Valve recession as a result of bench tests carried out isolating impact using a cast and a sintered seat insert material (both based on a tool steel matrix)

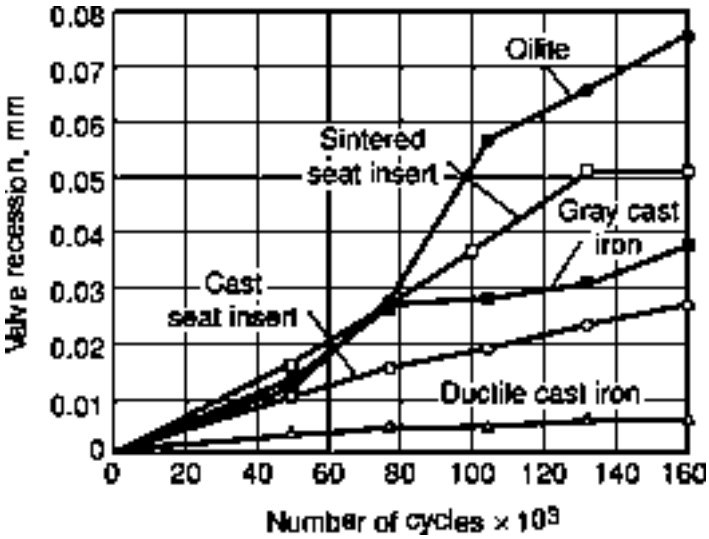


Fig. 27 Valve recession as a result of bench tests carried out isolating impact using a number of different seat materials

As shown in Fig. 28, recession due to impact wear is roughly proportional to valve closing velocity squared. This implies that wear is related to the energy of the valve on closing.

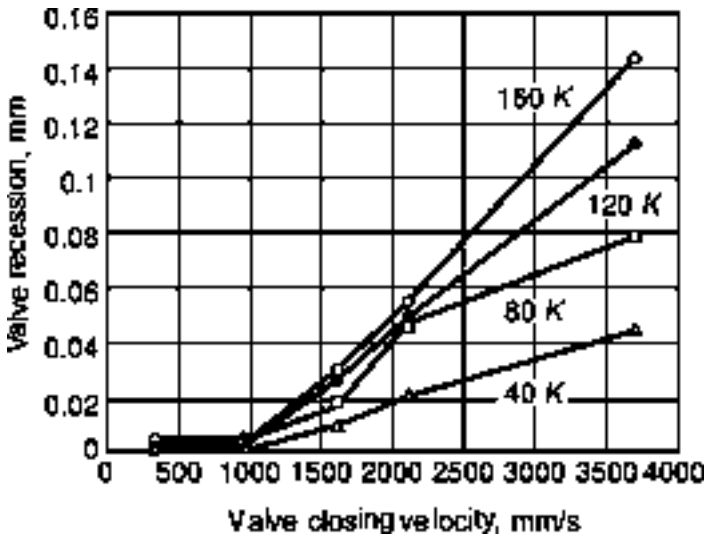


Fig. 28 Valve recession vs. valve closing velocity for a series of bench tests carried out isolating impact using a cast tool steel based seat insert material

As a result, a relationship of the same form as that used in erosion studies to model wear mass, *W*, (Ref 44) (based on impact energy) (Eq 8), was used to model wear due to the impact of the valve on the seat during valve closure (Ref 46). In this case, the impact energy (Eq 9) was determined using the mass of valve and follower added to half the mass of valve spring and the valve velocity at impact.

Values of *K* and *n* were derived using an iterative process to fit Eq 8 to experimental data. At each data point, the velocity was recalculated to take account of the change in lift due to recession. The results of this process for tests run with a cast seat insert (based on a tool steel matrix) are shown in Fig. 29. As can be seen, the model produces good correlation over a range of velocities.

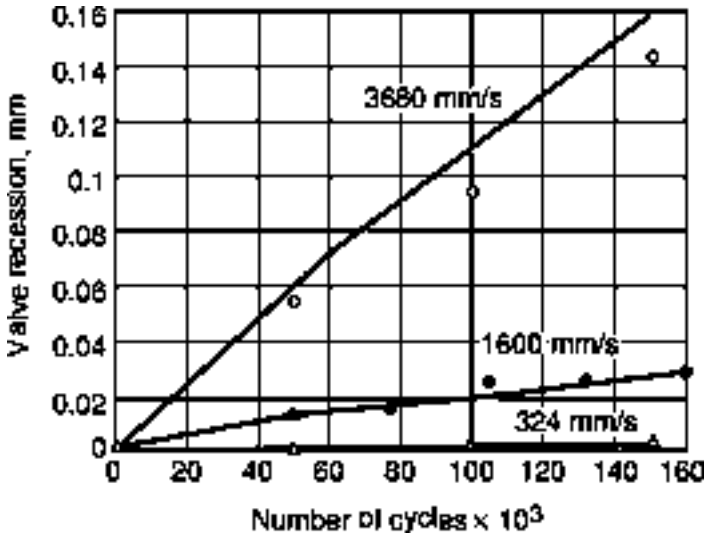


Fig. 29 Model predictions for valve recession caused by impact wear. Line represents model prediction; data points indicate bench test results.

The equation for impact wear was combined with an equation for sliding wear (based on an Archard type wear law, Ref 50) to give an overall wear model. In order to incorporate the change in pressure at the interface and any other effects likely to lead to a reduction in the wear rate with time, such as work hardening, a term consisting of the ratio of the initial valve/seat contact area, A_i, to the contact area after N cycles, A, to the power of a constant, j, was included. A value of 10 for j was determined empirically using bench and engine test data.

$$V = \left(\frac{k\bar{P}N\delta}{H} + KNe^{n} \right) \left(\frac{A_i}{A} \right)^{j} \qquad \text{(Eq 10)}$$

where $\overline{P}$ is the average load at the valve/seat interface during combustion in the cylinder, N, k is a sliding wear coefficient, δ is the slip at the valve/seat interface, and H is the hardness of the softer material.
Equation 10 gives a wear volume, which is then converted to a recession value, r, using equations derived from the seat geometry. Equation 11 gives r in terms of V for the case where valve and seat angles are equal:

$$r = \left(\sqrt{\frac{V}{\pi R_i \cos\theta_s \sin\theta_s} + w_i^2} - w_i \right) \sin\theta_s \qquad \text{(Eq 11)}$$

where R_i is the initial seat insert radius, θ_s is the seat insert seating face angle, and w_i is the initial seat insert seating face width (as measured). R_i can be calculated using w_i and the radius and seating face width as specified for the seat insert (R_d and w_d).
Results of model predictions for engine tests carried out using tool steel based cast and sintered seat inserts are shown in Fig. 30. As shown in Fig. 31, the contribution due to impact, as predicted by the model, is far greater than that of sliding.

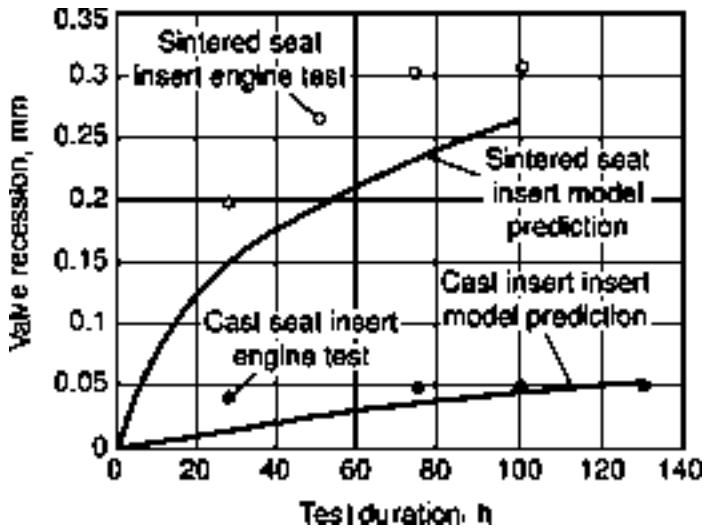


Fig. 30 Valve recession predictions using the combined impact-sliding model

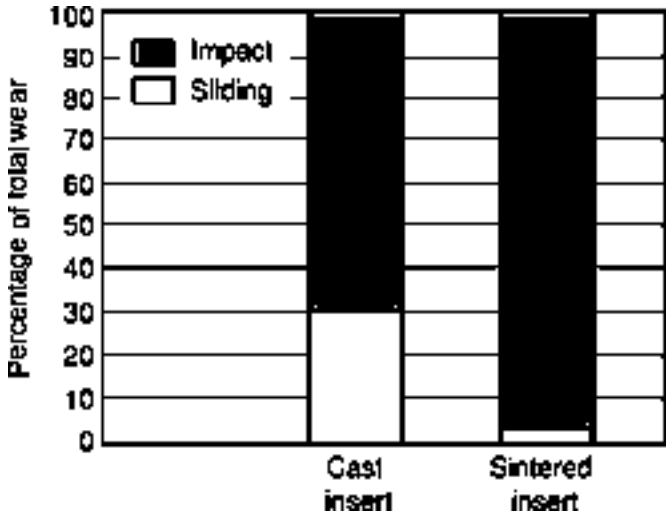


Fig. 31 Percentage of total valve recession caused by impact and sliding

References cited in this section

44. K. Wellinger and H. Breckel, Kenngrossen und Verscheiss Beim Stoss Metallischer Werkstoffe (Deformation and Wear in the Impact of Metallic Workpieces), *Wear*, Vol 13, 1969, p 257–281

46. R. Lewis and R.S. Dwyer-Joyce, "Design Tools for Predicting Inlet Valve Recession and Solving Valve Failure Problems," SAE Paper 2001-01-1987, presented at the SAE International Spring Fuel and Lubricants meeting (Orlando, FL), Society of Automotive Engineers, 7–9 May 2001, in SP-1624, 2001, p 105–114

47. R. Lewis, R.S. Dwyer-Joyce, and G. Josey, Investigation of Wear Mechanisms Occurring in Passenger Car Diesel Engine Inlet Valves and Seat Inserts, *J. Fuels Lubric., Transactions of the SAE 1999*, 2000, p 610–618

48. R. Lewis, R.S. Dwyer-Joyce, and G. Josey, Design and Development of a Bench Test-Rig for Investigating Diesel Engine Inlet Valve and Seat Wear, *Trans. Mech. Eng., Institution of Engineers Australia*, Vol ME24 (No. 1), 2000, p 39–46

49. R. Lewis and R.S. Dwyer-Joyce, An Experimental Approach to Solving Combustion Engine Valve and Seat Wear Problems, *Proc. of the 27th Leeds-Lyon Symposium on Tribology*, Elsevier, Tribology Series No. 39, 2001, p 629–640

50. J.F. Archard, Contact and Rubbing of Flat Surfaces, *J. Appl. Phys.*, Vol 24 (No. 8), 1953, p 981–988

Impact Wear Failures

Roger Lewis and Rob Dwyer-Joyce, The University of Sheffield

References

1. P.A. Engel, Impact Wear, *Friction, Lubrication and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 263–270

2. A.W.L. De Gee, C.P.L. Commissaris, and J.H. Zaat, The Wear of Sintered Aluminium Powder (SAP) Under Conditions of Vibrational Contact, *Wear,* Vol 7, 1964, p 535–550

3. E. Rabinowicz and K. Hozaki, Impact Wear of Ductile Metals, *Proc. JSLE International Tribology Conf.,* Tokyo, 1985, p 263–268

4. S.L. Rice, A Review of Wear Mechanisms and Related Topics, *Proc. International Conf. on the Fundamentals of Tribology,* Massachusetts Institute of Technology Press, 1978, p 469–476

5. S.L. Rice, The Role of Microstructure in the Impact Wear of Two Aluminium Alloys, *Wear,* Vol 54 (No. 2), 1978, p 291–301

6. S.L. Rice and S.F. Wayne, Wear of Two Titanium Alloys Under Repetitive Compound Impact, *Wear,* Vol 61, 1980, p 69–76

7. S.L. Rice and R. Solecki, "Wear of Homogeneous and Composite Materials Under Conditions of Repeated Normal and Sliding Impact," AFOSR Tech Rep. 1241, 1979

8. S.L. Rice, H. Nowotny, and S.L. Wayne, Formation of Subsurface Zones in Impact Wear, *ASLE Trans.,* Vol 24 (No. 2), 1981, p 264–268

9. S.L. Rice, H. Nowotny, and S.L. Wayne, Characteristics of Metallic Subsurface Zones in Sliding and Impact Wear, Wear, Vol 74, 1981, p 131–142

10. P.A. Engel, *Impact Wear of Materials,* Elsevier, Amsterdam, 1976

11. G.W. Stachowiak and A.W. Batchelor, *Engineering Tribology,* Butterworth-Heinemann, 2001

12. R.S. Montgomery, The Mechanism of Percussive Wear of Tungsten Carbide Composites, *Wear,* Vol 12, 1968, p 309–329

13. E. Rabinowicz, Metal Transfer During Static Loading and Impacting, *Proc. Phys. Soc.,* Vol 65B, 1952, p 630–640

14. R.G. Bayer, P.A. Engel, and S.L. Sirico, Impact Wear Testing Machine, *Wear,* Vol 19, 1972, p 343–354

15. E.B. Iturbe, I.G. Greenfield, and T.W. Chow, Surface Layer Hardening of Polycrystalline Copper by Multiple Impact, *J. Mater. Sci.,* Vol 15, 1980, p 2331–2334

16. G. Laird, Repetitive- and Single-Blow Impact Testing of Wear Resistant Alloys, *J. Test. Eval.,* Vol 23 (No. 5), 1995, p 333–340

17. C.J. Studman and F.E. Field, A Repeated Impact Testing Machine, *Wear,* Vol 41, 1977, p 373–381

18. I.R. Sare, Repeated Impact-Abrasion of Ore-Crushing Hammers, *Wear,* Vol 87, 1983, p 202–225

19. P.A. Engel, H.C. Lee, and J.L. Zable, Dynamic Response of a Print Belt System, *IBM J. Res. Dev.,* Vol 23 (No. 4), 1979, p 403–410

20. S.L. Rice, S.F. Wayne, and H. Nowotny, Material Transport Phenomena in the Impact Wear of Titanium Alloys, *Wear,* Vol 61, 1980, p 215–226

21. S.L. Rice, H. Nowotny, and S.F. Wayne, The Role of Specimen Stiffness in Sliding and Impact Wear, *Wear,* Vol 77 (No. 1), 1982, p 13–28

22. Y. Yang, H. Fang, Y. Zheng, Z. Yang, and Z. Jiang, The Failure Models Induced by White Layers During Impact Wear, *Wear,* Vol 185, 1995, p 17–22

23. B. Zhang, Y. Liu, W. Shen, Y. Wang, X. Tang, and X. Wang, A Study on the Behaviour of Adiabatic Shear Bands in Impact Wear, *Wear,* Vol 198, 1996, p 287–292

24. N.P. Suh, An Overview of the Delamination Theory of Wear, *Wear,* Vol 44, 1977, p 1–16

25. N.P. Suh, The Delamination Theory of Wear, *Wear,* Vol 25, 1973, p 111–124

26. H. Nowotny, S.L. Rice, and S.F. Wayne, Impact Wear of a Tool Steel, *Wear,* Vol 81 (No. 1), 1982, p 175–181

27. N. Saka, A. Eleiche, and N.P. Suh, Wear of Metals at High Sliding Speeds, *Wear,* Vol 44, 1977, p 109–126

28. S.L. Rice, H. Nowotny, and S.F. Wayne, Specimen Counterface Bulk Hardness Effects in Impact Wear of 17-4 PH Steel Pairs, *Wear,* Vol 103 (No. 2), 1985, p 175–185

29. P.A. Engel and J.L. Siroco, Impact Wear Study of Lubricated Contacts, *Trans. ASLE,* Vol 18, 1975, p 279–289

30. H. Nowotny, S.L. Rice, and S.F. Wayne, Characteristics of Wear Debris in Impact Sliding, *Wear,* Vol 68, 1981, p 159–167

31. E.S. Zanoria and P.J. Blau, Effect of Incidence Angle on the Impact-Wear Behaviour of Silicon Nitride, *J. Am. Ceram. Soc.,* Vol 81 (No. 4), 1998, p 901–909

32. S.R. Srinivasan and P.J. Blau, Effect of Relative Humidity on Repetitive Impact Behaviour of Machined Silicon Nitride, *J. Am. Ceram. Soc.,* Vol 77 (No. 3), 1994, p 683–688

33. E.S. Zanoria and P.J. Blau, Effects of Machined Surface Condition on the Repeated Impact Behaviour of Silicon Nitride, *Wear,* Vol 218, 1998, p 66–77

34. S.R. Srinivasan, P.J. Blau, and J.L. Bjerke, Effect of Machining Residual Stresses on the Repetitive Impact Behaviour of Silicon Nitride, *J. Mater. Res.,* Vol 10 (No. 1), 1995, p 95–100

35. R. Bantle and A. Matthews, Investigation into the Impact Wear Behaviour of Ceramic Coatings, *Surf. Coat. Technol.,* Vol 74–75, 1995, p 857–868

36. R.G. Bayer, Impact Wear of Elastomers, *Wear,* Vol 112, 1986, p 105–120

37. S.L. Rice, Reciprocating Impact Wear Testing Apparatus, *Wear,* Vol 45, 1977, p 85–95

38. R.G. Bayer and E. Sacher, The Effect of Voids on the Sliding Friction of Polyimide Film, *Wear,* Vol 37, 1976, p 15–20

39. G.E. Johnson, H.E. Bair, E.W. Anderson, and J.H. Daane, The Nature of Water in Polyethylene and its Relationship to Dielectric Loss, *Proc. National Academy of Sciences National Research Council Conf. On Electrical Insulation and Dielectric Phenomena,* 1976

40. R.G. Bayer, P.A. Engel, and E. Sacher, Impact Wear Phenomena in Thin Polymer Films, *Wear,* Vol 32, 1975, p 181–194

41. P.J. Blau and T.A. Hanft, Quantitative Measurement of Repetitive Impact Damage on Ground Silicon Nitride Surfaces, *Tribol. Int.,* Vol 27 (No. 2), 1994, p 109–118

42. N.J. Mahoney, R.J. Grieve, and T. Ellis, A Simple Experimental Method for Studying the Impact Wear of Metals, *Wear,* Vol 98 (No. 1–3), 1984, p 79–87

43. P.A. Engel, Percussive Impact Wear, *Tribol. Int.,* Vol 11, 1978, p 169–176

44. K. Wellinger and H. Breckel, Kenngrossen und Verscheiss Beim Stoss Metallischer Werkstoffe (Deformation and Wear in the Impact of Metallic Workpieces), *Wear,* Vol 13, 1969, p 257–281

45. R.W. Fricke and C. Allen, Repetitive Impact-Wear of Steels, *Wear,* Vol 163, 1993, p 837–847

46. R. Lewis and R.S. Dwyer-Joyce, "Design Tools for Predicting Inlet Valve Recession and Solving Valve Failure Problems," SAE Paper 2001-01-1987, presented at the SAE International Spring Fuel and Lubricants meeting (Orlando, FL), Society of Automotive Engineers, 7–9 May 2001, in SP-1624, 2001, p 105–114

47. R. Lewis, R.S. Dwyer-Joyce, and G. Josey, Investigation of Wear Mechanisms Occurring in Passenger Car Diesel Engine Inlet Valves and Seat Inserts, *J. Fuels Lubric., Transactions of the SAE 1999,* 2000, p 610–618

48. R. Lewis, R.S. Dwyer-Joyce, and G. Josey, Design and Development of a Bench Test-Rig for Investigating Diesel Engine Inlet Valve and Seat Wear, *Trans. Mech. Eng., Institution of Engineers Australia,* Vol ME24 (No. 1), 2000, p 39–46

49. R. Lewis and R.S. Dwyer-Joyce, An Experimental Approach to Solving Combustion Engine Valve and Seat Wear Problems, *Proc. of the 27th Leeds-Lyon Symposium on Tribology,* Elsevier, Tribology Series No. 39, 2001, p 629–640

50. J.F. Archard, Contact and Rubbing of Flat Surfaces, *J. Appl. Phys.,* Vol 24 (No. 8), 1953, p 981–988

Spalling from Impact Events

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

---

Introduction

SPALLING is a type of surface damage where material separates from the surface in the form of flakes or chips. The depth and area of material separation from spalling is usually more extensive than pitting damage. Spalling occurs under various conditions, and specialized terms are sometimes used to characterize spalling damage under specific conditions. For example, in railway engineering, the term "shelling" is used to describe an advanced stage of spalling.

Spalling is a type of wear mode that typically occurs from rolling contact on bearing materials or gear teeth. Failure of moving surfaces that are in contact can be due to a number of causes, related primarily to excessive

wear. Pitting occurs at points of maximum Hertzian contact stress. A failure of this type depends on the number of stress cycles and results in small surface fatigue cracks. Small pits are formed, usually 0.5 to 1.0 mm (20 to 40 mils) in diameter. Large areas of material removal are known as spalling, and this form of surface degradation often has its origins as subsurface fatigue damage from cyclic rolling contact. Hence, the spalling of gear teeth and bearing materials is also known as fatigue wear or rolling contact fatigue.

Spalling damage on a surface can also occur from impact events. For example, the spalling of striking/struck tools is of considerable interest from the engineering, economic, and safety standpoints. The early investigations of spalling by A.H. Burn, by H.O. McIntire and G.K. Manning, and by J.W. Lodge made significant contributions to our knowledge of this metallurgical phenomenon. However, many of the results of these investigations were not published in the open literature, and an awareness of these works often times resulted simply from one's involvement in product liability litigation.

This article reviews the significant contributions of these earlier investigations. In particular, the field tests conducted by A.H. Burn and the laboratory tests of McIntire and Manning and of J.W. Lodge are considered in detail. The remainder of this article is concerned with the metallography and fractography of spalling. The macrostructure and microstructure of spall cavities are described, along with some aspects of the numerous specifications for striking/struck tools. The availability of new, more spall-resistant metals is described, and the safety aspects of striking/struck tools in railway applications are reviewed.

Spalling from Impact Events

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

---

Development of Testing and Analysis Methods for Spalling of Striking Tools

As a result of eye accidents that occurred during the use of hammers in the 1950s, A.H. Burn (Ref 1), a metallurgist with Imperial Chemical Industries Ltd. (ICI), England, had eighty-nine 0.7 to 0.9 kg (1.5 to 2 lb) ball-peen hammers withdrawn from service for examination and tests to determine if the hammers were in compliance with BS 876-1949, the British standard then in effect for hand hammers. Burn found that hammers with hardness levels within, over, and below specification had spalled, indicating that hammers of any hardness could spall. Subsequently, in an attempt to find a more damage-resistant steel for the manufacture of hammers, he placed into service hammers made from a variety of steels and followed their performance over a period of time. The steels tested included plain carbon and low-alloy steels, and the hardness of each hammer complied with the hardness limits required by BS 876 (520 to 640 HV; 50.5 to 57.5 HRC). Although this type of testing yields useful information, it suffers from a lack of control, the inability to acquire quantitative data, and a rather prolonged test period. Relative to these shortcomings, Burn stated in an ICI report dated April 1959 that “Laboratory testing of potentially useful materials rather than a further practical trial is required at this stage. Some form of falling weight test is envisaged” (Ref 2).

In 1958, H.O. McIntire and G.K. Manning of Battelle Memorial Institute coauthored a report entitled “A Study of the Spalling of Hand Hammers” (Ref 3); the research was supported by four manufacturers of striking/struck tools. Using high-speed photography, these researchers determined that the velocity of a hammer head striking 16-penny nails being driven into wood by an experienced carpenter was 14 m/s (45 ft/s), and, furthermore, when striking a hammer against a steel anvil in a deliberate attempt to cause spalling, velocities up to 33 m/s (107 ft/s) were recorded. Therefore, to conduct spalling tests of hammers, McIntire and Manning designed the device shown in Fig. 1, which is capable of producing hammer-head velocities in the range of 14 to 33 m/s (45 to 107 ft/s). In the words of McIntire and Manning, “The hammer handles were held in a rubber-lined clamp attached at one end of a rotating shaft. A torque spring was connected through a lever arm at the opposite end of the shaft. To wind up the spring and place the hammer in position to begin the blow, a ratiomotor drive rotated the shaft. A trigger mechanism released the torque spring when it had been wound to a predetermined angle. The trigger could be attached at several positions, allowing the spring to be torqued to angles of from 105 to 280°. Upon release, the energy stored in the spring rotated the shaft and hammer until the hammer struck the anvil. The anvil was 114 mm ($4\frac{1}{2}$ in.) in diameter and 89 mm ($3\frac{1}{2}$ in.) high and was made of tool steel. In the

photograph, it is shown tilted at an angle of 25° from the horizontal, so that the edge of the hammer face would be struck against the face of the anvil. Prior to making the tests, the anvil height was adjusted so that, when horizontal (0°), the hammer face and the surface of the anvil would meet squarely. The energy imparted to the hammers, when the spring was wound to different positions and then released, was determined experimentally." Four settings, corresponding to energies of 30, 43, 95, and 190 J (22, 32, 70 and 140 ft · lbf), were used for the spalling tests.

Fig. 1 Hammer-testing device used by McIntire and Manning. Source: Ref 3

Joseph W. Lodge, who, at one time, was the chief research engineer for Youngstown Sheet and Tube Company, was actively involved in hammer-related testing in the 1960s and 1970s. He performed spalling tests using two significantly different experimental setups (Ref 4). The first fixture used by Lodge is shown in Fig. 2. The test specimen was a 25 mm (1 in.) diameter by 76 mm (3 in.) cylinder whose ends were heat treated to a specified hardness within the range 38 to 60 HRC using the Jominy end-quench technique. One end of the cylinder rested on three anvils heat treated to a hardness of 61 HRC. The anvils were positioned so that they had equal overlaps with the end of the test sample, the test areas being separated by a 120° arc along the periphery of the end of the specimen. There was a stem on the bottom of each anvil that fitted into a radial, greased slot. This design feature allowed the anvil to move in a radial direction as the chip formed. The test fixture was placed in a mechanical testing machine, and the test specimen was pressed against the anvils using a crosshead speed of 2.5 mm/min (0.1 in./min); load versus displacement was recorded as the test proceeded. The disk in the upper part of Fig. 2 was guided by the three vertical posts, and the upper end of the test specimen fitted into a blind hole in the bottom of the disk. Because the spalling of striking tools normally occurs under impact loading, Lodge also performed tests by striking the disk with a sledge hammer.

Fig. 2 First fixture used by Lodge to simulate spalling. Source: Ref 4

The second test fixture designed and constructed by Lodge is shown in Fig. 3, 4, and 5. A 25 mm (1 in.) diameter by 76 mm (3 in.) specimen also was used for these tests. As described by Lodge, “The specimen is cradled close to the end of the balanced arm on the near end of the shaft. The arm is rotated counterclockwise at constant speeds through a V-belt drive from a constant-speed electric motor. The anvil that the specimen strikes protrudes from the left end of the massive steel block, hinged as shown in the photograph. The anvil block is cocked by a prop under the right rear corner that allows the specimen, while rotating at constant speed, to just clear the top edge of the shaped anvil. The anvil is triggered into striking position while the arm is rotating by an electric solenoid that pulls out the prop. The vertical distance that the anvil moves into the path of the rotating specimen is controlled by the calibrated screw stop located at the right front corner of the massive block. Triggering the device while it is rotating at constant speed is initiated by manually pressing a button, but it is timed electronically. A timing disk with a short slot in its periphery is attached to the far end of the rotating shaft. A solid-state infrared emitter and detector scans the rim of the disk, and when the slot passes between them, the solenoid is actuated, but only when the manual firing button is pressed. In the photograph, the slot can be seen at the bottom edge of the disk just leaving the detector as the arm carrying the specimen has just cleared the anvil. This allows the time of one complete revolution of the arm for the coil springs to overcome the inertia of the massive block and snap into a fixed position before impact. For instance, at 600 rpm, this time interval amounts to approximately one-tenth of a second. A box of fine sand is placed in the hatch under the anvil to catch chips as they are knocked off of the specimens. The chips are retrieved from the sand with a small permanent magnet.

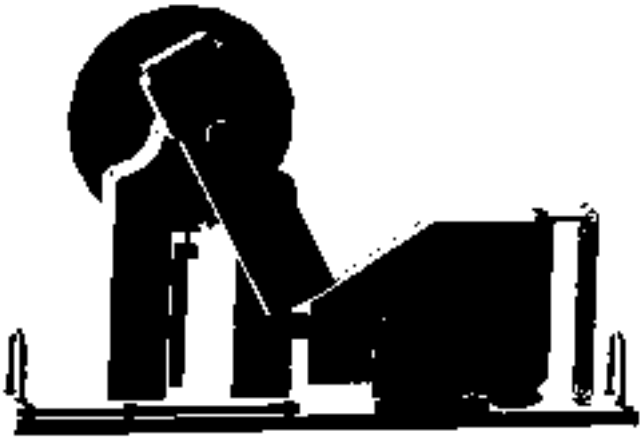


Fig. 3 Second fixture used by Lodge to simulate spalling. Courtesy of J.E. McGarry

Fig. 4 Trailing side of the balanced arm of Lodge's fixture

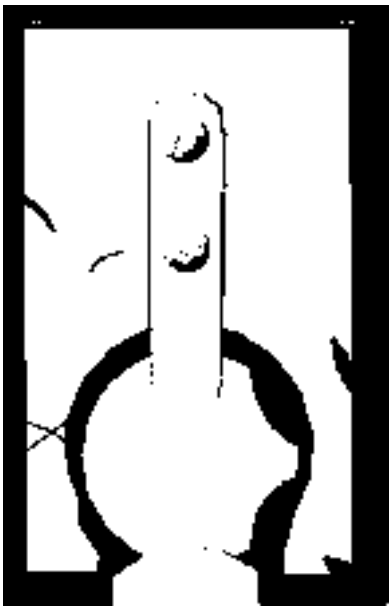


Fig. 5 Leading edge of Lodge's specimen and holder

A small lever arm on the back of the rotating arm is necessary to hold the specimen in place at the startup until it reaches a constant testing speed, reached in about two revolutions. At a constant speed, centrifugal force holds the specimen in place. Simultaneously, centrifugal force and air pressure rotate the small lever to clear the slot and allow the specimen to be knocked out unrestricted on impact.

The specimen was carried in a slot in the end of the arm, and a narrow contact strip on the top center of the specimen ensured consistency in positioning the specimen, because any forward movement would change the

overlap on contact with the anvil. The slot was oversized and had two line contacts with the specimen to minimize friction and make certain that only the momentum of the specimen and none of the arm is involved in the test." A similar, if not identical, device also has been used for studies of spalling at Stanley Tools.

Velez and Powell (Ref 5) investigated the spalling of 1060 steel using the relatively simple device shown in Fig. 6. A cylindrical steel projectile, 9.5 mm (0.375 in.) diameter by 19 mm (0.75 in.) with a hardness of 55 HRC, was fired at the periphery of a cylindrical target (test specimen) that was machined from 1060 steel heat treated to various hardnesses. The overlap of the projectile on the face of the target was 4.7 mm (0.187 in.), that is, half the diameter of the projectile. The velocity of the projectile after exiting the steel tube was measured with a timer, the separation between the detection points being 152 mm (6 in.). The experiments were performed with helium pressures ranging from 345 to 1380 kPa (50 to 200 psi), corresponding to projectile kinetic energies of 16.0 and 63.3 J (11.8 and 46.7 ft · lbf), respectively. A cardboard box filled with sand was positioned to collect the metal fragment from the target and also the rebounded projectile, both of which leave the impact area with appreciable velocities.

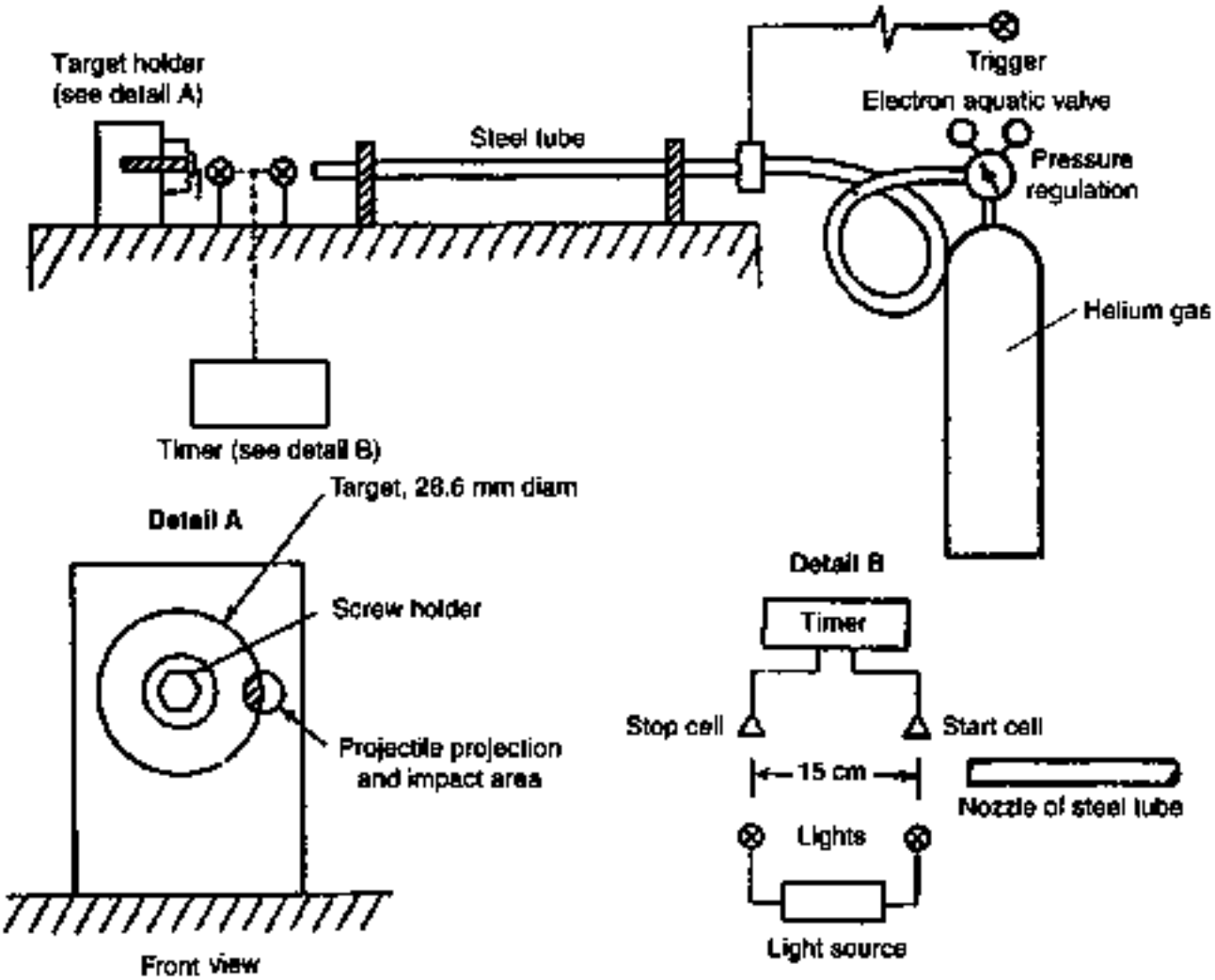


Fig. 6 Device used by Velez and Powell to investigate the spalling of 1060 steel. Source: Ref 5

A hammer-testing machine, called the hammer impact tester (HIT), of more recent vintage has been designed for Woodings-Verona Tool Works by Korvick Drafting & Design Services of Pittsburgh, PA (see Fig. 7). Woodings-Verona originally used this machine to automate the hammer-testing procedure specified in ANSI/HTI B173.3, the specification for heavy striking tools. Another of these testing machines was used by Conrail for its testing program, the objectives of which were to compare the relative resistance to striking/struck-face deformation and spalling of various railroad percussion tools. The ultimate goal of the program at Conrail was to reduce maintenance-of-way employee injuries. The HIT is a simple electropneumatic machine that is able to repeatedly strike an anvil with a hammer or spike maul. The end of the striking-tool handle is fitted and secured into a specially designed grip assembly. When air pressure in the cylinder is released, the tool is pivoted upward in the grip assembly until an electrical trigger switch is engaged. Air

pressure then is put into the cylinder as the driving force for the test, and the striking tool is pivoted downward onto the anvil, which is held rigidly in a fixture. A digital counter records how many blows have been struck. Primary machine settings include the height of the trigger switch and the air pressure supplied to the cylinder. A higher trigger-switch setting means that the tool is raised higher, has more potential energy, and, consequently, impacts the anvil with greater force. Increasing the air pressure in the cylinder also increases the striking force. The machine can accommodate a range of hammer handle lengths and any type of struck tool, that is, the anvil.

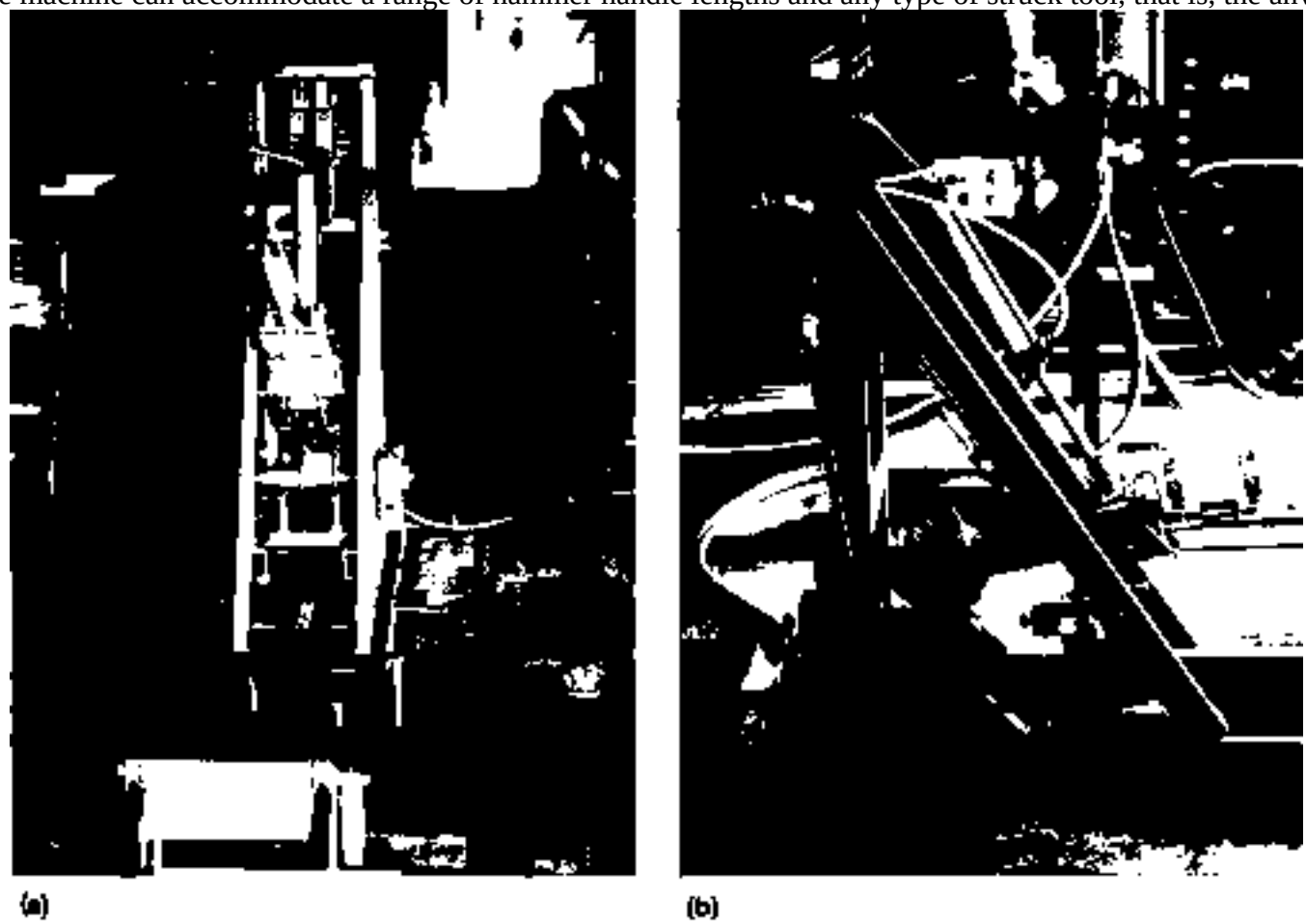


Fig. 7 Front (a) and side (b) views of the hammer impact tester

At Conrail, the most challenging and time-consuming aspect of the setup of the HIT machine was the impact calibration, because it was necessary to ensure that the blow delivered to the anvil was similar to that delivered by a man. If the impacts were not similar, then there would be some question concerning the relevance of the results of the tests to the in-the-field use of the tools. To calibrate the HIT, the struck end of a drift pin was instrumented with strain gages. Various trackmen and laboratory employees struck the drift pin with a sledge hammer, and the output waveform (voltage versus time) was recorded with a data acquisition system (Somat 2100). The HIT then was used to strike the drift pin, using various pressures and trigger-switch settings, and then the waveforms were compared. The waveforms were quite different. The HIT waveform appeared to be more like a “spike,” with a very fast rising/falling voltage peak. The typical curve for a man striking the drift pin displayed a lower peak voltage over a longer period of time. After installing a new, reinforced concrete floor and industrial rubber mounts under the HIT to dampen the impact, the waveforms for the man and the machine were similar. A reduced air pressure setting was also used to keep the impact force at a lower, more realistic value. The research and test department of Norfolk Southern Railroad now has possession of the HIT machine.

The testing devices described previously were designed to simulate the spalling of striking/struck tools around the periphery of the striking/struck surface, the usual failure mode for tools of these kind. Lodge's two testing devices produced an aligned edge-to-edge impact, such as that shown schematically in Fig. 8 where the surfaces of the striking/struck objects are parallel to one another. In the case of Velez and Powell's research, the overlap was greater than indicated in Fig. 8 for the aligned edge-to-edge impact. Each of the three types of impacts shown in Fig. 8 could have been investigated with the testing device designed by McIntire and Manning, but

they limited their studies to edge-to-surface and aligned edge-to-edge impacts. The edge-to-surface impact corresponds to that situation in which one of the tools is impacted on the periphery of the working face and the other is impacted at a site more centrally located on the working face. Thus far, in the tests conducted at Conrail with the HIT apparatus, the HIT has been set up to produce a glancing impact, such as that shown in Fig. 9. Without definitive supporting evidence, the authors opine that spalling occurs most frequently as a result of edge-to-surface impacts, simply because this type of impact would seem to have a greater probability of occurrence than the edge-to-edge impacts. However, certain situations, such as a spike maul driving spikes into wooden railroad ties, can lead to edge-to-edge contact more frequently than situations involving other tool combinations. The user of a striking tool attempts assumedly to impact the striking/struck surfaces at their centers, and, therefore, edge-to-surface impacts should be more probable than the other types of impacts. Consequently, on the basis of this line of reasoning, spalling tests should be performed preferably with equipment that, at a minimum, permits edge-to-surface impacts over a range of angles between the striking and struck surfaces.

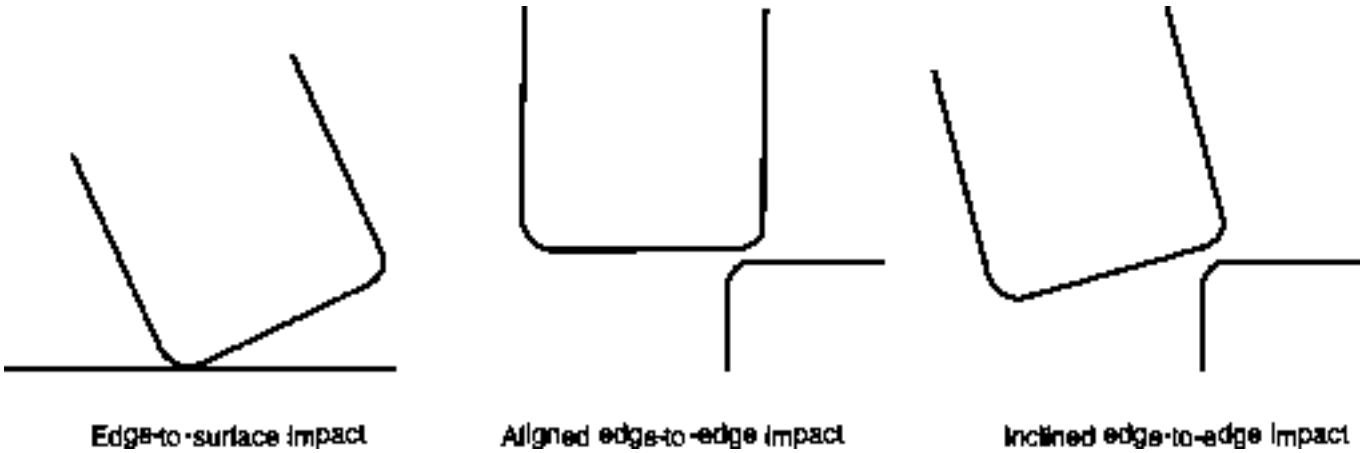


Fig. 8 Striking/struck tool impact geometries

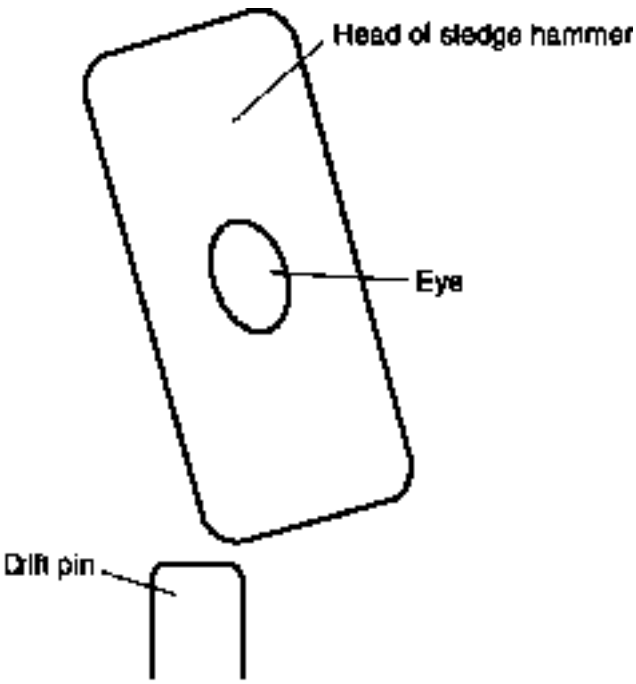


Fig. 9 Glancing impact used for the HIT tests at Conrail

References cited in this section

1. A.H. Burn, “Chipping of Hand Hammers,” *Engineering,* 12 July 1957

2. A.H. Burn, “The Resistance to Chipping of 1% Cr-Mo Steel Hand Hammers,” internal report, Imperial Chemical Industries Ltd., April 1959

3. H.O. McIntire and G.K. Manning, “A Study of the Spalling of Hand Hammers,” Battelle Memorial Institute, unpublished report, 19 May 1958

4. J.W. Lodge, “Tests Simulating the Edge Chipping of Hammers,” Youngstown Sheet and Tube Company, unpublished report, and presented at meeting of the Hand Tool Institute (North Key Largo, FL), 1973–1975

5. J.F. Velez and G.W. Powell, Some Metallographic Observations on the Spalling of AISI 1060 Steel by the Formation of Adiabatic Shear Bands, *Wear*, Vol 65–66, 1980–1981, p 367

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

Conclusions Drawn by Spalling Studies

The hammers used for Burn's field tests were made from a variety of steels: plain carbon steels having a carbon content in the range of 0.45 to 0.65 wt%, a 1% Cr-Mo steel having the composition 0.43C-0.17Ni-0.98Cr-0.27Mo, and a more highly alloyed steel having the composition 0.65C-1Si-3Ni. Again, each of the hammers had a hardness in the range of 50.5 to 57.5 HRC, as required by BS 876. The tests also included Stanley “rim-tempered” hammers. The 1% Cr-Mo steel was selected for the field tests, because, at a given hardness, it possessed greater ductility than the standard 0.55% C steel (BS 876). On completion of the field tests, Burn concluded that “The results indicate that there is no significant difference between the hammers either in chipping or in rate of wear. The inference has been drawn that to improve chipping resistance, a radical and probably expensive change of material is required” (Ref 6). Subsequently, as a result of spalling tests conducted with a variety of steels (1045, 1060, 1072, 1075, 1080, 1552, 9260, and 4340), J.W. Lodge reached the same conclusion. In a letter dated June 5, 1982, to Mr. Robert Burke, president of Warwood Tool, he stated, “My tests didn't show any superiority for silicon or other alloy steels,” echoing the same conclusion he had voiced at a meeting of the Hand Tools Institute in 1973 (Ref 4). The question as to why the spalling resistance of this wide range of steels is essentially the same is considered later on in the discussion. Statistically, Burn found that chipping occurred once every 200 hammer weeks. Another interesting observation made by Burn was that the strain hardening of the striking face that occurs during the use of a hammer increased the hardness 1.5 to 2.5 points HRC (Ref 2). A fact not mentioned earlier about Burn's field tests was that the entire hammer head of each hammer was austenitized, quenched, and tempered. This was contrary to the normal practice at that time and one that is still used today of austenitizing only the striking end of the hammer in a lead or salt bath and essentially end quenching the striking face. However, the entire head of sledge hammers and spike mauls currently manufactured according to American Railway Engineering and Maintenance-of-Way Association (AREMA) specifications is austenitized, quenched, and tempered. The subject of whole-head versus partial-head heat treatment is discussed further in a later section of this article.

Effects of Tool Geometry. Failing to find a ferritic plain carbon or low-alloy steel with superior resistance to chipping, Burn briefly considered the matter of tool geometry. He concluded, in part, that “Chipping of hammers is not a rare phenomenon. What is unusual is consequential damage. This may be only a matter of geometry” (Ref 6). The consequential damage to which he refers is undoubtedly eye and other bodily injuries. The implication of Burn's statement is that some control over the incidence of these injuries could be effected by modification of the tool geometry. Precisely what Burn had in mind is unknown, but he undoubtedly had reference at least to the chamfered or radiused edge around the periphery of the striking surface.

The research carried out by McIntire and Manning yielded information about this particular aspect of tool geometry. McIntire and Manning used commercial, off-the-shelf nail, ball-peen, and tinner's-setting hammers for their spalling tests. The nail and ball-peen hammers had a chamfer, whereas the tinner's-setting hammers did

not. The dimensions used by McIntire and Manning to specify the chamfer are shown schematically in Fig. 10. The chamfers on the nail and ball-peen hammers ranged in width from 1.6 to 3.2 mm ($\frac{1}{16}$ to $\frac{1}{8}$in.) and the chamfer angle from 30 to 54°, the ball-peen hammers consistently having the widest chamfer (3.2 mm, or $\frac{1}{8}$in.). The ball-peen hammers were more resistant to spalling when subjected to edge-to-surface impacts (Fig. 8) than the nail hammers, a result McIntire and Manning attributed to the wider chamfer and also to a more slender handle possibly making the tool more shock resistant. The tinner's-setting hammers were the least resistant to spalling, lacking any chamfer at all.

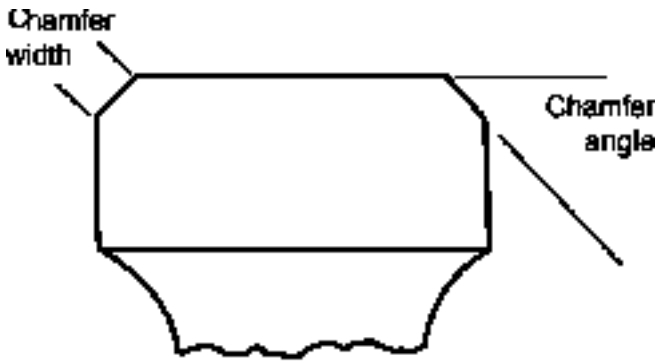


Fig. 10 Chamfer dimensions specified by McIntire and Manning

The research by Lodge also provided information about the effects of tool geometry. The geometry of the chamfered and radiused edges used by Lodge for his edge-to-edge tests with the device shown in Fig. 2 is shown schematically in Fig. 11. He used two chamfer sizes, 1.6 and 3.2 mm ($\frac{1}{16}$ and $\frac{1}{8}$ in.), and edge radii of 1.6 and 3.2 mm ($\frac{1}{16}$ and $\frac{1}{8}$ in.). A tool-post grinder was used to machine the chamfered or radiused edge on his specimens. On the basis of his test results, Lodge concluded that “Either chamfered or radiused edges are beneficial in that they prevent firm bites with overlaps equal to or less than their widths. The shape of the edge has a pronounced effect on the size and shape of the chip. Under the conditions of these tests, the radiused edge is at least equal to the chamfered edge for small overlaps and superior for a 4.8 mm ($\frac{3}{16}$ in.) overlap” (Ref 7). As suggested by Fig. 12, to obtain a firm bite of one surface on another, the overlap must increase when the edges of the impacting surfaces are chamfered or radiused. Therefore, the area of the sheared surface over which fracture propagates when spalling occurs is thereby also increased, causing an increase in the striking force required to produce spalling.

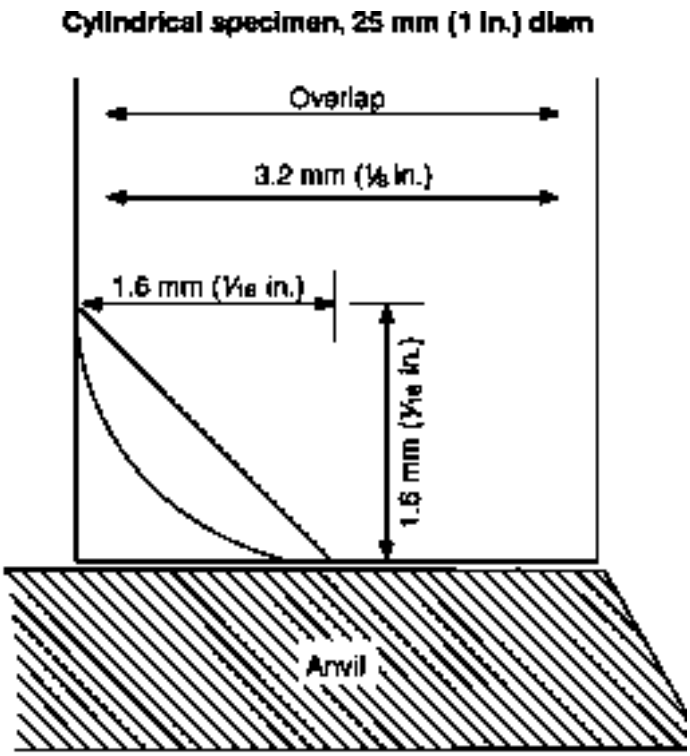


Fig. 11 Geometry of the chamfered and radiused edges used by Lodge

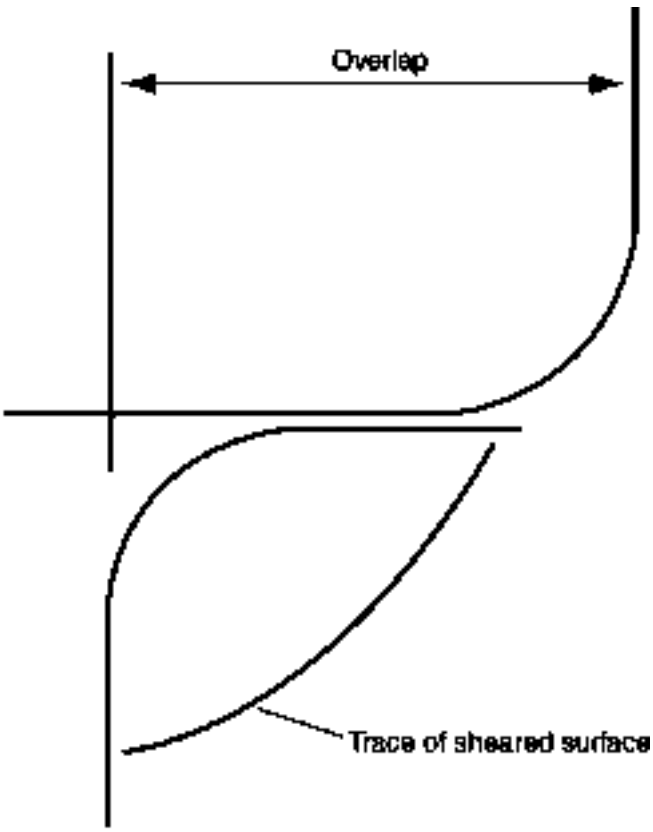


Fig. 12 Position of the sheared surface as determined by the extent of overlap

Thus, the research of McIntire and Manning and of Lodge established firmly the fact that chamfered or radiused edges of striking/struck surfaces increase the resistance to spalling for both edge-to-surface and edge-to-edge impacts. However, over a period of time, the shape of a chamfered or radiused edge of a striking/struck tool may be altered as a result of heavy use. The striking/struck surface contiguous to the chamfered/radiused edge may be plastically deformed. The surface layer of the tool is displaced radially outward, thereby reducing the effectiveness of the chamfered/radiused edge. The ultimate effect of such plastic deformation can be

mushrooming around the periphery of the tool face. The presence of mushrooming around the periphery of the striking/struck face should serve as a warning to the user of the tool that spalling is imminent.

Effects of Impact Angle. For the case of edge-to-surface impacts (Fig. 8), McIntire and Manning determined that spalling can occur only over a limited angular range between the striking and struck surfaces. For example, nail hammers did not spall if the angular separation between the hammer/struck surfaces was less than 20° but would usually spall when the angle was 25°. Spalling was less likely to occur at an angular separation of 35°. Thus, the optimal conditions for spalling occur in a narrow angular range centered about an angle of 25°. Tinner's-setting hammers, on the other hand, spalled readily at an angular separation of only 7.5°, obviously because of the 90° corner around the periphery of the striking surface. None of the hammers could be made to spall with flush blows, that is, an angular separation of 0°.

The locations from which spalls are propelled from the periphery of the face of a striking tool are referred to with a clock-face designation, as illustrated in Fig. 13 for nail, ball-peen, and tinner's-setting hammers. In the case of edge-to-surface impacts, spalls are not propelled at random angles from the periphery of the striking face, but rather the flight paths are confined to a limited angular range. The flight paths of spalls from nail, ball-peen, and tinner's-setting hammers as determined by McIntire and Manning are shown in Fig. 14 (Ref 3). Similar results were obtained from spalling tests performed at Stanley Tool using the Lodge-type testing device (Fig. 3). According to J. Olson (Ref 8), the flight paths of chips resulting from an edge-to-surface impact are confined within a circular cone having an apex angle of 20°. It is doubtful that the flight paths of spalls resulting from edge-to-edge impacts are limited to the same angular range as edge-to-surface impacts, because the geometry of an edge-to-edge impact is subject to more variation than that of an edge-to-surface impact.

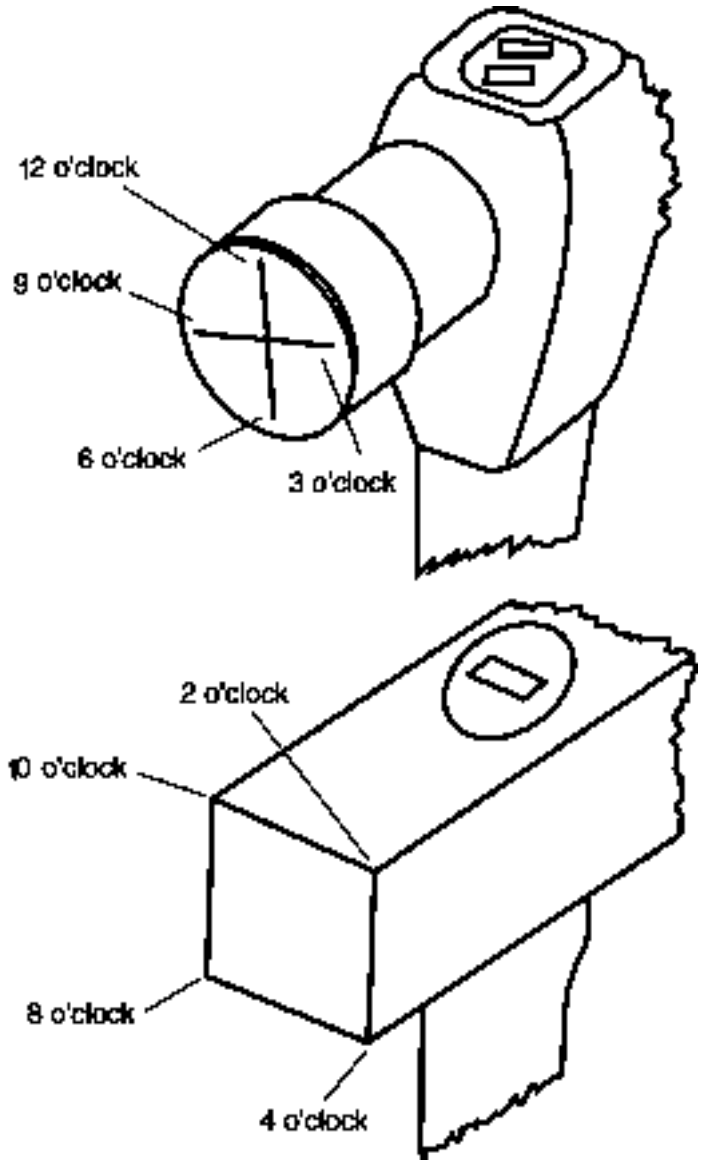


Fig. 13 Clock-face designation of the peripheral sites on the striking face

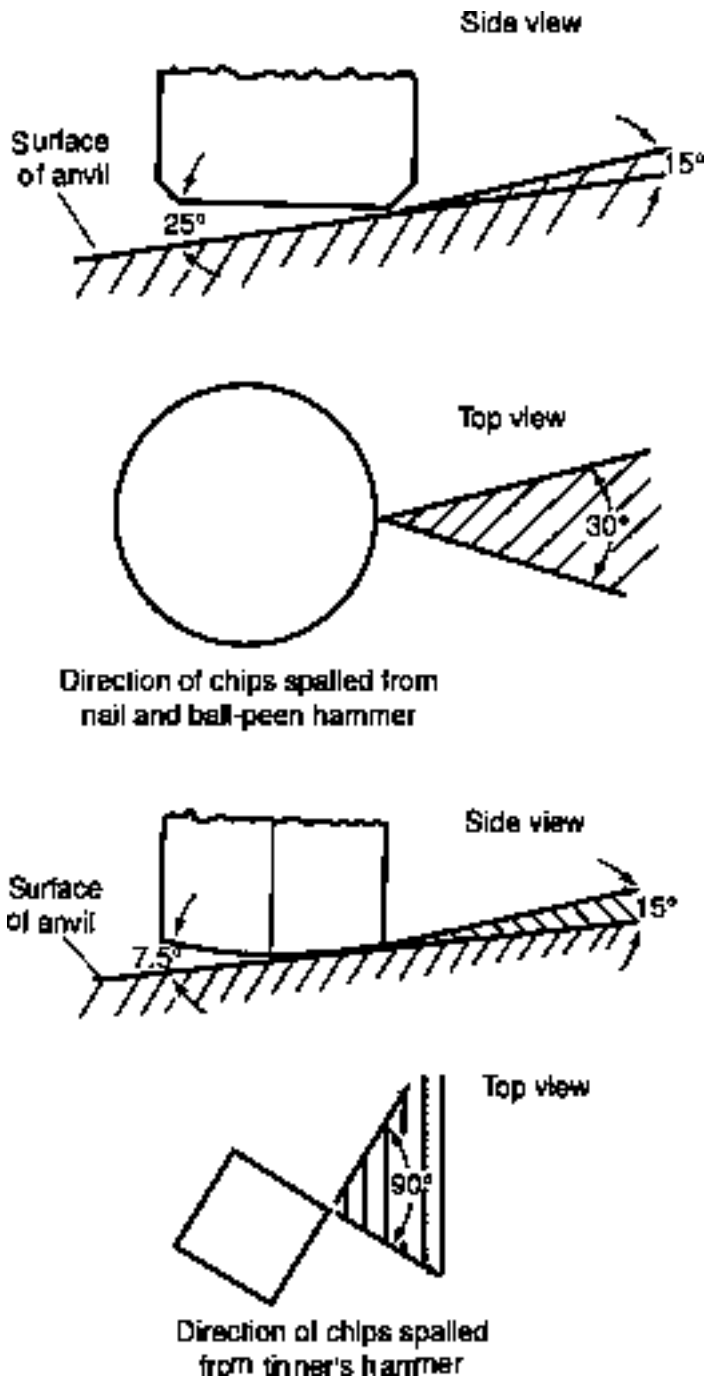


Fig. 14 Spatial distribution of the spall flight paths

The limited angular range of spall flight paths resulting from edge-to-surface impacts has a direct relationship to personal injuries that occur as a result of the spalling of striking/struck tools. For example, if the line of sight of the user of a hammer makes an acute angle with the normal to the struck surface, and a spall hits the user in the upper body, the spall usually originates from a site between the 5 and 7 o'clock locations on the striking face of the hammer. A spall that strikes the lower body under these conditions invariably originates from the struck tool. However, a spall from the struck tool can hit the upper body if the line of sight is roughly parallel to the struck surface. An actual example of such an occurrence involved a man who was lying on his back and looking upward at the end of a jammed, horizontal pin that he was striking with a hammer. A spall from the periphery of the end of the pin penetrated the man's eye, causing loss of eyesight. Spalls that originate from the vicinity of the 3 and 9 o'clock locations put bystanders in jeopardy; accidents of this kind are a matter of record. McIntire and Manning observed that it is more difficult to produce spalling at the 12 o'clock location than at the 6 o'clock location. This observation is supported by the distribution of damage on the striking face of hammers returned from the field; the incidence of spalling at the 12 o'clock location is relatively rare. As noted previously, the spalls that can cause eye injuries to the user of a hammer originate from a site within the 5 to 7 o'clock

locations, and, as also noted earlier, a chamfered/radiused edge increases the resistance to spalling. Therefore, accidents of this kind could be reduced significantly by increasing the size of the chamfer/radius along the periphery of the striking face from the 5 to the 7 o'clock locations. Lodge essentially made this suggestion when he stated, “Well, maybe you can increase the radius near the handle…” during a speech to the Hand Tools Institute Hammer Subcommittee in 1973 (Ref 9). Such a localized change to the geometry of the striking face was never incorporated into the manufacturing process for hammers. However, since about 1987, tool manufacturers have increased the size of the chamfer/radius on both striking and struck tools. The result is that the shape of the striking/struck end of some tools resembles a truncated cone, but with a convex contour to the striking/struck face. The AREMA has established detailed specifications for the contour of the faces of its striking/struck percussion tools. The AREMA specification for the contour of striking faces is given in Fig. 15 (Ref 10).

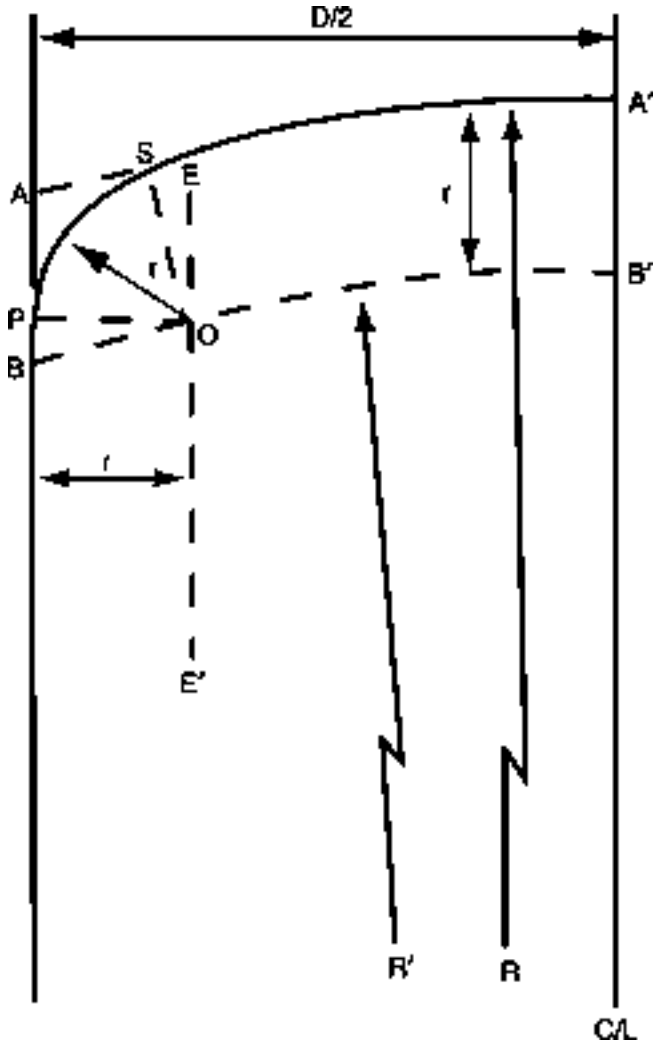


Fig. 15 The AREMA specification for the contour of the striking face. D, stock diameter or, with hexagons and octagons, the distance across the flats; R, crown radius; r, corner radius; R′, crown radius minus corner radius. (1) Draw tool head centerline (C/L). (2) Draw line AC, representing the side of the tool head. Except in tapered tools, this line is parallel to C/L at a distance of D/2. (3) Draw arc AA′ with a radius of R. (4) Draw line EE′ parallel to line AC at a distance (r) equal to the corner radius to be used. (5) Draw arc BB′ parallel to arc AA′ at a distance equal to the corner radius. The radius of this arc (R′) is equal to the crown radius minus the corner radius. (6) Draw arc SP with a radius r from point O, the intersection of line EE′ and arc BB′. The resultant arc is tangent to arc AA′ and line AC at points S and P, respectively.

Metallurgical Factors. As described earlier, the field tests of A.H. Burn and the laboratory tests of J.W. Lodge demonstrated that the spalling resistance of a wide range of plain carbon and low-alloy steels is essentially the same. The metallurgical basis for this result, plus other metallurgical factors associated with spalling, is now considered. The tensile properties of low-alloy steels quenched and tempered to a range of hardness values

were investigated by W.G. Patton (Ref 11) and by E.J. Janitzky and M. Baeyertz (Ref 12). The experimental results of Patton are shown in Fig. 16 and those of Janitzky and Baeyertz in Fig. 17. Patton's data were obtained using low-alloy steels with 0.30 to 0.50 wt% C and those of Janitzky and Baeyertz using low-alloy steels with 0.30 to 0.45 wt% C. Figure 18 contains additional data of this kind for Ni-Cr-Mo steels (Ref 13) and also for an unspecified plain carbon steel. The yield/tensile strengths of the latter steel fall within the property bands for the Ni-Cr-Mo low-alloy steels, but the elongation is lower than that of low-alloy steels. As a final example of the marked similarity of the tensile properties of plain carbon and low-alloy steels quenched and tempered to the same hardness, the relationship between the tensile strength and hardness for quenched and tempered and also as-rolled, annealed, or normalized carbon and alloy constructional steels is shown in Fig. 19 (Ref 14).

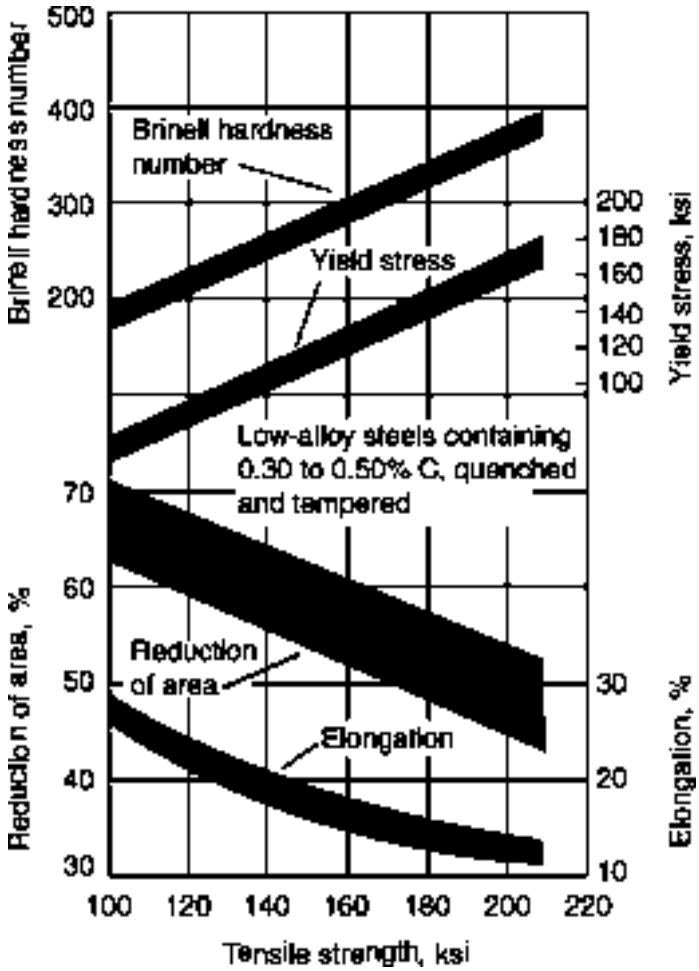


Fig. 16 Mechanical properties of quenched and tempered low-alloy steel (0.30–0.50 wt% C) as determined by Patton. Source: Ref 11

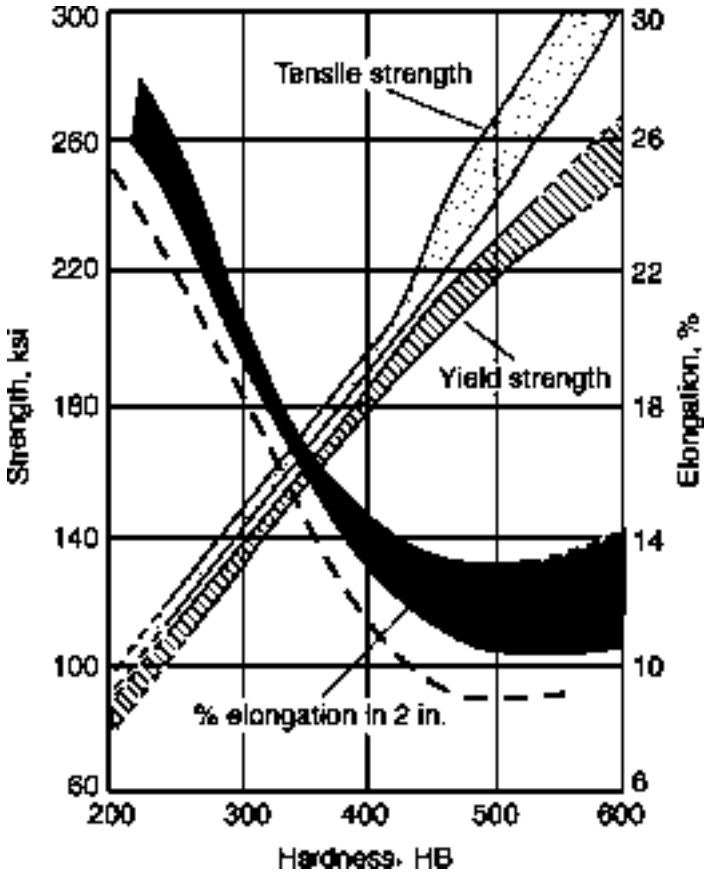


Fig. 17 Mechanical properties of quenched and tempered low-alloy steel (0.30–0.45 wt% C) as determined by Janitsky and Baeyertz. Source: Ref 12

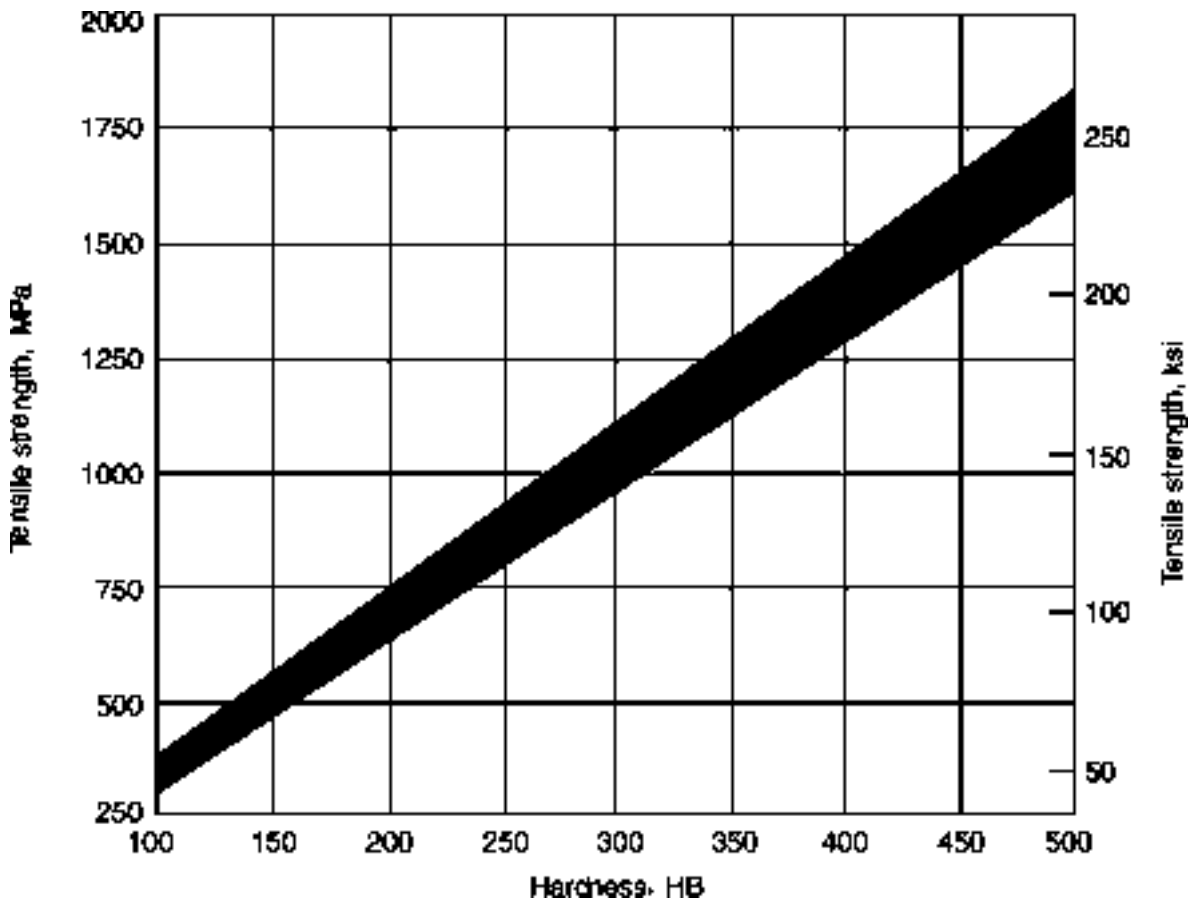


Fig. 18 Mechanical properties of quenched and tempered plain carbon and Ni-Cr-Mo steels. Source: Ref 13

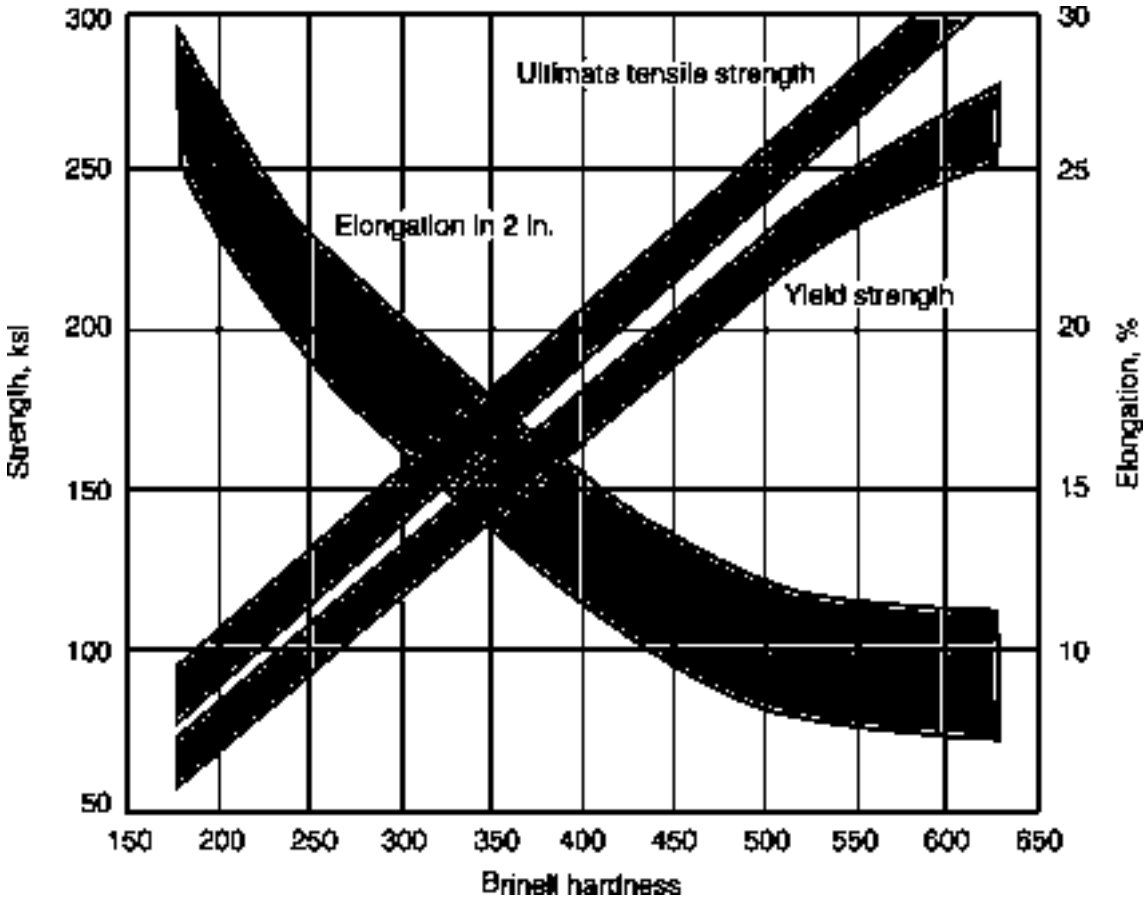


Fig. 19 Strength-hardness correlation for carbon and low-alloy steels. Source: Ref 14

Clearly, the marked similarity of the tensile properties of quenched-and-tempered, plain carbon and low-alloy steels explains why Burn and Lodge did not find any significant difference in the spalling resistance of the wide variety of constructional steels tested in their research. That the spalling resistance of striking/struck tools bears a strong relationship to the tensile properties is suggested by the radial section through an edge-to-edge impact, as shown in Fig. 20 (Ref 5). By analogy with the compression of a plate between hard platens (Ref 15), a wedge-shaped region of large strains in the target should be produced by the impact of the projectile. This description of the plastic deformation associated with the spalling process is oversimplified, as is evident from subsequent discussion of actual spall cavities. As a further simplification, the stress system that produces spalling (at hardnesses below approximately 60 HRC) can be visualized as consisting essentially of a shear stress with a compressive stress acting on the sheared region. J.W. Lodge conducted some spalling tests at -73 °C (-100 °F) and found that spalling still occurred in a ductile manner, that is, by shear. He attributed this result to the nature of the stress system associated with edge-to-edge impacts. As a result of the impact loading associated with spalling, the shear strain rate is usually very high, resulting in localized adiabatic heating and concomitant microstructural changes that are the subject of subsequent further discussion.

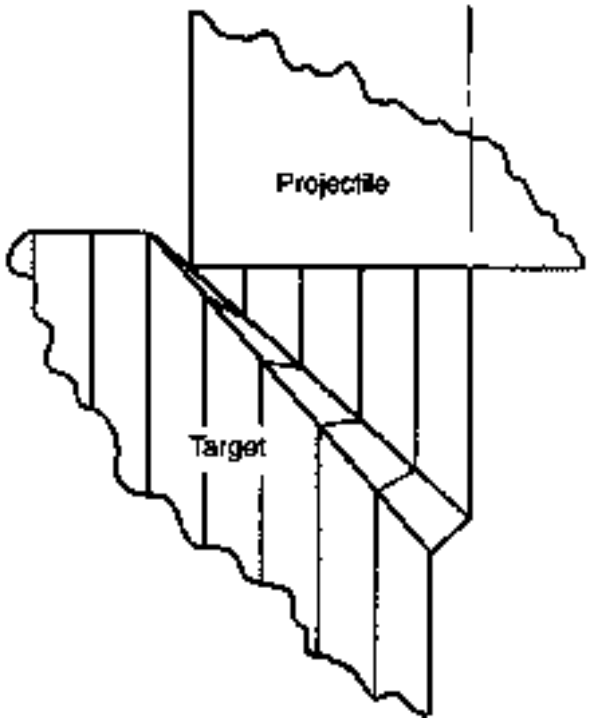


Fig. 20 Band of concentrated shear produced by an edge-to-edge impact

Additional metallurgical factors associated with spalling are described as follows.

For a given overlap and a given impact energy, the resistance to spalling is greater at higher hardness levels. McIntire and Hall (Ref 16) found that 5 kg (12 lb) sledge hammers fabricated from AIS 9260 steel and quenched and tempered to a hardness in the range of 55 to 59 HRC are less likely to spall than are softer hammers when struck against an anvil with a hardness of 51 to 52 HRC. The experimental conditions that yielded this result were single, edge-to-edge impacts with an overlap of 2.5 mm (0.1 in.) and an impact energy in the range of 339 to 407 J (250 to 300 ft · lbf). Supporting this result is the observation by McIntire and Manning (Ref 3) that extra-hard hammers, that is, hammers with a hardness in the range of 63 to 66 HRC, are more difficult to spall than softer hammers, but the extra-hard hammers could be spalled in relatively few blows. Relative to this matter of hardness, Lodge (Ref 4) noted that "The harder specimens show little evidence of damage as testing variables approach those necessary for chipping. Therefore, the harder specimens are both less likely to chip or to establish potentially dangerous conditions short of chipping." In the higher-hardness condition, the flow stress is high and the force generated by the impact must produce a stress on the potential shear surface that exceeds the flow stress, if plastic deformation is to occur. Once the flow stress is exceeded, localized flow is likely to continue, because the strain-hardening coefficient, *n*, at high hardness is low, approximately 0.1. In the case of lower hardness levels, the strain-hardening coefficient is greater, and, consequently, the plastic flow produced by an impact is distributed over a larger volume of the microstructure, delaying the onset of localized deformation and, in all likelihood, requiring more impacts to produce spalling.

The tendency of a hammer to spall is decreased as the hardness of the object being struck is reduced. In fact, McIntire and Manning determined that hammers struck against an anvil with a hardness of 24.5 HRC did not spall, even with heavy blows of 190 J (140 ft · lbf); the heads came off and the handles broke. Clearly, a softer object (tool, anvil, etc.) being struck by a harder hammer deforms plastically and, therefore, the contact area between the two increases, which thereby has the effect of decreasing the stress in the impacted region of the hammer.

If the hardness of a striking tool exceeds 60 HRC by 3 to 5 points, then fracture occurs by a brittle mode rather than shear.

The concept has often been expressed that the ideal microstructure for a hammer consists of a hard surface layer (2.5 to 3.8 mm, or 0.1 to 0.15 in. thick) supported by an underlying softer, tough core; this same concept is also applied to power-transmission gears. However, as pointed out by Lodge, spalling occurs in the hardened periphery of the striking face, with the core not playing any role in the process. Consequently, the use of high-hardenability steels for hammers should not pose any problems, and, in fact, as noted earlier, the tests of Burn

and of Lodge demonstrated that the resistance to spalling of hammers made from low-alloy steels and from plain carbon steels is essentially the same.

References cited in this section

2. A.H. Burn, "The Resistance to Chipping of 1% Cr-Mo Steel Hand Hammers," internal report, Imperial Chemical Industries Ltd., April 1959

3. H.O. McIntire and G.K. Manning, "A Study of the Spalling of Hand Hammers," Battelle Memorial Institute, unpublished report, 19 May 1958

4. J.W. Lodge, "Tests Simulating the Edge Chipping of Hammers," Youngstown Sheet and Tube Company, unpublished report, and presented at meeting of the Hand Tool Institute (North Key Largo, FL), 1973–1975

5. J.F. Velez and G.W. Powell, Some Metallographic Observations on the Spalling of AISI 1060 Steel by the Formation of Adiabatic Shear Bands, *Wear*, Vol 65–66, 1980–1981, p 367

6. A.H. Burn, "Practical Trial of $1\frac{1}{2}$ lb Pein Hand Hammers in Engineering Workshops," internal report, Imperial Chemical Industries Ltd., July 1962

7. J.W. Lodge, "An Investigation of the Mode of Failure by Tests Simulating the Chipping Hand Hammers," unpublished report

8. J.D. Olson, Stanley Works, private communication, 1 Dec 1988

9. J.W. Lodge, speech given to the Hand Tools Institute Hammer Subcommittee, (Cleveland, OH), 8 Nov 1973

10. American Railway Engineering and Maintenance-of-Way Association, "Manual for Railway Engineering," 2000, p 5, 6, 54

11. W.G. Patton, *Met. Prog.*, Vol 43, 1943, p 726

12. E.J. Janitzky and M. Baeyertz, *Alloying Elements in Steel*, E.C. Bain and H.W. Paxton, Ed., American Society for Metals, 1948, p 225

13. *Metals Handbook, 1948 Edition*, American Society for Metals, 1948, p 456

14. *Properties and Selection: Irons and Steels*, Vol 1, *Metals Handbook*, 9th ed., 1978, p 461

15. A. Nadai, *Theory of Flow and Fracture of Solids*, McGraw-Hill, New York, 1959, p 541

16. H.O. McIntire and A.M. Hall, "The Performance of Hand-Striking Tools Under Simulated Service Conditions," Battelle Memorial Institute, unpublished report, 30 June 1962

Spalling from Impact Events

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

Metallography and Fractography of Spalling

As is well known, light-etching bands of untempered martensite are very often produced by the rapid, concentrated shear strain associated with spalling. The adiabatic heating within the shear band, so-called adiabatic shear bands, raises the temperature locally to a level sufficient to austenitize the band, which subsequently transforms to martensite as a result of the quenching action of the colder surroundings. The microstructural evolution of shear bands in annealed (95 HRB) 1060 steel is shown in Fig. 21 (Ref 5). Adiabatic shear bands in the same steel but with a hardness of 45 HRC are shown in Fig. 22. The metallographic samples were nickel plated for good edge retention. Targets (see Fig. 6) of these hardnesses were selected because multiple impacts were required to produce spalling. A given target was struck in one location once, a second location twice, and so on until spalling occurred; the projectile in each case had an impact energy of 49 J (36 ft · lbf). Each of the impacted areas was sectioned centrally along a radial plane. Annealed targets also were used, because the pattern of deformation in the ferrite-plus-pearlite microstructure is more evident than it is in a quenched-and-tempered microstructure. The micrographs displayed in Fig. 21 and 22 were taken in the immediate vicinity of the reentrant corner formed by the edge of the projectile where it impacts the target (Fig. 20).

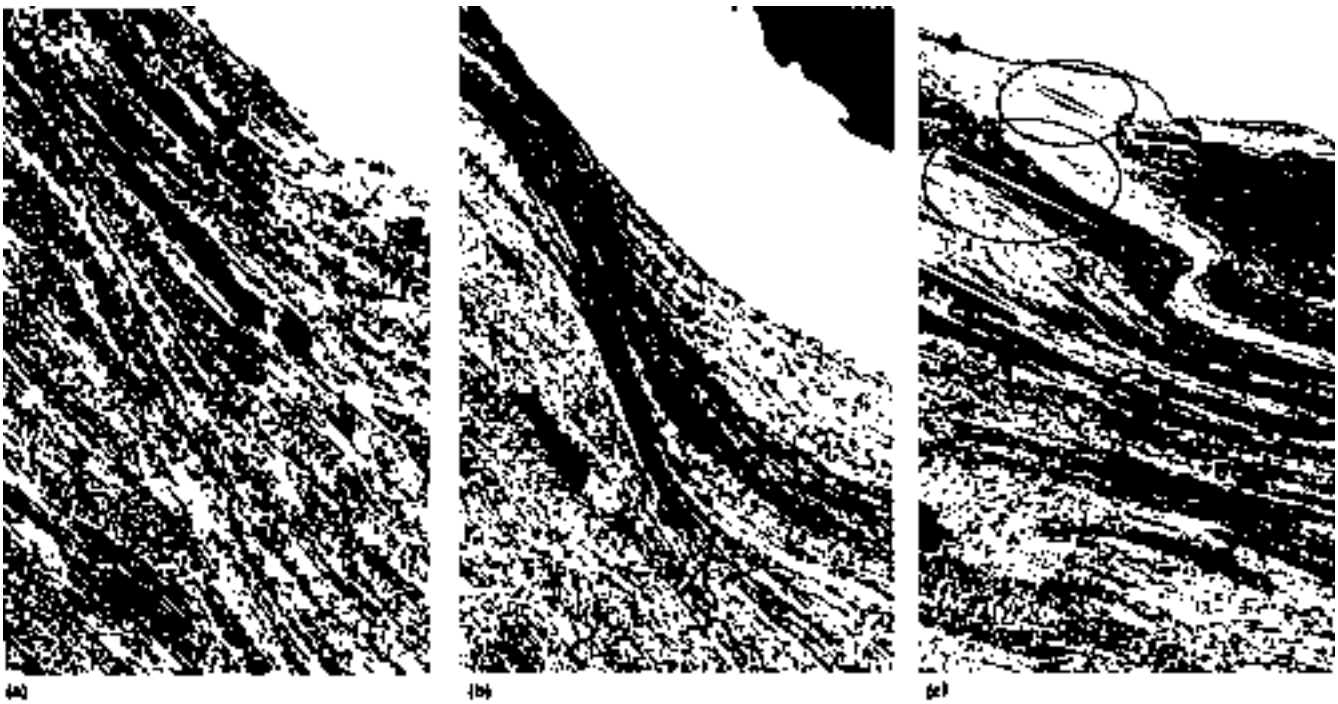


Fig. 21 Evolution of adiabatic shear bands in annealed (95 HRB) 1060 steel with increasing work input. (a) After 2 impacts. (b) After 4 impacts. (c) After 5 impacts. All 660×

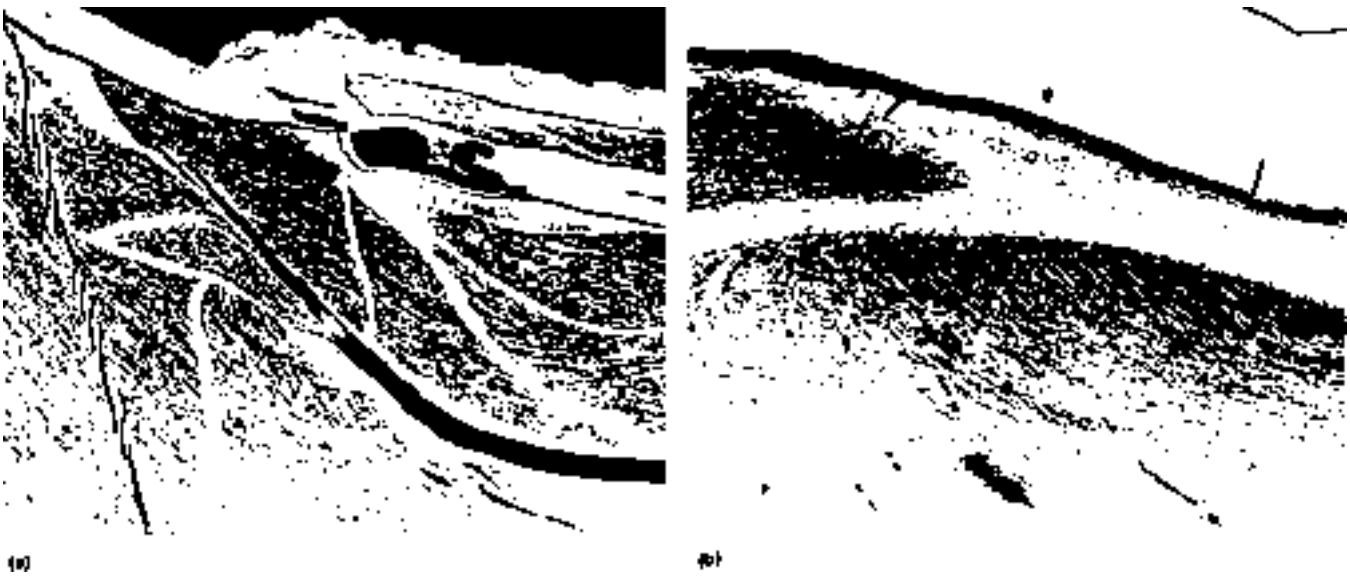


Fig. 22 Transformation shear bands in quenched and tempered (45 HRC) 1060 steel. (a) 330×. (b) 660×

The microstructure of the annealed 1060 after two, four, and five impacts is shown at a magnification of 660× in micrographs (a), (b), and (c), respectively, of Fig. 21. The pattern of plastic flow is relatively diffuse but distinguishable after two impacts and well defined after four impacts. A light-etching band with a convoluted geometry formed during the fifth impact. The microhardness impressions in the upper left corner of micrograph (c) clearly demonstrate the high hardness of the light-etching band relative to that of the surrounding matrix.
In contrast to the response of the annealed microstructure, the deformation pattern in the quenched-and-tempered microstructure is not as diffuse, and light-etching bands had formed during the second impact (Fig. 22(a)), 330×. However, because the light-etching bands appear to be oriented at random and some of the bands are cracked, the first impact may have produced a number of the light-etching bands. Seven impacts were required to spall the target, and the band of untempered martensite through which final fracture occurred is shown in Fig. 22(b) at 660×. Because cracks were present in the shear bands after two impacts at one site on this particular target, it is somewhat surprising that seven impacts were required to produce spalling at another site, suggesting, perhaps, that the geometric disposition of the projectile relative to the target immediately before impact is subject to significant variability. The prespalled configuration of an impacted area is shown in Fig. 23 (330×). Note the light-etching band of concentrated shear and also the separation that has occurred at the impacted end of the band.

Fig. 23 Shear displacement of a spall prior to complete separation along the shear band. 330×

A spall cavity and the sheared surface of the spall that was ejected from the cavity are shown in Fig. 24 at a magnification of 12×. The macroscopic topographical features of the sheared surface tend to run parallel to the

shear direction. Figure 25 contains higher-magnification scanning electron microscopy (SEM) micrographs of the sheared surface shown in Fig. 24. Two distinct areas can be discerned in the 1000× micrograph in Fig. 25(a): the sheared, striated area, and an area characterized by an irregular topography, which is shown at higher magnification (11,000×) in Fig. 25(b). The area of irregular topography resembles the knobbly structure observed by Stock and Thompson (Ref 17) on the fracture surface of aluminum alloys subjected to ballistic impact. This may be evidence of melting, as suggested by these researchers, or possibly, and perhaps in this case, surface oxidation.

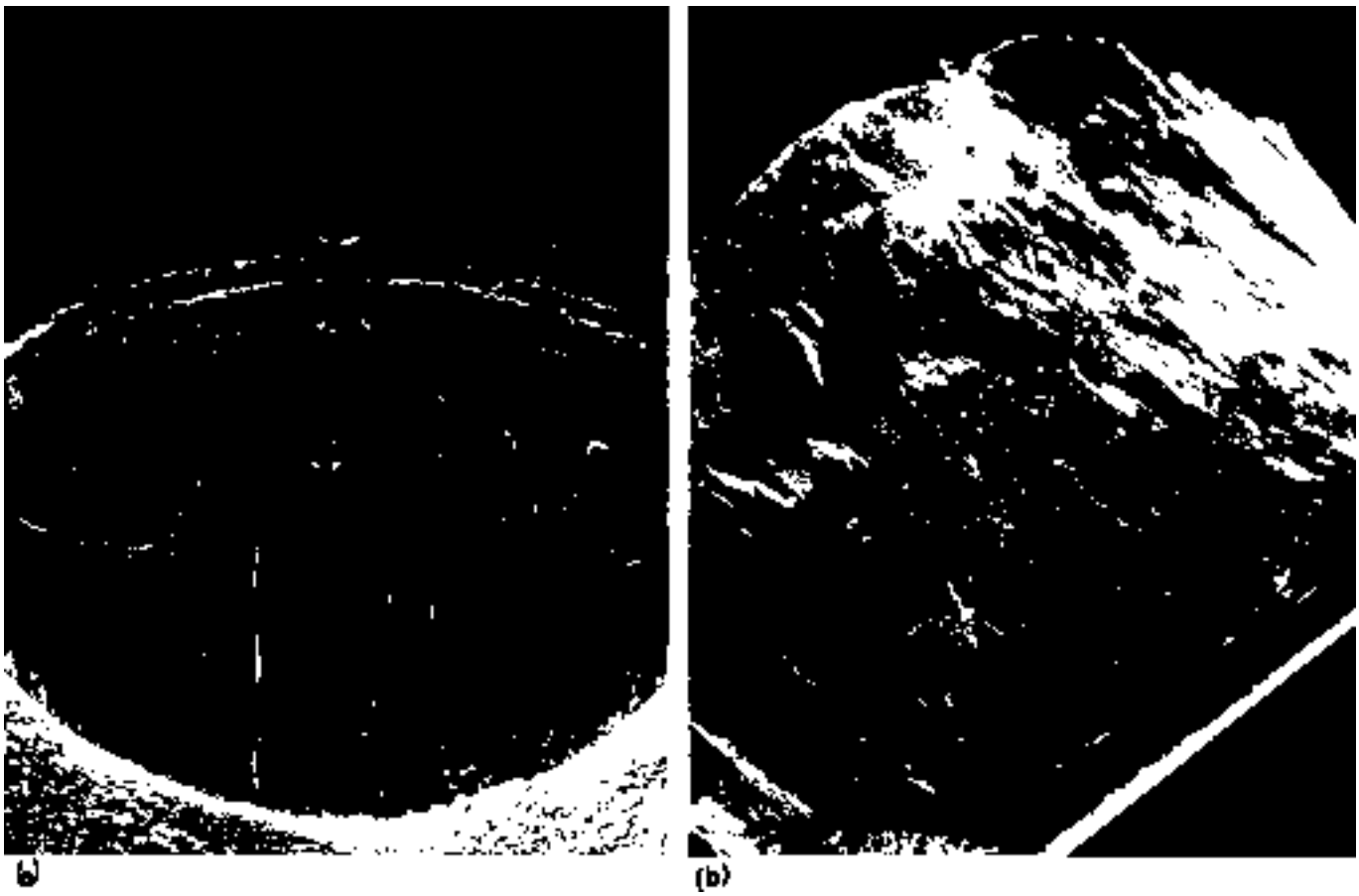


Fig. 24 Sheared surface of a spall cavity (a) and the spall (b). Both 12×

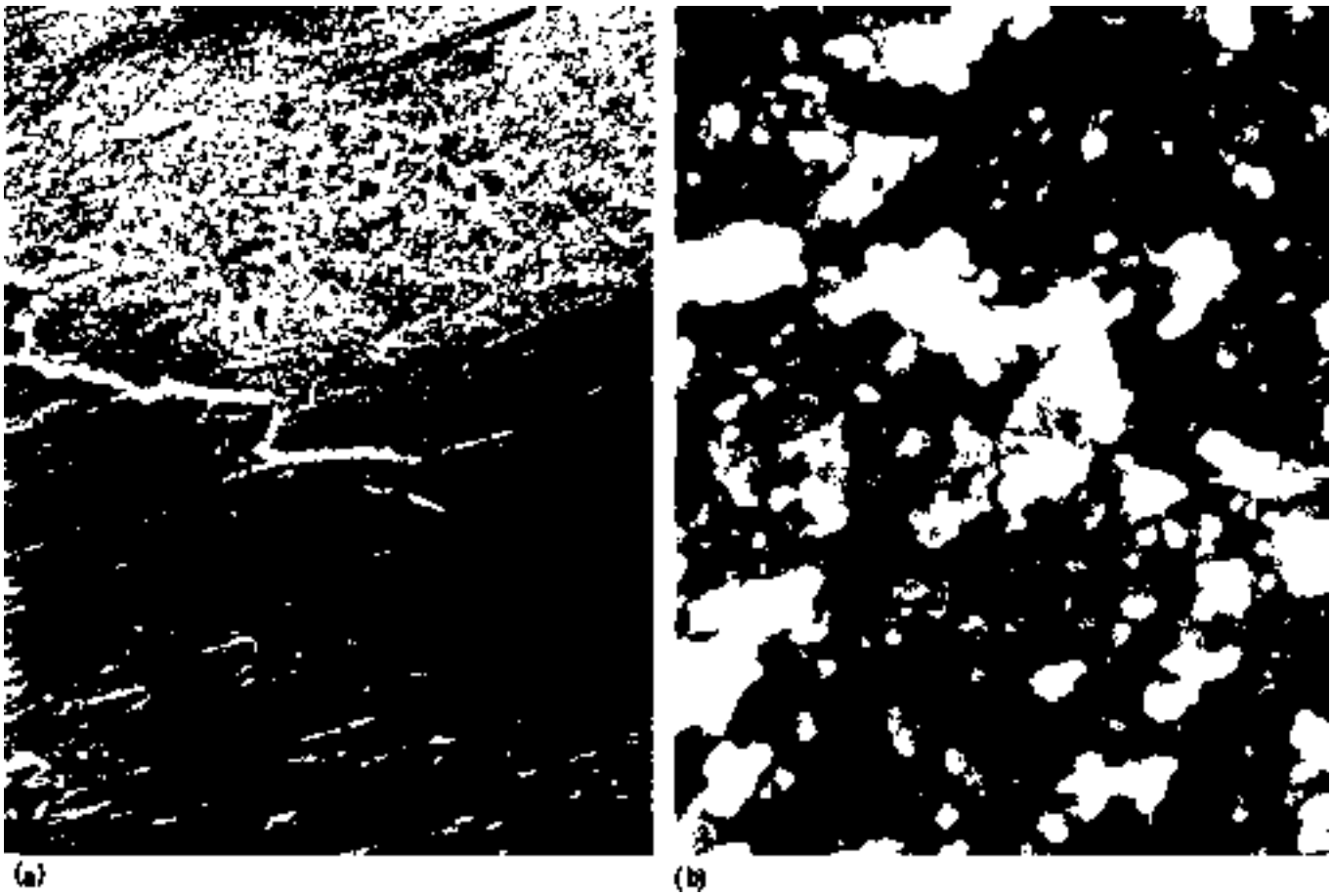


Fig. 25 A SEM micrograph of the sheared surface shown in Fig. 24. (a) 1000×. (b) 11,000×

The geometry of spall cavities is subject to considerable variation. The cavity shown in Fig. 26 has what might be described as a “simple” appearance. The actual geometry of this cavity is not apparent from an examination of the photograph. The portion of the cavity surface labeled “A” is roughly perpendicular to the striking surface of the hammer, whereas the portion labeled “B” approaches being parallel to the striking surface. Thus, the geometry of the surface of this cavity is significantly different than that suggested by Fig. 20. Three cavities that have a similar appearance and are distinctly different from the simple cavity in Fig. 26 are shown in Fig. 27, 28, and 29. The three cavities have in common a sheared or deformed area around the rim of the cavity on the striking surface, and two of the cavities have pieces of the striking surface still attached to the rim. Each of these cavities is now discussed in greater detail.

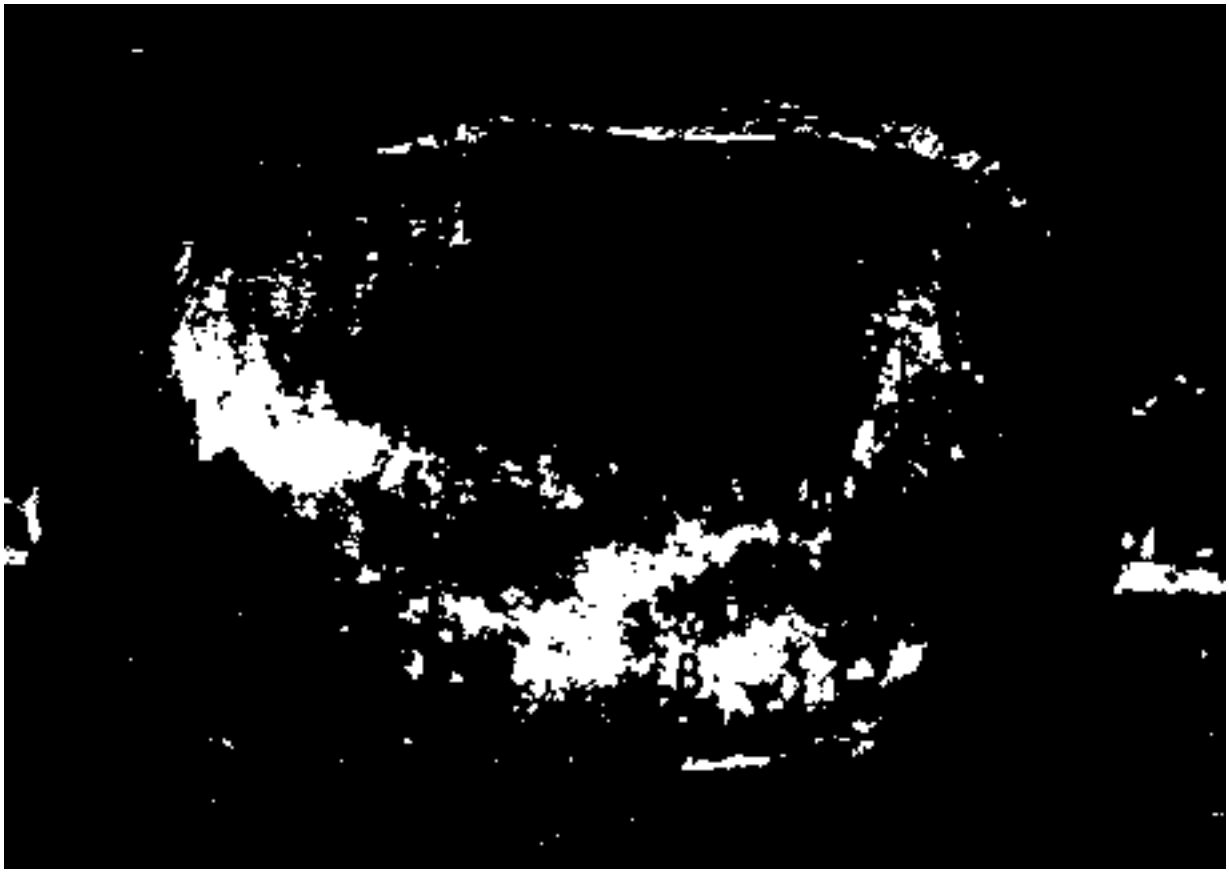


Fig. 26 “Simple” spall cavity, 18×. A, cavity surface roughly perpendicular to the striking surface of the hammer; B, cavity surface almost parallel to the striking surface

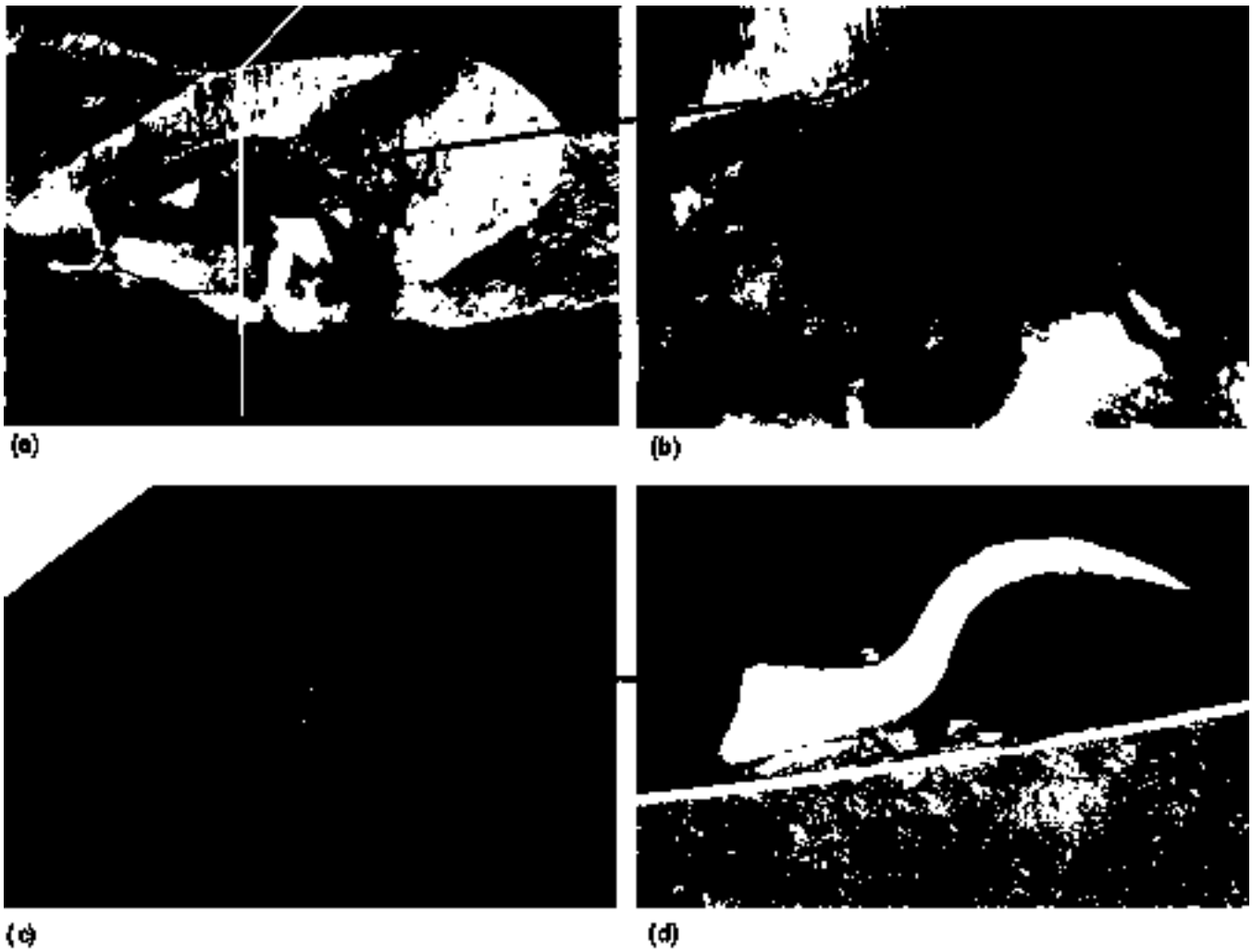


Fig. 27 Spall cavity in a small ball-peen hammer, 49 HRC. (a) 10×. (b) 28×. (c) 28×. (d) 224×

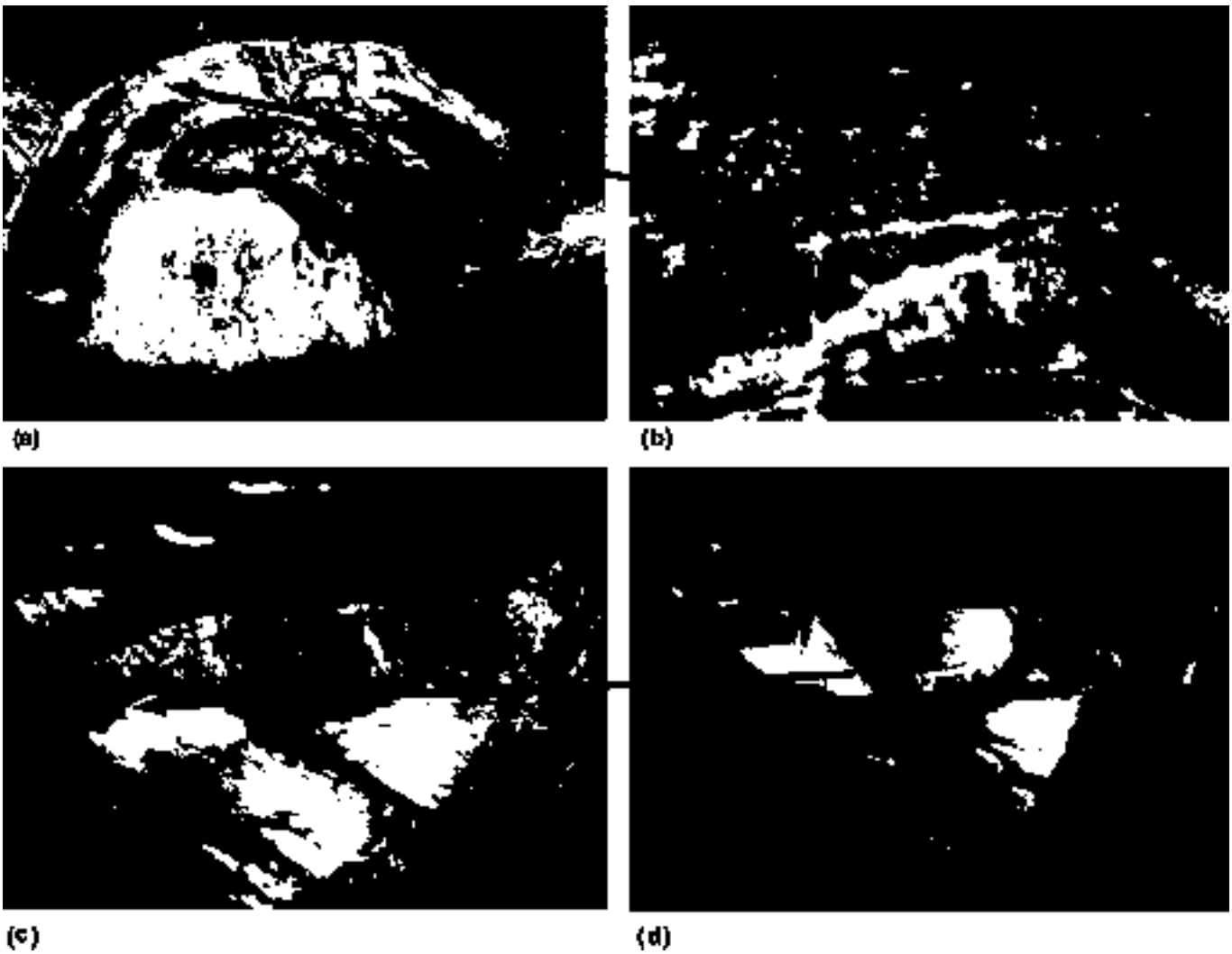


Fig. 28 Spall cavity formed by the action of several shear bands. (a) 7×. (b) 29.5×. (c) 11×. (d) 11×

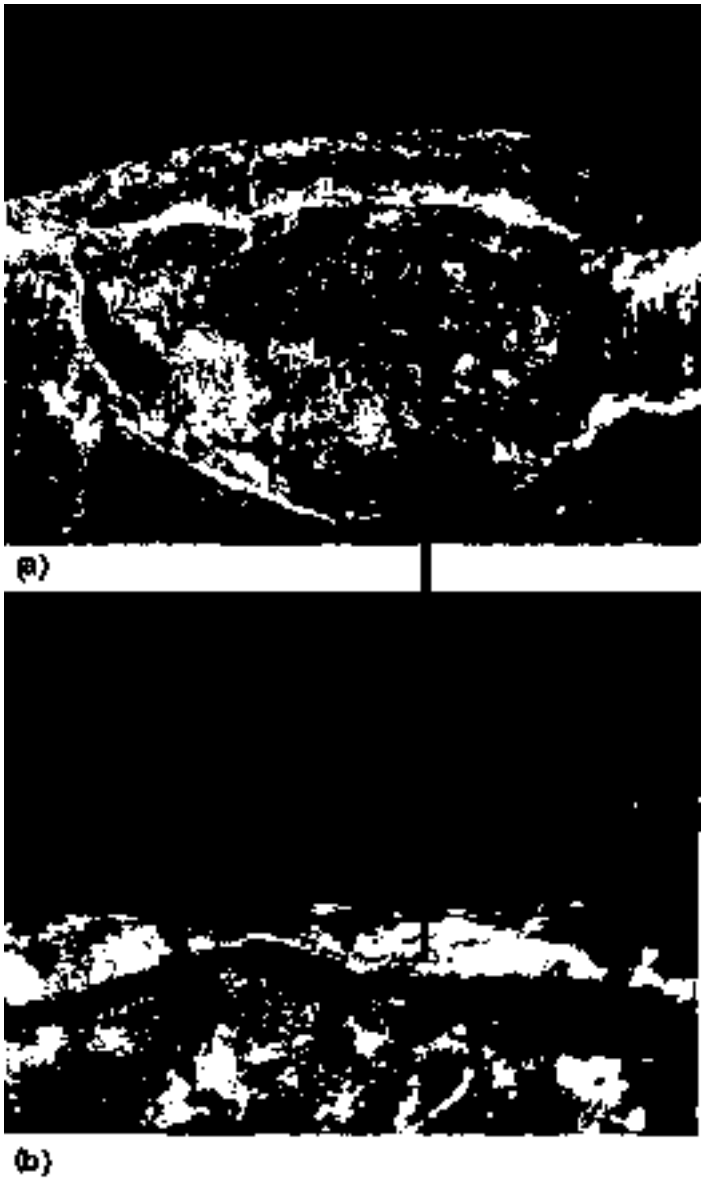


Fig. 29 Spall cavity whose formation involved indentation of the striking face. (a) 5×. (b) 13.5×

The spall cavity shown in Fig. 27 was produced in a small ball-peen hammer, with the diameter of the striking face being 25.4 mm (1 in.) and the hardness of the striking face being 49 HRC. The rim of the cavity, Fig. 27(a), has a sheared appearance and the piece of the striking surface still attached to the rim of the cavity; Fig. 27(b) is darkened by a thin oxide film on its surface. The latter effect was caused by the heat generated by the plastic deformation associated with the spalling process, the minimum temperature being at least in the 430 to 540 °C (800 to 1000 °F) range. It is not uncommon to observe small, oxidized areas in a spall cavity. The cavity was sectioned along the two lines in Fig. 27(a), with the plane of section cut through the end of the oxidized appendage. The polished and etched section through the cavity is shown in Fig. 27(c). An adiabatic shear band (ASB), the light-etching layer of untempered martensite, covers the entire surface of the cavity. The section of the ASB labeled “A” formed along the surface of the rim of the cavity (“A” in Fig. 27[a]) and the section of the ASB labeled “B” covers the major portion of the cavity surface. The profile of the ASB suggests that two separate shears were involved in the spalling process. Additional information on this particular matter could have been obtained if the spall had been available for examination. Note that the tip of the oxidized appendage, Fig. 27(d), was not attacked by the etchant, indicating that its microstructure also is untempered martensite.

The cavities in Fig. 27 and 28 have in common a sheared rim. Observe in Fig. 28(b) that the gouges or grinding marks in the striking surface terminate at the edge of the sheared rim. Two pieces of the striking surface remain attached to the edge of the sheared rim. One end of the larger of the two pieces, Fig. 28(d), was broken off and positioned so that its underside could be examined microscopically. The underside of the end piece is enclosed by the ellipse in Fig. 28(c), and it is apparent that the underside was produced by shear. The formation of this

cavity and the spall undoubtedly required the propagation of several shear bands through the affected region of the hammer head. It is quite possible that the spall may have contained an internal shear band, but the spall was not available for examination.

The spall cavity in Fig. 29(a) also has a deformed rim, but it formed in a different manner than the shear rims in Fig. 27 and 28. Note that the gouges or grinding marks on the striking surface are also visible on the rim; that is, the marks do not terminate at the edge of the rim (see, for example, the mark indicated by the arrow). Thus, the rim appears to be simply an indentation in the striking surface. As can be seen in Fig. 29(b), the rim overhangs the cavity. Consequently, this cavity may have formed by the initial ejection of the spall and subsequent plastic deformation of the striking surface contiguous to the edge of the cavity. Speculation regarding such details of spalling could be limited considerably if cavities could be sectioned for metallographic examination, and likewise with the spalls. Oftentimes, however, striking/struck tools are the subject of litigation, and destructive testing is usually not permitted. Returning for a moment to the simple cavity in Fig. 26, the absence of a sheared or deformed rim is now very apparent, and its absence suggests that the area of impact was confined within the projected area of the spall onto the striking surface. Here again, the availability of the spall for microscopic examination would provide valuable information about this particular matter.

The formation of large spalls is invariably associated with a defective material. Figures 30 and 31 are good examples of one type of defect, quench cracks. In Fig. 30, the area “PC” is a prior crack, in all probability a quench crack, whose surface has not been discolored by oxidation, because the crack did not intersect any of the hammer-head surfaces; that is, it is a totally internal crack. The original boundary of area “PC,” as can be seen in Fig. 30(b), resembles a parabola. However, the quench crack does extend under the sheared surface, “S,” in Fig. 30(b) and under the fracture surface at the left end of the cavity; the separation between the mating surfaces of the quench crack is indicated by the arrows in Fig. 30(b) and (c). Propagation of the original quench crack apparently occurred during use of the tool, because the fracture surface beyond the parabolic boundary of the quench crack does not display much, if any, evidence of shear deformation. Final separation of the spall certainly involved shear over the “S” areas in micrographs (b), (c), and (d). The left end of the cavity, micrograph (c), however, probably was cracked prior to the spalling event, because the fracture surface does not display any evidence of shear. Another example of a quench crack causing a large spall to be ejected from a hammer is shown in Fig. 31. In this case, it is obvious that the defect is a quench crack, because the surface of the crack oxidized during tempering. The contrast between the old and new areas of the fracture surface greatly simplifies the analysis of this kind of failure. Another, but less common, material defect responsible for the formation of large spalls is temper embrittlement. One of the authors (Gordon W. Powell) has had the opportunity to examine three ripping hammers that failed in much the same manner; that is, the chip geometries were essentially the same. These hammers were purchased in a single, boxed lot. The embrittled grain boundaries were revealed metallographically using Rucker's etchant.

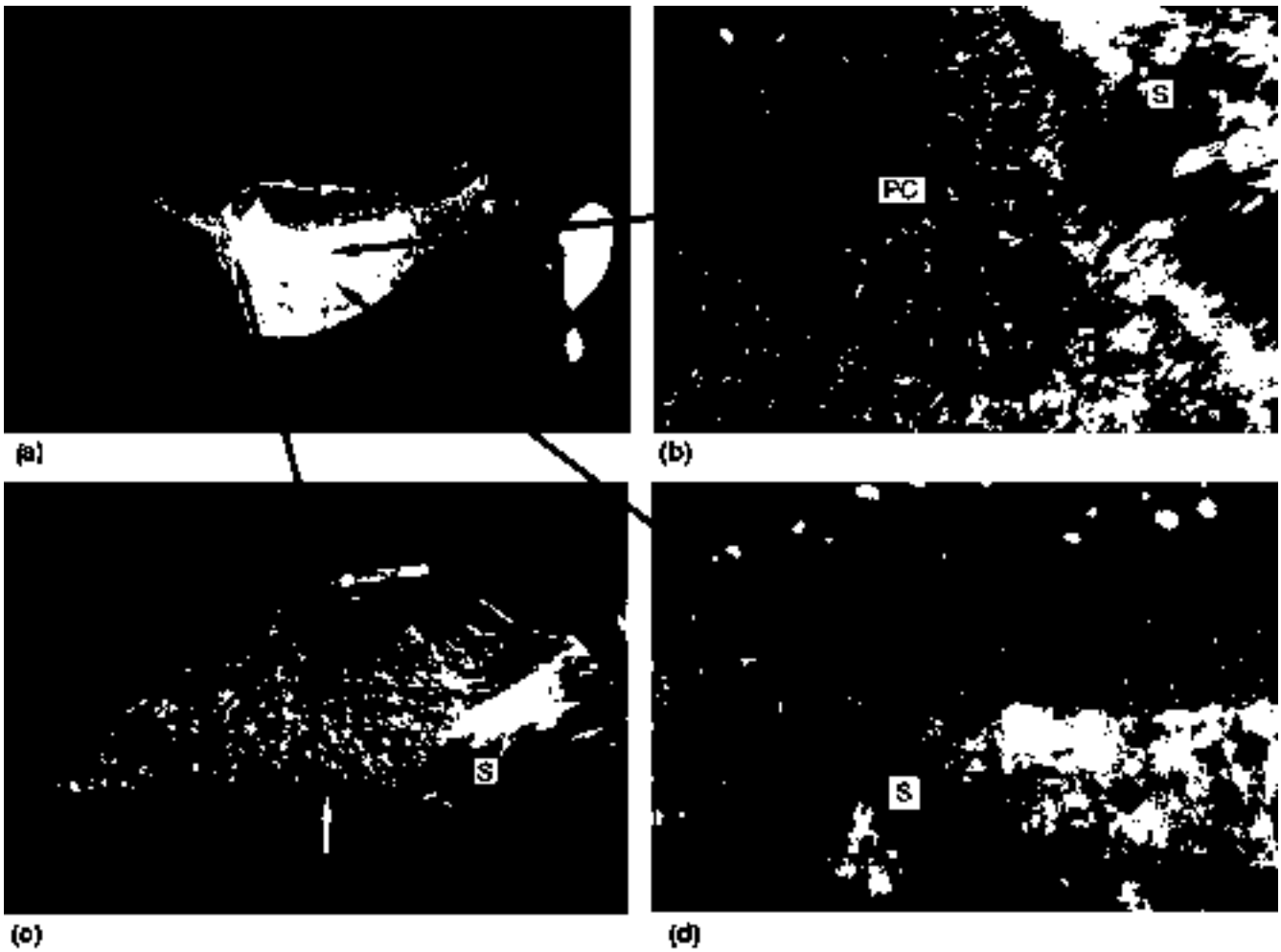


Fig. 30 Spall caused by quench crack. (a) View showing location of spall. (b) 7×. (c) 5×. (d) 28×

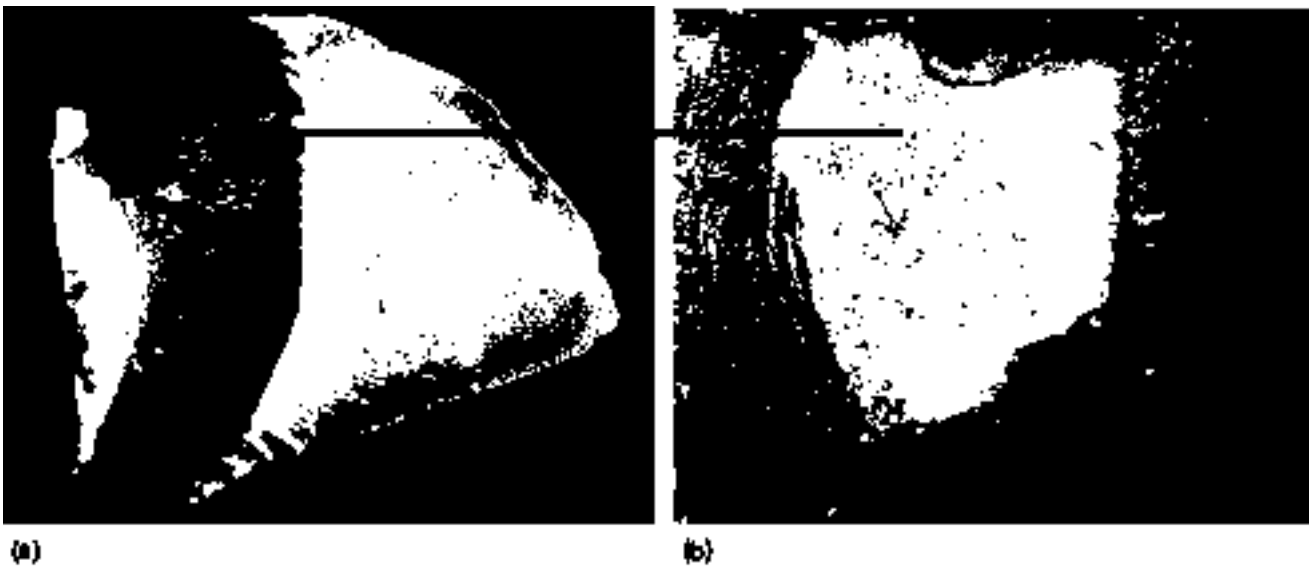


Fig. 31 Another example of spalling caused by a quench crack. (a) 7×. (b) 28×

References cited in this section

5. J.F. Velez and G.W. Powell, Some Metallographic Observations on the Spalling of AISI 1060 Steel by the Formation of Adiabatic Shear Bands, *Wear,* Vol 65–66, 1980–1981, p 367

17. T.A.C. Stock and K.R.L. Thompson, Penetration of Aluminum Alloys by Projectiles, *Metall. Trans.*, 1970, p 219

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

Comments on Specifications for Striking/Struck Tools

Table 1 is a summary of the hardness-related details of six striking/struck tool specifications. The maximum hardness permitted for the striking face by two of the specifications is 60 HRC. Brittle, cleavage fracture tends to occur at slightly higher hardness levels, and, therefore, a hardness of 60 HRC is very close, perhaps too close, to exceeding the safety limit for a striking tool. In principle, the boundary between a safe and an unsafe upper hardness limit is not discrete but more transitional in character. The lower limit of the hardness of the striking face is usually about 50 HRC, but, in the case where the tool can be used to strike another steel tool, such as a wedge, the lower limit is reduced to about 45 HRC. At this hardness, that is, 45 HRC, the strain-hardening exponent is greater, the deformation more wide spread, and thus the tool tends to mushroom, thereby providing the user of the tool with a warning that spalling may be imminent.

Table 1 Striking/struck tool specifications

Specification	Type of tool	Composition of steel [(a)]	Hardness, HRC	Striking test
AREMA	Percussion track tools	9260A: 0.56–64C-1.8–2.0Si-0.75–1.0Mn 9260B: 0.51–0.60C-1.8–2.2Si-0.75–1.0Mn-0.35–50Mo-0.45Si max	Striking face: 51–55 Struck face: 44–48 Cutting edge: 56–60 Point end: 52–56	None
ANSI B173.3-1991	Heavy striking tools	0.45–0.88C-0.3–1.2Mn-0.35 Si max (ASTM A576, A322, A681)	Striking face: 45–60 Wood chopper's maul: 44–55 Peen: 45–60	Block hardness: 92–105 HRB
BS 876-1981	Hammers	0.5–0.6C-0.5–0.9Mn-0.1–0.4Si 0.2–0.3Cr for head mass > 1.8 kg (4 lb)	Striking face: 50–58 3 mm below striking face: 46	Normalized steel block Same steel as hammer Six blows to periphery of face
GGG-H-86C-1963	Hammers	0.46–0.85C-0.2–0.9Mn-0.1–3Si	Double-face hammer: 44–55 Carpenter's hammer: 50–60 Ball-peen	<60 cm (25 in.) handle: 92–105 HRB >60 cm (25 in.) handle: 40–45 HRC

			hammer: 50–57	
ANSI B209.3-1990	Wood splitting wedges	Carbon steel (ASTM A29, A576, A681)	35 max	Drop-weight test, 4.5 kg (10 lb) from 1.5 m (5 ft)
ANSI B209.1-1991	Chisels	Carbon or alloy steel (ASTM A29, A322, A331, A576, A681)	Struck face: 45 max Cutting edge: 53–60	Drop-weight test

Note: All railroad striking/struck tools are currently manufactured from the grade B steel. Nonpercussion tools, such as claw bars, rail tongs, rail forks, etc., are manufactured using plain carbon steel.

(a) A29/A29M: Standard Specification for Steel Bars, Carbon and Alloy, Hot-Wrought and Cold-Finished; A322: Standard Specification for Steel Bars, Alloy, Standard Grades; A331: Standard Specification for Steel Bars, Alloy, Cold-Finished; A576: Standard Specification for Steel Bars, Carbon, Hot-Wrought, Special Quality; A681: Standard Specification for Tool Steels, Alloy

According to Table 1, the maximum allowable hardness of a struck tool is lower than that of a striking tool. Therefore, the struck tool is capable of undergoing more macroscopic plastic deformation, and this deformation increases the contact area between the struck and the striking tools during impact. The increased area of contact distributes the stresses in the striking tool over a greater volume, thereby reducing the probability of spalling. The struck tool is more susceptible to mushrooming, but, as suggested previously, the mushrooming provides a measure of safety in that it is a warning of imminent spalling. It is appropriate at this point to consider an aspect of the striking test cited in Table 1. Two of the specifications, ANSI B173.3-1991 and GGG-H-86C-1963, require that a striking tool be struck against a steel block having a hardness of 92 to 105 HRB, whereas BS 876-1981 states that a normalized steel block be used for the striking test. These hardness levels are quite low. Consequently, the steel blocks should undergo significant plastic deformation when impacted by the striking tool, and, according to the line of reasoning expressed earlier, the probability of the striking tool spalling is reduced significantly. Recall that McIntire and Manning were unable to spall hammers with an impact energy of 190 J (140 ft · lbf) struck against an anvil having a hardness of 24 HRC. Thus, even though ANSI B173.3 calls for “100 full swinging blows by a man of average build,” the striking tests using relatively soft anvils favor the striking tool. On the other hand, however, Federal Specification GGG-H-86C requires the use of an anvil with a hardness of 40 to 45 HRC if the length of the handle of the hammer is greater than 635 mm (25 in.). Clearly, this test represents a more severe test of the striking tool. However, each of the tools previously covered by this specification, that is, GGG-H-86C, is now the subject of a Commercial Item Description (CID); for example, the double-face hammer is covered by CID A-A-1293B. The CIDs require that the tools conform to an ANSI specification, B173.3, in the case of double-face hammers, and that the tools be subjected to magnetic-particle or ultrasonic testing. The striking test using an anvil with a hardness of 40 to 45 HRC has been replaced by the ANSI specifications that call for the use of a much softer anvil, 92 to 105 HRB.

An important aspect of the striking test that must be considered is the relative orientation of the striking and struck tools immediately prior to impact. The description of the striking test in ANSI B173.3-1991 does not indicate how the blows are to be delivered to the steel block. However, under the heading “Safety Requirements and Limitations of Use,” this specification states that “A blow of a heavy striking tool should always be struck squarely with the striking face parallel with the surface being struck. Glancing blows and overstrikes and understrikes should be avoided.” If the striking test is conducted in this manner, then it is highly unlikely that the striking tool will be damaged, unless the tool contains a quench crack or it has been temper embrittled. McIntire and Manning could not cause any of the hammers to spall when the tool impacted squarely against an anvil. More specifically, McIntire and Manning concluded that “Good quality hammers can be struck severe blows against surfaces as hard as 60 HRC without spalling, providing the face of the hammer meets the surface squarely or at a low angle (up to approximately 20° for nail and ball-peen hammers and 2° for tinner's hammers).” The striking test prescribed by British hammer standard BS 876-1981 is a considerably more realistic test in the sense that it calls for six full, sharp, consecutive blows followed by the instruction that “a further six blows shall be struck, varying the angle of delivery, to bring as much as possible of the periphery of the face or pein into contact with the bar.” However, it must be pointed out that this specification states that this striking test is intended to determine the security of the head to the handle and is apparently not intended to provide some measure of the resistance to spalling of the hammer head. A striking test performed by an

individual would undoubtedly have some degree of randomness in the relative orientation of the striking and struck tools prior to impact, but a striking test performed using a HIT machine, for example, would be subject to significantly more control. Thus, if the purpose(s) of a striking test is precisely defined, then the information sought would best be obtained using a device such as a HIT machine, because the geometry and the energy of the impact may be controlled and varied independently.

Finally, with regard to the heat treatment of striking tools, specifications BS 876-1991 and ANSI B173.3-1991 state that the working ends of the tool are to be heat treated separately, that is, a whole-head heat treatment is not to be used. According to the former specification, the hardened zone shall not extend into a neck or, where there is no neck, to more than halfway to the near edge of the eye, and, in the case of ANSI B173.3-1991, the poll (the steel immediately behind the hardened surface layer) shall be a toughened supporting core. However, in the case of railroad striking tools (sledges, spike mauls), the entire head is heat treated. The hard-surface/tough-core concept has been considered to be the ideal structure for a striking tool. As mentioned earlier, J.W. Lodge did not attach much significance to this concept, because, as he observed, spalling always occurs in the hardened region around the periphery of the striking face, with the core not being involved in the process. Furthermore, the AREMA alloy 9260B (see Table 1), which is a deep-hardening steel, is used for percussion track tools. The relatively high hardenability of this steel permits striking tools to be redressed without subsequent heat treatment. According to the AREMA manual, “All tools made with alloy steel to be redressed without subsequent heat treatment shall be initially heat treated so that the hardness specified in article 6.1.2.3.1 is maintained to a depth from the end not less than the average cross-sectional thickness.” Clearly, the use of this steel violates the hard-surface/tough-core concept, but field experience has demonstrated that the performance of these tools is quite satisfactory. As reported earlier, A.H. Burn's hammer tests included a 0.65C-1Si-3Ni steel, and J.W. Lodge's tests included 4340 steel; these steels, which are definitely deep-hardening steels, did not perform any better or worse than the other steels tested. Consequently, these results raise the issue as to whether or not the hard-surface/tough-core concept is a determinative factor for the performance of striking tools.

Spalling from Impact Events

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

---

New Materials for More Spall-Resistant Tools

As is evident in earlier discussion, testing different materials to evaluate their tendency for spalling would be a major undertaking. In order to quickly evaluate potential opportunities, a theoretical or predictive analysis was conducted. The basis for this analysis and the results are described subsequently.

Localized deformation is invariably associated with spall formation. This localized deformation mechanism (ASB) plays a major role in several other deformation processes. Adiabatic shear band formation has been identified as a cause of early failures in ballistic armor and has been extensively studied in order to develop methods to predict adiabatic shear formation from materials properties (Ref 18). Similar models have been used to predict the onset of adiabatic shear in machining operations (Ref 19). Unlike the detrimental effects for armor applications, adiabatic shear in machining is desirable, because it leads to reduction of work for metal chip formation.

The fundamental premise of these predictive models for the onset of adiabatic shear is that localized plastic flow initiates when thermal softening, caused by temperature increases associated with the work-energy release of plastic deformation, matches or surpasses work-hardening and strain-rate effects of higher strain rates associated with flow localization. A model to predict flow localization via shear band formation has been proposed (at least for plane-strain deformation) and is extensively discussed in Ref 20 with some supporting experimental data. The model includes, as a primary variable, the strain-rate hardening exponent, *m*. Shear band formation is predicted when the following parameter α becomes equal to 5:

$$\alpha = \frac{\delta \ln \dot{\varepsilon}}{\delta \bar{\varepsilon}} = \frac{\left(\frac{\partial \ln \bar{\sigma}}{\partial \bar{\varepsilon}}\right)\Big|_{\dot{\varepsilon}}}{m} = \frac{\gamma'}{m} = 5$$

where, γ' is a temperature-compensated strain-hardening term, m is the strain-rate sensitivity exponent, and $\bar{\sigma}$ and $\bar{\varepsilon}$ are average stress and strain, respectively.

The critical mechanical properties controlling increases in flow stress are the strain-hardening exponent (n) and the strain-rate sensitivity index (m). The critical physical and mechanical properties controlling thermal softening are density, specific heat, and change of flow stress with temperature. All of these physical and mechanical properties are very similar for steels quenched and tempered to the same hardness. Therefore, it is not surprising that neither Burn (Ref 3) nor Lodge (Ref 4) found any difference in spalling tendency for different steels with similar hardness.

A literature review of tests used to measure the resistance of materials to ASB formation was conducted in an attempt to identify new alloys with more resistance to the formation of adiabatic shear and thus, more resistance to spalling. The evaluation revealed two alloys with properties that suggested they would resist adiabatic shear: a stainless steel with unusually high work-hardening characteristics (TRIP steel), and a nickel-base superalloy (IN-718) with a very low rate of strength reduction with increased temperature and low thermal softening. Experimental evaluation of these two materials in ballistic tests concluded that they performed no better than quenched-and-tempered steel. This suggests that they would not resist spalling any better than conventional quenched-and-tempered steels.

Review of the literature for ballistic testing revealed a relatively new steel alloy that has some improvement over conventional quenched and tempered steel. This material is a high-purity, highly alloyed steel that depends on second-phase particle strengthening to attain high strength (AerMet 100). The ballistic response improvement of this alloy over conventional quenched-and-tempered steels was measurable but not outstanding. The potential improvement in a percussion application is not clear, due to the lack of any correlation between ballistic response and spall formation. A spalling test evaluation of this material may be warranted, based on its ballistic response. To date, however, the open literature does not describe any new materials that have clearly demonstrated spall resistance superior to the steels currently being used.

References cited in this section

3. H.O. McIntire and G.K. Manning, “A Study of the Spalling of Hand Hammers,” Battelle Memorial Institute, unpublished report, 19 May 1958

4. J.W. Lodge, “Tests Simulating the Edge Chipping of Hammers,” Youngstown Sheet and Tube Company, unpublished report, and presented at meeting of the Hand Tool Institute (North Key Largo, FL), 1973–1975

18. P.C. Chou, J. Heshemi, A. Chou, and H.C. Rogers, Experimentation and Finite Element Simulation of Adiabatic Shear Bands in Controlled Penetration Impact, *Int. J. Impact Eng.*, Vol. 11 (No. 3), 1991, p 305–321

19. J.Q. Xie, A.E. Bayoumi, and H.M. Zbib, Analytical and Experimental Study of Shear Localization in Chip Formation in Orthogonal Machining, *J. Mater. Eng. Perform.*, Vol 4 (No. 1), 1995, p 32–39

20. S. Semiatin and J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization*, American Society for Metals, 1984, p 71

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

Conclusions

The most significant conclusion derived from this review is that the spalling behavior of plain carbon and low-alloy steels quenched and tempered to the same hardness is essentially the same. That is, low-alloy steels are not more resistant to spalling than plain carbon steels. Consequently, tool geometry then becomes a very important factor in limiting the incidence of spalling.

Some of the specifications for striking tools call for a striking face with a relatively thin, hardened surface layer that is supported by a softer, tough core. However, the test results obtained by A.H. Burn and by J.W. Lodge with high-hardenability steels and also the current use of 9260B for railroad tools clearly indicate that this requirement is unnecessary and that high-hardenability steels can be used to manufacture good-quality tools. With regard to striking tests, as called for by some specifications, the work of McIntire and Manning demonstrates that the use of anvils of low hardness (<25 HRC) does not constitute a good test of the spalling resistance of a striking tool, and, furthermore, any requirement that the striking tool impact the anvil with a flush blow diminishes significantly the value of the test.

The use of highly alloyed steels to reduce or eliminate spalling does not appear to be warranted. A review of predictive models related to spalling suggests that highly alloyed or high-cleanliness steels will not significantly reduce spalling potential. However, a definitive investigation of spalling has not been performed using these alloys.

Spalling from Impact Events

Dennis L. McGarry, FTI/SEA Consulting, Inc.; Gordon W. Powell, The Ohio State University; Cameron Lonsdale, Standard Steel

---

References

1. A.H. Burn, “Chipping of Hand Hammers,” *Engineering,* 12 July 1957

2. A.H. Burn, “The Resistance to Chipping of 1% Cr-Mo Steel Hand Hammers,” internal report, Imperial Chemical Industries Ltd., April 1959

3. H.O. McIntire and G.K. Manning, “A Study of the Spalling of Hand Hammers,” Battelle Memorial Institute, unpublished report, 19 May 1958

4. J.W. Lodge, “Tests Simulating the Edge Chipping of Hammers,” Youngstown Sheet and Tube Company, unpublished report, and presented at meeting of the Hand Tool Institute (North Key Largo, FL), 1973–1975

5. J.F. Velez and G.W. Powell, Some Metallographic Observations on the Spalling of AISI 1060 Steel by the Formation of Adiabatic Shear Bands, *Wear,* Vol 65–66, 1980–1981, p 367

6. A.H. Burn, “Practical Trial of $1\frac{1}{2}$ lb Pein Hand Hammers in Engineering Workshops,” internal report, Imperial Chemical Industries Ltd., July 1962

7. J.W. Lodge, “An Investigation of the Mode of Failure by Tests Simulating the Chipping Hand Hammers,” unpublished report

8. J.D. Olson, Stanley Works, private communication, 1 Dec 1988

9. J.W. Lodge, speech given to the Hand Tools Institute Hammer Subcommittee, (Cleveland, OH), 8 Nov 1973

10. American Railway Engineering and Maintenance-of-Way Association, “Manual for Railway Engineering,” 2000, p 5, 6, 54

11. W.G. Patton, *Met. Prog.*, Vol 43, 1943, p 726

12. E.J. Janitzky and M. Baeyertz, *Alloying Elements in Steel*, E.C. Bain and H.W. Paxton, Ed., American Society for Metals, 1948, p 225

13. *Metals Handbook, 1948 Edition*, American Society for Metals, 1948, p 456

14. *Properties and Selection: Irons and Steels*, Vol 1, *Metals Handbook*, 9th ed., 1978, p 461

15. A. Nadai, *Theory of Flow and Fracture of Solids*, McGraw-Hill, New York, 1959, p 541

16. H.O. McIntire and A.M. Hall, “The Performance of Hand-Striking Tools Under Simulated Service Conditions,” Battelle Memorial Institute, unpublished report, 30 June 1962

17. T.A.C. Stock and K.R.L. Thompson, Penetration of Aluminum Alloys by Projectiles, *Metall. Trans.*, 1970, p 219

18. P.C. Chou, J. Heshemi, A. Chou, and H.C. Rogers, Experimentation and Finite Element Simulation of Adiabatic Shear Bands in Controlled Penetration Impact, *Int. J. Impact Eng.*, Vol. 11 (No. 3), 1991, p 305–321

19. J.Q. Xie, A.E. Bayoumi, and H.M. Zbib, Analytical and Experimental Study of Shear Localization in Chip Formation in Orthogonal Machining, *J. Mater. Eng. Perform.*, Vol 4 (No. 1), 1995, p 32–39

20. S. Semiatin and J. Jonas, *Formability and Workability of Metals: Plastic Instability and Flow Localization*, American Society for Metals, 1984, p 71

Corrosive Wear Failures

Introduction

CORROSIVE WEAR is defined as the degradation of materials in which both corrosion and wear mechanisms are involved. The combined effects of wear and corrosion can result in total material losses that are much greater than the additive effects of each process taken alone, which indicates synergism between the two processes. This article focuses on the corrosion-wear synergism in aqueous slurry and grinding environments. Examples of corrosive wear problems encountered in industry are provided, and the effects of environmental factors on corrosive wear are discussed. Experimental measurement of corrosive wear is discussed in more detail in the article “Corrosive Wear” in *Friction, Lubrication, and Wear Technology*, Volume 18 of *ASM Handbook.*

Occurrences in Practice

Corrosion and wear damage to materials, both directly and indirectly, cost the nation hundreds of billions of dollars annually. Although corrosion can often occur in the absence of mechanical wear, the converse is rarely true. Corrosion accompanies the wear process to some extent in all environments, except in a vacuum and inert atmospheres. Corrosion and wear often combine to cause aggressive damage in a number of industries, such as mining, mineral processing, chemical processing, and energy production. Corrosion and wear processes involve many mechanisms, the combined actions of which lead to the mutual reinforcement of their effectiveness. Seventeen synergistic relationships between abrasion, impact, and corrosion that could significantly increase wear damage in wet and aqueous environments have been identified (Ref 1).

Slurry Handling. The movement of materials as a slurry is an efficient means of transportation that is used in many industries. However, the movement of these slurries can cause significant corrosive wear. Pumps, elbows, tee junctions, valves, flotation cells, and hydrocyclones are parts of slurry-handling systems that are subject to corrosive wear. The periodic replacement of worn parts often results in costly process downtime. Pipe rotation has been considered good practice for those pipes that are subject to localized wear, but careful monitoring is necessary to prevent failure of the worn parts because of corrosion (Ref 2). Turbulence and eddies are promoted either where fittings or valves project into the mainstream flow or where the diameter of a pipe changes. The momentum of the particles is sufficient to cause abrasive damage to the side walls of the pipe in its curves or elbows, where the change in direction does not produce eddies. Slurry particles can impart damage to the bottom of the pipe by tumbling or sliding in straight pipe sections.

Sliding wear can occur, for example, in an ore chute in a minerals processing plant, where moist ore particles slide on an inclined platform, chute, or screen. Figure 1 shows a worn wire-cloth screen in a minerals processing plant. Another example is a submerged valve that opens and closes to control the flow of material. Here, the parts are subjected to metal rubbing, as well as to the abrasive and corrosive nature of a slurry. Other, more common corrosive wear problems occur in equipment such as automobile brakes and engines (Ref 3), fillings used in dental applications, and tapes rubbed by recording heads in electronic video and audio equipment. Each of these wear examples has a corrosive component that results from exposure to reactive environments.

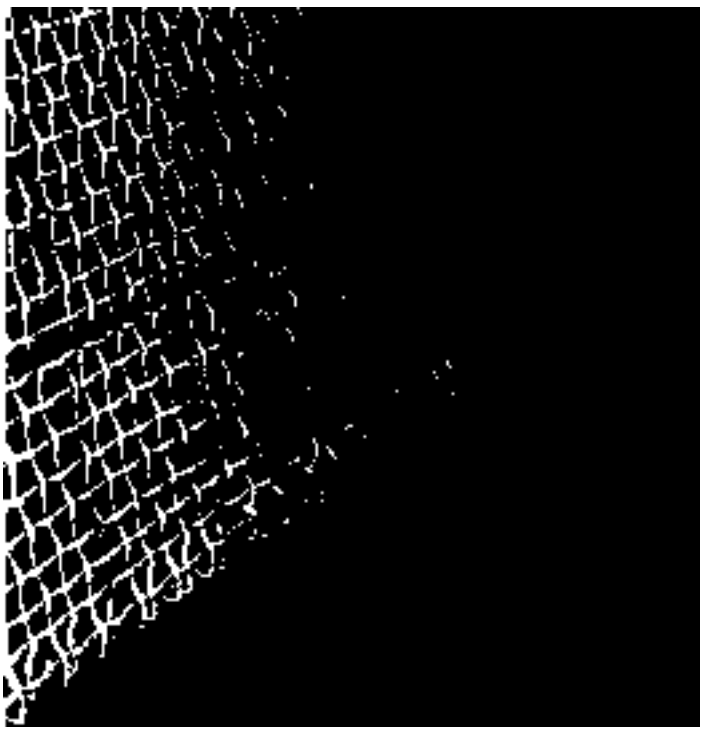


Fig. 1 Worn wire screen in minerals processing plant showing effects of sliding wear in moist environment

Crushing and Grinding. Corrosive wear is prevalent in the grinding of mineral ores. The annual cost of grinding media frequently approaches or exceeds the cost of energy in some processes (Ref 4). Figure 2 shows ball mill liner plates in both new and worn conditions. Figure 3 shows a cone crusher liner that wore completely through after only three months of use.

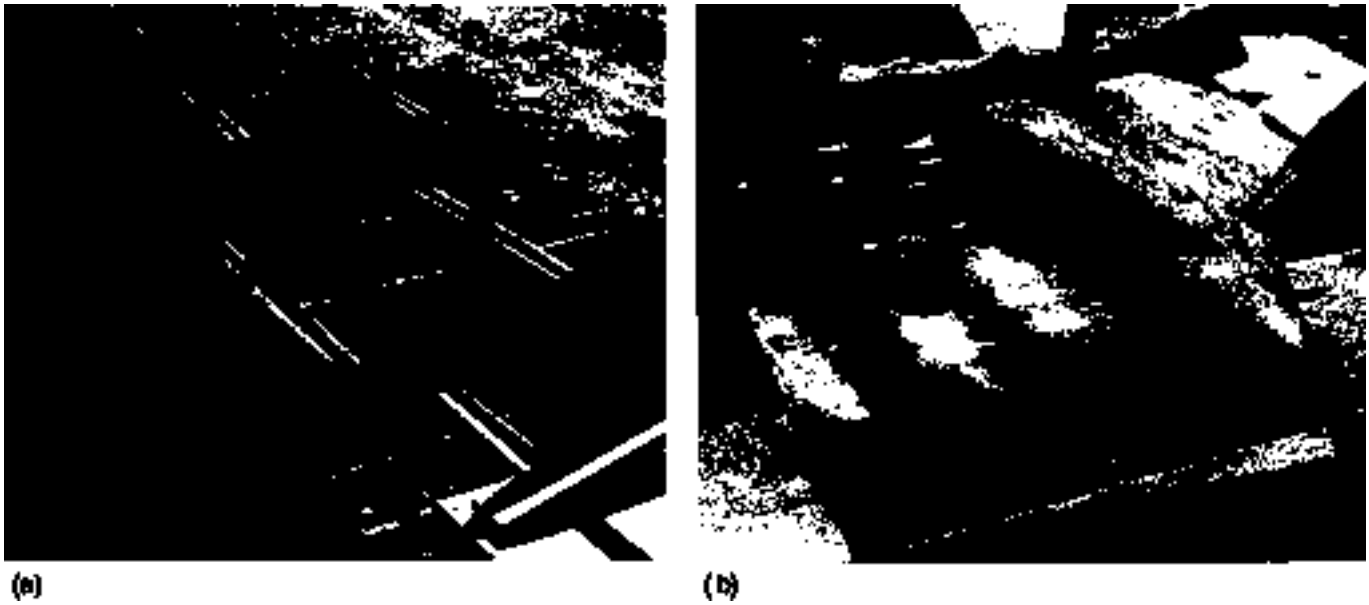


Fig. 2 Ball mill liner plates. (a) New condition, before installation. (b) Worn condition, with partially worn grinding belt

Fig. 3 Worn cone crusher liner after three months of use

Corrosion and wear damage arising from comminution is equivalent to 12% of more than 30 billion kW·h of energy consumed annually by U.S. industry for crushing and grinding purposes, whereas contributions from minerals account for 50%; cement, 25%; coal, 13%; and agricultural products, 12% (Ref 5). The development of more efficient mining, transportation, and processing systems has allowed exploitation of many large, low-grade ore bodies, which involves the crushing, grinding, and treatment of huge volumes of abrasive material to satisfy modern demand for materials (Ref 6). This trend will increase the future consumption of grinding media in an industry where over 0.9 Mg (1×10^9 tons) of ore are crushed annually (Ref 5).

Wear debris and corrosion products that are formed during comminution affect product quality and can adversely affect subsequent beneficiation by altering the chemical and electrochemical properties of the mineral system (Ref 4, 7, 8, 9). Electrochemical interactions between minerals and grinding media can occur, causing galvanic coupling that leads to increased corrosive wear. Corrosion and wear are also important in nonmining industries, such as the pulp and paper industry, where physical processing steps such as the grinding of wood chips are increasingly being substituted for chemical processing steps in order to reduce water-treatment requirements and fulfill regulatory standards.

High-Temperature Processes. Many chemical processes take place at elevated temperatures and involve corrosive wear. A process such as coal gasification involves hot gases with entrained solid particles that impinge on the containment vessel surface. Several studies have been conducted to describe the wear-corrosion

synergism that takes place during the oxide formation and subsequent removal by abrasive particles (Ref 10, 11, 12).

Power-generation plants use processes that occur without any particles present and involve only the transport of solutions or steam. In these cases, liquid forces on the solid surface, which are due to turbulence or droplet impacts, mechanically remove protective layers of corrosion products, thus exposing more base material to corrosive action. An example of this type of process is the erosion-corrosion of steam turbines. Other systems may involve the dissolution of the protective oxide layer, because of a continual flow of liquid past the surface. Cavitation is also a problem where the high-speed flow of liquid is present. This subject is discussed in the article "Cavitation Erosion" in this Volume.

References cited in this section

1. D.J. Dunn, Metal Removal Mechanisms Comprising Wear in Mineral Processing, *Wear of Materials*, K.C. Ludema, Ed., American Society of Mechanical Engineers, 1985, p 501–508

2. A.M.F. Carter and D. Howarth, "A Literature Review of the Factors That Influence the Wear of Slurry-Handling Systems," Report M319, Council for Mineral Technology, Sept 1987

3. F.H. Stott, J.E. Breakell, and R.C. Newman, The Corrosive Wear of Cast Iron Under Potentiostatically-Controlled Conditions in Sulfuric Acid Solutions, *Corros. Sci.*, Vol 30, 1990, p 813–830

4. R.L. Pozzo and I. Iwasaki, Pyrite-Pyrrhotite Grinding Media Interactions and Their Effects on Media Wear and Flotation, *J. Electrochem. Soc.*, Vol 136 (No. 6), 1989, p 1734–1740

5. National Materials Advisory Board, "Comminution and Energy Consumption," NMAB-364, National Academy Press, 1981

6. K. Adam, K.A. Natarajan, S.C. Riemer, and I. Iwasaki, Electrochemical Aspects of Grinding Media-Mineral Interaction in Sulfide Ore Grinding, *Corrosion*, Vol 42 (No. 8), 1986, p 440–446

7. R.L. Pozzo and I. Iwasaki, Effect of Pyrite and Pyrrhotite on the Corrosive Wear of Grinding Media, *Miner. Metall. Process.*, Aug 1987, p 166–171

8. K.A. Natarajan, S.C. Riemer, and I. Iwasaki, Influence of Pyrrhotite on the Corrosive Wear of Grinding Balls in Magnetite Ore Grinding, *Int. J. Miner. Process.*, Vol 13, 1984, p 73–81

9. R.L. Pozzo and I. Iwasaki, An Electrochemical Study of Pyrrhotite-Grinding Media Interaction Under Abrasive Conditions, *Corrosion*, Vol 43 (No. 3), 1987, p 159–169

10. A.V. Levy and Y. Man, "Erosion-Corrosion Mechanisms and Rates in Fe-Cr Steels," Paper 111, Corrosion 86 (Houston), National Association of Corrosion Engineers, 1986

11. S. Agarwal and M.A.H. Howes, Erosion/Corrosion of Materials in High-Temperature Environments, *Proc. AIME Conf. on High Temperature Corrosion in Energy Systems* (Detroit), American Institute of Mining, Metallurgical, and Petroleum Engineers, Sept 1984

12. T. Foley and A.V. Levy, The Erosion of Heat Treated Steels, *Wear*, Vol 48 (No. 1), 1983, p 181

Effect of Environmental Factors on Corrosive Wear

Environmental factors affect corrosive wear in materials-handling systems. Some of the more important factors that influence the material losses in slurry transport and grinding operations in aqueous media and ambient conditions are discussed subsequently.

Slurry Particle Impingement on Two-Body Corrosive Wear

Slurry particles that strike the target material under the influence of their own momentum and that of the carrier fluid impart two-body wear damage to component parts. They are not constrained by another solid, as distinguished from grinding, which is discussed later.

Abrasive particle characteristics, such as shape, density, size, hardness, and their relation to wear rate, are discussed subsequently.

Particle Shape. Until recently, no investigation had directly reported the effect of particle shape on wear in slurry-handling systems. However, it was commonly agreed that the angularity of the slurry abrasives was important in determining the wear rate of materials where corrosion was not dominant. When slurries are recycled in a closed slurry pot system, the wear rate can decrease dramatically with time (Ref 13, 14, 15). This is due to the microscopic rounding of the abrasive particles. The difference in wear rates for recycled and flow-through slurries is shown in Fig. 4 for type A514 steel in a water/silica sand slurry containing 50 × 70 mesh particles.

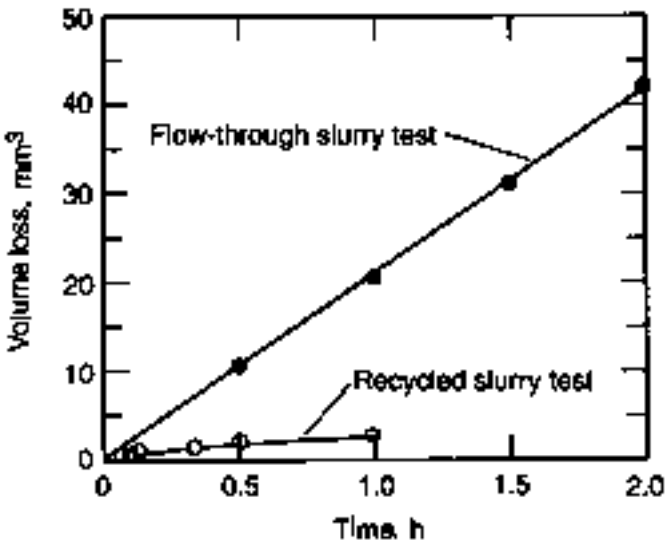


Fig. 4 Comparison of slurry wear for the flow-through and recycled slurry systems using a low-alloy steel in a 2% silica sand slurry. Source: Ref 13

Figure 5 shows the worn surfaces of the steel that resulted from the two tests. The specimen exposed for 1 h in the flow-through test (top photograph) had many fine grooves over its entire surface, typical of cutting wear. The bottom photograph shows the smooth, wave-type pattern that developed after 1.67 h with a recycled slurry, indicating a deformation wear mechanism.

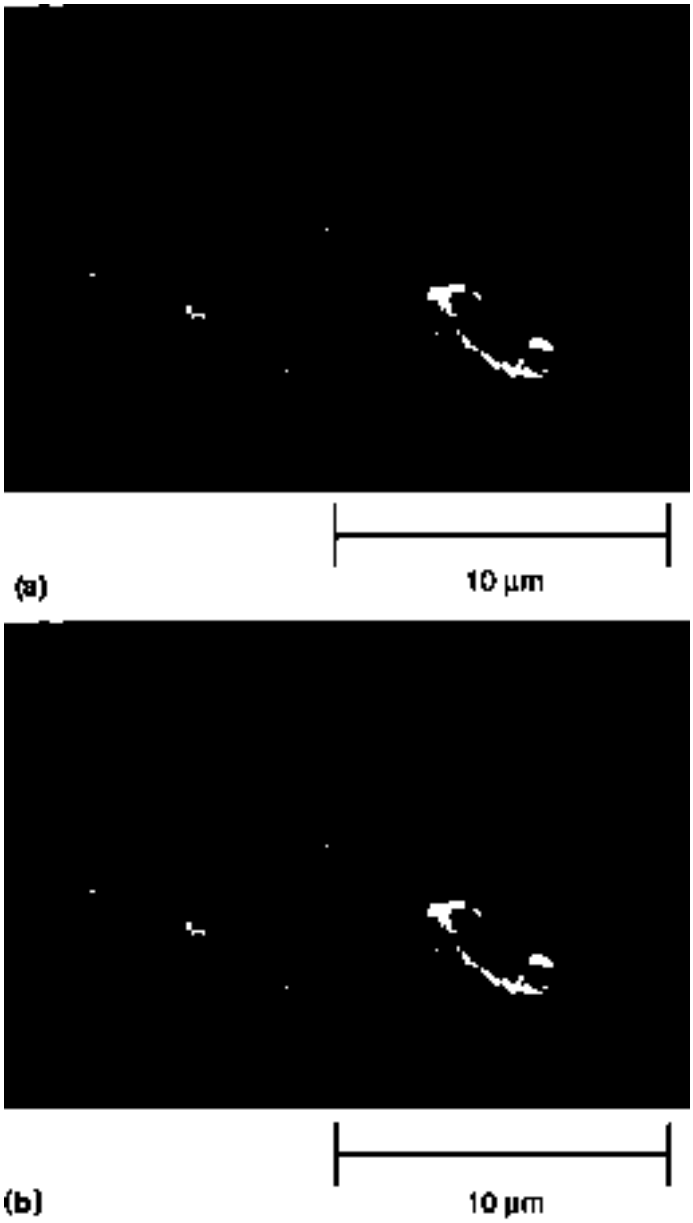


Fig. 5 Comparison of wear surfaces for low-alloy steel specimens worn in (a) flow-through and (b) recycled slurry tests for 1 h and 1.67 h, respectively. Source: Ref 13

The angularity of a particle depends on the radius of each protruding point on its surface, relative to its average radius. The more angular a particle, the higher is its apparent density (Ref 16) (Fig. 6). This effect results in a lower maximum velocity at which impact between the particles is purely elastic. If the predominant mode of material loss is corrosion, the angularity of the particles may not be such an important consideration.

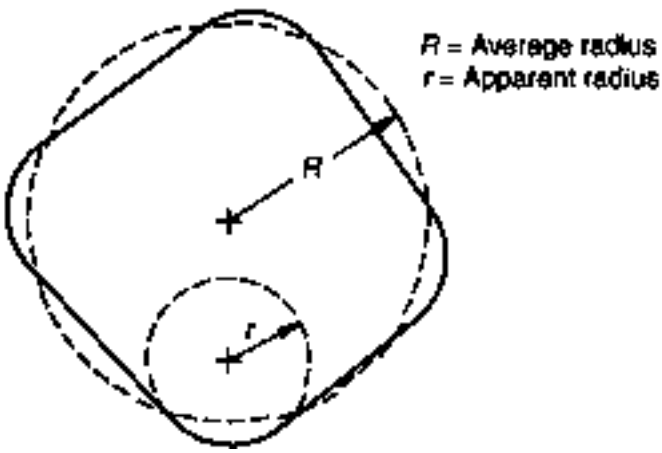


Fig. 6 Relationship between angularity and apparent density of a particle. Apparent density is $(R/r)^3$ × real reactive density. Source: Ref 16

Particle density is an important parameter that influences the corrosive wear caused by slurries. For a particle of a given angularity, the denser the particle, the more likely it is to cause either deformation or cutting wear on impact. This is a consequence of the dissipation of more energy in the same volume.

Particle Size. Because of their greater kinetic energy, particles of the same density and angularity in slurries with the same nominal velocity impart greater wear losses to materials as their size is increased. A recent study (Ref 17) has shown that larger particles have a higher collision efficiency and a higher impact velocity than smaller particles. These factors contribute to increased wear rates as slurry particle size increases.

Particle Hardness. The wear of materials has been shown to dramatically increase when the slurry particles are harder than the material being worn. However, increases in the hardness of the abrasive, above a critical hardness value, do not appreciably increase the wear rate. Likewise, when extremely hard abrasive particles are present in the slurry, moderate increases in the hardness of the worn material do little to retard the wear rate.

Slurry characteristics, such as velocity, angle of attack, solids concentration, and their relation to wear rate, are discussed as follows.

Velocity of the slurry is the single most important factor that controls the rate of wear. The wear rate is an exponential function of velocity; exponents are reported to range from 1.6 to 4.8 but are generally between 2 and 3 (Ref 18). The velocity of the slurry not only affects the rate of mechanical damage of a material but the corrosion rate as well. Above a threshold velocity, corrosion products can effectively be stripped from an alloy, thus making available a new surface that is susceptible to corrosive attack.

The angle of attack of slurry particles on a material determines, to a large extent, the role that the abrasives will play in corrosive wear. In straight pipelines where the flow of slurry particles is basically parallel to the walls of the pipe, little damage to the underlying metal results from direct contact between the abrasive and alloy (Ref 19). The abrasives, along with the solution, serve to slowly remove the corrosion products and enhance the corrosion rate. The extent of this removal process is controlled by the particle size, angularity, and slurry density.

On the other hand, when a curvature in the pipe is present, the slurry particles tend to wear the pipe, as depicted in Fig. 7. In this condition, the slurry particles are more likely to not only remove any corrosion products but to directly impart mechanical wear damage to the alloy. If the alloy is one that normally passivates (such as a stainless steel), the abrasive can continually remove the protective film and allow corrosion to proceed at a much higher rate than if no abrasive particles were present. The removal of the base metal depends on the type of material being worn.

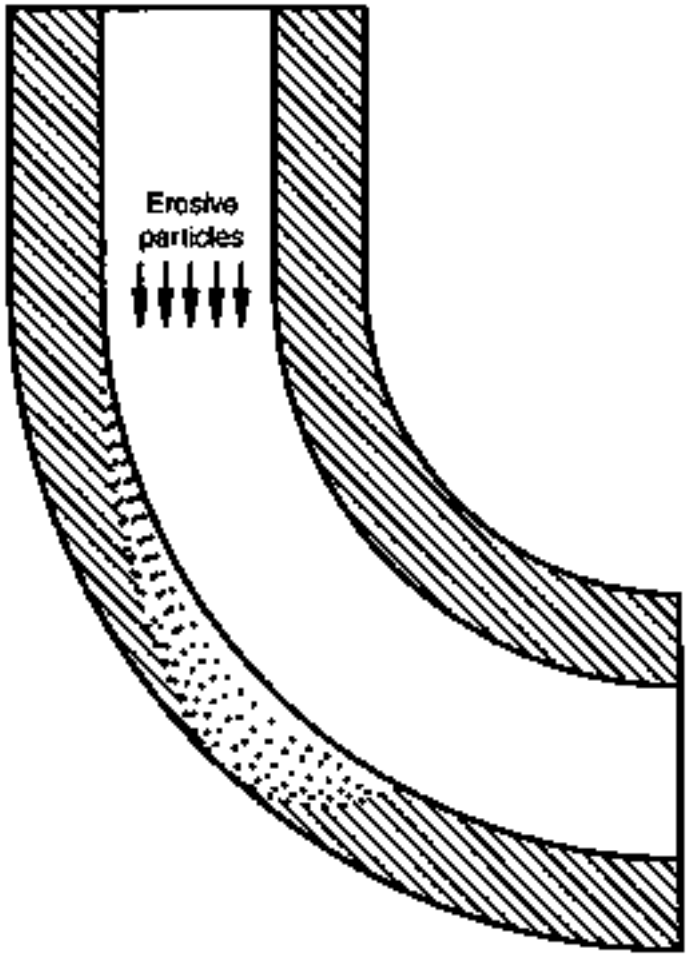


Fig. 7 Changing pattern of wear and, hence, changing angle of impingement, at bend in pipe. Source: Ref 16

Generally, ductile alloys wear greatest at shallow angles of attack, from 10 to 30°, whereas brittle materials wear greatest at 90°. The reason for these differences is the type of wear that occurs. Cutting wear is dominant for low angles of attack with ductile material, whereas deformation wear is greatest at larger angles of attack.

Solids Concentration. Increasing the solids concentration in a slurry generally increases the wear rate. The increase is only proportional to the solids concentration for dilute slurries. For denser slurries, particle-particle interaction tends to decrease the dependence of the wear rate on slurry density (Ref 14, 20).

Hydrodynamics. Recent work (Ref 21, 22) on the trajectories and impact velocities of particles during slurry erosion has shown that not all of the particles directed at a target will impact it. The fraction of slurry particles that impact the surface of the target material is controlled by fluid velocity and viscosity as well as the size, shape, and density of the slurry particles.

Flow-dependent corrosive wear typically occurs at geometrical irregularities, such as fittings, valves, and weld beads, where the flow separates from the wall of the containment vessel. Flow separation and reattachment produces high turbulence intensity and particle-wall interactions that lead to high corrosive wear rates (Ref 23, 24, 25). Figure 8 shows streamlines in a pipe for flow contraction and flow expansion.

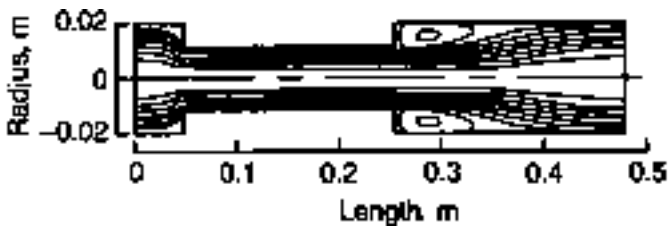


Fig. 8 Streamlines for water in pipe of variable diameter. Source: Ref 24

Material losses have been correlated with the local, near-wall intensity of turbulence in single-phase (liquid only) flows (Ref 23, 24). It was hypothesized that intensive near-wall turbulence disrupted the protective corrosion product and disturbed the mass transfer layer, thus enhancing oxygen transport to the metal surface and corrosion rates.

Corrosion products and the mass transfer of oxygen play an important role in the total material losses encountered in straight sections of a carbon steel pipeline. The material loss rate is under oxygen mass transfer control, with corrosion being the dominant mode of metal loss (Ref 19, 26). The magnitude of the erosion-corrosion can be estimated using well-established mass transfer correlations for oxygen diffusion.

In aqueous solutions of near-neutral pH, mechanical wear prevents the formation of a rust film that completely covers the interior of a pipe. By providing a barrier to oxygen, this rust film in carbon steel pipes is responsible for corrosion rates of <1 mm/year (<0.04 in./year). When abrasive solids are present, coverage of the metal by the rust film is incomplete, and the islands of bare metal that are present can act as efficient cathodes and result in substantially increased rates, such as >10 mm/year (>0.4 in./year).

References cited in this section

13. B.W. Madsen and R. Blickensderfer, A New Flow-Through Slurry Erosion Wear Test, *Slurry Erosion: Uses, Applications, and Test Methods,* STP 946, J.E. Miller and F.E. Schmidt, Jr., Ed., ASTM, 1987, p 169–184

14. B.W. Madsen, A Study of Parameters Using a New Constant-Wear-Rate Slurry Test, *Wear of Materials 1985,* K.C. Ludema, Ed., American Society of Mechanical Engineers, 1985, p 345

15. R. Blickensderfer, J.H. Tylczak, and B.W. Madsen, "Laboratory Wear Testing Capabilities of the Bureau of Mines," IC 9001, Bureau of Mines, 1985

16. J.D.A. Bitter, A Study of Erosion Phenomena—Part II, *Wear,* Vol 6, 1963, p 169–190

17. R.S. Lynn, K.K. Wong, and H. McI. Clark, On the Particle Size Effect in Slurry Erosion, *Wear of Materials 1991,* K.C. Ludema, Ed., American Society of Mechanical Engineers, 1991, p 77–82

18. H.McI. Clark, Slurry Erosion, *Proc. Conf. on Corrosion-Erosion-Wear of Material at Elevated Temperatures* (Berkeley), Electric Power Research Institute/National Association of Corrosion Engineers, Jan 1990

19. J. Postlethwaite, The Control of Erosion-Corrosion in Slurry Pipelines, *Mater. Perform.,* Dec 1987, p 41–45

20. L.D.A. Jackson, Slurry Abrasion, *Trans. Can. Inst. Min. Metall.,* Vol 70, 1967, p 219–224

21. H. Hojo, K. Tsuda, and T. Yabu, Erosion Damage of Polymeric Material by Slurry, *Wear,* Vol 112, 1986, p 17–28

22. H. Hojo, K. Tsuda, and T. Yabu, Erosion Behavior of Plastics by Slurry Jet Method, *Kagaku Ronbunshu,* Vol 14, 1988, p 161–166

23. W. Blatt, T. Kohley, U. Lotz, and E. Heitz, The Influence of Hydrodynamics on Erosion-Corrosion in Two-Phase Liquid-Particle Flow, *Corrosion,* Vol 45 (No. 10), 1989, p 793–804

24. S. Nesic and J. Postlethwaite, Relationship Between the Structure of Disturbed Flow and Erosion-Corrosion, *Corrosion,* Vol 46 (No. 11), 1990, p 874–880

25. U. Lotz and J. Postlethwaite, Erosion-Corrosion in Disturbed Two Phase Liquid/Particle Flow, *Corros. Sci.,* Vol 30 (No. 1), 1990, p 95–106

26. J. Postlethwaite, M.H. Dobbin, and K. Bergevin, The Role of Oxygen Mass Transfer in the Erosion-Corrosion of Slurry Pipelines, *Corrosion,* Vol 42 (No. 9), 1986, p 514–521

Grinding Wear: Impact and Three- Body Abrasive-Corrosive Wear

The wear of grinding media and crushers in mineral processing systems is caused by the combination of abrasion, corrosion, and high-energy impact of ore and metal components. The increased size of modern crushing and grinding equipment greatly increases the kinetic energy at metal/ore interfaces, which results in high wear rates. The crushed ore particles participate in three-body abrasion by virtue of being confined between two solid surfaces.

Dependence on Force. Impact severity has a strong effect on the grinding media wear rate for dynamic loading conditions (Ref 1). The wear rate was four times greater in an 8.5 m (28 ft) semiautogenous grinding mill (8.30 kgf, or 18.3 lbf, balls) than that attained in a 2.9 m (9.5 ft) ball mill (1.81 kgf, or 4 lbf, balls). The ball velocities in the two mills were calculated to be 9.14 m/s (30 ft/s) and 5.33 m/s (17.5 ft/s), respectively. However, under conditions of a static load, the total wear rate of a metal alloy was approximately linear with load (Ref 27). A rotating metal alloy specimen was used between two ceramic anvils submerged in an abrasive slurry.

Dependence on Abrasive Type. Abrasive wear rates for steels and irons generally increase with an increase in mineral hardness. When the metal-to-mineral hardness ratio is greater than 0.6, marked improvement in the ability to resist abrasive wear is shown (Ref 28). In order to achieve this favorable hardness ratio, hard $(Fe,Cr)_3C$ carbides found in low-alloy nickel-chromium irons and harder $(Fe,Cr)_7C_3$ carbides found in high-chromium white cast irons are used.

Dependence on Galvanic Interaction between Minerals and Metal Alloys. Electrochemical interactions between ore minerals and grinding media can occur. Mineral particles with potentials that are nobler than those of steel grinding media become galvanically coupled and result in accelerated corrosive wear (Ref 6). In recent studies (Ref 29, 30), it has been reported that the relative contributions from corrosive and abrasive wear are highly variable from one mineral-metal system to another and are largely dependent on the mineral slurry characteristics as well as the properties of the grinding media. In the presence of surface abrasion, the corrosion resistance, hardness of the grinding media, pH and conductivity of the pulp, and oxygen content influence the magnitude of galvanic currents between minerals and grinding media. Galvanic currents exist between ferritic and martensitic phases in a microstructure of steel grinding media as well as between sulfide minerals and these two phases (Ref 31).

Role of Localized Corrosion in Grinding. In the study of high-carbon low-alloy steel grinding media, localized electrochemical cells existed between strained and unstrained regions of the alloy (Ref 27). This resulted in preferential anodic dissolution in microgrooves, which promoted pitting. The overall corrosive wear rates of the steels were found to be strongly dependent on the localized corrosion.

References cited in this section

1. D.J. Dunn, Metal Removal Mechanisms Comprising Wear in Mineral Processing, *Wear of Materials,* K.C. Ludema, Ed., American Society of Mechanical Engineers, 1985, p 501–508

6. K. Adam, K.A. Natarajan, S.C. Riemer, and I. Iwasaki, Electrochemical Aspects of Grinding Media-Mineral Interaction in Sulfide Ore Grinding, *Corrosion,* Vol 42 (No. 8), 1986, p 440–446

27. D. Kotlyar and M.E. Wadsworth, “The Role of Localized Corrosion on Corrosive Abrasive Wear of High Carbon Low Alloy Steel Grinding Media,” Paper 246, Corrosion 88 (Houston), National Association of Corrosion Engineers, 1988

28. J.D. Watson, P.J. Mutton, and I.R. Sare, Abrasive Wear of White Cast Irons, *Met. Forum,* Vol 3 (No. 1), 1980

29. R. Perez and J.J. Moore, The Influence of Grinding Ball Composition and Wet Grinding Conditions on Metal Wear, *Wear of Materials 1983,* American Society of Mechanical Engineers, 1983, p 67–78

30. K.A. Natarajan and I. Iwasaki, Electrochemical Aspects of Grinding Media-Mineral Interactions in Magnetite Ore Grinding, *Int. J. Miner. Process.,* Vol 13, 1984, p 53–71

31. J.W. Jang, I. Iwasaki, and J.J. Moore, The Effect of Galvanic Interaction Between Martensite and Ferrite in Grinding Media Wear, *Corrosion,* Vol 45 (No. 5), 1989, p 402–407

Means for Combating Corrosive Wear

Research and engineering efforts in wear, corrosion, and materials science are being pursued to combat corrosive wear in aqueous environments. The areas of investigation described subsequently include materials selection, surface treatments, and handling-environment modifications.

Materials Selection. The selection of the right material for a particular corrosive wear environment can lead to extended life of component parts, less costly downtime, and other economic advantages.

One approach is to place a number of materials in actual service and to compare the material losses of each over a given time. This technique can give the best solution to materials selection but is time-consuming. It also limits the number of materials that can be tried and does not often result in the application of scientific principles to obtain the most cost-effective material for the given life of a process. However, actual field service should be used as a concluding step before final choices of materials are made.

Another approach is to adapt laboratory tests to field situations. Electrodes were mounted at the ends of mill liner bolts, and polarization curves of the media material electrodes were obtained by a telemetry-radio system (Ref 32). Although the data were often noisy, correlations with laboratory corrosive wear tests were made.

In a test developed in the laboratory in a field experiment, a laboratory slurry pot test was converted to a portable model that could be transported to the field for use with slurries there (Ref 33, 34). This approach enabled the testing of a large number of candidate materials in a relatively short time, because this slurry pot design allowed for the simultaneous comparison of 16 specimens.

The equipment could be useful in screening materials for further evaluation in actual service, thus shortening the time needed to determine the better material while offering a wider selection than actual field use could. This test is most useful when the hydrodynamic and geometric parameters in the test chamber are similar to those experienced in the actual service conditions.

A more fundamental and scientific approach is to study the wear and corrosion characteristics of various materials in the laboratory and extrapolate these results to actual field use. A great deal of knowledge and modeling of wear, corrosion, and wear-corrosion synergism is needed, along with a detailed model of the hydrodynamics. This research is still in its infancy but is making progress as the number of researchers who can recognize wear-corrosion synergism increases.

Surface Treatment. Rather than use the more-corrosive-resistant material to make the entire component, one may choose to use a thinner layer of a costly material on the surface of the more economical substrate material. Claddings, surface treatments such as hard facings and patching with welds, or replaceable liners have all been used. In pipelines where corrosion is a problem, it is sometimes economical to line the pipes with a smaller-diameter polymer that is more corrosion resistant but lacks the strength to withstand high pressures on its own.

Modification of the materials-handling environment can be effective in controlling the corrosive wear of component parts. Solution conditioning, such as adjusting the pH and deaeration, can reduce the amount of material losses in a corrosive-wear environment. Slurry conditioning is not economical for the short slurry lines used in mining operations unless some method of water separation and recirculation is used (Ref 19).

Use of Corrosion Inhibitors. Oxidizing inhibitors, such as chromates and nitrites, have been used to raise the potential of an alloy into passive regions and to lower their corrosion rates (Ref 35). These inhibitors form a surface film when used in high concentrations. Chromates also act as effective inhibitors when used at concentrations much lower than those required to produce active-passive transitions. At these lower

concentrations, chromates act as cathodic inhibitors in neutral solutions. However, because chromates are toxic, alternative nontoxic inhibitors with a similar action should be considered. Many governments restrict the effluent limit to levels as low as 0.05 ppm chromate, which makes the choice of chromate inhibitors unacceptable. Progress in developing nonchromate inhibitors has been slow (Ref 19).

Slurry Parameters. Reduction of the slurry velocity is a major factor in controlling the rate of material losses if the mechanical wear, *W*, is important, because the wear rate generally varies with the velocity raised to exponents of 2, 3, or 4. For slurry pipelines that carry 20% silica sand (30 × 50 mesh), very little abrasive wear occurs below a velocity of 6 m/s (20 ft/s) (threshold value). Above this velocity, *W* becomes more dominant as a means of material degradation.

Particle size can also be a factor in the wear of slurry-handling equipment. The mechanical wear, *W*, will not be a problem if the particle size is sufficiently reduced so that the particles are fine enough to follow the streamlines of the solution, rather than impact the walls of the containment part. However, fine slurry particles and low velocities may result in conditions mild enough to permit the growth of rust films and scale, which can lead to pitting (Ref 19).

The particle size and velocity combination should be maintained below the threshold value, where *W* is not a factor, but with a particle size that causes enough abrasion to maintain a rust- and scale-free pipe surface. In this condition, however, the pipe should be protected from corrosion by inhibitors and/or deaeration. This would eliminate the need for unnecessary size reduction.

Cathodic protection is a useful method for protecting short sections of slurry pipelines, pumps, elbows, and other equipment (Ref 36). However, the length of the protection must be kept sufficiently short to prevent overprotection and subsequent hydrogen blistering of the protected surface. The throwing power of cathodic protection, either by impressed current or galvanic anodes, is insufficient for this method to be used to protect the inside of slurry pipelines.

Sacrificial electrodes, such as zinc plugs in water or slurry pumps, have been used to cathodically protect the pump casing. This means of protection may be applicable to other instances where it is economically feasible.

Design of Pumps, Valves, and Elbows. Research has shown that computer modeling is becoming a tool for the design of wear-resistant slurry pumps (Ref 37, 38, 39, 40, 41). A computer code is used to carry out the numerical flow simulations within the pump channels, and an energy-based, wear-predictive model has allowed for the prediction of wear rates for various geometries.

The use of these types of models can extend the wear life of pumps by altering the geometry of the interior of the pump. This effectively removes the areas of high wear rates and spreads the energy absorbed by the pump head evenly around its interior. This same type of modeling has been done for areas of disturbed flow (Ref 42, 43, 44, 45).

Other means of controlling corrosive wear are available, including:

- Increasing the pipe diameter in order to decrease the slurry velocity and help ensure laminar flow
- Increasing the thickness of materials in critical areas
- Inserting impingement plates or baffles to shield critical areas from high wear
- Directing the inlet flow of materials to avoid high particle velocities at the wall of containment vessels

References cited in this section

19. J. Postlethwaite, The Control of Erosion-Corrosion in Slurry Pipelines, *Mater. Perform.*, Dec 1987, p 41–45

32. A.E. Isaacson, P.J. McDonough, and J.H. Maysilles, “Corrosion Rate Determination in Industrial Ore Grinding Environments,” Paper 232, Corrosion 88 (Houston), National Association of Corrosion Engineers, 1988

33. B.W. Madsen, A Portable Slurry Wear Test for the Field, *Trans. ASME—J. Fluids Eng.*, Vol 111, 1989, p 324–340

34. B.W. Madsen, A Comparison of the Wear of Polymers and Metal Alloys in Laboratory and Field Slurries, *Wear*, Vol 134, 1989, p 59–79

35. G.R. Hoey, W. Dingley, and C. Freeman, Corrosion Inhibitors Reduce Ball Wear in Grinding Sulfide Ore, *CIM Bull.,* Vol 68, 1975, p 120–123

36. J. Postlethwaite, B.J. Brady, M.W. Hawrylak, and E.B. Tinker, Effects of Corrosion on the Wear Patterns in Horizontal Slurry Pipelines, *Corrosion,* Vol 34 (No. 7), 1978, p 245

37. M.C. Roco, P. Nair, and G.R. Addie, Test Approach for Dense Slurry Erosion, *Slurry, Erosion: Uses, Applications, and Test Methods,* STP 946, J.E. Miller and F.E. Schmidt, Jr., Ed., ASTM, 1987, p 185–210

38. M.C. Roco, "Wear Patterns in Centrifugal Slurry Pumps," Paper 244, Corrosion 88 (Houston), National Association of Corrosion Engineers, 1988

39. M.C. Roco, Optimum Wearing High Efficiency Design of Phosphate Slurry Pumps, *Proc. 11th International Conference of Slurry Technology,* March 1986, p 277–285

40. M.C. Roco and P. Nair, Erosion of Concentrated Slurries in Turbulent Flow, *J. Pipelines,* Vol 4, 1984, p 213–221

41. K. Ahmad, R.C. Baker, and A. Goulas, Computation and Experimental Results of Wear in a Slurry Pump Impeller, *Proc. Institution of Mechanical Engineers,* Vol 200 (No. C6), 1986

42. J. Postlethwaite, "Influence of Flow System Geometry on Erosion-Corrosion," Corrosion 91 (Houston), National Association of Corrosion Engineers, 1991

43. S. Nesic and J. Postlethwaite, Hydrodynamics of Disturbed Flow and Erosion-Corrosion, Part I—A Single-Phase Flow Study, *Can. J. Chem. Eng.,* Vol 69, 1991, p 698–703

44. S. Nesic and J. Postlethwaite, Hydrodynamics of Disturbed Flow and Erosion-Corrosion, Part II—A Two-Phase Flow Study, *Can. J. Chem. Eng.,* Vol 69, 1991, p 704

45. S. Nesic and J. Postlethwaite, A Predictive Model for Localized Erosion-Corrosion, *Corrosion,* Vol 47 (No. 8), 1991, p 582–589

Acknowledgment

This article was adapted from B.W. Madsen, Corrosive Wear, *Friction, Lubrication, and Wear Technology,* Volume 18, *ASM Handbook,* p 271–279.

Corrosive Wear Failures

References

1. D.J. Dunn, Metal Removal Mechanisms Comprising Wear in Mineral Processing, *Wear of Materials,* K.C. Ludema, Ed., American Society of Mechanical Engineers, 1985, p 501–508

2. A.M.F. Carter and D. Howarth, “A Literature Review of the Factors That Influence the Wear of Slurry-Handling Systems,” Report M319, Council for Mineral Technology, Sept 1987

3. F.H. Stott, J.E. Breakell, and R.C. Newman, The Corrosive Wear of Cast Iron Under Potentiostatically-Controlled Conditions in Sulfuric Acid Solutions, *Corros. Sci.,* Vol 30, 1990, p 813–830

4. R.L. Pozzo and I. Iwasaki, Pyrite-Pyrrhotite Grinding Media Interactions and Their Effects on Media Wear and Flotation, *J. Electrochem. Soc.,* Vol 136 (No. 6), 1989, p 1734–1740

5. National Materials Advisory Board, “Comminution and Energy Consumption,” NMAB-364, National Academy Press, 1981

6. K. Adam, K.A. Natarajan, S.C. Riemer, and I. Iwasaki, Electrochemical Aspects of Grinding Media-Mineral Interaction in Sulfide Ore Grinding, *Corrosion,* Vol 42 (No. 8), 1986, p 440–446

7. R.L. Pozzo and I. Iwasaki, Effect of Pyrite and Pyrrhotite on the Corrosive Wear of Grinding Media, *Miner. Metall. Process.,* Aug 1987, p 166–171

8. K.A. Natarajan, S.C. Riemer, and I. Iwasaki, Influence of Pyrrhotite on the Corrosive Wear of Grinding Balls in Magnetite Ore Grinding, *Int. J. Miner. Process.,* Vol 13, 1984, p 73–81

9. R.L. Pozzo and I. Iwasaki, An Electrochemical Study of Pyrrhotite-Grinding Media Interaction Under Abrasive Conditions, *Corrosion,* Vol 43 (No. 3), 1987, p 159–169

10. A.V. Levy and Y. Man, “Erosion-Corrosion Mechanisms and Rates in Fe-Cr Steels,” Paper 111, Corrosion 86 (Houston), National Association of Corrosion Engineers, 1986

11. S. Agarwal and M.A.H. Howes, Erosion/Corrosion of Materials in High-Temperature Environments, *Proc. AIME Conf. on High Temperature Corrosion in Energy Systems* (Detroit), American Institute of Mining, Metallurgical, and Petroleum Engineers, Sept 1984

12. T. Foley and A.V. Levy, The Erosion of Heat Treated Steels, *Wear,* Vol 48 (No. 1), 1983, p 181

13. B.W. Madsen and R. Blickensderfer, A New Flow-Through Slurry Erosion Wear Test, *Slurry Erosion: Uses, Applications, and Test Methods,* STP 946, J.E. Miller and F.E. Schmidt, Jr., Ed., ASTM, 1987, p 169–184

14. B.W. Madsen, A Study of Parameters Using a New Constant-Wear-Rate Slurry Test, *Wear of Materials 1985,* K.C. Ludema, Ed., American Society of Mechanical Engineers, 1985, p 345

15. R. Blickensderfer, J.H. Tylczak, and B.W. Madsen, “Laboratory Wear Testing Capabilities of the Bureau of Mines,” IC 9001, Bureau of Mines, 1985

16. J.D.A. Bitter, A Study of Erosion Phenomena—Part II, *Wear,* Vol 6, 1963, p 169–190

17. R.S. Lynn, K.K. Wong, and H. McI. Clark, On the Particle Size Effect in Slurry Erosion, *Wear of Materials 1991,* K.C. Ludema, Ed., American Society of Mechanical Engineers, 1991, p 77–82

18. H.McI. Clark, Slurry Erosion, *Proc. Conf. on Corrosion-Erosion-Wear of Material at Elevated Temperatures* (Berkeley), Electric Power Research Institute/National Association of Corrosion Engineers, Jan 1990

19. J. Postlethwaite, The Control of Erosion-Corrosion in Slurry Pipelines, *Mater. Perform.,* Dec 1987, p 41–45

20. L.D.A. Jackson, Slurry Abrasion, *Trans. Can. Inst. Min. Metall.,* Vol 70, 1967, p 219–224

21. H. Hojo, K. Tsuda, and T. Yabu, Erosion Damage of Polymeric Material by Slurry, *Wear,* Vol 112, 1986, p 17–28

22. H. Hojo, K. Tsuda, and T. Yabu, Erosion Behavior of Plastics by Slurry Jet Method, *Kagaku Ronbunshu,* Vol 14, 1988, p 161–166

23. W. Blatt, T. Kohley, U. Lotz, and E. Heitz, The Influence of Hydrodynamics on Erosion-Corrosion in Two-Phase Liquid-Particle Flow, *Corrosion,* Vol 45 (No. 10), 1989, p 793–804

24. S. Nesic and J. Postlethwaite, Relationship Between the Structure of Disturbed Flow and Erosion-Corrosion, *Corrosion,* Vol 46 (No. 11), 1990, p 874–880

25. U. Lotz and J. Postlethwaite, Erosion-Corrosion in Disturbed Two Phase Liquid/Particle Flow, *Corros. Sci.,* Vol 30 (No. 1), 1990, p 95–106

26. J. Postlethwaite, M.H. Dobbin, and K. Bergevin, The Role of Oxygen Mass Transfer in the Erosion-Corrosion of Slurry Pipelines, *Corrosion,* Vol 42 (No. 9), 1986, p 514–521

27. D. Kotlyar and M.E. Wadsworth, "The Role of Localized Corrosion on Corrosive Abrasive Wear of High Carbon Low Alloy Steel Grinding Media," Paper 246, Corrosion 88 (Houston), National Association of Corrosion Engineers, 1988

28. J.D. Watson, P.J. Mutton, and I.R. Sare, Abrasive Wear of White Cast Irons, *Met. Forum,* Vol 3 (No. 1), 1980

29. R. Perez and J.J. Moore, The Influence of Grinding Ball Composition and Wet Grinding Conditions on Metal Wear, *Wear of Materials 1983,* American Society of Mechanical Engineers, 1983, p 67–78

30. K.A. Natarajan and I. Iwasaki, Electrochemical Aspects of Grinding Media-Mineral Interactions in Magnetite Ore Grinding, *Int. J. Miner. Process.,* Vol 13, 1984, p 53–71

31. J.W. Jang, I. Iwasaki, and J.J. Moore, The Effect of Galvanic Interaction Between Martensite and Ferrite in Grinding Media Wear, *Corrosion,* Vol 45 (No. 5), 1989, p 402–407

32. A.E. Isaacson, P.J. McDonough, and J.H. Maysilles, "Corrosion Rate Determination in Industrial Ore Grinding Environments," Paper 232, Corrosion 88 (Houston), National Association of Corrosion Engineers, 1988

33. B.W. Madsen, A Portable Slurry Wear Test for the Field, *Trans. ASME—J. Fluids Eng.,* Vol 111, 1989, p 324–340

34. B.W. Madsen, A Comparison of the Wear of Polymers and Metal Alloys in Laboratory and Field Slurries, *Wear,* Vol 134, 1989, p 59–79

35. G.R. Hoey, W. Dingley, and C. Freeman, Corrosion Inhibitors Reduce Ball Wear in Grinding Sulfide Ore, *CIM Bull.,* Vol 68, 1975, p 120–123

36. J. Postlethwaite, B.J. Brady, M.W. Hawrylak, and E.B. Tinker, Effects of Corrosion on the Wear Patterns in Horizontal Slurry Pipelines, *Corrosion,* Vol 34 (No. 7), 1978, p 245

37. M.C. Roco, P. Nair, and G.R. Addie, Test Approach for Dense Slurry Erosion, *Slurry, Erosion: Uses, Applications, and Test Methods,* STP 946, J.E. Miller and F.E. Schmidt, Jr., Ed., ASTM, 1987, p 185–210

38. M.C. Roco, "Wear Patterns in Centrifugal Slurry Pumps," Paper 244, Corrosion 88 (Houston), National Association of Corrosion Engineers, 1988

39. M.C. Roco, Optimum Wearing High Efficiency Design of Phosphate Slurry Pumps, *Proc. 11th International Conference of Slurry Technology,* March 1986, p 277–285

40. M.C. Roco and P. Nair, Erosion of Concentrated Slurries in Turbulent Flow, *J. Pipelines,* Vol 4, 1984, p 213–221

41. K. Ahmad, R.C. Baker, and A. Goulas, Computation and Experimental Results of Wear in a Slurry Pump Impeller, *Proc. Institution of Mechanical Engineers,* Vol 200 (No. C6), 1986

42. J. Postlethwaite, "Influence of Flow System Geometry on Erosion-Corrosion," Corrosion 91 (Houston), National Association of Corrosion Engineers, 1991

43. S. Nesic and J. Postlethwaite, Hydrodynamics of Disturbed Flow and Erosion-Corrosion, Part I—A Single-Phase Flow Study, *Can. J. Chem. Eng.,* Vol 69, 1991, p 698–703

44. S. Nesic and J. Postlethwaite, Hydrodynamics of Disturbed Flow and Erosion-Corrosion, Part II—A Two-Phase Flow Study, *Can. J. Chem. Eng.,* Vol 69, 1991, p 704

45. S. Nesic and J. Postlethwaite, A Predictive Model for Localized Erosion-Corrosion, *Corrosion,* Vol 47 (No. 8), 1991, p 582–589

Erosive Wear Failures

Thomas A. Adler, U.S. Department of Energy

Introduction

EROSION is the progressive loss of original material from a solid surface due to mechanical interaction between that surface and a fluid, a multicomponent fluid, an impinging liquid, or impinging solid particles (Ref 1). Erosion is a rather broad term and can be further classified into a number of more specific terms. These include cavitation erosion, liquid impingement erosion, and solid particle erosion. Cavitation erosion is due to the formation and collapse, within a liquid, of cavities or bubbles that contain vapor or gas or both. Liquid impingement erosion is due to impacts by liquid drops or jets. Solid impingement erosion is due to impacts by solid particles. "Impingement" also connotes that the particles are smaller than the solid surface and that the impacts are distributed over that surface or a portion of it. Erosion can occur in combination with other forms of degradation, such as corrosion. This is referred to as erosion-corrosion.

The detrimental effects of erosion have caused problems in a number of industries. In the power-generation industry, erosion damage has occurred to boiler tubes and water walls (Ref 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17), steam turbines (Ref 18, 19, 20, 21), land-based gas turbines (Ref 21, 22, 23), solids transport systems (Ref 17, 24, 25, 26, 27, 28), cyclones (Ref 17, 28, 29, 30), valves (Ref 17, 31, 32), and inlet nozzles (Ref 33). In the oil and gas industry, choke valves are damaged by erosion (Ref 34, 35, 36, 37). In the petrochemical industry, process control valves (Ref 37) and power recovery turbines (Ref 18, 38) are damaged by erosion. In chemical process industries, solids transport systems are damaged by erosion (Ref 25, 39, 40, 41). In the aircraft industry, damage has occurred to aircraft engines (Ref 19, 21, 42, 43, 44).

Erosion occurs as the result of a number of different mechanisms, depending on the composition, size, and shape of the eroding particles, their velocity and angle of impact, and the composition of the surface being eroded. The sensitivity of brittle and ductile materials to the angle of impact generally follows the behavior shown in Fig. 1, with ductile materials experiencing the maximum rate at approximately 20 to 30° and brittle materials at approximately 90° (Ref 45). This tendency of maximum erosion for a ductile material at approximately 20 to 30° is not universally observed and is dependent on factors such as particle shape and fragmentation (Ref 46, 47, 48). When ductile materials are impacted by spherical particles that do not fragment, the maximum erosion rate can occur at 90°. Erosion of a given material as a function of time often follows a pattern consisting of an incubation period with little or no material removal, followed by an increasing rate, and finally a steady state. Some materials exhibit peaks with decelerating rate with or without a final steady state (Fig. 2, Ref. 49). Most mechanistic models are concerned only with the steady state.

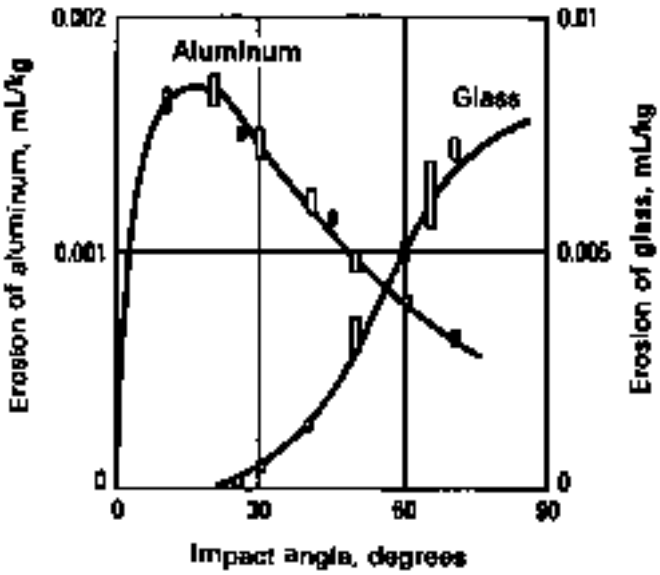


Fig. 1 Effect of impact angle on erosion of aluminum and glass by 300 µm iron spheres at 10 m/s (34 ft/s)

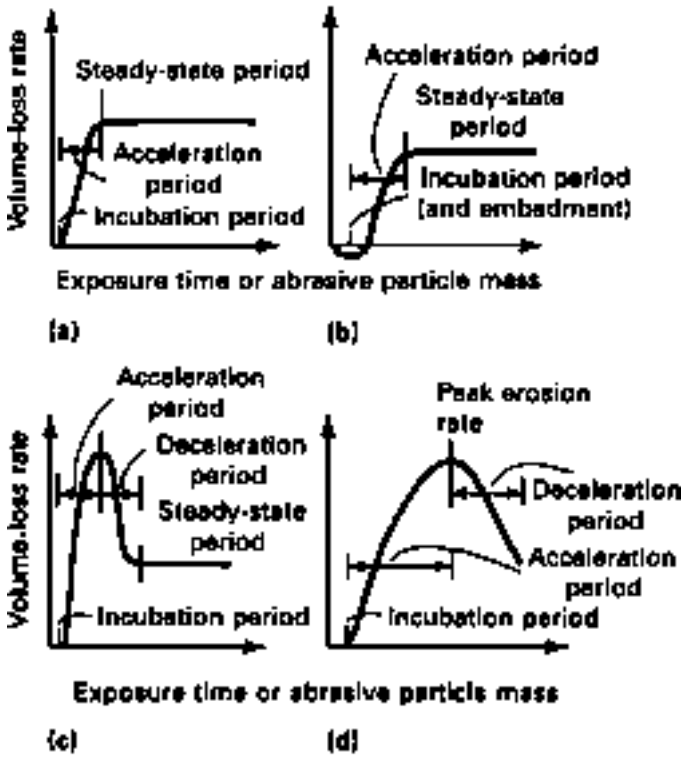


Fig. 2 Characteristic volume-loss rate versus time curves. (a) Type I. (b) Type II. (c) Type III. (d) Type IV

The erosion of materials has been attributed to a number of mechanisms, including cutting (Ref 45, 50, 51, 52, 53, 54), fragmentation (Ref 55), elastic and elastic-plastic fracture (Ref 56, 57, 58), extrusion (Ref 59, 60), fatigue (Ref 61, 62), delamination (Ref 63, 64), deformation localization (Ref 47, 54, 65, 66), and melting (Ref 44, 59, 65). An even greater number of analytical models have been proposed, but none has been completely satisfactory. To show the general trends in modeling erosion, it is probably best to start with a generic, empirical formula for erosion (Ref 5, 36, 67) and later proceed to more specific models:

$$E = K_0 V^p f(\alpha) \qquad \text{(Eq 1)}$$

where E is the erosion rate, V is the particle velocity, p is the velocity exponent, f is a function of the impact angle of the particles, α, and K_0 is a constant that includes all other effects. The velocity exponent, p, varies from 2 to 3 with a mean of approximately 2.4 for ductile materials (Ref 66, 67), while it varies from 2 to 6 with a mean value of approximately 3 for brittle materials (Ref 56, 68, 69). It is perhaps easiest to consider the erosion of ductile and brittle materials separately, but it must be understood that, in reality, there is a continuum of materials. In addition, some materials, such as composites and cermets, have both brittle and ductile components and exhibit mixed behavior.

References cited in this section

1. "Standard Terminology Relating to Wear and Erosion," G 40, *Annual Book of ASTM Standards,* ASTM

2. J. Stringer, Practical Experience with Wastage at Elevated Temperatures in Coal Combustion Systems, *Wear,* Vol 186–187, 1995, p 11–27

3. B.-E. Lee, C.A.J. Fletcher, and M. Behnia, Computational Study of Solid Particle Erosion for a Single Tube in Cross Flow, *Wear,* Vol 240, 2000, p 95–99

4. B.E. Lee, C.A.J. Fletcher, and M. Behnia, Computational Prediction of Tube Erosion in Coal Fired Power Utility Boilers, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 121, Oct 1999, p 746–750

5. J.F. Drennen and J.G. McGowan, Prediction of Flyash Erosion in Convective Pass Sections of Coal Fired Systems, *Proc. Seventh Int. Conf. Erosion by Liquid and Solid Impacts,* University of Cambridge, 1987, p 75-1 to 75-6

6. J.J. Xie and P.M. Walsh, Erosion-Oxidation of Carbon Steel in the Convection Section of an Industrial Boiler Cofiring Coal-Water Fuel and Natural Gas, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 119, July 1997, p 717–722

7. M. Sucking and C. Allen, Critical Variables in High Temperature Erosive Wear, *Wear,* Vol 203–204, 1997, p 528–536

8. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 122–128

9. A.V. Levy, B.Q. Wang, G.Q. Geng, and W. Mack, Erosion-Corrosion of Steels in the Convection Pass Region of a Commercial Circulating Fluidized Bed Combustor, *Wear of Materials 1991,* American Society of Mechanical Engineers, 1991, p 697–701

10. A.V. Levy, B.Q. Wang, G.Q. Geng, and W. Mack, "Erosivity of Particles in Circulating Fluidized Bed Combustors," Paper 543, Corrosion 89, National Association of Corrosion Engineers, 1989

11. A. Levy and Y.-F. Man, Erosion-Corrosion of Chromium Steels, *Proc. Corrosion-Erosion-Wear of Materials at Elevated Temperatures, 1986,* National Association of Corrosion Engineers, 1987, p 168–203

12. A.V. Levy, B.-Q. Wang, Y.-F. Man, and N. Jee, Erosion-Corrosion of Steels in Simulated and Actual Fluidized Bed Combustor Environments, *Wear,* Vol 131, 1989, p 85–103

13. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 98–102

14. H.H. Krause and I.G. Wright, Boiler Tube Failures in Municipal Waste-to-Energy Plants, *Mater. Perform.* Vol 35, Jan 1996, p 46–54

15. N.G. Solomon, Erosion-Resistant Coatings for Fluidized Bed Boilers, *Mater. Perform.* Vol 37, 1998, p 38–43

16. P.M. Rogers, I.M. Hutchings, J.A. Little, and F. Kara, Microstructural Characterization of Surface Layers Formed on Alloy Steel During Erosion-Oxidation in a Fluidized Bed, *J. Mater. Sci.* Vol 32, 1997, p 4575–4583

17. M.S. Crowley, Know Causes of, Solutions to, CFB-Boiler Refractory Problems, *Power,* Vol 136, Jan 1992, p 54–57

18. R.C. Tucker and A.A. Ashari, The Structure Property Relationship of Erosion Resistant Thermal Spray Coatings, *Proc. 15th Int. Thermal Spray Conf.,* ASM International, 1998, p 259–262

19. P.N. Walsh, J.M. Quets, and R.C. Tucker, Jr., Coatings for the Protection of Turbine Blades from Erosion, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 117, Jan 1995, p 152–155

20. I.A. Diaz-Tous, A.H. Khan, and T.H. McCloskey, Solid Particle Erosion Technology Assessment, *Advances in Steam Turbine Technology for the Power Generation Industry, PWR,* Vol 26, American Society of Mechanical Engineers, 1994, p 247–254

21. W. Tabakoff, M. Metwally, and A. Hamed, High-Temperature Coatings for Protection against Turbine Deterioration, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 117, Jan 1995, p 146–151

22. J.E. Restall and D.J. Stephenson, High Temperature Erosion of Coated Superalloys for Gas Turbines, *Mater. Sci. Eng.,* Vol 88, 1987, p 273–282

23. M. Menguturk and E.F. Sverdrup, Calculated Tolerance of a Large Electric Utility Gas Turbine to Erosion Damage by Coal Gas Ash Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 193–224

24. S. Seetharamu, P. Sampathkumaran, and R.K. Kumar, Erosion Resistance of Permanent Moulded High Chromium Iron, *Wear,* Vol 186–187, 1995, p 159–167

25. C.R. Dimond, The Specification and Installation of Alumina Ceramics in Industry, *Mater. Des.,* Vol 8 May/June 1987, p 152–155

26. T.D. Johnson, M.B.J. Low, M.T. Parry, and D.J. Wall, Erosion in Pulverized Fuel Distribution Systems: Materials Selection and Plant Performance, *Proc. Seventh Int. Conf. Erosion by Liquid and Solid Impacts,* University of Cambridge, 1987, p 74-1 to 74-8

27. R.G. Wellman and C. Allen, The Effects of Angle of Impact and Materials Properties on the Erosion Rates of Ceramics, *Wear,* Vol 186–187, 1995, p 117–122

28. S.J. Dapkunas, A Summary of Erosion Experience in the U.S. Department of Energy's Coal Conversion Plants, *Proc. Fifth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1979, p 43-1 to 43-14

29. S. Danyluk, W.J. Shack, J.Y. Park, and M.M. Mamoun, Solid Particle Erosion of a Cyclone from a Coal Gasification Pilot Plant, *Wear of Material 1981,* American Society of Mechanical Engineers, 1981, p 619–624

30. C.A. Youngdahl, K. Pater, and M.J. Gorski, Erosive Wear and Design Evaluation of a Stainless Steel Cyclone on the Gasification Pilot Plant Facility at Morgantown, *The Properties and Performance of Materials in the Coal Gasification Environment,* American Society for Metals, 1981, p 709–734

31. J.S. Hansen, Relative Erosion Resistance of Several Materials, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, pp 148–162

32. J.S. Hansen, J.E. Kelly, and F.W. Wood, "Erosion Testing of Potential Valve Materials for Coal Gasification Systems," Report of Investigations 8335, U.S. Department of the Interior, Bureau of Mines, 1979

33. A. Drotlew, P. Christodoulou, and V. Gutowski, Erosion of Ferritic Fe-Cr-C Cast Alloys at Elevated Temperatures, *Wear,* Vol 211, 1997, p 120–128

34. R. Hamzah, D.J. Stephensen, and J.E. Strutt, Erosion of Material Used in Petroleum Production, *Wear,* Vol 186–187, 1995, p 493–496

35. L. Nøkleberg and T. Søntvedt, Erosion in Choke Valves—Oil and Gas Industry Applications, *Wear,* Vol 186–187, 1995, p 401–412

36. K. Haugen, O. Kvernvold, A. Ronold, and R. Sandberg, Sand Erosion of Wear-Resistant Materials: Erosion in Choke Valves, *Wear,* Vol 186–187, 1995, p 179–188

37. R.J.K. Wood, and D.W. Wheeler, Design and Performance of a High Velocity Air-Sand Jet Impingement Erosion Facility, *Wear,* Vol 220, 1998, p 95–112

38. N. Gladys and V. Laura, Analysis of Rotor-Blade Failure due to High Temperature Corrosion/Erosion, *Surf. Coat. Technol.,* Vol 120–121, 1999, p 145–150

39. K.G. Budinski, Erosion of 316 Stainless Steel by Dicalcium Phosphate, *Wear,* Vol 186–187, 1995, p 145–149

40. E. Deaquino, D. Martínez, A. Perez, A. Velasco, M.A. Flores, and R. Viramontes, *Powder Technol.,* Vol 95, 1998, p 55–60

41. D.A. Brosnan, M.S. Crowley, and R.C. Johnson, CPI Drive Refractory Advances, *Chem. Eng.,* Vol 105, Oct 1998, p 110–120

42. G.P. Tilly, Sand Erosion of Metals and Plastics: A Brief Review, *Wear,* Vol 14, 1969, p 241–248

43. R.W. Bruce, Development of 1232 °C (2250 °F) Erosion and Impact Tests for Thermal Barrier Coatings, *Tribol. Trans.,* Vol 41, 1998, p 399–410

44. C.E. Smeltzer, M.E. Gulden, and W.A. Compton, Mechanism of Metal Removal by Impacting Dust Particles, *J. of Basic Eng. (Trans. ASME*), Vol 92, Sept 1970, p 639–654

45. J.G.A. Bitter, A Study of Erosion Phenomena, Parts I and II, *Wear,* Vol 6, 1963, p 5–21, 169–190.

46. A.K. Cousens and I.M. Hutchings, Influence of Erodent Shape on the Erosion of Mild Steel, *Proc. Sixth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1983, p 41-1 to 41-8

47. G. Sundararajan, A Comprehensive Model for the Solid Particle Erosion of Ductile Materials, *Wear of Materials 1991,* American Society of Mechanical Engineers, 1991, p 503–511

48. M. Roy, Y. Tirupatataiah, and G. Sundararajan, Effect of Particle Shape on the Erosion of Cu and Its Alloys, *Mater. Sci. Eng.,* Vol A165, 1993, p 51–63

49. P.V. Rao and D.H. Buckley, Time Effects of Erosion by Solid Particle Impingement on Ductile Materials, *Proc. Sixth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1983, p 38-1 to 38-10

50. I. Finnie, The Mechanism of Erosion of Ductile Metals, *Proc. Third U. S. National Congress of Applied Mechanics,* American Society of Mechanical Engineers, 1958, p 527–532

51. I. Finnie, Erosion of Surfaces by Solid Particles, *Wear,* Vol 3, 1960, p 87–103

52. I. Finnie, Erosion by Solid Particles in a Fluid Stream, *Symposium on Erosion and Cavitation,* STP 307, ASTM, 1962, p 70–82

53. J.H. Neilson and A. Gilchrist, Erosion by a Stream of Solid Particles, *Wear,* Vol 11, 1968, p 111–122

54. R.E. Winter and I.M. Hutchings, Solid Particle Erosion Studies Using Single Angular Particles, *Wear,* Vol 29, 1974, p 181–194

55. G.P. Tilly, A Two Stage Mechanism of Ductile Erosion, *Wear,* Vol 23, 1973, p 87–96

56. S.M. Wiederhorn and B.J. Hockey, Effect of Material Parameters on the Erosion Resistance of Brittle Materials, *J. Mater. Sci.,* Vol 18, 1983, p 766–780

57. A.W. Ruff and S.M. Wiederhorn, Erosion by Solid Particle Impact, *Treatise on Material Science and Technology,* Vol 16, Academic Press, 1979, p 69–126

58. A.G. Evans, M.E. Gulden, and M. Rosenblatt, Impact Damage in Brittle Materials in the Elastic-Plastic Response Regime, *Proc. R. Soc. (London) A,* Vol 361, 1978, p 343–365

59. R. Bellman, Jr. and A. Levy, Erosion Mechanism in Ductile Metals, *Wear,* Vol 70, 1981, p 1–27

60. A.V. Levy, The Platelet Mechanism of Erosion of Ductile Metals, *Wear,* Vol 108, 1986, p 1–21

61. I.M. Hutchings, A Model for the Erosion of Metals by Spherical Particles at Normal Incidence, *Wear,* Vol 70, 1981, p 269–281

62. P.S. Follansbee, G.B. Sinclair, and J.C. Williams, Modelling of Low Velocity Particulate Erosion in Ductile Materials by Spherical Particles, *Wear,* Vol 74, 1982, p 107–122

63. A.V. Levy and S. Jahanmir, The Effects of the Microstructure of Ductile Alloys on Solid Particle Erosion Behavior, *Corrosion-Erosion-Behavior of Materials,* TMS-AIME, 1980, p 177–189

64. S. Jahanmir, The Mechanics of Subsurface Damage in Solid Particle Erosion, *Wear,* Vol 61, 1980, p 309–324

65. T. Christman and P.G. Shewmon, Adiabatic Shear Localization and Erosion of Strong Aluminum Alloys, *Wear,* Vol 54, 1979, p 145–155

66. G. Sundararajan and P.G. Shewmon, A New Model for the Erosion of Metals at Normal Incidence, *Wear,* Vol 84, 1983, p 237–258

67. I.M. Hutchings, Mechanical and Metallurgical Aspects of the Erosion of Metals, *Proc. Corrosion/Erosion of Coal Conversion Systems Materials Conf.,* National Association of Corrosion Engineers, 1979, p 393–428

68. P.H. Shipway and I.M. Hutchings, The Role of Particle Properties in the Erosion of Brittle Materials, *Wear,* Vol 193, 1996, p 105–113

69. G. Sundararajan and M. Roy, Solid Particle Erosion Behavior of Metallic Materials at Room and Elevated Temperatures, *Tribol. Int.,* Vol 30, 1997, p 339–359

Thomas A. Adler, U.S. Department of Energy

Erosion of Ductile Materials

The erosion of ductile materials by solid particles can be separated into two components, a cutting component and a deformation component, based on the angle of impact of the particles. The component of the particle velocity parallel to the surface produces the cutting component of erosion. Erosion by cutting can be considered similar to chip formation in machining, in that pieces of material are removed by the impact of a sharp-edged particle. This component dominates the wear rate at low impact angles. A model for the cutting component was introduced by Finnie (Ref 50, 51, 52) and was developed by Bitter (Ref 45) and Neilson and Gilchrist (Ref 53):

$$f_c(\alpha) = A \cos^2 \alpha \sin n\alpha \qquad \text{(Eq 2)}$$

where n is $\pi/2\alpha_0$ for $\alpha < \alpha_0$, n is $\pi/2\alpha$ for $\alpha > \alpha_0$, and α_0 and A are constants. This function is illustrated in Fig. 3 (Ref 70). The model originally developed by Finnie overestimated the erosion rate, due to material piling up on the sides of the impact site instead of the groove being totally removed as wear debris (Ref 71, 72). However, the model does correctly account for the shape of the erosion curve as a function of impact angle.

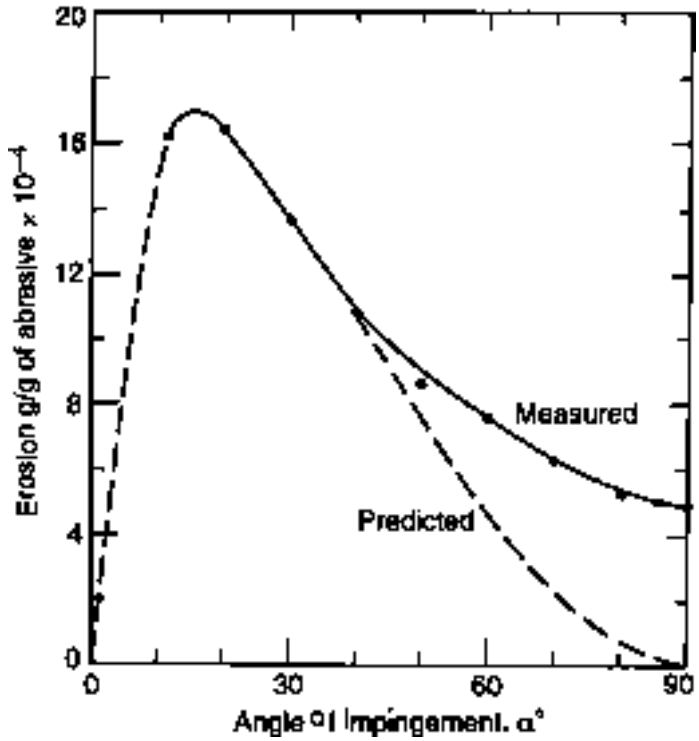


Fig. 3 Experimental and predicted curves of erosion of aluminum by 120-mesh silicon carbide particles at 152 m/s (500 ft/s)

The other component of erosion is produced by the component of velocity perpendicular to the surface. Bitter called this component *deformation wear* (Ref 45). This component involves repeated battering of the surface, leading to the eventual fracture and removal of wear debris (Ref 72). This component was modeled by Bitter (Ref 45) and was developed by Neilson and Gilchrist (Ref 53):

$$f_d(\alpha) = B \sin^2 \alpha \qquad \text{(Eq 3)}$$

where B is a constant. This component has the shape of the glass curve in Fig. 1. Many mechanisms have been suggested for the deformation component, including an extrusion mechanism with subsequent lip removal (Ref 59, 60), a low-cycle fatigue mechanism (Ref 61, 62), a delamination mechanism (Ref 63, 64), a localization of plastic deformation mechanism (Ref 47, 66), a work hardening with embrittlement mechanism (Ref 45), and a mechanism involving melting (Ref 44). Of these, most studies seem to favor an extrusion with subsequent lip formation, delamination, low-cycle fatigue, or localization of plastic deformation mechanism. Combining the two components leads to the complete angular term:

$$f(\alpha) = A \cos^2 \alpha \sin n\alpha + B \sin^2 \alpha \qquad \text{(Eq 4)}$$

Particle shape plays a significant role in erosion. Angular particles produce much larger erosion rates than spherical particles, especially at low impact angles (Ref 46, 47, 48, 73, 74, 75, 76). The change in erosion rate due to particle shape depends mostly on the cutting component (although the particle shape appears to play a role in the deformation component also [Ref 47]). Spherical particles produce little or no cutting wear, which results in $B > A$ in Eq 4. The appearance of a surface eroded by spherical particles is shown in Fig. 4 (Ref 77). This surface illustrates the extruded lip formation or platelet mechanism of the deformation component of erosion. Angular particles usually produce large quantities of cutting wear, which results in $A > B$ in Eq 4. The appearance of a surface eroded by angular particles is shown in Fig. 5 (Ref 78). This surface illustrates the cutting mechanism of erosion. The role of particle shape on the erosion rate has been illustrated a number of times, but it is difficult to quantify, except empirically, because it depends on a large number of properties of both the eroding particle and the surface being eroded (Ref 3, 4, 5, 23, 33, 36, 53, 76).

Fig. 4 Surface of 1100-O aluminum eroded by 700 μm diameter spherical steel shot at 60° impact angle

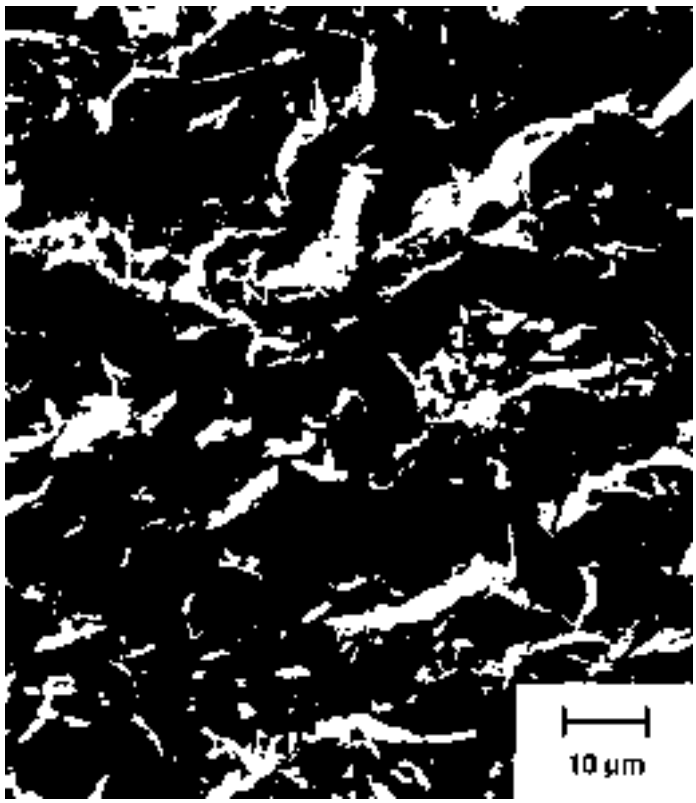


Fig. 5 Surface of 1020 steel eroded by SiC at 80 m/s (260 ft/s) and 30° impact angle

The constant K_0 in Eq. 1 includes the effects of a number of other factors, such as the size of the impacting particles, the hardness of the particles relative to the hardness of the eroded surface, and temperature effects. The first of these factors is the size of the particles (Ref 55, 73, 79). For very small particles (less than about 5 μm), there is no measurable erosion. As the particle size increases from this lower limit, the erosion rate increases until an upper limit is reached. Above this upper limit of particle size, the erosion rate is independent of particle size. This upper limit is approximately 50 to 100 μm. The particle size factor, K_s, can be modeled as (Ref 55):

$$K_s = \left[1 - \left(\frac{d_0}{d}\right)^{3/2} \left(\frac{V_0}{V}\right)\right]^2 \qquad \text{(Eq 5)}$$

where d is the particle size, V is the velocity, and d_0 and V_0 are constants.

The second factor is the hardness of the particles relative to the hardness of the surface (Ref 74, 76, 80). This factor is shown in Fig. 6. When particles are more than twice as hard as the surface, the erosion rate is

independent of the particle hardness. When particles are less than half the hardness of the surface, no erosion is measurable. Between these limits, there is a continuous transition. This factor, K_h, has been modeled as (Ref 76):

$$K_h = 1 - \exp\left[-\ln 2\left(\frac{H_p}{mH}\right)^{1-2m}\right] \qquad \text{(Eq 6)}$$

where H_p is the hardness of the particle, H is the hardness of the surface, and m is a constant.

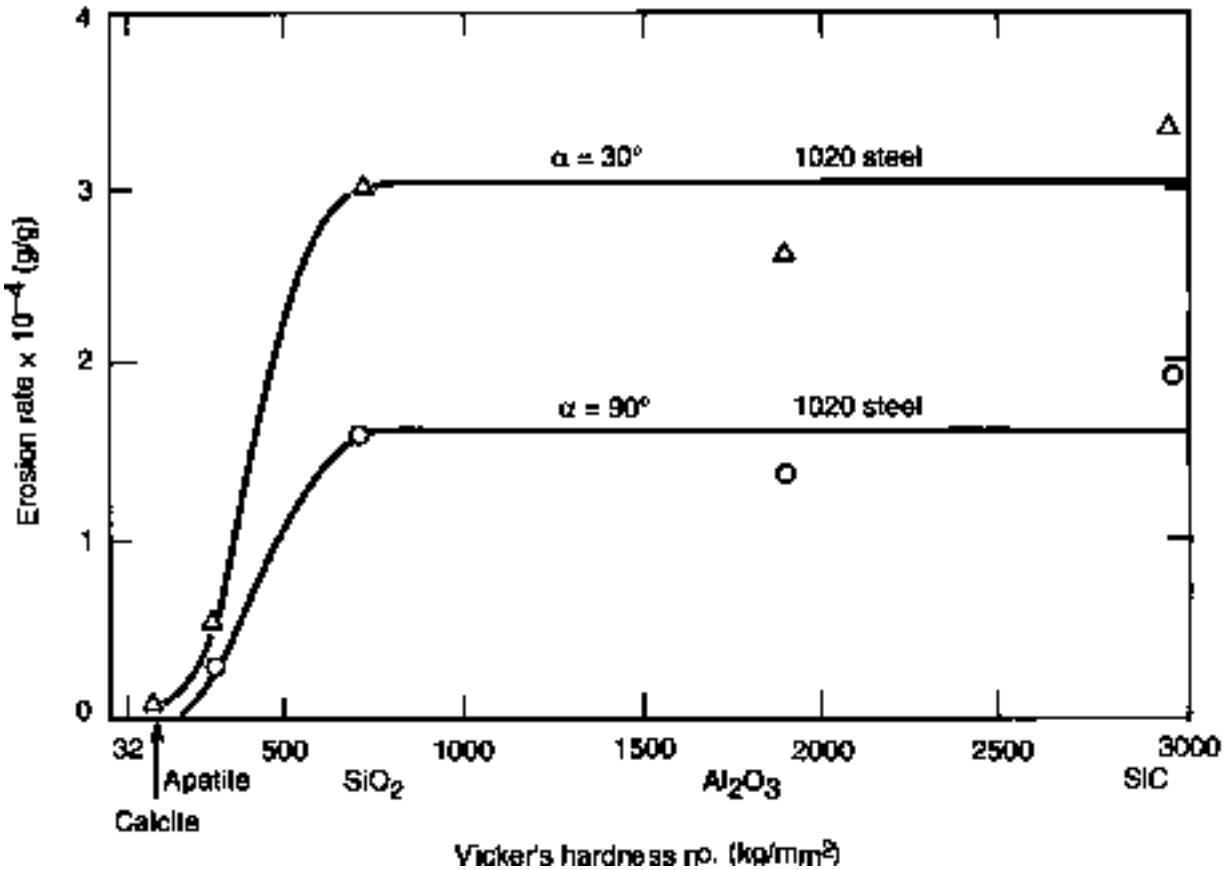


Fig. 6 Erosion rates of 1020 steel by 180 to 250 μm (7 to 10 mil) particles at 80 m/s (260 ft/s)

Increasing the temperature has a mixed effect on the erosion rate for ductile materials (Ref 69, 81). A substantial amount of high-temperature erosion testing has been done in support of the gas turbine and coal gasification industries. This testing has provided a substantial amount of results on temperature effects.

Corrosion may increase or decrease the apparent erosion rate, depending on the rate of oxide or other corrosion product formation and the resistance of the product to erosion as compared to the normal surface material (Ref 69, 81). If the corrosion product is slow growing and more erosion resistant than the underlying metal, it may effectively protect the metal from erosion. Similarly, deposition of some process material on the surface may also protect it. These protective mechanisms have been observed on some boiler heat-exchanger tubes, but are probably more the exception than the rule. More often, oxidation of the underlying metal with rapid removal of the fragile and brittle oxide (particularly at high angles of impact) accelerates erosion.

The three factors—particle size, particle hardness, and temperature (K_t)—can be combined as a product:

$$K_0 = K_s K_h K_t \qquad \text{(Eq 7)}$$

Any additional factors, such as fragmentation of the particles, could also be included in Eq 7.

References cited in this section

3. B.-E. Lee, C.A.J. Fletcher, and M. Behnia, Computational Study of Solid Particle Erosion for a Single Tube in Cross Flow, *Wear,* Vol 240, 2000, p 95–99

4. B.E. Lee, C.A.J. Fletcher, and M. Behnia, Computational Prediction of Tube Erosion in Coal Fired Power Utility Boilers, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 121, Oct 1999, p 746–750

5. J.F. Drennen and J.G. McGowan, Prediction of Flyash Erosion in Convective Pass Sections of Coal Fired Systems, *Proc. Seventh Int. Conf. Erosion by Liquid and Solid Impacts,* University of Cambridge, 1987, p 75-1 to 75-6

23. M. Menguturk and E.F. Sverdrup, Calculated Tolerance of a Large Electric Utility Gas Turbine to Erosion Damage by Coal Gas Ash Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 193–224

33. A. Drotlew, P. Christodoulou, and V. Gutowski, Erosion of Ferritic Fe-Cr-C Cast Alloys at Elevated Temperatures, *Wear,* Vol 211, 1997, p 120–128

36. K. Haugen, O. Kvernvold, A. Ronold, and R. Sandberg, Sand Erosion of Wear-Resistant Materials: Erosion in Choke Valves, *Wear,* Vol 186–187, 1995, p 179–188

44. C.E. Smeltzer, M.E. Gulden, and W.A. Compton, Mechanism of Metal Removal by Impacting Dust Particles, *J. of Basic Eng. (Trans. ASME*), Vol 92, Sept 1970, p 639–654

45. J.G.A. Bitter, A Study of Erosion Phenomena, Parts I and II, *Wear,* Vol 6, 1963, p 5–21, 169–190.

46. A.K. Cousens and I.M. Hutchings, Influence of Erodent Shape on the Erosion of Mild Steel, *Proc. Sixth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1983, p 41-1 to 41-8

47. G. Sundararajan, A Comprehensive Model for the Solid Particle Erosion of Ductile Materials, *Wear of Materials 1991,* American Society of Mechanical Engineers, 1991, p 503–511

48. M. Roy, Y. Tirupatataiah, and G. Sundararajan, Effect of Particle Shape on the Erosion of Cu and Its Alloys, *Mater. Sci. Eng.,* Vol A165, 1993, p 51–63

50. I. Finnie, The Mechanism of Erosion of Ductile Metals, *Proc. Third U. S. National Congress of Applied Mechanics,* American Society of Mechanical Engineers, 1958, p 527–532

51. I. Finnie, Erosion of Surfaces by Solid Particles, *Wear,* Vol 3, 1960, p 87–103

52. I. Finnie, Erosion by Solid Particles in a Fluid Stream, *Symposium on Erosion and Cavitation,* STP 307, ASTM, 1962, p 70–82

53. J.H. Neilson and A. Gilchrist, Erosion by a Stream of Solid Particles, *Wear,* Vol 11, 1968, p 111–122

55. G.P. Tilly, A Two Stage Mechanism of Ductile Erosion, *Wear,* Vol 23, 1973, p 87–96

59. R. Bellman, Jr. and A. Levy, Erosion Mechanism in Ductile Metals, *Wear,* Vol 70, 1981, p 1–27

60. A.V. Levy, The Platelet Mechanism of Erosion of Ductile Metals, *Wear,* Vol 108, 1986, p 1–21

61. I.M. Hutchings, A Model for the Erosion of Metals by Spherical Particles at Normal Incidence, *Wear,* Vol 70, 1981, p 269–281

62. P.S. Follansbee, G.B. Sinclair, and J.C. Williams, Modelling of Low Velocity Particulate Erosion in Ductile Materials by Spherical Particles, *Wear,* Vol 74, 1982, p 107–122

63. A.V. Levy and S. Jahanmir, The Effects of the Microstructure of Ductile Alloys on Solid Particle Erosion Behavior, *Corrosion-Erosion-Behavior of Materials,* TMS-AIME, 1980, p 177–189

64. S. Jahanmir, The Mechanics of Subsurface Damage in Solid Particle Erosion, *Wear,* Vol 61, 1980, p 309–324

66. G. Sundararajan and P.G. Shewmon, A New Model for the Erosion of Metals at Normal Incidence, *Wear,* Vol 84, 1983, p 237–258

69. G. Sundararajan and M. Roy, Solid Particle Erosion Behavior of Metallic Materials at Room and Elevated Temperatures, *Tribol. Int.,* Vol 30, 1997, p 339–359

70. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 7

71. I.M. Hutchings, Mechanisms of the Erosion of Metals by Solid Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 59–76

72. I. Finnie, A. Levy, and D.H. McFadden, Fundamental Mechanisms of the Erosive Wear of Ductile Metals by Solid Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 36–58

73. S. Bahadur and R. Badruddin, Erodent Particle Characterization and the Effect of Particle Size and Shape on Erosion, *Wear of Materials 1989,* Vol 1, American Society of Mechanical Engineers, 1989, p 143–153

74. A.V. Levy and P. Chik, The Effects of Erodent Composition and Shape on the Erosion of Steel, *Wear,* Vol 89, 1983, p 151–162

75. I. Kleis, Probleme der Bestimmung des Strahlverschleisses bei Metallen (The Problem of the Determination of the Wear of Metals by a Beam (of Abrasive Particles)), *Wear,* Vol 13, 1969, p 199–215

76. A. Magnée, Generalized Law of Erosion: Application to Various Alloys and Intermetallics, *Wear,* Vol 181–183, 1995, p 500–510

77. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 15

78. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 63

79. A. Mishra and I. Finnie, On the Size Effect in Abrasive and Erosive Wear, *Wear,* Vol 65, 1981, p 359–373

80. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 62

81. J.R. Nicholls, Laboratory Studies of Erosion-Corrosion Processes under Oxidizing and Oxidizing/Sulphidising Conditions, *Mater. High Temp.,* Vol 14, 1997, p 289–306

Erosive Wear Failures

Thomas A. Adler, U.S. Department of Energy

Erosion of Brittle Materials

Erosion of brittle materials by hard, solid particles involves a process in which material is lost from the target by brittle fracture (Ref 27, 56, 57, 58, 68). The size and shape of the cracks that form in the surface depend on several factors. These include particle shape, mass, and velocity and target material hardness and toughness. The erosion rate has been found to correlate with the fracture toughness and hardness, as in the following equation (Ref 56):

$$E = K_2 V^p R^a \rho^b K_c^c H^d f(\alpha) \qquad \text{(Eq 8)}$$

where V is the velocity, R is the particle radius, ρ is the particle density, K_c is the surface material fracture toughness, H is the surface material hardness, and p, a, b, c, d, and K_2 are constants. A regression analysis found the following for the exponents: p = 2.8, a = 3.9, b = 1.4, c = -1.9, and d = 0.48. This indicates that the erosion rate increases as the hardness of the surface material increases and the fracture toughness of the surface material decreases. These exponents were compared to theoretical models based on elastic-plastic fracture (Ref 57, 58), and rough agreement was found (Ref 56). The variation of erosion as a function of impact angle for brittle materials is shown in Fig. 1 for glass. Equation 3 can be used to model this curve (Ref 45, 57).

References cited in this section

27. R.G. Wellman and C. Allen, The Effects of Angle of Impact and Materials Properties on the Erosion Rates of Ceramics, *Wear,* Vol 186–187, 1995, p 117–122

45. J.G.A. Bitter, A Study of Erosion Phenomena, Parts I and II, *Wear,* Vol 6, 1963, p 5–21, 169–190.

56. S.M. Wiederhorn and B.J. Hockey, Effect of Material Parameters on the Erosion Resistance of Brittle Materials, *J. Mater. Sci.,* Vol 18, 1983, p 766–780

57. A.W. Ruff and S.M. Wiederhorn, Erosion by Solid Particle Impact, *Treatise on Material Science and Technology,* Vol 16, Academic Press, 1979, p 69–126

58. A.G. Evans, M.E. Gulden, and M. Rosenblatt, Impact Damage in Brittle Materials in the Elastic-Plastic Response Regime, *Proc. R. Soc. (London) A,* Vol 361, 1978, p 343–365

68. P.H. Shipway and I.M. Hutchings, The Role of Particle Properties in the Erosion of Brittle Materials, *Wear,* Vol 193, 1996, p 105–113

Erosive Wear Failures

Thomas A. Adler, U.S. Department of Energy

Examples of Erosive Wear Failures

Erosive wear occurs in many situations where a large number of small solid or liquid particles impact against a surface or where the collapse of gas-filled bubbles in a cavitating liquid causes surface damage. Some examples of erosive wear failures are given in the following sections on abrasive erosion, liquid impingement erosion, cavitation, and erosion-corrosion. Liquid impingement erosion and cavitation are also discussed further in separate articles in this Volume.

Abrasive Erosion. The term *abrasive erosion* is sometimes used to describe erosion in which the solid particles move nearly parallel to the solid surface. For example, abrasive erosion in pump components generally occurs due to the impingement of particles at very oblique angles to the surface. Such abrasion attacks ductile materials much more readily than it does hard, relatively brittle materials, as pointed out by Finnie (Ref 82). Particulate erosion of a ductile steel impeller by a catalytic fluid that was not expected to be erosive is shown in Fig. 7.

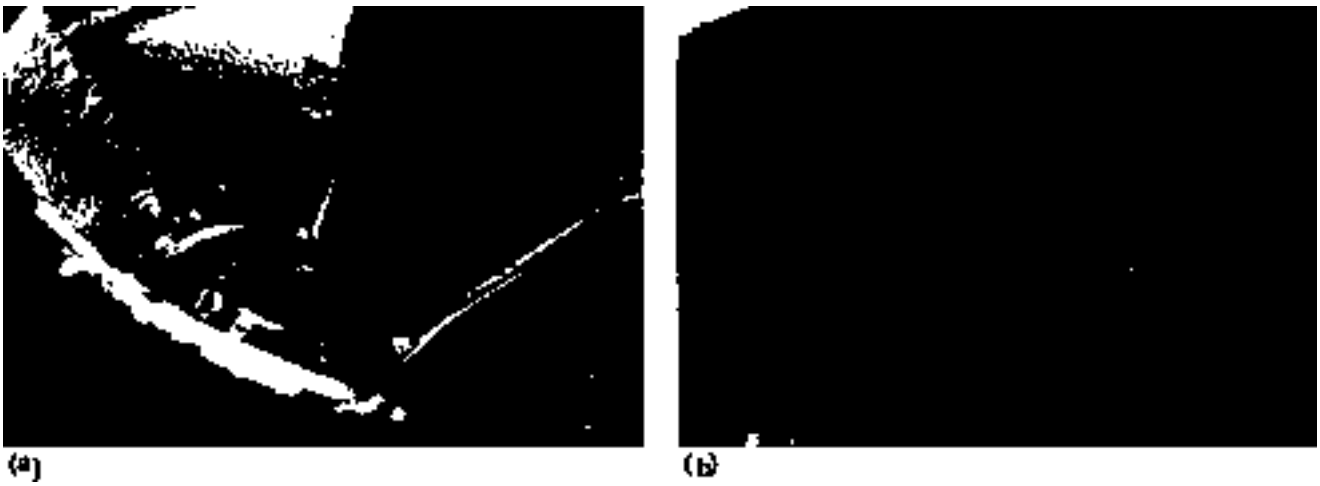


Fig. 7 Particulate erosion of a ductile steel impeller by an abrasive catalyst. (a) 0.25×. (b) 1×

Erosion-resistant metals include white cast iron (standard gray or ductile cast irons are poor), high-chromium (that is, 13 to 28%) alloy steel, cobalt-base superalloys such as Stellite, and special nickel-base alloys such as Ni-Hard. These materials are useful not only for the flow-path surfaces but also as sleeves in sealing areas, particularly if packing is used. It is good practice to use hard metal or ceramic expellers near seal entrances to keep particles out of the sealing clearance. If thorough flushing of packing presents a design problem, abrasion-resistant SiC ceramic mechanical seals together with an expeller or minimal flushing can be used. Some of the newest designs of mechanical seals for abrasive service are even axially split for ease of replacement.

Another example of abrasive erosion is the impact of fly ash entrained in the flue gases in screen tubes or superheater tubes of boiler systems. Erosion is enhanced by high flow velocities; thus, partial fouling of gas passages in tube bands by deposition of fly ash can lead to erosion by forcing the flue gases to flow through smaller passages at higher velocity. This effect, sometimes called laning, exposes tube surfaces to a greater probability of impact by particles having higher kinetic energy, thus increasing the rate of damage. Erosion by fly ash causes polishing, flat spots, wall thinning, and eventual tube rupture. Fly-ash erosion can be controlled by coating tube surfaces with refractory cements or other hard, wear-resistant materials, although this reduces the heat-transfer capability of the surface. An alternative method is the use of baffles to channel gas flow away from critical areas.

Liquid Impingement Erosion. Erosion can take place in a liquid medium even without the presence of solid abrasive particles in that medium. Liquid droplets impacting a solid surface collision at high speed is the form of liquid erosion called liquid impingement erosion or liquid impact erosion, as discussed in more detail in the article “Liquid-Impact Erosion” in this Volume and in the article “Liquid Impingement Erosion” in *Friction, Lubrication, and Wear Technology*, Volume 18 of *ASM Handbook*. Corrosion may or may not occur simultaneously with liquid impingement erosion. When corrosion is involved, the term *erosion-corrosion* is often used.

A simplified schematic of liquid impingement erosion processes is illustrated in Fig. 8. The impact-pressure history (and resulting stress waves) can produce circumferential cracks or deformation patterns around the initial area of impact (Fig. 8a), depending on the properties of the surface material and the energy of the impact. Following impact, the liquid flows away radially at high velocity. The spreading liquid may hit nearby surface asperities or the surface steps resulting from plastic deformation caused by the initial impact pressure. The force of this impact stresses the asperity or surface step at its base and may produce a crack (Fig. 8b). Subsequent impacts by other drops may widen the crack or detach a particle entirely (Fig. 8c). Direct hits on existing cracks, pits, or other deep depressions can produce accelerated damage by a microjet impingement mechanism (Fig. 8d). Eventually, the pits and secondary cracks intersect, and larger pieces of the surface become detached.

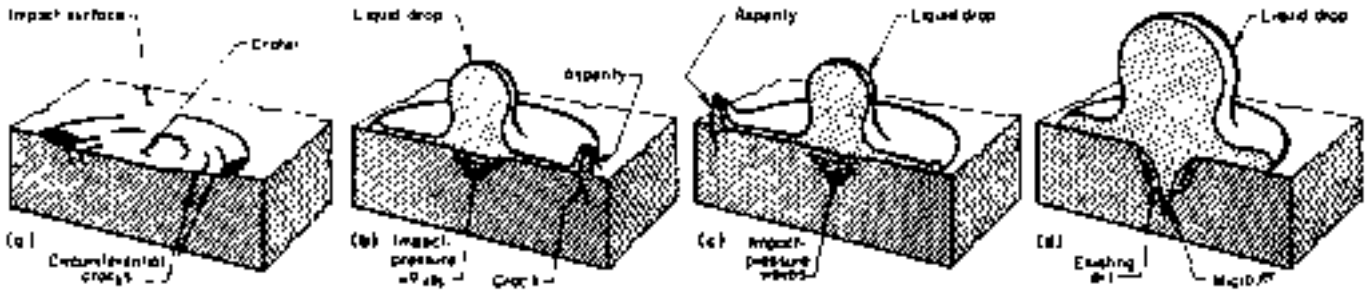


Fig. 8 Processes by which a material is damaged by liquid impingement erosion. (a) Solid surface showing initial impact of a drop of liquid that produces circumferential cracks in the area of impact or produces shallow craters in very ductile materials. (b) High-velocity radial flow of liquid away from the impact area is arrested by a nearby surface asperity, which cracks at its base. (c) Subsequent impact by another drop of liquid breaks the asperity. (d) Direct hit on a deep pit results in accelerated damage, because shock waves bouncing off the sides of the pit cause the formation of a high-energy microjet within the pit.

Among the components most susceptible to liquid impingement erosion are low-pressure turbine blades, low-temperature steam piping, and condenser or other heat-exchanger tubes that are subjected to direct impingement by wet steam. Liquid impingement erosion in tubing or piping is most likely when fluid velocities exceed 2.1 m/s (7 ft/s). Damage occurs first at locations where direction of flow changes, such as elbows or U-bends. Large-radius bends are less susceptible to such damage; however, use of erosion-resistant materials, such as austenitic stainless steel, is often more effective. Erosion of heat-exchanger tubes caused by impingement of wet steam can sometimes be eliminated by redirecting flow with baffles.

Moisture erosion of low-pressure blades has been a problem throughout steam turbine history and remains a concern today. For example, Fig. 9 shows two portions of the leading edge of a blade from the last stage of a low-pressure steam turbine. Liquid impingement erosion in its advanced stages is often characterized by a surface that appears jagged, and composed of sharp peaks and pits. One portion (at left, Fig. 9) was protected by an erosion shield of Stellite 6B; the other portion (at right, Fig. 9) was unprotected type 403 (modified) stainless steel. The shield, made of 1 mm (0.04 in.) thick rolled strip and brazed onto the leading edge, resisted erosion quite effectively, but the unprotected base metal did not.

Fig. 9 Two portions of a modified type 403 stainless steel steam turbine blade damaged by liquid impingement erosion. The portion at left was protected by a shield of 1 mm (0.04 in.) thick rolled Stellite 6B brazed onto the leading edge of the blade; the portion at right was unprotected. Compare amounts of metal lost from protected and unprotected portions. Both 2.5×

Both blade portions shown in Fig. 9 also illustrate the dependence of erosion on hydrodynamic intensity. Damage was most severe at the leading edge, where hydrodynamic intensity was greatest. Away from the

leading edge, impacting droplets were smaller, impact velocity was lower, and impact occurred at oblique rather than right angles; therefore, damage was progressively less severe.

It may be impossible to reduce the hydrodynamic intensity significantly without seriously degrading performance. In such instances, the use of erosion-resistant metals may be the only practical solution to a problem of liquid erosion. Many of the erosion-resistant metals can be applied as welded overlays; this makes salvage or repair of damaged surfaces easier or surface treatment of new components less costly than would be possible if the component had to be made entirely of erosion-resistant metal. Because liquid erosion is basically a surface phenomenon, the use of erosion-resistant overlays is effective in combating damage. The Stellite alloys and stainless steels are the alloys most widely used as overlays.

However, many small parts are not amenable to protection by the use of erosion-resistant overlays. Therefore, the most effective means of combatting erosion of small parts is to increase the hardness of the metal or to specify a more erosion-resistant metal.

Erosion-Corrosion. When movement of a corrodent over a metal surface increases the rate of attack due to mechanical wear and corrosion, the attack is called *erosion-corrosion.* It is encountered when particles in a liquid impinge on a metal surface, causing the removal of protective surface films, such as air-formed protective oxide films or adherent corrosion products, and exposing new reactive surfaces that are anodic to uneroded neighboring areas on the surface. This results in rapid localized corrosion of the exposed areas in the form of smooth-bottomed shallow recesses.

Nearly all flowing or turbulent corrosive media can cause erosion-corrosion. The attack may exhibit a directional pattern related to the path taken by the corrodent as it moves over the surface of the metal. Figure 10 shows the interior of a 50 mm (2 in.) copper river waterline that has suffered pitting and general erosion due to excessive velocity of the water. The brackish river water contained some suspended solids that caused the polishing of the copper pipe surface. The horseshoe-shaped pits (facing upstream) are typical of the damage caused by localized turbulence. The copper pipe was replaced with fiberglass-reinforced plastic piping.

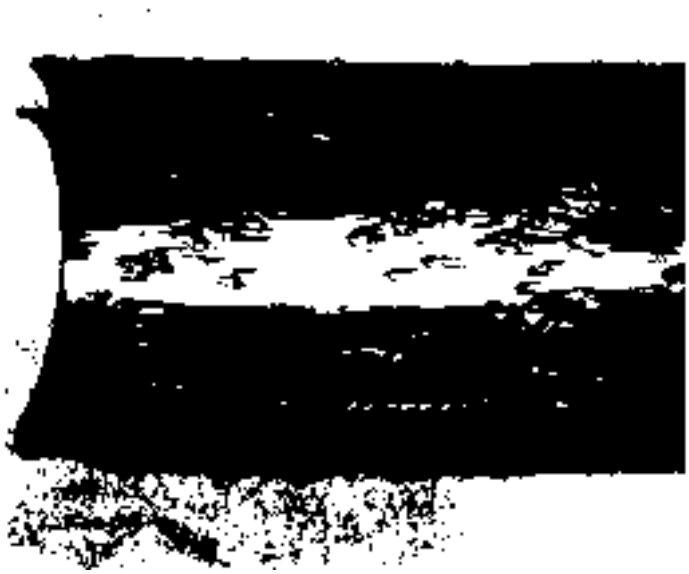


Fig. 10 Erosion pitting caused by turbulent river water flowing through copper pipe. The typical horseshoe-shaped pits point upstream. 0.5×

In piping systems, erosion-corrosion can be reduced by increasing the diameter of the pipe, thus decreasing velocity and turbulence. The streamlining of bends is useful in minimizing the effects of impingement. Inlet pipes should not be directed onto the vessel walls, if this can be avoided. Flared tubing can be used to reduce problems at the inlet tubes in a tube bundle. Erosion-corrosion presents a distinct appearance. The metal surface usually exhibits severe weight loss, with many hollowed-out regions; the overall surface has a carved appearance along the pipe or tube wall.

Impingement attack or corrosion is a term used to describe some severe form of erosion-corrosion, where impingement is encountered and erosion is more intense. It occurs frequently in turns or elbows of tubes or

pipes or on surfaces of impellers or turbines. An example of this type of erosion-corrosion is shown in Fig. 11. The combined effects of high velocities and corrosion are clearly visible on the pump impeller. Impingement corrosion attack can also occur as the result of partial blockage of a tube. The impinging stream can rapidly perforate tube walls, especially if silt or mud have an additional erosive effect. Steam erosion is another form of impingement corrosion. It occurs when high-velocity wet steam contacts a metal surface. The resulting attack usually produces a roughened surface showing a large number of small cones with the points facing in the direction of flow.

Fig. 11 The classic appearance of erosion-corrosion in a CF-8M pump impeller

Impingement attack often produces a horseshoe-shaped pit, where the deep undercut end points in the direction of flow. Impingement corrosion attack can also occur as the result of partial blockage of a tube. This form of attack frequently occurs in condenser systems handling seawater or brackish water that contains entrained air or solid particles and circulates through the system at relatively high velocities and with turbulent flow. The pits are elongated in the direction of flow and are undercut on the downstream side. When the condition becomes serious, a series of horseshoe-shaped grooves with the open ends on the downstream side may be formed. As the attack progresses, the pits may join one another, forming fairly large patches of undercut pits.

When impingement attack occurs in heat-exchanger or condenser tubing, it is usually confined to a short distance on the inlet end of the tube where the fluid flow is turbulent. Impingement attack can be controlled by use of corrosion-resistant alloys, such as copper alloy C71500 (copper nickel, 30%). Impingement attack also has been somewhat relieved by inserting relatively short, tapered plastic or alloy sleeves inside the tube on the inlet end of the tube bundle. Other corrective measures include decreasing the velocity, streamlining the fluid flow, and decreasing the amount of entrained air or solid particles. These may be attained by a streamlined design of water boxes, injector nozzles, and piping. Abrupt angular changes in direction of fluid flow, low-pressure pockets, obstruction to smooth flow, and any other feature that can cause localized high velocities or turbulence of the circulating water should be minimized as much as possible. Sacrificial impingement baffles can minimize damage to the shell-side surfaces of tubes.

Cavitation erosion is the most severe form of erosion-corrosion, although cavitation damage may occur without active corrosion. Cavitation principally occurs when relative motion between a metal surface and a liquid environment causes vapor bubbles to appear. When the bubbles collapse, they impose hammerlike blows

simultaneously with the initiation of tearing action that appears to pull away portions of the surface. Although the tearing action can remove any protective oxide film that exists on the surface of a metal, exposing active metal to the corrosive influence of the liquid environment, corrosion is not essential to cavitation erosion. Whenever high velocities give rise to extremely low-pressure areas, as in a jet or rotary pump, vapor bubbles collapse at high-pressure areas and destroy the protective film on the metal surface or disrupt the metal itself.

Cavitation erosion occurs typically on rotors or pumps, on the trailing faces of propellers and of water turbine blades, and on the water-cooled side of diesel engine cylinders. Damage can be reduced by operating rotary pumps at the highest possible head of pressure to avoid formation of bubbles. For turbine blades, aeration of water serves to cushion the damage caused by the collapse of bubbles. Neoprene or similar elastomer coatings on metals are reasonably resistant to damage from this cause.

Cavitation erosion occurs in regions of a system where a combination of temperature and flow velocity causes growth and subsequent collapse of large quantities of vapor-filled bubbles in a flowing stream of liquid. Damage occurs at the points of bubble collapse. In steam plants, the components most often damaged by cavitation erosion are feed lines, pump casings, and pump impellers. Figure 12 shows the typical appearance of a part extensively damaged by cavitation erosion; the surface is deeply and irregularly pitted.

Fig. 12 Cast iron suction bell, 455 mm (18 in.) in diameter, from a low-pressure general service water pump that failed by cavitation erosion after about 5 years of service. Note the deeply pitted surface and the irregular shape of the erosion pattern, both of which are typical characteristics of cavitation damage.

Cavitation erosion is most effectively avoided by reduction of either flow velocity or temperature (reduction of flow velocity is preferable). Other measures include selection of material and the use of erosion-resistant welded overlays.

Reference cited in this section

82. I. Finnie, Some Observations on the Erosion of Ductile Metals, *Wear,* Vol 19, 1972, p 81–90

Erosive Wear Failures

Thomas A. Adler, U.S. Department of Energy

References

1. "Standard Terminology Relating to Wear and Erosion," G 40, *Annual Book of ASTM Standards,* ASTM

2. J. Stringer, Practical Experience with Wastage at Elevated Temperatures in Coal Combustion Systems, *Wear,* Vol 186–187, 1995, p 11–27

3. B.-E. Lee, C.A.J. Fletcher, and M. Behnia, Computational Study of Solid Particle Erosion for a Single Tube in Cross Flow, *Wear,* Vol 240, 2000, p 95–99

4. B.E. Lee, C.A.J. Fletcher, and M. Behnia, Computational Prediction of Tube Erosion in Coal Fired Power Utility Boilers, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 121, Oct 1999, p 746–750

5. J.F. Drennen and J.G. McGowan, Prediction of Flyash Erosion in Convective Pass Sections of Coal Fired Systems, *Proc. Seventh Int. Conf. Erosion by Liquid and Solid Impacts,* University of Cambridge, 1987, p 75-1 to 75-6

6. J.J. Xie and P.M. Walsh, Erosion-Oxidation of Carbon Steel in the Convection Section of an Industrial Boiler Cofiring Coal-Water Fuel and Natural Gas, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 119, July 1997, p 717–722

7. M. Sucking and C. Allen, Critical Variables in High Temperature Erosive Wear, *Wear,* Vol 203–204, 1997, p 528–536

8. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 122–128

9. A.V. Levy, B.Q. Wang, G.Q. Geng, and W. Mack, Erosion-Corrosion of Steels in the Convection Pass Region of a Commercial Circulating Fluidized Bed Combustor, *Wear of Materials 1991,* American Society of Mechanical Engineers, 1991, p 697–701

10. A.V. Levy, B.Q. Wang, G.Q. Geng, and W. Mack, "Erosivity of Particles in Circulating Fluidized Bed Combustors," Paper 543, Corrosion 89, National Association of Corrosion Engineers, 1989

11. A. Levy and Y.-F. Man, Erosion-Corrosion of Chromium Steels, *Proc. Corrosion-Erosion-Wear of Materials at Elevated Temperatures, 1986,* National Association of Corrosion Engineers, 1987, p 168–203

12. A.V. Levy, B.-Q. Wang, Y.-F. Man, and N. Jee, Erosion-Corrosion of Steels in Simulated and Actual Fluidized Bed Combustor Environments, *Wear,* Vol 131, 1989, p 85–103

13. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 98–102

14. H.H. Krause and I.G. Wright, Boiler Tube Failures in Municipal Waste-to-Energy Plants, *Mater. Perform.* Vol 35, Jan 1996, p 46–54

15. N.G. Solomon, Erosion-Resistant Coatings for Fluidized Bed Boilers, *Mater. Perform.* Vol 37, 1998, p 38–43

16. P.M. Rogers, I.M. Hutchings, J.A. Little, and F. Kara, Microstructural Characterization of Surface Layers Formed on Alloy Steel During Erosion-Oxidation in a Fluidized Bed, *J. Mater. Sci.* Vol 32, 1997, p 4575–4583

17. M.S. Crowley, Know Causes of, Solutions to, CFB-Boiler Refractory Problems, *Power,* Vol 136, Jan 1992, p 54–57

18. R.C. Tucker and A.A. Ashari, The Structure Property Relationship of Erosion Resistant Thermal Spray Coatings, *Proc. 15th Int. Thermal Spray Conf.,* ASM International, 1998, p 259–262

19. P.N. Walsh, J.M. Quets, and R.C. Tucker, Jr., Coatings for the Protection of Turbine Blades from Erosion, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 117, Jan 1995, p 152–155

20. I.A. Diaz-Tous, A.H. Khan, and T.H. McCloskey, Solid Particle Erosion Technology Assessment, *Advances in Steam Turbine Technology for the Power Generation Industry, PWR,* Vol 26, American Society of Mechanical Engineers, 1994, p 247–254

21. W. Tabakoff, M. Metwally, and A. Hamed, High-Temperature Coatings for Protection against Turbine Deterioration, *J. Eng. Gas Turbines Power (Trans. ASME),* Vol 117, Jan 1995, p 146–151

22. J.E. Restall and D.J. Stephenson, High Temperature Erosion of Coated Superalloys for Gas Turbines, *Mater. Sci. Eng.,* Vol 88, 1987, p 273–282

23. M. Menguturk and E.F. Sverdrup, Calculated Tolerance of a Large Electric Utility Gas Turbine to Erosion Damage by Coal Gas Ash Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 193–224

24. S. Seetharamu, P. Sampathkumaran, and R.K. Kumar, Erosion Resistance of Permanent Moulded High Chromium Iron, *Wear,* Vol 186–187, 1995, p 159–167

25. C.R. Dimond, The Specification and Installation of Alumina Ceramics in Industry, *Mater. Des.,* Vol 8 May/June 1987, p 152–155

26. T.D. Johnson, M.B.J. Low, M.T. Parry, and D.J. Wall, Erosion in Pulverized Fuel Distribution Systems: Materials Selection and Plant Performance, *Proc. Seventh Int. Conf. Erosion by Liquid and Solid Impacts,* University of Cambridge, 1987, p 74-1 to 74-8

27. R.G. Wellman and C. Allen, The Effects of Angle of Impact and Materials Properties on the Erosion Rates of Ceramics, *Wear,* Vol 186–187, 1995, p 117–122

28. S.J. Dapkunas, A Summary of Erosion Experience in the U.S. Department of Energy's Coal Conversion Plants, *Proc. Fifth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1979, p 43-1 to 43-14

29. S. Danyluk, W.J. Shack, J.Y. Park, and M.M. Mamoun, Solid Particle Erosion of a Cyclone from a Coal Gasification Pilot Plant, *Wear of Material 1981,* American Society of Mechanical Engineers, 1981, p 619–624

30. C.A. Youngdahl, K. Pater, and M.J. Gorski, Erosive Wear and Design Evaluation of a Stainless Steel Cyclone on the Gasification Pilot Plant Facility at Morgantown, *The Properties and Performance of Materials in the Coal Gasification Environment,* American Society for Metals, 1981, p 709–734

31. J.S. Hansen, Relative Erosion Resistance of Several Materials, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, pp 148–162

32. J.S. Hansen, J.E. Kelly, and F.W. Wood, “Erosion Testing of Potential Valve Materials for Coal Gasification Systems,” Report of Investigations 8335, U.S. Department of the Interior, Bureau of Mines, 1979

33. A. Drotlew, P. Christodoulou, and V. Gutowski, Erosion of Ferritic Fe-Cr-C Cast Alloys at Elevated Temperatures, *Wear,* Vol 211, 1997, p 120–128

34. R. Hamzah, D.J. Stephensen, and J.E. Strutt, Erosion of Material Used in Petroleum Production, *Wear,* Vol 186–187, 1995, p 493–496

35. L. Nøkleberg and T. Søntvedt, Erosion in Choke Valves—Oil and Gas Industry Applications, *Wear,* Vol 186–187, 1995, p 401–412

36. K. Haugen, O. Kvernvold, A. Ronold, and R. Sandberg, Sand Erosion of Wear-Resistant Materials: Erosion in Choke Valves, *Wear,* Vol 186–187, 1995, p 179–188

37. R.J.K. Wood, and D.W. Wheeler, Design and Performance of a High Velocity Air-Sand Jet Impingement Erosion Facility, *Wear,* Vol 220, 1998, p 95–112

38. N. Gladys and V. Laura, Analysis of Rotor-Blade Failure due to High Temperature Corrosion/Erosion, *Surf. Coat. Technol.,* Vol 120–121, 1999, p 145–150

39. K.G. Budinski, Erosion of 316 Stainless Steel by Dicalcium Phosphate, *Wear,* Vol 186–187, 1995, p 145–149

40. E. Deaquino, D. Martínez, A. Perez, A. Velasco, M.A. Flores, and R. Viramontes, *Powder Technol.,* Vol 95, 1998, p 55–60

41. D.A. Brosnan, M.S. Crowley, and R.C. Johnson, CPI Drive Refractory Advances, *Chem. Eng.,* Vol 105, Oct 1998, p 110–120

42. G.P. Tilly, Sand Erosion of Metals and Plastics: A Brief Review, *Wear,* Vol 14, 1969, p 241–248

43. R.W. Bruce, Development of 1232 °C (2250 °F) Erosion and Impact Tests for Thermal Barrier Coatings, *Tribol. Trans.,* Vol 41, 1998, p 399–410

44. C.E. Smeltzer, M.E. Gulden, and W.A. Compton, Mechanism of Metal Removal by Impacting Dust Particles, *J. of Basic Eng. (Trans. ASME*), Vol 92, Sept 1970, p 639–654

45. J.G.A. Bitter, A Study of Erosion Phenomena, Parts I and II, *Wear,* Vol 6, 1963, p 5–21, 169–190.

46. A.K. Cousens and I.M. Hutchings, Influence of Erodent Shape on the Erosion of Mild Steel, *Proc. Sixth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1983, p 41-1 to 41-8

47. G. Sundararajan, A Comprehensive Model for the Solid Particle Erosion of Ductile Materials, *Wear of Materials 1991,* American Society of Mechanical Engineers, 1991, p 503–511

48. M. Roy, Y. Tirupatataiah, and G. Sundararajan, Effect of Particle Shape on the Erosion of Cu and Its Alloys, *Mater. Sci. Eng.,* Vol A165, 1993, p 51–63

49. P.V. Rao and D.H. Buckley, Time Effects of Erosion by Solid Particle Impingement on Ductile Materials, *Proc. Sixth Int. Conf. Erosion by Liquid and Solid Impact,* University of Cambridge, 1983, p 38-1 to 38-10

50. I. Finnie, The Mechanism of Erosion of Ductile Metals, *Proc. Third U. S. National Congress of Applied Mechanics,* American Society of Mechanical Engineers, 1958, p 527–532

51. I. Finnie, Erosion of Surfaces by Solid Particles, *Wear,* Vol 3, 1960, p 87–103

52. I. Finnie, Erosion by Solid Particles in a Fluid Stream, *Symposium on Erosion and Cavitation,* STP 307, ASTM, 1962, p 70–82

53. J.H. Neilson and A. Gilchrist, Erosion by a Stream of Solid Particles, *Wear,* Vol 11, 1968, p 111–122

54. R.E. Winter and I.M. Hutchings, Solid Particle Erosion Studies Using Single Angular Particles, *Wear,* Vol 29, 1974, p 181–194

55. G.P. Tilly, A Two Stage Mechanism of Ductile Erosion, *Wear,* Vol 23, 1973, p 87–96

56. S.M. Wiederhorn and B.J. Hockey, Effect of Material Parameters on the Erosion Resistance of Brittle Materials, *J. Mater. Sci.,* Vol 18, 1983, p 766–780

57. A.W. Ruff and S.M. Wiederhorn, Erosion by Solid Particle Impact, *Treatise on Material Science and Technology,* Vol 16, Academic Press, 1979, p 69–126

58. A.G. Evans, M.E. Gulden, and M. Rosenblatt, Impact Damage in Brittle Materials in the Elastic-Plastic Response Regime, *Proc. R. Soc. (London) A,* Vol 361, 1978, p 343–365

59. R. Bellman, Jr. and A. Levy, Erosion Mechanism in Ductile Metals, *Wear,* Vol 70, 1981, p 1–27

60. A.V. Levy, The Platelet Mechanism of Erosion of Ductile Metals, *Wear,* Vol 108, 1986, p 1–21

61. I.M. Hutchings, A Model for the Erosion of Metals by Spherical Particles at Normal Incidence, *Wear,* Vol 70, 1981, p 269–281

62. P.S. Follansbee, G.B. Sinclair, and J.C. Williams, Modelling of Low Velocity Particulate Erosion in Ductile Materials by Spherical Particles, *Wear,* Vol 74, 1982, p 107–122

63. A.V. Levy and S. Jahanmir, The Effects of the Microstructure of Ductile Alloys on Solid Particle Erosion Behavior, *Corrosion-Erosion-Behavior of Materials,* TMS-AIME, 1980, p 177–189

64. S. Jahanmir, The Mechanics of Subsurface Damage in Solid Particle Erosion, *Wear,* Vol 61, 1980, p 309–324

65. T. Christman and P.G. Shewmon, Adiabatic Shear Localization and Erosion of Strong Aluminum Alloys, *Wear,* Vol 54, 1979, p 145–155

66. G. Sundararajan and P.G. Shewmon, A New Model for the Erosion of Metals at Normal Incidence, *Wear,* Vol 84, 1983, p 237–258

67. I.M. Hutchings, Mechanical and Metallurgical Aspects of the Erosion of Metals, *Proc. Corrosion/Erosion of Coal Conversion Systems Materials Conf.,* National Association of Corrosion Engineers, 1979, p 393–428

68. P.H. Shipway and I.M. Hutchings, The Role of Particle Properties in the Erosion of Brittle Materials, *Wear,* Vol 193, 1996, p 105–113

69. G. Sundararajan and M. Roy, Solid Particle Erosion Behavior of Metallic Materials at Room and Elevated Temperatures, *Tribol. Int.,* Vol 30, 1997, p 339–359

70. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 7

71. I.M. Hutchings, Mechanisms of the Erosion of Metals by Solid Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 59–76

72. I. Finnie, A. Levy, and D.H. McFadden, Fundamental Mechanisms of the Erosive Wear of Ductile Metals by Solid Particles, *Erosion: Prevention and Useful Applications,* STP 664, ASTM, 1979, p 36–58

73. S. Bahadur and R. Badruddin, Erodent Particle Characterization and the Effect of Particle Size and Shape on Erosion, *Wear of Materials 1989,* Vol 1, American Society of Mechanical Engineers, 1989, p 143–153

74. A.V. Levy and P. Chik, The Effects of Erodent Composition and Shape on the Erosion of Steel, *Wear,* Vol 89, 1983, p 151–162

75. I. Kleis, Probleme der Bestimmung des Strahlverschleisses bei Metallen (The Problem of the Determination of the Wear of Metals by a Beam (of Abrasive Particles)), *Wear,* Vol 13, 1969, p 199–215

76. A. Magnée, Generalized Law of Erosion: Application to Various Alloys and Intermetallics, *Wear,* Vol 181–183, 1995, p 500–510

77. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 15

78. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 63

79. A. Mishra and I. Finnie, On the Size Effect in Abrasive and Erosive Wear, *Wear,* Vol 65, 1981, p 359–373

80. A. Levy, *Solid Particle Erosion and Erosion-Corrosion of Materials,* ASM International, 1995, p 62

81. J.R. Nicholls, Laboratory Studies of Erosion-Corrosion Processes under Oxidizing and Oxidizing/Sulphidising Conditions, *Mater. High Temp.,* Vol 14, 1997, p 289–306

82. I. Finnie, Some Observations on the Erosion of Ductile Metals, *Wear,* Vol 19, 1972, p 81–90

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

Introduction

CAVITATION EROSION is a type of wear in hydraulic turbines, on pump impellers, on ship propellers, valves, heat-exchanger tubes, and other hydraulic structures in contact with high-velocity liquids subjected to pressure changes. This type of wear also has been observed in mechanical devices such as plain bearings, seals, and orifices in which fluid goes through severe restrictions. Cavitation damage may occur when surfaces in contact with fluid are subjected to vibrations (e.g., water-cooled diesel-engine cylinder liner). Sometimes, cavitation can initiate other types of wear such as adhesion or abrasion because of wear particles produced at the friction interface, making the failure analysis more difficult.

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

Cavitation Mechanisms

Cavitation can occur if the pressure on a liquid is reduced sufficiently to cause formation of vapor-filled voids, or cavities (bubbles), even at room temperature. When the liquid that contains cavities is subsequently subjected to higher hydrostatic pressure, the bubbles can collapse suddenly. This collapse can cause surface damage through shock waves and microjets.

Shock waves. Surrounding liquid that fills the cavity may have very high speed. When the cavity is reduced to a smaller dimension, pressure in the cavity is so high and the liquid speed of the interface is so rapid that residual vapor in the cavity has no time to condense. This cavity is so reduced on forming a very high pressure bubble. When the pressure reaches the maximum level, a compression wave front forms at the interface. This front spreads in the liquid when the bubble becomes bigger again (bounce). If this collapse-bounce is repeated, solid surface in the liquid is subjected to a cyclic pressure, which can cause surface fatigue.

Microjets. When the cavity is near a solid surface, liquid coming from the solid surface side will be slower than that from inside the fluid. An initial spherical bubble will reduce asymmetrically as shown in Fig. 1, and a microjet of liquid will be projected toward the solid surface. The speed of the microjet depends on many factors, most importantly, the pressure. Some calculation results show that the speed is generally over 100 m/s (330 ft/s). This speed can produce a pressure of several hundred MPa, sufficient to damage most common materials (Ref 1).

Fig. 1 Numerical simulation of a bubble collapse near a solid surface showing evolution of the bubble during collapse

Characteristics of the Cavitation Process. The cavitation process has the following four characteristics:

- *High pressure:* The pressure peak can reach several hundred MPa or even climb to several GPa. This pressure is higher than the elastic limit of most engineering materials.
- *Small dimension:* Dimensions of the microjet are very small (from a few micrometers up to several hundred micrometers). Thus, each impact on the solid surface concerns only a very small area.
- *Short time:* The duration of the impact is about several microseconds.
- *High temperature:* Because of the localized dissipation of energy during collapse, local temperature can be very high (several thousand °C).

While the preceding information gives a general view of the cavitation process, many observed phenomena cannot be completely explained by shock wave or single-cavity collapse. When a large number of cavities are formed at the same time, it may produce a cluster of cavities. Interactions between bubble collapses in a cluster may be very complex. Energy generated by the collapse of outer cavities in the cluster may be transferred to the cluster center, leading to an increase of local hydrostatic pressure higher than that due to an individual collapse.

After the microjet impact, liquid flow on the solid surface could also have some effect on material removal. This is the reason why some surface-crack formation observed could not be attributed only to microjet impact on the solid surface (Ref 2, 3, 4, 5, 6, 7, 8).

References cited in this section

1. J.P. Franc, F. Avellan, B. Belahadji, D. Fréchou, D.H. Fruman, L. Briançon-Marjollet, A. Karimi, J.L. Kuency, and J.M. Michel, *La cavitation, mécanismes physiques et aspects industriels, (Cavitation, Physics Mechanism and Industry Aspects),* Presses Universitaires de Grenoble, 1995

2. G. Lesaffre and Y.M. Chen, "Etude sur les mécanismes de cavitation" ("Mechanism of Cavitation"), CETIM Report 1D1230, Centre Technique des Industries Mécaniques, 2001

3. C.M. Hansson and I.L.H. Hansson, Cavitation Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992

4. C.M. Preece, Cavitation Erosion, *Treatise on Materials Science and Technology,* Vol 16, Academic Press, 1979, p 249–305

5. F. Erdmann-Jesnitzer, Louis, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 171–196

6. B. Vyas and C.M. Preece, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 77–105

7. B. Vyas and I.L.H. Hansson, The Cavitation Erosion-Corrosion of Stainless Steel, *Corros. Sci.,* Vol 130 (No. 8-9), 1991, p 761–770

8. H. Grein, De la cavitation: une vue d'ensemble, (Cavitation: An Overview), *Rev. Tech. Sulzer,* NR Recherche, 1974, p 87–112

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

Cavitation Erosion Analysis

Failure Mechanisms. Cavitation erosion is material removal or plastic deformation of a solid surface in contact with a fluid subjected to cavity collapse. Shock wave and microjet result in a mechanical loading of the solid surface. In terms of contact mechanics, this type of loading could be considered as solid-solid contact with or without friction. If stresses in the material resulting from the hydrostatic pressure at the solid surface are over the material elastic limit, it will lead to plastic deformation at or near the surface. It is similar to a material removal mechanism by abrasion. If stresses are lower than the material elastic limit, cyclic loading by a large number of collapses also can damage the material by a surface fatigue mechanism. However, the surface fatigue mechanism does not take into account high-speed and small-volume effects. Due to high-speed deformation, material behavior, as in the case of solid-solid impact, is governed by strain-rate sensitivity. Due to the small volume, a scale effect can be considered, and the work-hardening effect cannot be interpreted as it can be in a classical abrasion mechanism.

In terms of energy dissipation, impact energy can be absorbed by elastic and plastic deformation or fracture. The capability of energy absorption by deformation without material removal is linked directly to cavitation resistance of materials.

The failure mode by cavitation is influenced markedly by strain-rate sensitivity of materials. In practice, two mechanisms of cavitation failure are considered: those for ductile materials and those for brittle materials.

Ductile failure mechanism is observed for most engineering metallic materials that are not very sensitive to strain rate (Fig. 2). Metals with a face-centered cubic (fcc) structure generally are not sensitive to strain rate. Their response to cavitation often is similar to their static mechanical behavior. Main damage is caused by plastic deformation or ductile rupture, which can be attributed to microjet impact close to the surface. For metals with a body-centered cubic (bcc) structure, deformation generally is strain-rate sensitive. Thus, their reaction to cavitation is always a competition between flow and fracture. When pure iron is subjected to cavitation, it exhibits both brittle and ductile failure mechanisms. For hexagonal close-packed (hcp) structure, response to an applied stress can be either strain sensitive or not. In case of multiphase alloys, the size and dispersion of the different phases also can have an influence on the cavitation mechanism.

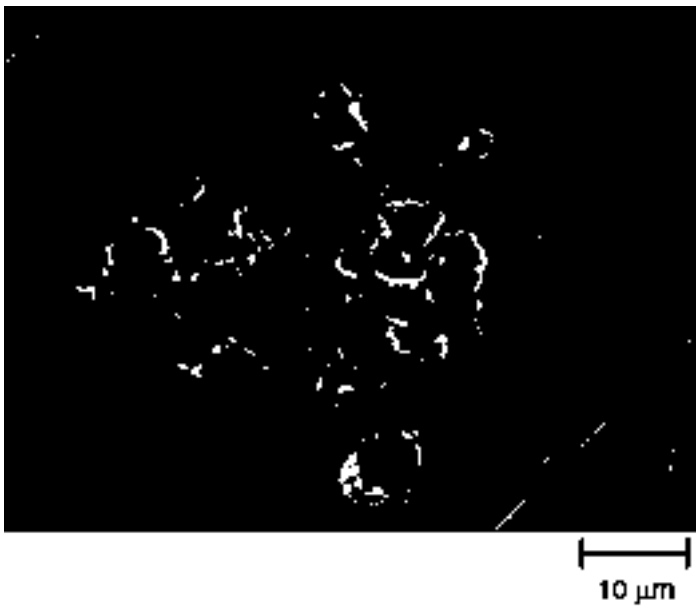


Fig. 2 Wear surface by cavitation of copper-base alloy in a lubricated gearbox. Courtesy of CETIM

Brittle failure mechanism is exhibited by materials such as ceramics and plexiglass that have a weak capability of energy absorption during deformation by impact (Fig. 3). Material removal is caused by propagation of cracks at the surface or at grain boundaries. In this case, most impact energy is dissipated by crack formation, which leads to more frequent material removal. The absorbed energy to remove a given volume of material by cracking is much less than that necessary to remove the same volume of material through plastic deformation. That is why brittle materials generally are less resistant to cavitation.

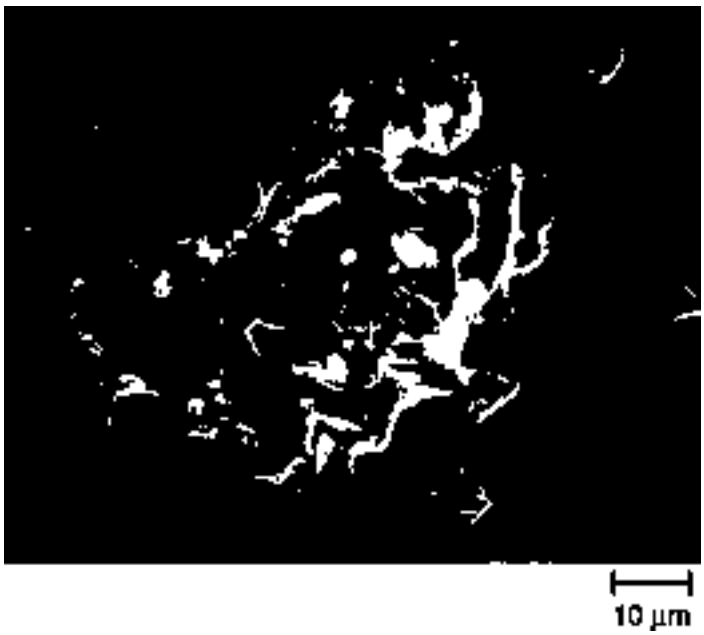


Fig. 3 Wear surface of Al_2O_3 after vibratory cavitation test. Courtesy of CETIM

Different Stages of Cavitation Erosion. Evolution of cavitation erosion depends on many parameters, such as material, surface shape, liquid, and cavitation conditions. There is no universal law for erosion rate (mass loss per unit time) evolution with period of exposure to cavitation. In most cases, however, a little mass loss is observed in the early stage of cavitation (incubation stage). This stage often is followed by a period of great increase of erosion rate (accumulation stage) or of a constant erosion rate (steady stage). After that, a decrease of erosion rate often is observed (attenuation stage)

Incubation Stage. This stage corresponds to an undetectable weight-loss period. Impact energy is dissipated by surface elastic or plastic deformation or by cracking for most metals with sometimes work-hardening effect at the surface. The surface exhibits some modification such as plastic flow, indentation traces, undulation, delineation of grain boundaries, coarse slip band, or cracking (Fig. 4). The determination of duration of this period depends on accuracy of weight measurements. Therefore, it is only a system response and not an intrinsic materials property. It is not reasonable to compare the durations of incubation between different types of tests or industry structures.

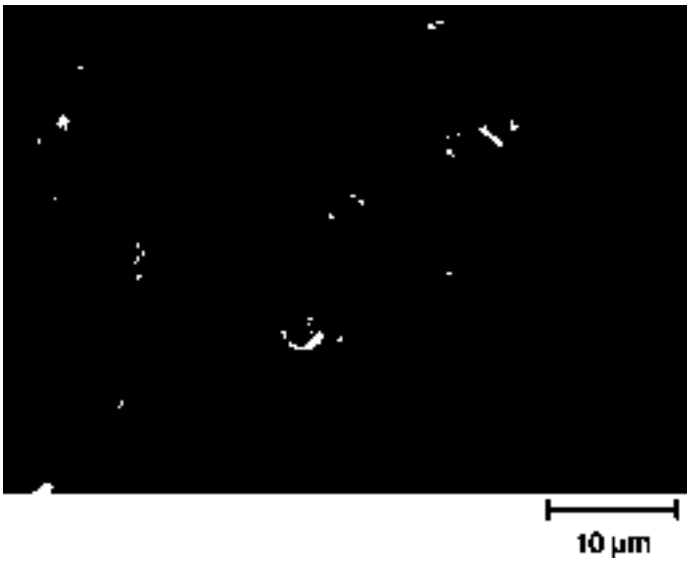


Fig. 4 Cavitation erosion: incubation stage of Ti-6Al-4V on vibratory cavitation test. Courtesy of CETIM

Accumulation or Steady Stage. When the work-hardening limit is reached, continuous plastic deformation leads to detachment of material and propagation of cracks near the surface, which results in an acceleration of material-removal rate. The worn surface becomes rougher with a large number of small pits and deep craters (Fig. 5). The erosion rate can increase or remain constant depending on material type and cavitation conditions. Neither the craters nor the pits are associated in any way with material nature, grain boundaries, slip lines, or any other structure feature. The duration of this stage can vary with cavitation resistance of material. There is no significant modification of surface morphology during this period. Mass loss can be very large for certain materials, such as aluminum.

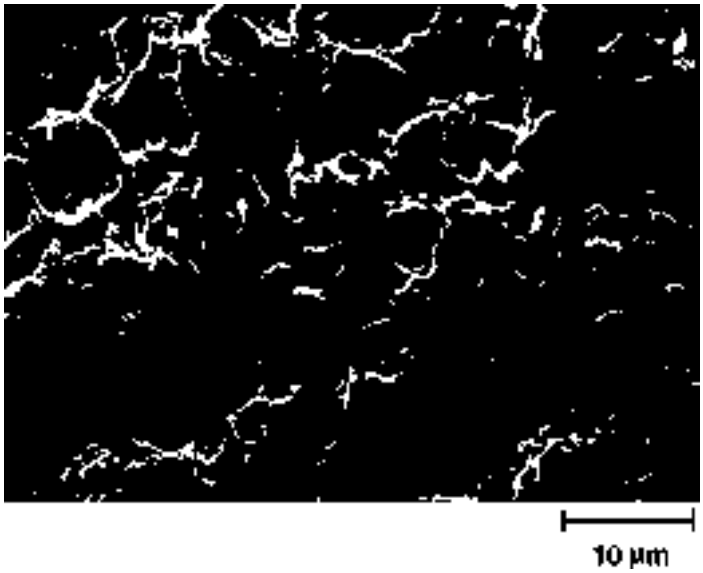


Fig. 5 Wear surface of 304 stainless steel after vibratory cavitation test. Courtesy of CETIM

Attenuation stage. The decrease of erosion rate during this period depends on many factors, such as materials properties, interactions between liquid flow and worn surface by an accommodation process. Residual air or gas bubbles in the deep craters also can act as a cushion and absorb a part of the impact energy. The worn surface corresponding to this period generally is rougher with craters more expanded. This stage of cavitation occurs only under certain conditions. During vibratory cavitation (ASTM G 32), no significant attenuation stage has been observed for aluminum alloys, copper alloys, carbon steel, stainless steel, and titanium alloys (Ref 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19).

References cited in this section

2. G. Lesaffre and Y.M. Chen, “Etude sur les mécanismes de cavitation” (“Mechanism of Cavitation”), CETIM Report 1D1230, Centre Technique des Industries Mécaniques, 2001

3. C.M. Hansson and I.L.H. Hansson, Cavitation Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992

4. C.M. Preece, Cavitation Erosion, *Treatise on Materials Science and Technology,* Vol 16, Academic Press, 1979, p 249–305

5. F. Erdmann-Jesnitzer, Louis, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 171–196

6. B. Vyas and C.M. Preece, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 77–105

7. B. Vyas and I.L.H. Hansson, The Cavitation Erosion-Corrosion of Stainless Steel, *Corros. Sci.,* Vol 130 (No. 8-9), 1991, p 761–770

8. H. Grein, De la cavitation: une vue d'ensemble, (Cavitation: An Overview), *Rev. Tech. Sulzer,* NR Recherche, 1974, p 87–112

9. K. Steller, T. Krzysztofowicz, and Z. Reymann, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 152–170

10. D. Berthe, Cavitation and Film Rupture in a Lubrificated Contact and Its Relation to Surface Degradations, *Cavitation and Related Phenomena in Lubrification,* congrès de Leeds, Sept 1974, p 185–188

11. D. Summers-Smith, R.W. Wilson, F.T. Barwell, P.J. Lidgitt, D. Berthe, J.M. Conway-Jones, and K.L. Johnson, Cavitation Damage, *Cavitation and Related Phenomena in Lubrification,* congrès de Leeds, Sept 1974, p 198–217

12. T. Akada, J. Iwamoto, and K. Sano, Fundamental Studies on Cavitation Erosion, *Bull. JSME,* Vol 20 (No. 147), 1977, p 1067–1075

13. R.K. Palhan, Effect of the Stress Level on Cavitation Erosion Rate, *Wear,* Vol 45 (No. 2), 1977, p 151–160

14. P.V. Rao and D. Buckley, Cavitation Erosion Size Scale Effects, *Wear,* Vol 96, 1984, p 239–253

15. A. Karimi, and F. Avellan, Comparison of Erosion Mechanisms in Different Types of Cavitation, *Wear,* Vol 113 (No. 3), 1986, p 305–322

16. A. Karimi and M. Maamouri, Microscopic Study of Cavitation Behaviour in Copper and Cu-5,7 wt% Al since Crystals, *Wear,* Vol 139 (No. 1), 1990, p 149–169

17. E.P. Rood, Review Mechanisms of Cavitation Inception, *J. Fluids Eng. (ASME),* Vol 113 (No. 2), 1991, p 163–165

18. M. Ahmed, Investigation of the Temperature Effects on Induced Impact Pressure and Cavitation Erosion, *Wear,* Vol 218, 1998, p 119–127

19. R.H. Richmann and W.P. MacNaughton, Correlation of Cavitation Behavior with Mechanical Properties of Metals, *Wear,* Vol 140 (No. 1), 1990, p 63–82

Yan-Ming Chen, CETIM Senlis, France

Industry Examples of Cavitation Failure

Cavitation in Bearings. Cavitation damage in oil-lubricated bearings currently is observed both in medium/slow speed diesel engines for marine or power station applications and in high-speed automotive engines. Because of fluctuation of radial force from the crankshaft, instability of lubricant flow, variation of oil pressure can be sufficient to produce bubble inception, collapse, and microjet formation process. In practice, however, the fluctuations of radial force vary with engine design and operation parameters such as load and speed. It is not always easy to determine the real cause of cavitation damage.

Four main cavitation mechanisms in plain bearings are now generally recognized; these are described in the following paragraphs. Wear particles produced by cavitation also can accelerate bearing damage.

Suction. Since most cases of suction erosion are found in the upper half of the main bearing, it is easy to imagine the damage mechanism with cyclic movement of crankshaft in radial direction. When the crankshaft moves from the upper to the lower half of the bearing, the sudden decrease of pressure due to firing load produces an aspiration effect in oil, so bubbles can be formed in this region where the pressure is lowest in the circumference. Collapse of the bubbles forms microjets, which are responsible for the surface damage observed (Fig. 6). High-speed diesel engines can have more risk of suction because of high radial moving speed of the shaft.

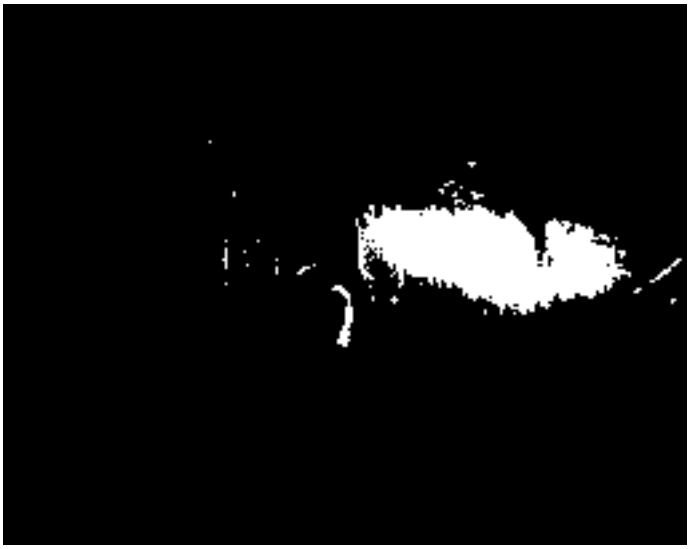


Fig. 6 Cavitation erosion of main bearing of diesel engine. Courtesy of CETIM

Discharge. Oil in the clearance escapes at high speed under radial action of the shaft due to firing load. This type of cavitation takes place in the lower half of the bearing. High-speed oil jet, leading to a drop of pressure, can produce cavities, and their collapse can cause damage on the adjacent surface.

Flow Instability. Variations in bearing geometry (oil grooves, edges of oil hole) often are responsible for flow instability. When oil flow goes from a grooved region to an ungrooved section, an area of low-pressure oil is created at the end of drilling. Vapor bubbles can be formed in this region and their collapse can cause cavitation damage. This phenomenon often is observed in medium/low speed engines because an ungrooved bearing surface is needed to increase carrying capacity (Fig. 7).

Fig. 7 Wear traces on a hydrodynamic bearing surface. Courtesy of CETIM

High-Pressure Impact Mechanism. When oil flow through a shaft drilling into a bearing is interrupted by the edge of the oil outlet as the drilling rotates or by inertia forces, high pressure can be produced. This variation of pressure can generate cavitation damage (Ref 20, 21, 22, 23).

Wear particles produced by cavitation can accelerate bearing damage. If wear particles from cavitation erosion are not removed systematically from the contact surface, these particles can cause other types of wear, such as solid-particle erosion or third-body abrasion (Fig. 8).

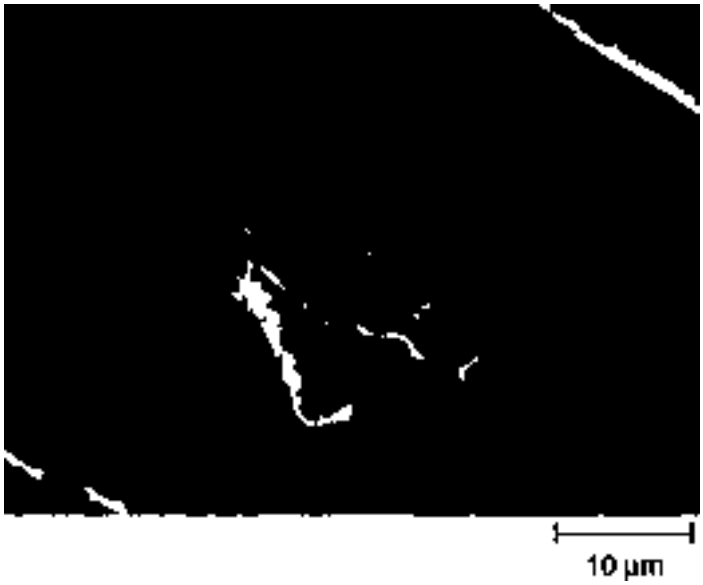


Fig. 8 Wear traces on a plain bearing surface initiated by cavitation. Courtesy of CETIM

Cavitation erosion in centrifugal pumps has been studied intensely by means of both analytical calculation and experimental tests. But this problem still has not been resolved entirely because cavitation erosion in this type of pump is a very complex physical process that depends on many parameters. Generation and collapse of bubbles are influenced by pump geometry, liquid characteristics, materials properties, and hydrodynamic and thermodynamic parameter variations (Ref 24, 25, 26, 27, 28).

Most cavitation erosion damage takes place on the suction surface of the impeller because of suction instabilities. With decreasing flow rate, the fluid approaches the impeller blades with larger and larger angles of incidence. This leads to great variations of pressure and velocity field inside the impeller. Vapor bubbles can be generated, and collapses occur near the suction surface (Fig. 9). Suction instabilities can be related to fluid characteristic changes and pipe and impeller geometries.

Fig. 9 Wear on suction surface of centrifugal pump impeller by cavitation and solid particle erosion. Courtesy of CETIM

Cavitation erosion also can be found in the pressure side of impeller blades. (Fig. 10). In this area, erosion damage can be caused by excessive high-flow rates. Interaction of suction recirculation and incoming flow form vortices that result in cavitation vapor.

Fig. 10 Wear on pressure surface of centrifugal pump impeller by cavitation and solid particle erosion. Courtesy of CETIM

Cavitation in gearboxes. Pitting on the contact surface of gear teeth generally is attributed to contact fatigue, a pure mechanical phenomena. In certain cases, however, cavitation also can cause pitting damage at the tooth surface. According to high-speed photography (16,000 to 30,000 frames per second) observation, bubbles and collapses can take place at the line of contact, as founded by theoretical analysis. This conclusion was confirmed by experimental study on a vibrating plate test rig. Gear vibration is the main cause of this type of cavitation damage (Ref 29). Similar phenomena were observed on tooth surfaces in a gearbox subjected to vibrations (Fig. 11) and on the side surface of a gear pump (Fig. 12).

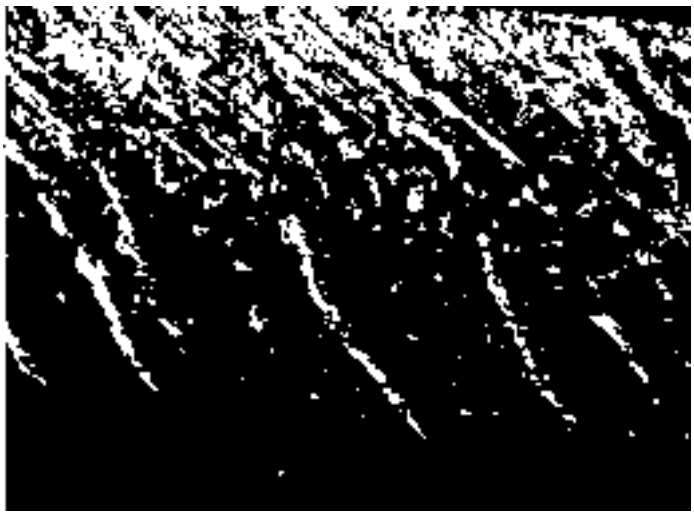


Fig. 11 Cavitation erosion on tooth surfaces subject to vibrations. Courtesy of CETIM

Fig. 12 Cavitation erosion on side surface of a gear pump subjected to vibrations. Courtesy of CETIM

References cited in this section

20. R.W. Wilson, Cavitation Damage in Plain Bearings, *Cavitation and Related Phenomena in Lubrification,* congrès de Leeds, Sept 1974, p 177–183

21. R.D. James, Erosion Damage in Engine Bearings, *Tribol. Int.,* Vol 8 (No. 4), 1976, p 161–170

22. D.R. Garner, R.D. James, and J.F. Warriner, Cavitation Erosion Damage in Engine Bearings: Theory and Practice, *J. Eng. Power (Trans. ASME),* Vol 102 (No. 4), 1980, p 847–857

23. G.N. Blount and R.D. James, Cavitation Erosion Damage in Diesel Engine Main Bearings, *Tribol. Reciprocating Engines,* congrès de Leeds, 1982, p 297–303

24. J.F. Gülich and A. Rösch, L'érosion de cavitation dans les pompes centrifuges (Cavitation Erosion in Centrifugal Pumps), *Rev. Tech. Sulzer,* Vol 70 (No. 1), 1988, p 28–33

25. J.F. Gülich, Diagnosis of Cavitation in Centrifugal Pumps, *World Pumps,* Vol 308, 1992, p 15–20

26. R. Canavelis and P. Grison, Aspects industriels de la cavitation dans les pompes (Industrial Aspects of Cavitation in Pumps), *Rev. Fr. Méc.,* Vol 4, 1986, p 171–181

27. D.P. Sloteman, Avoiding Cavitation in the Suction Stage of High-Energy Pumps, *World Pumps,* Vol 348, 1995, p 10–48

28. W.K. Chan, Correlation between Cavitation Type and Cavitation Erosion in Centrifugal Pumps, *Int. J. Heat Fluid Flow,* Vol 11 (No. 3), 1990, p 269–271

29. J.B. Hunt, A.J. Ryde-Weller, and F.A.H. Asmead, Cavitation between Meshing Gear Teeth, *Wear,* Vol 71 (No. 1), 1981, p 65–78

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

Cavitation Resistance of Materials

Cavitation resistance of materials (Ref 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53) is a key factor in the prevention of cavitation failures. In order to reduce cavitation erosion wear, one of the most intuitive reactions of mechanical engineers is to look for better materials for the same design. Materials scientists endeavor to create new materials, especially surface coatings and treatments, to answer these needs.

However, cavitation resistance is not an intrinsic property of material, but a system response. In addition to the nature of the material (composition, structure, heat treatment, geometry, surface roughness, residual stress, etc.), cavitation resistance depends largely on liquid property, flow speed, vibration characteristics, temperature, hydrostatic pressure, and so on. For certain materials, cavitation resistance is related to hardness, but for most metallic materials, it is related more closely to fatigue strength of materials (Ref 19). It is very difficult, therefore, to establish a universal rule for materials selection to minimize cavitation erosion wear.

In practical applications, laboratory tests often are used to evaluate cavitation resistance of a group of candidate materials. If the test conditions are sufficiently close to those of real applications, the test data can be very useful for the first choice of materials. Some recent laboratory data also can give new ideas for materials selection.

Laboratory tests used to evaluate erosion and cavitation damage in metals include:

- High-velocity flow tests, including venturi tubes, rotating disks, and ducts containing specimens in throat sections
- High-frequency vibratory tests using either magnetostriction devices or piezoelectric devices
- Impinging jet tests using either stationary or rotating specimens exposed to high-speed jet or droplet impact

These tests generally are designed to provide high-erosion intensities on small specimens in relatively short times. These methods may not closely simulate service conditions, but they are useful for ranking candidate materials. The test methods are described subsequently.

Metals and Alloys. Various metals and alloys are used widely in mechanical engineering, but their degrees of cavitation erosion resistance are quite different. Generally, hard metals have a good resistance, although the hardness is not always a good indicator of cavitation resistance, especially for alloys.

Several high-purity metals and some commercially available metals and alloys were tested by means of vibratory apparatus (Fig. 13). The specimen is opposite the vibration horn with a distance of 0.9 mm (0.04 in.) Tungsten, molybdenum, cobalt, and their alloys are much more resistant to cavitation erosion than magnesium, aluminum, zinc, copper, nickel, and iron and their alloys.

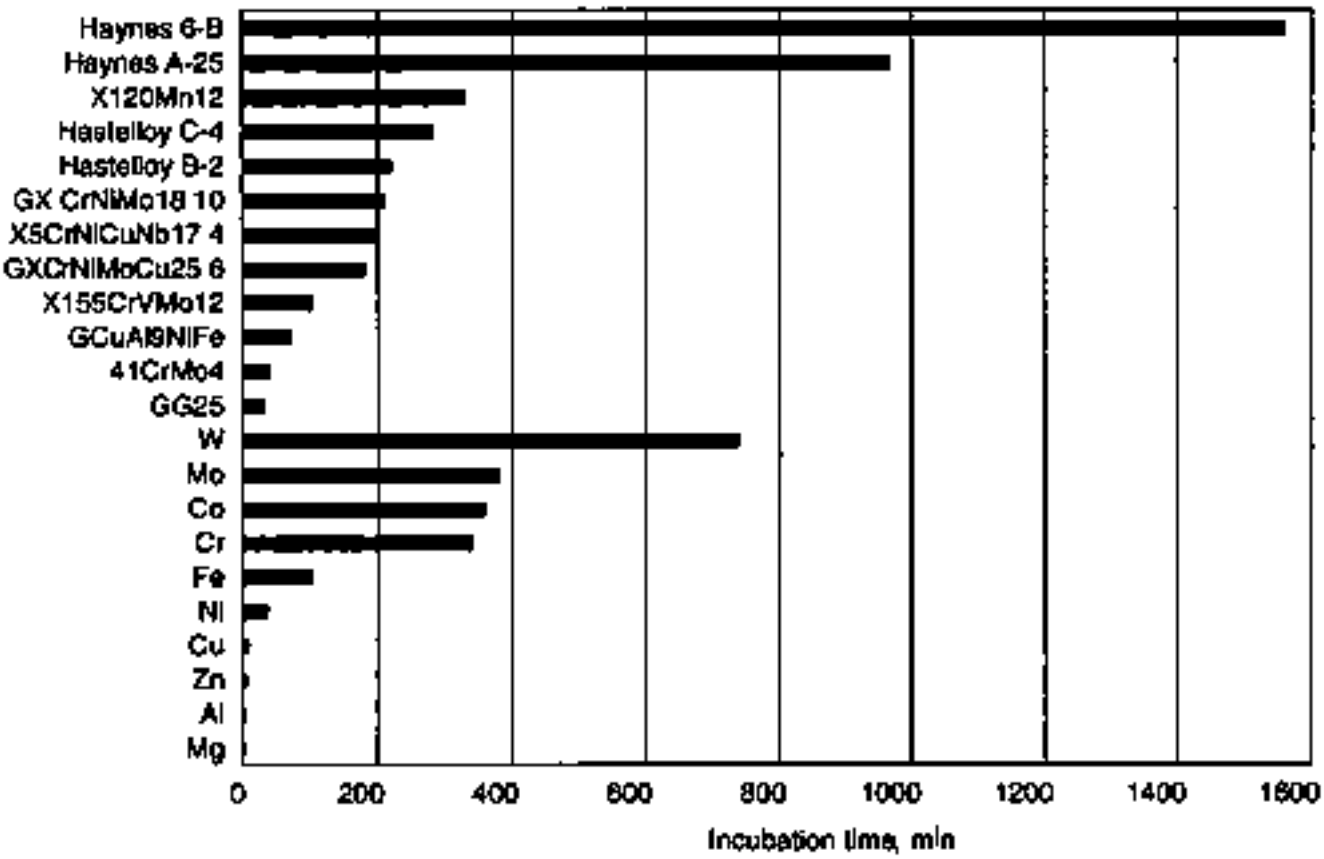


Fig. 13 Incubation time of different metals and alloys (frequency = 21.1 kHz; distance between specimen and vibration horn = 0.9 mm; vibration amplitude = 35 µm; temperature = 20 °C; liquid: water). Source: Ref 30

Similar ranking results are obtained with the specimen directly mounted in the vibration horn (Fig. 14), but cavitation wear is evaluated by erosion rate of steady stage (weight loss per unit time, mg/h).

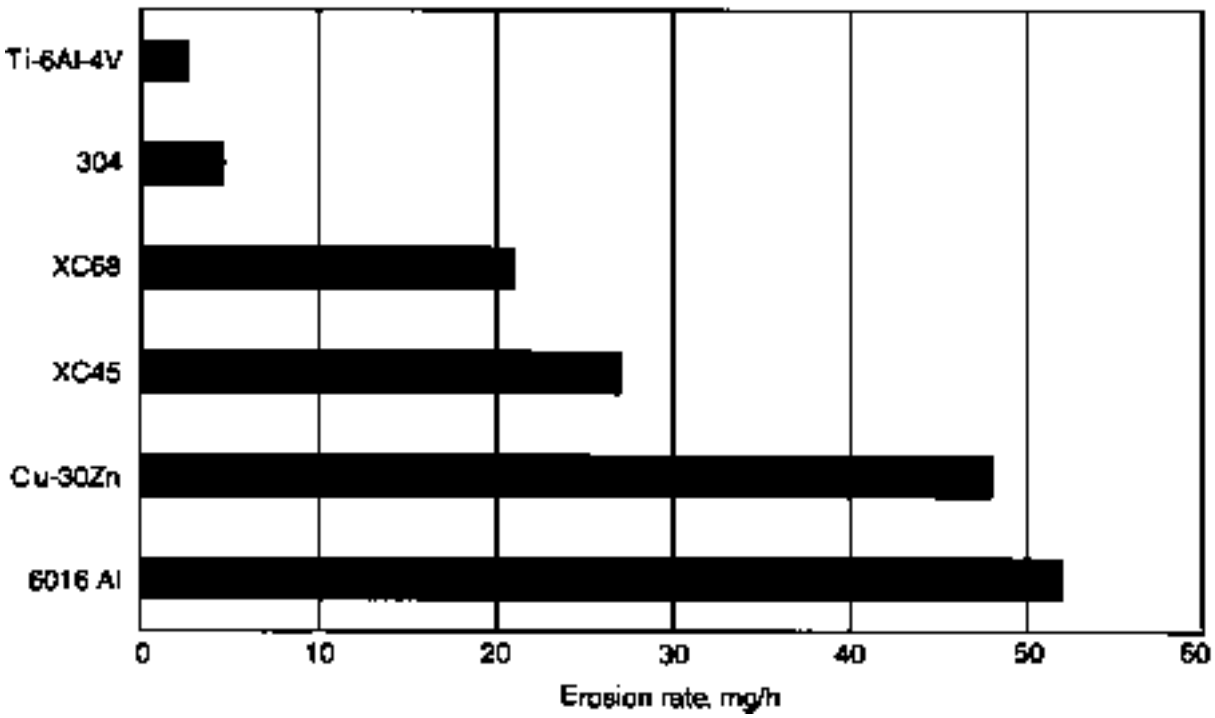


Fig. 14 Erosion rate of different metals and alloys (frequency = 20 kHz; specimen mounted in vibration horn; vibration amplitude = 50 µm; temperature = 20 °C; liquid: distilled water). Source: Ref 2

A large range of materials were tested by means of a vortex test apparatus. Good cavitation erosion resistance is noted with Stellite, polyamide 12, and Ti-6A1-4V (Fig. 15).

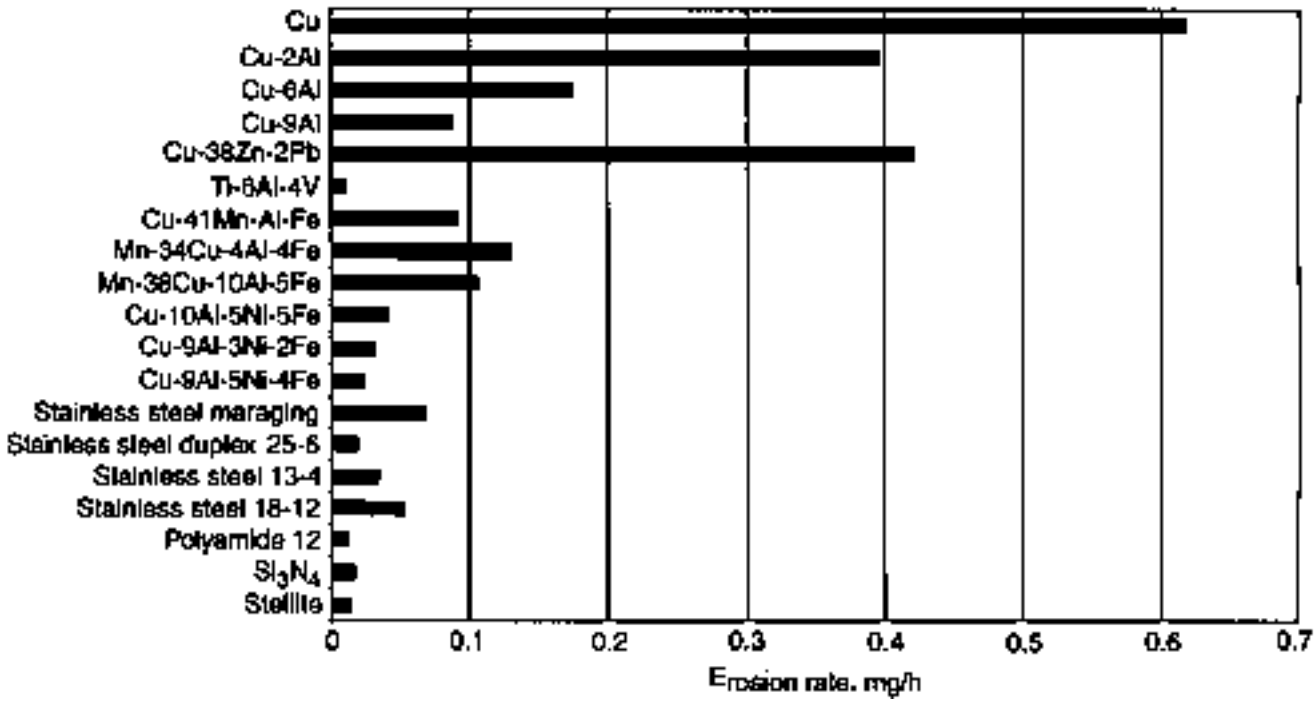


Fig. 15 Erosion rate of different materials in a vortex erosion apparatus (input pressure = 7.7 bars; output pressure = 1 bar; flow: 3.3 mm^3/h; input speed = 18 m/s; rotation speed: 400 turns/min; liquid: distilled water; temperature = 18). Source: Ref 1

Aluminum generally has a weak cavitation erosion resistance compared with steels. This resistance can be considerably improved, however, by aging and alloying (Fig. 16).

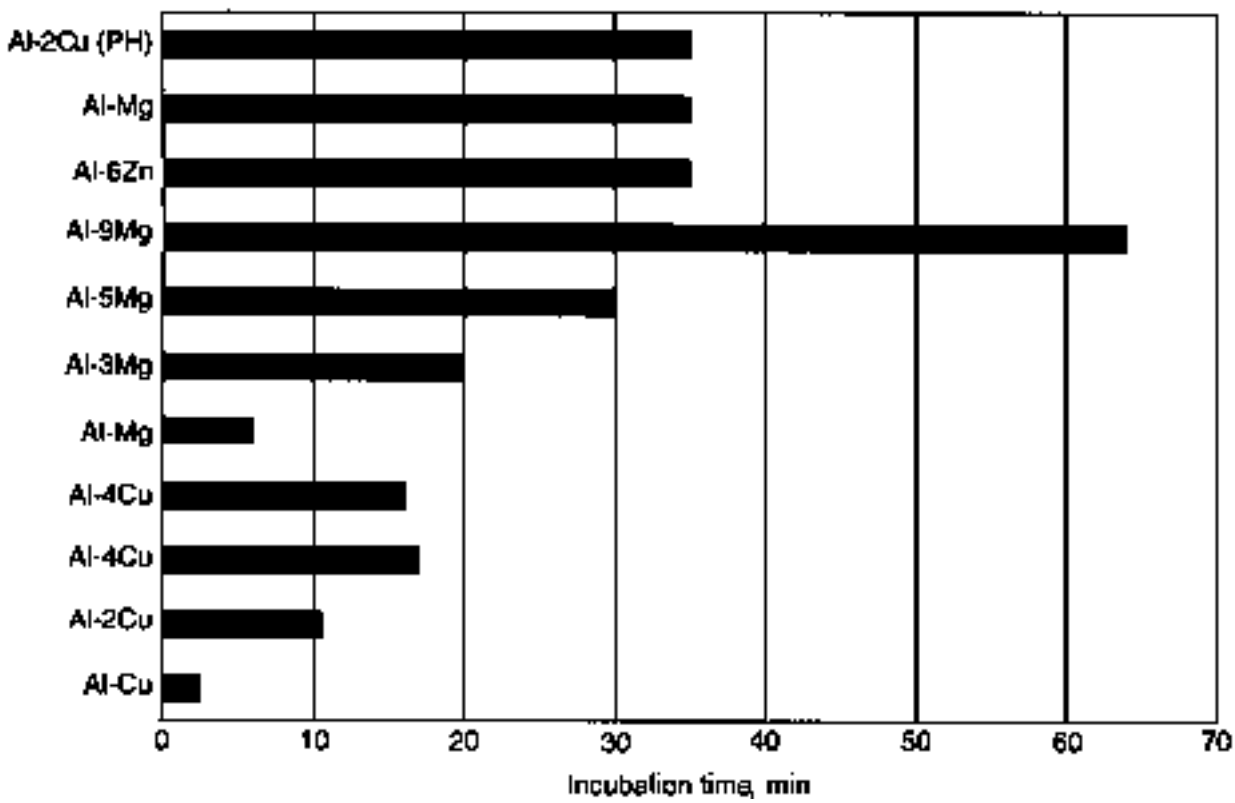


Fig. 16 Erosion rate of aluminum treated in different conditions (AQ: as-quenched; OA: over-aged; PH: peak-hardened; frequency = 20 kHz; distance between specimen and vibration horn = 1.25 mm; vibration amplitude = 40 µm; temperature = 15 °C; liquid: distilled water). Source: Ref 31

Recent results show that TiNi has a very good cavitation erosion resistance. In 3% NaCl aqueous solution, the erosion rate of TiNi can be 38 times higher than that of 304 stainless steel (Fig. 17).

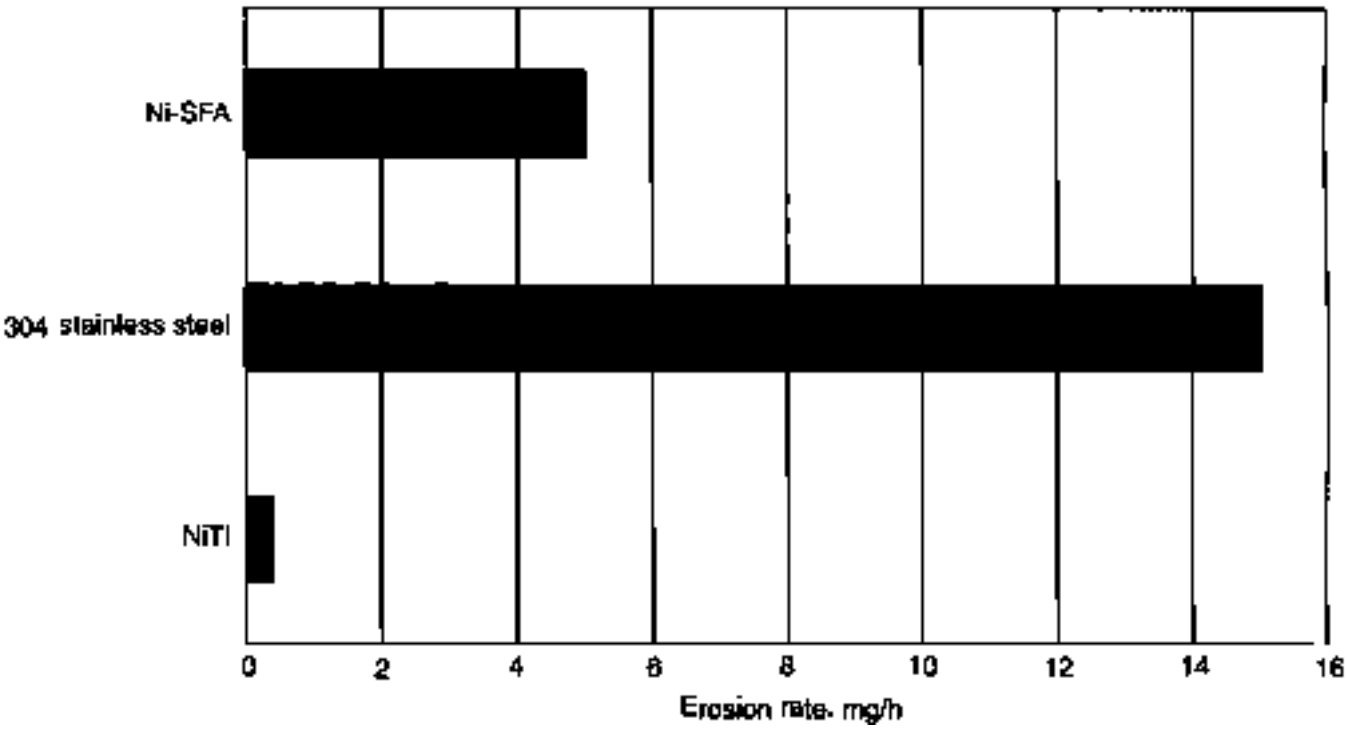


Fig. 17 Erosion rate of TiNi, 304 stainless steel, and Ni-base alloys (frequency = 20 kHz; specimen mounted in vibration horn; vibration amplitude = 50 µm; temperature = 20 °C; liquid: 3.5% NaCl aqueous solution). Source: Ref 34

Hydraulic components working in aqueous conditions suffer from cavitation. However, more and more cavitation phenomena are observed in sealing and lubrication systems working in oil. Experimental results show that brittle materials generally have higher erosion rates than ductile materials have. Long incubation periods and important attenuation stages have been observed in titanium alloys. (Fig. 18).

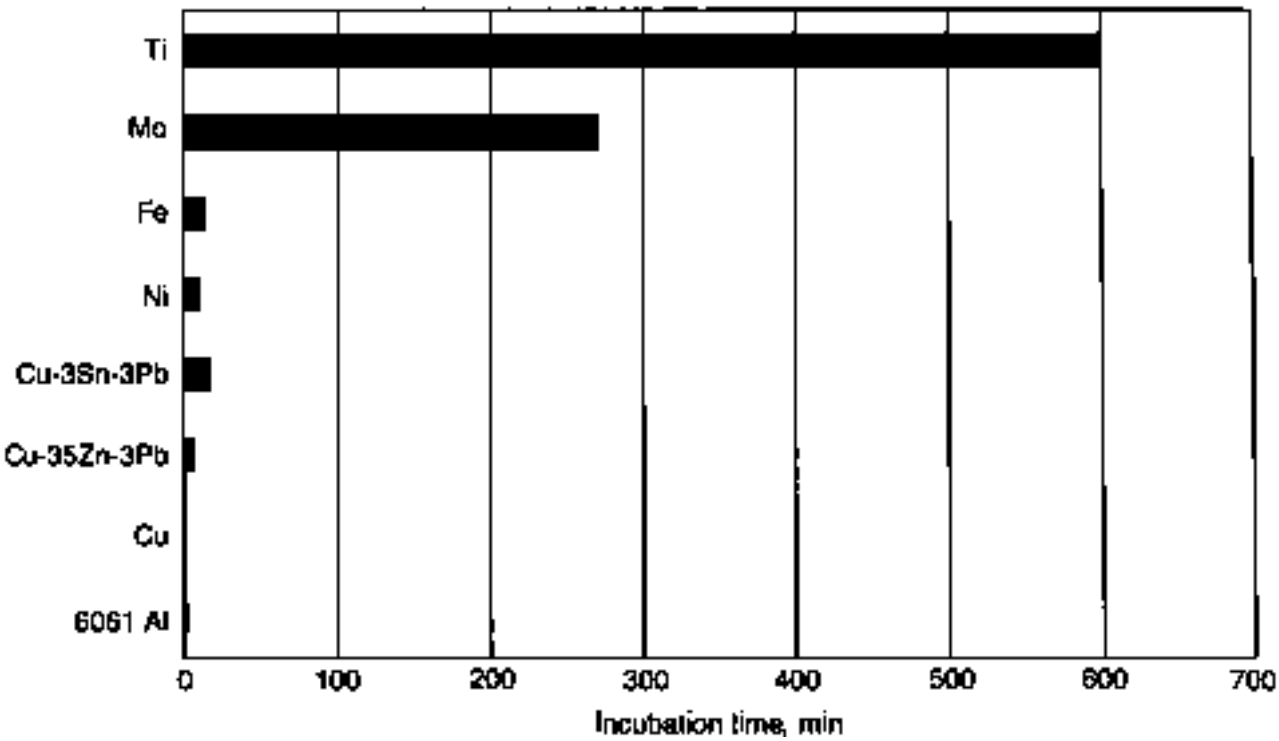


Fig. 18 Erosion rate of metals in mineral oil (frequency = 20 kHz; specimen mounted in vibration horn; vibration amplitude = 50 µm; liquid: mineral oil, viscosity at 20 °C = 110 cS). Source: Ref 35

Ceramics are used as components of hydraulic machinery and valves because of their superior corrosion, abrasion, and solid-particle erosion resistance. However, damage can occur by cavitation erosion. Some ceramics are more sensitive to cavitation than others. Experimental results show that bending resistance has no significant influence on cavitation. Cavitation erosion resistance depends largely on the microstructure of ceramics (e.g., grain size, open pores, and holes, Fig. 19). Microjet impact leads to intergranular fracture around defects of alumina surface generating granular particle removal. Better cavitation erosion resistance can be obtained for alumina with smaller grain size.

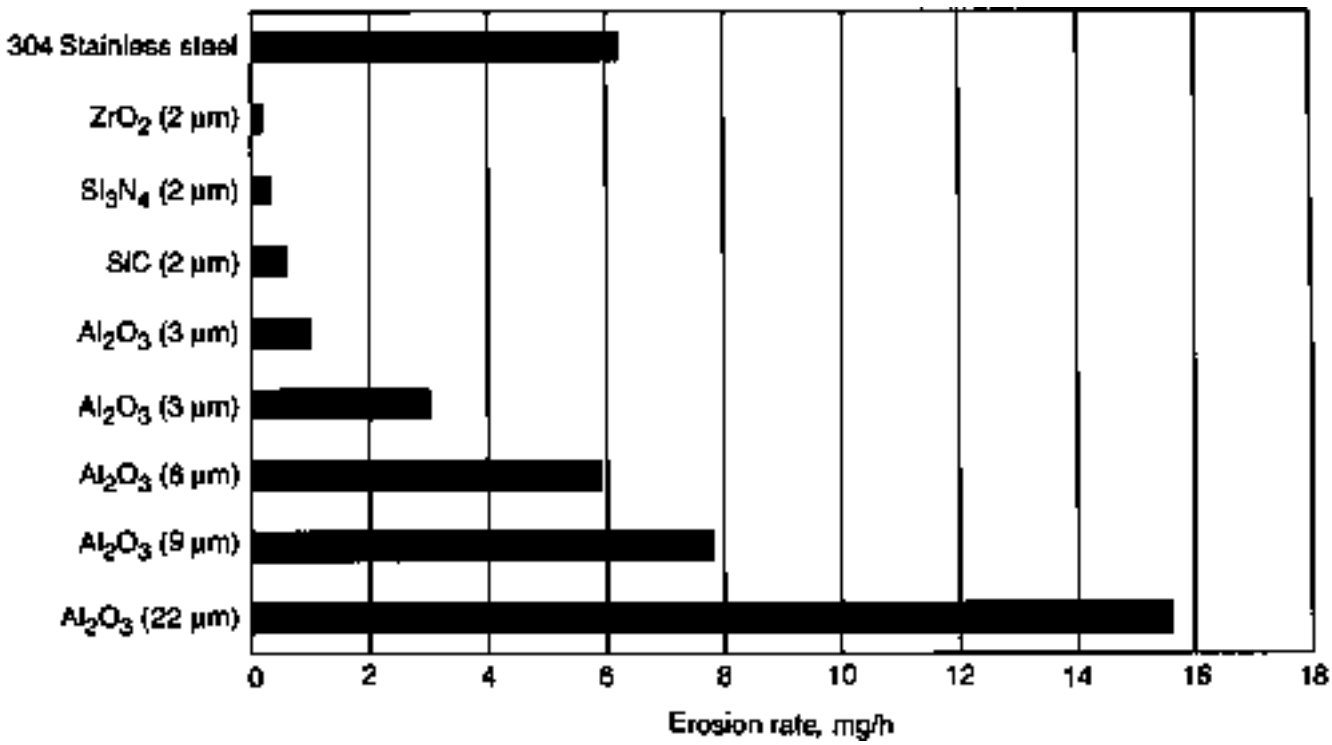


Fig. 19 Erosion rate of ceramics with different grain size (frequency = 20 kHz; distance between specimen and vibration horn = 1 mm; vibration amplitude = 50 µm; temperature = 25 °C; liquid: ion-exchanged water). Source: Ref 37

Surface coatings and treatments include metallic coatings and laser surface modification.
Metallic Coatings. Interactions of cavitation and corrosion lead to much more material removal for the components working in corrosive liquid. When choosing corrosion-resistant plating, the cavitation resistance of these parts must be taken into account. Thickness of coatings is often one of the most important parameters for hard coatings such as chromium and nickel. When the coating is removed, electrochemical reaction becomes more important. For soft, less-noble coatings such as zinc, erosion can be reduced by cathodic protection of the coating remaining outside (Fig. 20).

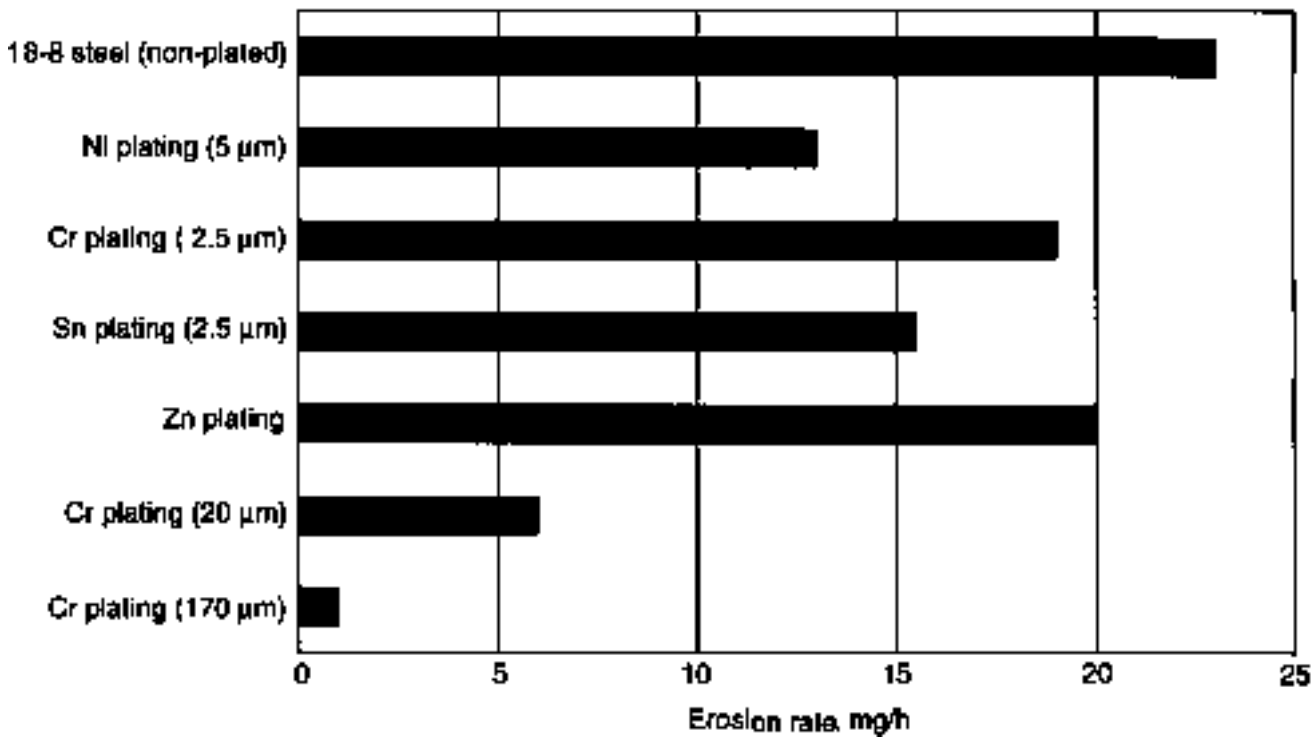


Fig. 20 Erosion rate of metallic coatings (frequency = 20 kHz; distance between specimen and vibration horn = 1 mm; vibration amplitude = 50 µm; temperature = 25 °C; liquid: 3% NaCl aqueous solution). Source: Ref 38

Laser Surface Modification. To improve wear resistance of components working under corrosion and cavitation conditions, laser surface modification is desirable in engineering applications because they consume only a small amount of precious materials at the surface. Modified surfaces can provide specific properties conducive to corrosion and cavitation resistance. Laser alloyed surface with carbon, and Al + Si can improve cavitation erosion resistance of UNS 31603 stainless steel 10 times. (Fig. 21). Significant improvement in corrosion and cavitation resistance, however, cannot be achieved simultaneously with current solutions.

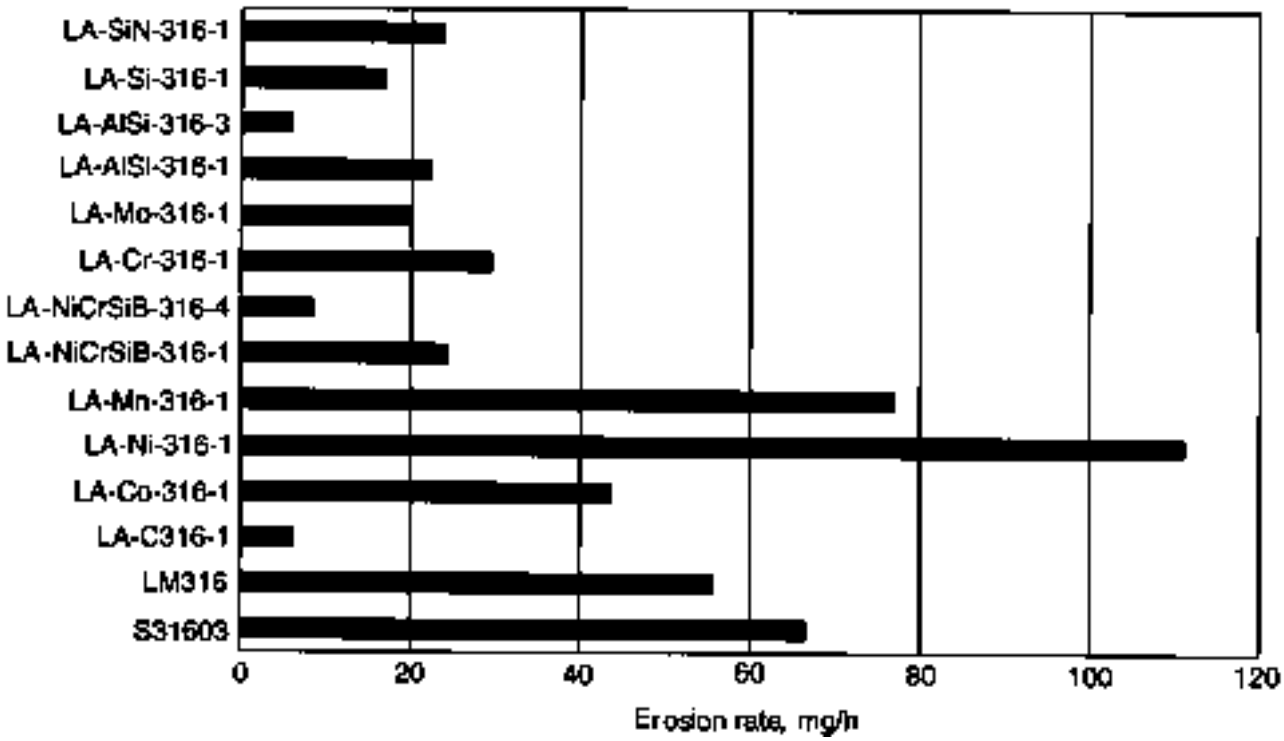


Fig. 21 Erosion rate of laser-modified 316 stainless steel (UNS 31603) 31603 stainless steel (frequency = 20 kHz; specimen mounted in vibration horn; vibration amplitude = 30 µm; temperature =23 °C; liquid: 3.5% NaCl aqueous solution). LA, laser alloyed; LM, laser modified. Source: Ref 42

References cited in this section

1. J.P. Franc, F. Avellan, B. Belahadji, D. Fréchou, D.H. Fruman, L. Briançon-Marjollet, A. Karimi, J.L. Kuency, and J.M. Michel, *La cavitation, mécanismes physiques et aspects industriels, (Cavitation, Physics Mechanism and Industry Aspects)*, Presses Universitaires de Grenoble, 1995

2. G. Lesaffre and Y.M. Chen, "Etude sur les mécanismes de cavitation" ("Mechanism of Cavitation"), CETIM Report 1D1230, Centre Technique des Industries Mécaniques, 2001

19. R.H. Richmann and W.P. MacNaughton, Correlation of Cavitation Behavior with Mechanical Properties of Metals, *Wear*, Vol 140 (No. 1), 1990, p 63–82

30. H.G. Feller and A.J. Kharrazi, Cavitation Erosion of Metal and Alloys, *Wear*, Vol 93, 1984, p 249–260

31. S. Vaidya and C.M. Preece, Cavitation Erosion of Age-Hardenable Aluminium Alloys, *Metall. Trans. A*, Vol 9, 1978, p 299–307

32. W.T. Fu, Y.B. Yang, T.F. Jing, Y.Z. Zhang, and M. Yao, The Resistance to Cavitation Erosion of CrMnN Stainless Steels, *J. Mater. Eng. Perform.*, Vol 7 (No. 6), 1998, p 801–804

33. Y. Iwai, T. Okada, and F.G. Hammitt, Effect of Temperature on the Cavitation Erosion of Cast Iron, *Wear*, Vol 96, 1983, p 149–253

34. S.K. Wu, H.C. Lin, and C.H. Yeh, A Comparison of the Cavitation Erosion Resistance of TiNi Alloys, SUS304 Stainless Steel and Ni-Based Self-Fluxing Alloy, *Wear*, Vol 244, 2000, p 85

35. B.C.S. Rao and D.H. Buckley, Characterisation of Erosion of Metallic Materials under Cavitation Attack in a Mineral Oil, *ASLE Trans.*, Vol 29 (No. 4), 1985, p 473–480

36. T. Okada and S. Hattori, Cavitation-Erosion of Aluminium Oxide and Silicon Carbide, *JSME Int. J.*, Vol 33 (No. 3), 1990, p 382–387

37. T. Okada, Cavitation Erosion of Ceramics, *Jpn. J. Tribol.*, Vol 41 (No. 8), 1996, p 803–814

38. T. Okada, Y. Iwai, and K. Awazu, Effects of Plating on Cavitation Erosion, *Wear*, Vol 124, 1988, p 21–31

39. W.J. Tomlinson, R.T. Moule, J.H.P.C. Megaw, and A.S. Brandsden, Cavitation Wear of Untreated and Laser-Processed Hardfaced Coatings, *Wear*, Vol 117 (No. 1), 1987, p 103–107

40. W.J. Tomlinson, J.H.P.C. Megaw, A.S. Brandsden, and M. Girardi, The Effect of Laser Melting on the Cavitation Wear of Grey Cast Iron in Distilled and 3% Salt Water, *Wear*, Vol 116 (No. 2), 1987, p 249–260

41. Y. Iwai, T. Okada, T. Fujieda, and K. Awazu, Effects of Hard Chromium Plating on Cavitation Erosion, *Wear*, Vol 128, 1988, p 189–200

42. C.T. Kwok, F.T. Cheng, and H.C. Man, Laser Surface Modifications of UNS S31603 Stainless Steel, Part II: Cavitation Erosion Characteristics, *Mater. Sci. Eng.*, Vol A290, 2000, p 74–88

43. C.T. Kwok, F.T. Cheng, and H.C. Man, Cavitation Erosion and Pitting Corrosion of Laser Surface Melted Stainless Steels, *Surf. Coat. Technol.*, Vol 99 (No. 3), 1998, p 295–304

44. S.M. Ahmed, K. Hokkirigawa, and P. Oba, Fatigue Failure of SUS 304 Caused by Vibratory Cavitation Erosion, *Wear*, Vol 177, 1994, p 129–137

45. J.L. Heuze, F. Gaillard, and B. Dubois, Etude de l'endommagement par érosion de cavitation de matériaux métalliques à structures CFC (Study on Cavitation Erosion Damage of Metallic Materials with CFC), *Mém. Sci. Rev. Métall.*, Vol 85 (No. 10), 1988, p 581–594

46. C. Morton, Combating Cavitation Erosion by Elastomeric-Type Coating, *Corros. Prev. Control*, Vol 36 (No. 2), 1989, p 31–33

47. A. Karimi, Cavitation Erosion of a Duplex Stainless Steel, *Mater. Sci. Eng.*, Vol 86 (No. 1/2), 1987, p 191–203

48. T. Okada, K. Awazu, and H. Kawasaki, Influence of Cavitation Erosion on Corrosion Fatigue and the Effect of Surface Coatings on Resistance, *Wear*, Vol 63 (No. 1), 1980, p 51–70

49. C.J. Heathcock, A. Ball, D. Yamey, and B.E. Protheroe, Cavitation Erosion of Cobalt Based Stellite Alloys, Cemented Carbides and Surface Treated Low Alloy Steels, *Wear of Metal*, congrès de San Francisco, 1981, p 597–606

50. C.M. Preece and C.W. Drapper, The Effect of Laser Quenching the Surfaces of Steels on Their Cavitation Erosion Resistance, *Wear*, Vol 67 (No. 3) 1981, p 321–328

51. G.S. Mikaelyan and S.Y. Nozdrev, The Resistance to Breakdown by Erosion of Certain Compositions and Materials Used for Surfacing, *Automat. Weld.*, Vol 27 (No. 12), 1974, p 31–33

52. M.G. Fontana, *Corrosion Engineering*, 3rd ed., McGraw-Hill, 1986

53. J.Z. Lichtman, D.H. Kallas, and A. Rufolo, Evaluating Erosion (Cavitation) Damage, *Handbook of Corrosion Testing and Evaluation*, W.H. Ailor, Ed., John Wiley & Sons, 1971, p 453–472

Yan-Ming Chen, CETIM Senlis, France

Other Prevention Parameters

Design Optimization. Optimized geometry can reduce variation of pressure, which is responsible for cavitation. To reduce cavitation in a diesel engine main bearing, for example, optimized groove ending geometry can reduce cavitation erosion effectively (Ref 22). A similar principle of modification can be applied in a centrifugal pump impeller.

Reducing Clearance. Reduction of diametrical clearance limits the journal velocity of the shaft in the bearing and the flow of oil. Cavitation erosion resulting from flow erosion, depending on flow velocity, thus can be minimized. The use of reduced diametrical clearance also is effective in suppressing the crankshaft, depending on engine noise. It suggests an eventual correlation between noise and cavitation erosion. In practice, clearance reduction differs from one engine to another and depends fundamentally on the limits of size and alignment to which components can be produced and assembled (Ref 21).

Reducing Vibration. Vibration is responsible for significant erosion damage in sealing and hydraulic circuit systems. The pressure variation due to the vibrations can result in bubble forming and collapse near the surface. The vibrations can have many sources: firing of engine, mechanical vibrations of other surrounding components, friction, and fluctuation of liquid pressure, and so on. Suppression of vibrations must start with a careful analysis of vibration sources. Sometimes, a very simple method such as mass change or isolating layers is sufficient to eliminate the vibration.

References cited in this section

21. R.D. James, Erosion Damage in Engine Bearings, *Tribol. Int.,* Vol 8 (No. 4), 1976, p 161–170

22. D.R. Garner, R.D. James, and J.F. Warriner, Cavitation Erosion Damage in Engine Bearings: Theory and Practice, *J. Eng. Power (Trans. ASME),* Vol 102 (No. 4), 1980, p 847–857

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

Cavitation Tests

There are two American Society for Testing and Materials (ASTM) standards for the evaluation of erosion and cavitation: G 32 (Ref 54), which is a vibratory test method using either a magnetostrictive or piezoelectric device that vibrates at a frequency of 20 kHz, and G 73 (Ref 55), which is much broader in scope than G 32. Both of these standards include definitions of terms related to cavitation and erosion, as well as information on specimen preparation, test conditions and procedures, and data interpretation. In both cases, the test apparatus is calibrated, and comparisons of the relative resistances of materials are made by the use of reference materials.

ASTM Standard G 32 is based on the generation and collapse of cavitation bubbles on a specimen surface vibrating at high frequency (Ref 54). It is used to evaluate the relative resistances of different materials to cavitation erosion.

Test Specimen. Reference 54 specifies a test specimen measuring 15.9 mm ± 0.05 mm (0.625 ± 0.002 in.) in diameter and not less than 3.2 mm (0.125 in.) in thickness. More details on specimen requirements are provided in Ref 54.

The test apparatus (Fig. 22) produces axial oscillation of the test specimen, which is partially immersed in the test liquid. This test is performed by either a magnetostrictive or piezoelectric transducer driven by an

electronic oscillator or amplifier. The transducer vibrates at a frequency of 20 kHz. The apparatus must include some method of measuring the displacement amplitude of the transducer as well as a means of maintaining the specified test temperature.

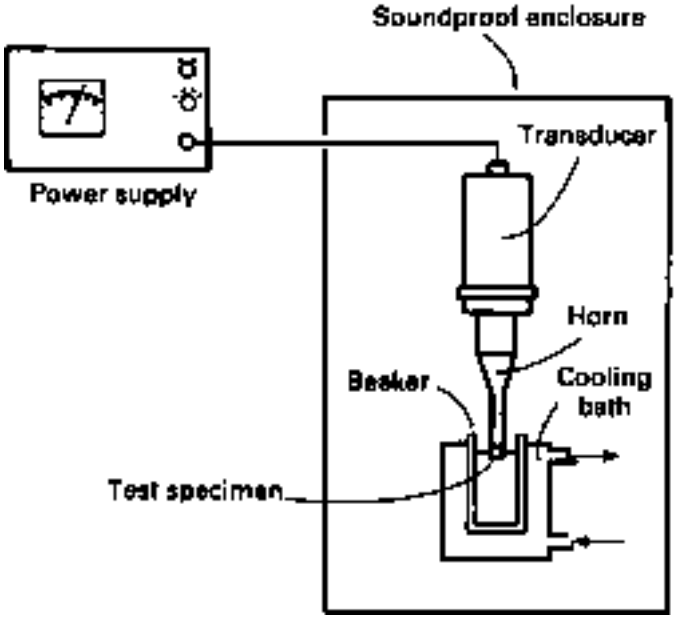


Fig. 22 Schematic of a typical vibratory erosion/cavitation test apparatus. Source: Ref 54

Test Conditions. The standard test liquid is distilled or other reagent water meeting the specifications of ASTM D 1193 (“Standard Specification for Reagent Water”) maintained at a temperature of 22 ± 1 °C (72 ± 2 °F). Air is maintained over the test liquid at a pressure of 96 ± 12 kPa (28.4 ± 3.5 in. of mercury). A peak-to-peak displacement amplitude of 0.05 mm (0.002 in.) ± 5% is specified for the duration of the test.

Test conditions can be varied by using different test liquids, temperatures, and pressures. Any such variations must be noted when results are reported. Liquids other than reagent water that have been used include petroleum derivatives, glycerin, and liquid metals. Water has been used at temperatures other than 22 °C (72 °F), and these tests have shown that the erosion rate peaks strongly at a temperature about midway between the freezing point and the boiling point of the test liquid. Therefore, the erosion rate of a material in water at normal atmospheric pressure peaks at about 50 °C (122 °F) and falls off on either side of this value. The decrease in erosion rate is most dramatic as the temperature rises above 50 °C (122 °F) (see the section “Effects of Test Parameters” in this article).

Test Procedure. The specimen must be cleaned and accurately weighed before testing, then immersed in the test liquid to a depth between 3.2 mm (0.125 in.) and 12.7 mm (0.5 in.). The peak-to-peak displacement amplitude should be monitored constantly. The test should be interrupted periodically to determine the mass loss of the specimen, which should be cleaned and dried carefully before reweighing. The time between measurements should be such that a plot of cumulative mass loss-versus exposure time can be established. The time between measurements, therefore, depends on the specimen material; softer materials (for example, aluminum alloys) can be examined every 15 min, while the interval between measurements for a harder material, such as Stellite alloy 6B, may range from 8 to 10 h. Testing should be continued at least until the erosion rate reaches a maximum and begins to diminish. When several materials are being compared, all should be tested until they reach comparable mean depths of erosion.

Data Acquisition and Reporting. The mean depth of erosion of the specimen should be calculated based on the full area of the test surface of the specimen. The report should include the following information:

- Alloy type, metallurgical condition, composition, and mechanical properties (including hardness)
- Specimen preparation, preferably including initial surface roughness measurement
- The number of specimens tested
- Tabulation of cumulative mass losses and cumulative exposure time for each specimen
- Plots of cumulative mean depth versus cumulative exposure time for each specimen

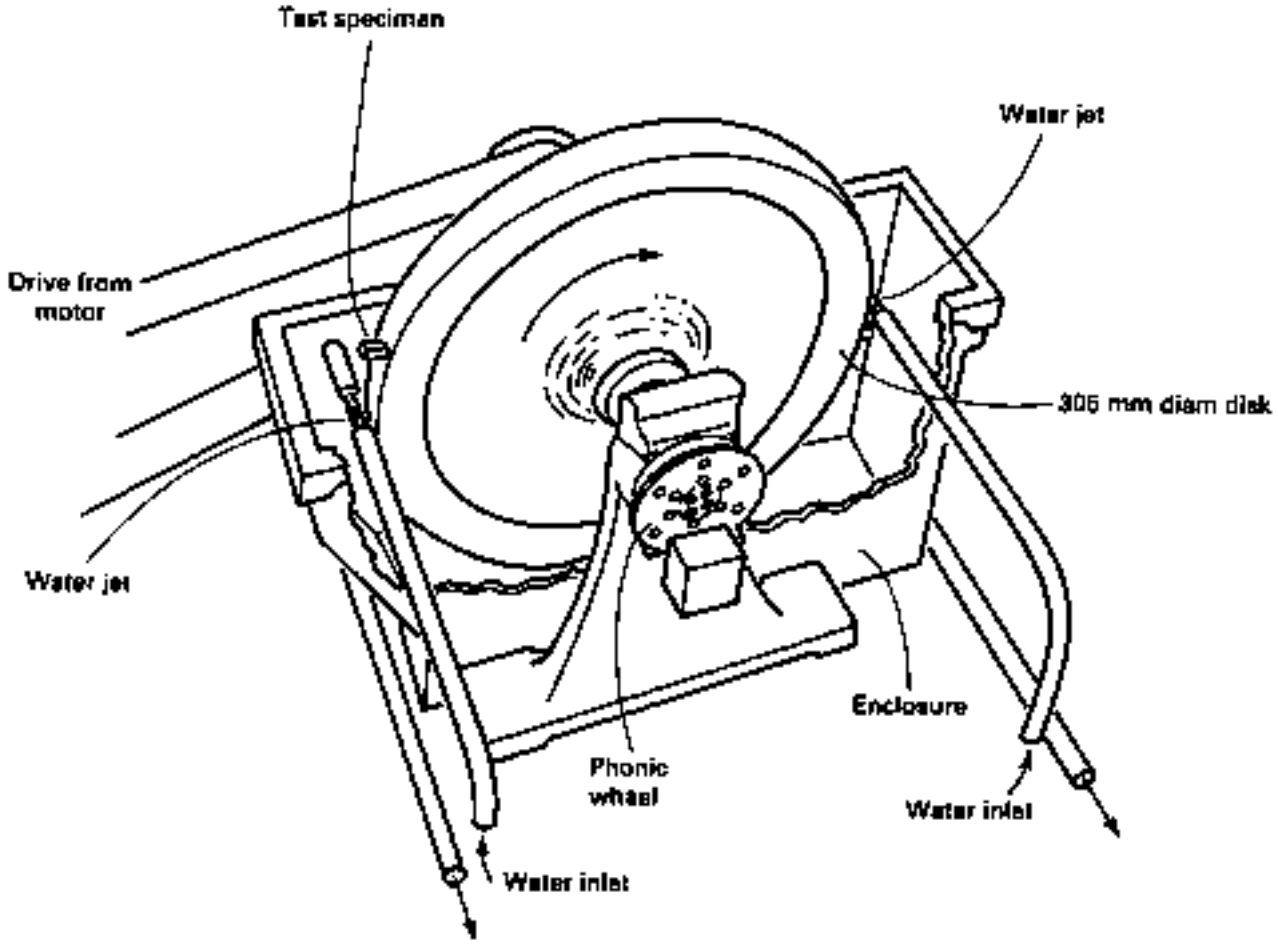
To facilitate comparisons between different materials, results often are expressed in terms of a single number based on the erosion resistance of the test specimen relative to that of a standard reference material. These types of ratings are discussed more fully in Ref 55.

The report also should include information on any deviations from the standard test procedure (for example, test liquid, temperature, or pressure, as discussed earlier) as well as any unusual occurrences or observations. In addition, each test should include at least one of the standard reference materials to facilitate calculation of the normalized erosion resistance of the test materials. Reference materials include aluminum alloy 1100-O, commercially pure annealed nickel (for example, Nickel 270), and type 316 stainless steel with a hardness of 150 to 175 HV. As mentioned, previously these materials are used also to calibrate the testing apparatus.

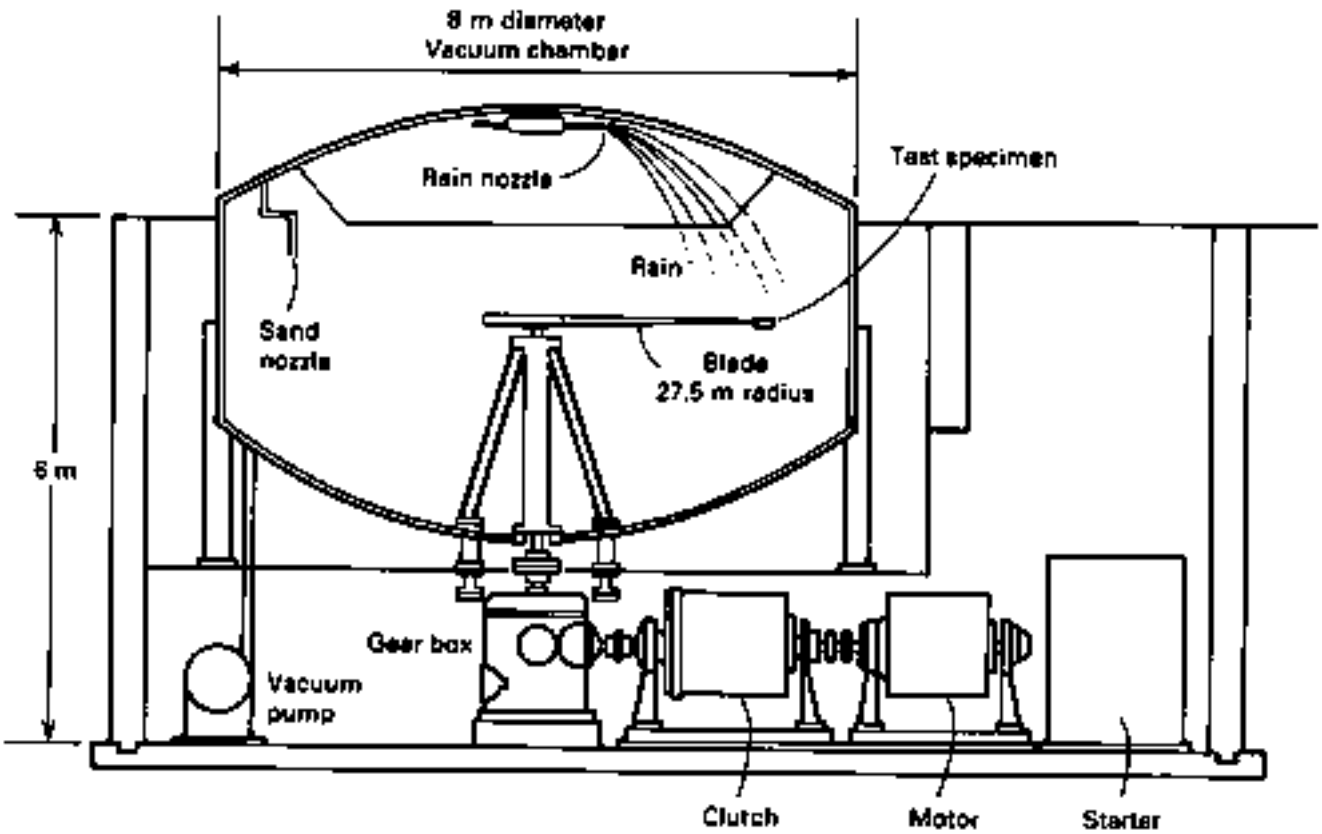
ASTM Standard G 73 provides guidelines for liquid drop impingement testing that, in addition to evaluating the resistance of metals, can be used to evaluate the degradation of optical properties of window materials and the destruction of coatings by the impinging liquid (Ref 55). The standard does not outline a single test method or apparatus but, instead, gives guidelines for setting up tests and specifies test and analysis procedures and the reporting of data.

Test Specimens. Metallic test specimens can be chosen to present a curved (airfoil or cylindrical) or flat surface to the impinging liquid. They can be machined from a solid bar or cut from sheet. If specimens are machined, care should be taken to avoid work hardening of the surface. Surface finish should range from 0.4 to 1.6 μm (16 to 63 μin.) rms (root mean square). If another surface finish is employed, it should be noted in the test report.

Test Apparatus. ASTM Standard G 73 mainly addresses erosion test devices of the rotating disk type. In these tests, the specimen or specimens are attached to a rotating disk or arm, and their circular path passes through one or more liquid jets or sprays, resulting in discrete impacts between the specimen and the droplets or the cylindrical surface of the jets. Peripheral velocities (and, therefore, impact velocities) ranging from about 50 m/s (165 ft/s) to as high as 1000 m/s (3280 ft/s) may be obtained, depending on the particular apparatus used. Figure 23 illustrates two examples of rotating disk test apparatuses.


Test specimen
Water jet
Drive from motor
305 mm diam disk
Water jet
Enclosure
Phonic wheel
Water inlet
Water inlet

(a)


8 m diameter
Vacuum chamber
Rain nozzle
Test specimen
Rain
Sand nozzle
Blade
27.5 m radius
6 m
Gear box
Vacuum pump
Clutch
Motor
Starter

(b)

Fig. 23 Examples of rotating disk and rotating arm erosion: cavitation test apparatuses. (a) Small, relatively low-speed rotating disk and jet apparatus. (b) Large, high-speed rotating arm spray apparatus. Source: Ref 55

Droplet or jet diameters also vary from about 0.1 to 5 mm (0.004 to 0.2 in.). Droplets can be generated by spray nozzles, vibrating hollow needles, or rotating disks with water fed onto their surface. The typical droplet or jet diameter and the volume of liquid impacting the specimen per unit time should be determined within 10%. Jet diameter usually is assumed to be the same as the nozzle diameter. Other types of liquid impact erosion test devices also are described briefly in Ref 55.

Test procedures for various materials vary to some extent. The standard gives procedures for structural materials and coatings, including metals, structural plastics, composites, and metals with ceramic or metallic coatings; elastomeric coatings; window materials; and transparent thin-film coatings on window materials. Details on these procedures are outlined in the standard.

Calculation of Erosion Resistance. Because it currently is not possible to define “absolute” erosion resistance, most investigators use comparative evaluations of different materials to quantify relative erosion resistance. ASTM Standard G 73 details several approaches for calculating and presenting relative or normalized measures of erosion resistance for the various classes of materials considered. These methods include evaluation based on time to failure, total material loss, erosion rate-time patterns, incubation period and maximum erosion rate, and terminal erosion rate. Results for different materials can also be normalized by direct comparison to one of the designated reference materials (for metals, aluminum alloys 1100-O or 6061-T6; annealed commercially pure nickel, such as Nickel 270; and type 316 stainless steel with a hardness between 155 and 170 HV) or by indirect comparison to a standardized reference scale. Details on these procedures, as well as on the calculation of an apparatus severity factor, are given in ASTM G 73.

Other Tests. Numerous other techniques are available for evaluating cavitation and erosion in the laboratory. Most of these practices are variations on the methods already described, employing rotating disks, vibrating devices, or venturi tubes to achieve the required fluid velocities (Ref 56). Another device used to produce cavitation and erosion is the liquid gun, which projects short, discrete slugs of liquid out of a nozzle onto the specimen (Ref 57). Details of other nonstandard test procedures are available in Ref 58, 59, 60.

Effects of Test Parameters (Ref 56). Test conditions—including the temperature, pressure, and flow velocity of the test liquid and the frequency, amplitude, and distance between the specimen and the vibrating surface in a vibrating fluid—all influence the incubation and intensity of cavitation. The effects of varying these test parameters are described next.

Variations in velocity and pressure should not be considered separately, because the intensity of erosion cavitation is a function of both parameters. Generally, erosion rates increase with velocity. A common technique used to evaluate the effect of velocity is to determine the number and size of pits produced in a soft material, such as aluminum, per unit time as velocity is varied (Ref 61). Although the number of pits observed increases over a wide range with velocity, the variation in pit size is relatively insignificant.

The influence of pressure has been investigated for both flow and vibratory cavitation (Ref 62, 63). In the former investigation, velocity was held constant and pressure was increased. Cavitation damage increased to a maximum and then decreased to zero at the pressure corresponding to cavitation inception (Ref 62). In vibratory testing, the erosion rate increased steadily with increasing pressure up to 4 atm, the maximum pressure obtainable in the device used (Ref 63). The authors noted, however, that at sufficient pressure the cavitation rate would, in fact, drop to zero.

Temperature. As mentioned earlier, erosion rate increases with temperature to a maximum, then decreases near the boiling point of the test liquid. The decrease in erosion rate near the boiling point is attributed to the increase in vapor pressure of the test liquid (Ref 62). The effect of temperature on the erosion rate of plain carbon steel in three aqueous environments is shown in Fig. 24.

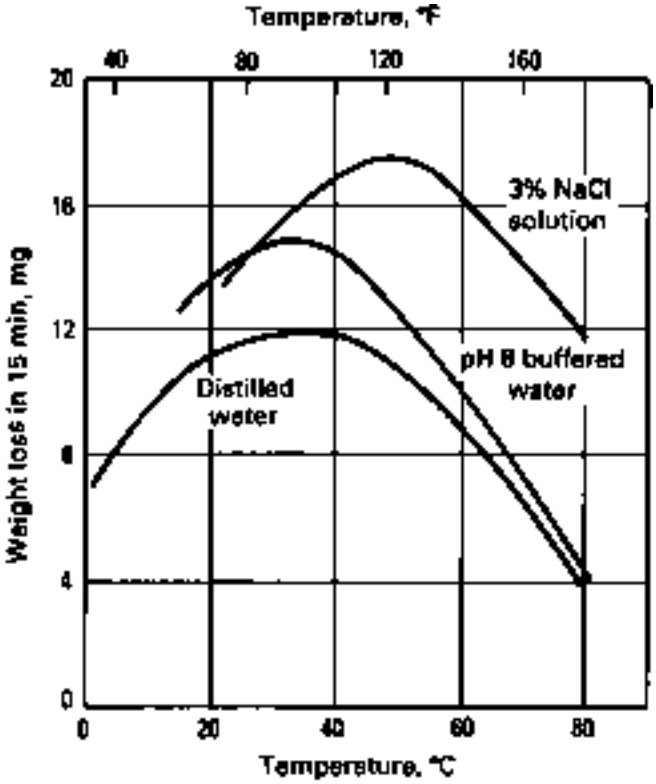


Fig. 24 Influence of temperature on the erosion rate of plain carbon steel in a vibratory cavitation device. Source: Ref 62

Amplitude and frequency in vibratory devices influence the degree of damage. In tests at constant frequency, erosion rate increased slightly with increasing amplitude (Ref 64). The incubation period for damage decreased with increasing amplitude.

Most vibratory devices operate at the the resonant frequency of a piezoelectric crystal. Therefore, frequency cannot be varied easily over a significant range. In one investigation, three different devices with frequencies of 10, 20, and 30 kHz were used to determine the effect of frequency on erosion rate (Ref 64). The stress produced by the cavitation increased as frequency decreased, but the effect was not as significant as that caused by changes in amplitude.

Distance between Specimen and Vibrating Surface. The amount of erosion damage to a specimen rises to a maximum as the distance between the specimen and the vibrating surface decreases. At a frequency of 20 kHz, the position that results in maximum erosion rates is approximately 0.5 mm (0.02 in.) from the vibrating surface (Ref 65). The position of the maximum is independent of amplitude, but no data are available on the influence of frequency on this position.

The influence of fluid properties has been studied by several investigators. Unfortunately, most of them made little attempt to identify those properties of the fluid that affect erosion rates. One investigator, however, considered fluid viscosity, vapor pressure, surface tension, and specific gravity for water, gasoline, and the sodium-potassium eutectic (Ref 66). The temperature dependence of erosion rates was attributed to the effects of velocity and surface tension at low temperatures and to the increase in vapor pressure at high temperatures. By comparing the temperature dependence of properties of the different fluids, he was able to predict the variation in erosion damage in the three fluids. The resulting equation, however, cannot be used to predict erosion rates in other fluids.

Correlations among Test Methods (Ref 53). Investigations have indicated that there are considerable differences in the erosion intensities produced by different test techniques (that is, high-velocity flow, vibratory, and impingement) and test parameters (vibratory frequency; specimen size and shape). There is, however, general agreement on the ranking of test materials. When the erosion resistance of a material is expressed as a normalized erosion resistance (NER), which is the ratio of the erosion rate of the test material to that of a standard material, results from different test methods and different laboratories can be directly compared and are reasonably consistent (Ref 66). Calculation of NER values and statistical analysis of erosion rate data are

discussed in detail in Ref 55, which also reports on the results of interlaboratory testing using a liquid impingement testing method at three specified velocities.

Correlations between Laboratory Results and Service (Ref 53). The erosion rates of materials that can be anticipated in service cannot be quantified at this time; however, methods have been developed to identify service erosion intensities under conditions of controlled operation. These methods include attaching aluminum panels to hydraulic turbine vanes (Ref 67), a strain-gage technique for measuring the relative intensities of laboratory and service devices (Ref 68), and the use of a radioactive paint to correlate the erosion intensity of an operating turbine with an erosion rate index (Ref 69). This latter investigation led to the establishment of relative resistance scales, in which laboratory exposures of 1 to 2 h in a vibratory apparatus can be compared with 16 to 100 h exposures in a venturi apparatus and with months of operation in full-scale turbines.

References cited in this section

53. J.Z. Lichtman, D.H. Kallas, and A. Rufolo, Evaluating Erosion (Cavitation) Damage, *Handbook of Corrosion Testing and Evaluation,* W.H. Ailor, Ed., John Wiley & Sons, 1971, p 453–472

54. "Standard Method of Vibratory Cavitation Erosion Test," G 32, *Annual Book of ASTM Standards,* ASTM

55. "Standard Practice for Liquid Impingement Erosion Testing," G 73, *Annual Book of ASTM Standards,* ASTM

56. C.M. Preece, Ed., Erosion, *Treatise on Materials Science and Engineering,* Vol 16, Academic Press, 1979

57. M.C. Rochester and J.H. Brunton, Influence of Physical Properties of the Liquid on the Erosion of Solids, *Erosion, Wear, and Interfaces with Corrosion,* STP 567, ASTM, 1974, p 128–147

58. M. Matsumura, Y. Oka, S. Okumoto, and H. Furuya, Jet-in-Slit Test for Studying Erosion-Corrosion, *Laboratory Corrosion Tests and Standards,* STP 866, G.S. Haynes and R. Baboian, Ed., ASTM, 1985, p 358–372

59. J.M. Hobbs, Report 69, National Engineering Laboratory, 1962

60. J.W. Holt and G.M. Wood, Ed., *Cavitation Research Facilities and Techniques,* American Society of Mechanical Engineers, 1964

61. R.T. Knapp, J.W. Daly, and F.G. Hammitt, *Cavitation,* McGraw-Hill, 1970

62. M.J. Robinson and F.G. Hammitt, *Trans. ASME D,* Vol 89, 1967, p 161

63. F.G. Hammitt and D.O. Rogers, *J. Mech. Eng. Sci.,* Vol 12, 1970, p 432

64. B. Vyas and C.M. Preece, *J. Appl. Phys.,* Vol 47, 1976, p 5133

65. F.J. Heymann, "Toward Quantitative Prediction of Erosion Damage," Report E-1463, Westinghouse Electric Corporation, Development Engineering Dept., June 1969

66. W.C. Leith, *Proc. ASTM,* Vol 65, 1965, p 789

67. R.T. Knapp, "Accelerated Field Tests of Cavitation Intensities," Paper 56-A-57, American Society of Mechanical Engineers, 1956

68. R. Canavelis, The Investigation of Cavitation Damage by Means of Resistance Gages, *Cavitation Research Facilities and Techniques,* J.W. Holt and G.M. Wood, Ed., American Society of Mechanical Engineers, 1964

69. S.L. Kerr and K. Rosenberg, An Index of Cavitation Erosion by Means of Radioisotopes, *Trans. ASME,* Vol 80 (No. 6), Aug 1958, p 1308–1314

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

Acknowledgments

The section on test method is adapted from *Corrosion,* Volume 13 of the *ASM Handbook,* 1987, p 311–313.

Cavitation Erosion

Yan-Ming Chen, CETIM Senlis, France

References

1. J.P. Franc, F. Avellan, B. Belahadji, D. Fréchou, D.H. Fruman, L. Briançon-Marjollet, A. Karimi, J.L. Kuency, and J.M. Michel, *La cavitation, mécanismes physiques et aspects industriels, (Cavitation, Physics Mechanism and Industry Aspects),* Presses Universitaires de Grenoble, 1995

2. G. Lesaffre and Y.M. Chen, “Etude sur les mécanismes de cavitation” (“Mechanism of Cavitation”), CETIM Report 1D1230, Centre Technique des Industries Mécaniques, 2001

3. C.M. Hansson and I.L.H. Hansson, Cavitation Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992

4. C.M. Preece, Cavitation Erosion, *Treatise on Materials Science and Technology,* Vol 16, Academic Press, 1979, p 249–305

5. F. Erdmann-Jesnitzer, Louis, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 171–196

6. B. Vyas and C.M. Preece, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 77–105

7. B. Vyas and I.L.H. Hansson, The Cavitation Erosion-Corrosion of Stainless Steel, *Corros. Sci.,* Vol 130 (No. 8-9), 1991, p 761–770

8. H. Grein, De la cavitation: une vue d'ensemble, (Cavitation: An Overview), *Rev. Tech. Sulzer,* NR Recherche, 1974, p 87–112

9. K. Steller, T. Krzysztofowicz, and Z. Reymann, *Erosion, Wear and Interfaces with Corrosion,* ASTM, 1974, p 152–170

10. D. Berthe, Cavitation and Film Rupture in a Lubrificated Contact and Its Relation to Surface Degradations, *Cavitation and Related Phenomena in Lubrification,* congrès de Leeds, Sept 1974, p 185–188

11. D. Summers-Smith, R.W. Wilson, F.T. Barwell, P.J. Lidgitt, D. Berthe, J.M. Conway-Jones, and K.L. Johnson, Cavitation Damage, *Cavitation and Related Phenomena in Lubrification,* congrès de Leeds, Sept 1974, p 198–217

12. T. Akada, J. Iwamoto, and K. Sano, Fundamental Studies on Cavitation Erosion, *Bull. JSME,* Vol 20 (No. 147), 1977, p 1067–1075

13. R.K. Palhan, Effect of the Stress Level on Cavitation Erosion Rate, *Wear,* Vol 45 (No. 2), 1977, p 151–160

14. P.V. Rao and D. Buckley, Cavitation Erosion Size Scale Effects, *Wear,* Vol 96, 1984, p 239–253

15. A. Karimi, and F. Avellan, Comparison of Erosion Mechanisms in Different Types of Cavitation, *Wear,* Vol 113 (No. 3), 1986, p 305–322

16. A. Karimi and M. Maamouri, Microscopic Study of Cavitation Behaviour in Copper and Cu-5,7 wt% Al since Crystals, *Wear,* Vol 139 (No. 1), 1990, p 149–169

17. E.P. Rood, Review Mechanisms of Cavitation Inception, *J. Fluids Eng. (ASME),* Vol 113 (No. 2), 1991, p 163–165

18. M. Ahmed, Investigation of the Temperature Effects on Induced Impact Pressure and Cavitation Erosion, *Wear,* Vol 218, 1998, p 119–127

19. R.H. Richmann and W.P. MacNaughton, Correlation of Cavitation Behavior with Mechanical Properties of Metals, *Wear,* Vol 140 (No. 1), 1990, p 63–82

20. R.W. Wilson, Cavitation Damage in Plain Bearings, *Cavitation and Related Phenomena in Lubrification,* congrès de Leeds, Sept 1974, p 177–183

21. R.D. James, Erosion Damage in Engine Bearings, *Tribol. Int.,* Vol 8 (No. 4), 1976, p 161–170

22. D.R. Garner, R.D. James, and J.F. Warriner, Cavitation Erosion Damage in Engine Bearings: Theory and Practice, *J. Eng. Power (Trans. ASME),* Vol 102 (No. 4), 1980, p 847–857

23. G.N. Blount and R.D. James, Cavitation Erosion Damage in Diesel Engine Main Bearings, *Tribol. Reciprocating Engines,* congrès de Leeds, 1982, p 297–303

24. J.F. Gülich and A. Rösch, L'érosion de cavitation dans les pompes centrifuges (Cavitation Erosion in Centrifugal Pumps), *Rev. Tech. Sulzer,* Vol 70 (No. 1), 1988, p 28–33

25. J.F. Gülich, Diagnosis of Cavitation in Centrifugal Pumps, *World Pumps,* Vol 308, 1992, p 15–20

26. R. Canavelis and P. Grison, Aspects industriels de la cavitation dans les pompes (Industrial Aspects of Cavitation in Pumps), *Rev. Fr. Méc.,* Vol 4, 1986, p 171–181

27. D.P. Sloteman, Avoiding Cavitation in the Suction Stage of High-Energy Pumps, *World Pumps,* Vol 348, 1995, p 10–48

28. W.K. Chan, Correlation between Cavitation Type and Cavitation Erosion in Centrifugal Pumps, *Int. J. Heat Fluid Flow,* Vol 11 (No. 3), 1990, p 269–271

29. J.B. Hunt, A.J. Ryde-Weller, and F.A.H. Asmead, Cavitation between Meshing Gear Teeth, *Wear,* Vol 71 (No. 1), 1981, p 65–78

30. H.G. Feller and A.J. Kharrazi, Cavitation Erosion of Metal and Alloys, *Wear,* Vol 93, 1984, p 249–260

31. S. Vaidya and C.M. Preece, Cavitation Erosion of Age-Hardenable Aluminium Alloys, *Metall. Trans. A,* Vol 9, 1978, p 299–307

32. W.T. Fu, Y.B. Yang, T.F. Jing, Y.Z. Zhang, and M. Yao, The Resistance to Cavitation Erosion of CrMnN Stainless Steels, *J. Mater. Eng. Perform.,* Vol 7 (No. 6), 1998, p 801–804

33. Y. Iwai, T. Okada, and F.G. Hammitt, Effect of Temperature on the Cavitation Erosion of Cast Iron, *Wear,* Vol 96, 1983, p 149–253

34. S.K. Wu, H.C. Lin, and C.H. Yeh, A Comparison of the Cavitation Erosion Resistance of TiNi Alloys, SUS304 Stainless Steel and Ni-Based Self-Fluxing Alloy, *Wear,* Vol 244, 2000, p 85

35. B.C.S. Rao and D.H. Buckley, Characterisation of Erosion of Metallic Materials under Cavitation Attack in a Mineral Oil, *ASLE Trans.,* Vol 29 (No. 4), 1985, p 473–480

36. T. Okada and S. Hattori, Cavitation-Erosion of Aluminium Oxide and Silicon Carbide, *JSME Int. J.,* Vol 33 (No. 3), 1990, p 382–387

37. T. Okada, Cavitation Erosion of Ceramics, *Jpn. J. Tribol.,* Vol 41 (No. 8), 1996, p 803–814

38. T. Okada, Y. Iwai, and K. Awazu, Effects of Plating on Cavitation Erosion, *Wear,* Vol 124, 1988, p 21–31

39. W.J. Tomlinson, R.T. Moule, J.H.P.C. Megaw, and A.S. Brandsden, Cavitation Wear of Untreated and Laser-Processed Hardfaced Coatings, *Wear,* Vol 117 (No. 1), 1987, p 103–107

40. W.J. Tomlinson, J.H.P.C. Megaw, A.S. Brandsden, and M. Girardi, The Effect of Laser Melting on the Cavitation Wear of Grey Cast Iron in Distilled and 3% Salt Water, *Wear,* Vol 116 (No. 2), 1987, p 249–260

41. Y. Iwai, T. Okada, T. Fujieda, and K. Awazu, Effects of Hard Chromium Plating on Cavitation Erosion, *Wear,* Vol 128, 1988, p 189–200

42. C.T. Kwok, F.T. Cheng, and H.C. Man, Laser Surface Modifications of UNS S31603 Stainless Steel, Part II: Cavitation Erosion Characteristics, *Mater. Sci. Eng.,* Vol A290, 2000, p 74–88

43. C.T. Kwok, F.T. Cheng, and H.C. Man, Cavitation Erosion and Pitting Corrosion of Laser Surface Melted Stainless Steels, *Surf. Coat. Technol.,* Vol 99 (No. 3), 1998, p 295–304

44. S.M. Ahmed, K. Hokkirigawa, and P. Oba, Fatigue Failure of SUS 304 Caused by Vibratory Cavitation Erosion, *Wear,* Vol 177, 1994, p 129–137

45. J.L. Heuze, F. Gaillard, and B. Dubois, Etude de l'endommagement par érosion de cavitation de matériaux métalliques à structures CFC (Study on Cavitation Erosion Damage of Metallic Materials with CFC), *Mém. Sci. Rev. Métall.,* Vol 85 (No. 10), 1988, p 581–594

46. C. Morton, Combating Cavitation Erosion by Elastomeric-Type Coating, *Corros. Prev. Control,* Vol 36 (No. 2), 1989, p 31–33

47. A. Karimi, Cavitation Erosion of a Duplex Stainless Steel, *Mater. Sci. Eng.,* Vol 86 (No. 1/2), 1987, p 191–203

48. T. Okada, K. Awazu, and H. Kawasaki, Influence of Cavitation Erosion on Corrosion Fatigue and the Effect of Surface Coatings on Resistance, *Wear,* Vol 63 (No. 1), 1980, p 51–70

49. C.J. Heathcock, A. Ball, D. Yamey, and B.E. Protheroe, Cavitation Erosion of Cobalt Based Stellite Alloys, Cemented Carbides and Surface Treated Low Alloy Steels, *Wear of Metal,* congrès de San Francisco, 1981, p 597–606

50. C.M. Preece and C.W. Drapper, The Effect of Laser Quenching the Surfaces of Steels on Their Cavitation Erosion Resistance, *Wear,* Vol 67 (No. 3) 1981, p 321–328

51. G.S. Mikaelyan and S.Y. Nozdrev, The Resistance to Breakdown by Erosion of Certain Compositions and Materials Used for Surfacing, *Automat. Weld.,* Vol 27 (No. 12), 1974, p 31–33

52. M.G. Fontana, *Corrosion Engineering,* 3rd ed., McGraw-Hill, 1986

53. J.Z. Lichtman, D.H. Kallas, and A. Rufolo, Evaluating Erosion (Cavitation) Damage, *Handbook of Corrosion Testing and Evaluation,* W.H. Ailor, Ed., John Wiley & Sons, 1971, p 453–472

54. "Standard Method of Vibratory Cavitation Erosion Test," G 32, *Annual Book of ASTM Standards,* ASTM

55. "Standard Practice for Liquid Impingement Erosion Testing," G 73, *Annual Book of ASTM Standards,* ASTM

56. C.M. Preece, Ed., Erosion, *Treatise on Materials Science and Engineering,* Vol 16, Academic Press, 1979

57. M.C. Rochester and J.H. Brunton, Influence of Physical Properties of the Liquid on the Erosion of Solids, *Erosion, Wear, and Interfaces with Corrosion,* STP 567, ASTM, 1974, p 128–147

58. M. Matsumura, Y. Oka, S. Okumoto, and H. Furuya, Jet-in-Slit Test for Studying Erosion-Corrosion, *Laboratory Corrosion Tests and Standards,* STP 866, G.S. Haynes and R. Baboian, Ed., ASTM, 1985, p 358–372

59. J.M. Hobbs, Report 69, National Engineering Laboratory, 1962

60. J.W. Holt and G.M. Wood, Ed., *Cavitation Research Facilities and Techniques,* American Society of Mechanical Engineers, 1964

61. R.T. Knapp, J.W. Daly, and F.G. Hammitt, *Cavitation,* McGraw-Hill, 1970

62. M.J. Robinson and F.G. Hammitt, *Trans. ASME D,* Vol 89, 1967, p 161

63. F.G. Hammitt and D.O. Rogers, *J. Mech. Eng. Sci.,* Vol 12, 1970, p 432

64. B. Vyas and C.M. Preece, *J. Appl. Phys.,* Vol 47, 1976, p 5133

65. F.J. Heymann, “Toward Quantitative Prediction of Erosion Damage,” Report E-1463, Westinghouse Electric Corporation, Development Engineering Dept., June 1969

66. W.C. Leith, *Proc. ASTM,* Vol 65, 1965, p 789

67. R.T. Knapp, “Accelerated Field Tests of Cavitation Intensities,” Paper 56-A-57, American Society of Mechanical Engineers, 1956

68. R. Canavelis, The Investigation of Cavitation Damage by Means of Resistance Gages, *Cavitation Research Facilities and Techniques,* J.W. Holt and G.M. Wood, Ed., American Society of Mechanical Engineers, 1964

69. S.L. Kerr and K. Rosenberg, An Index of Cavitation Erosion by Means of Radioisotopes, *Trans. ASME,* Vol 80 (No. 6), Aug 1958, p 1308–1314

Liquid-Impact Erosion

R.H. Richman, Daedalus Associates, Inc.

Introduction

EROSION of solid surfaces can be brought about solely by liquids in two ways: from damage induced by formation and subsequent collapse of voids or cavities within the liquid, and from high-velocity impacts between a solid surface and liquid droplets. The former process is called cavitation erosion and the latter is liquid-droplet erosion. Both forms of liquid-impact erosion are discussed in detail in *Friction, Lubrication, and Wear Technology,* Volume 18 of *ASM Handbook* (Ref 1, 2). Most of what is presented in those articles is not repeated here; instead, emphasis in this article is on manifestations of damage and on ways to minimize or repair liquid-impact damage.

References cited in this section

1. C.M. Hansson and I.L.H. Hansson, Cavitation Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 214–220

2. F.J. Heymann, Liquid Impingement Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 221–232

Liquid-Impact Erosion

R.H. Richman, Daedalus Associates, Inc.

Cavitation Erosion

Hydrodynamic effects in flowing liquids or in liquids exposed to an acoustical field can reduce the pressure in particular locations to a critical value that is related to the vapor pressure of the liquid. In those locations, cavities can be nucleated in the liquid, grow to a stable size, and be swept downstream. As the cavities reach regions of higher pressure, they become unstable and collapse violently. Cavity collapses not near a solid boundary tend to be symmetrical, and shock waves are emitted into the surrounding liquid. In contrast, cavities in contact with or very close to a solid surface collapse asymmetrically; one portion of the cavity boundary

folds inward, forming a jet of liquid directed toward the solid (Ref 1). A consensus of theoretical modeling and experimental measurements is that microjets achieve velocities up to about 500 m/s (1640 ft/s) for durations of 2 to 3 μs (Ref 3); associated collapse pressures are of the order of 1 to 1.4 GPa (145 to 203 ksi). (See Ref 4 for a compilation of past investigations of collapse pressures.) The impacts of these microjets, perhaps in conjunction with simultaneous shock waves, result in erosion of solid surfaces.

Affected Systems and Materials. Cavity formation and collapse in liquids have been studied extensively over the years (Ref 5, 6). Machines and components that suffer cavitation damage include hydraulic pumps, valves, hydroturbine runners and guide vanes, ship propellers and hydrofoils, mechanical devices with restricted fluid passages (e.g., seals and bearings), heat-exchanger tubes, and hydraulic structures such as spillways, gates, and tunnels that are associated with dams. Consequently, cavitation erosion is still a very active research topic worldwide.

Materials selection for hydraulic machines, especially massive ones, is dictated, in large part, by economics and fabricability. Thus, most large hydroturbines and many large centrifugal pumps are constructed of carbon-manganese steels, typically 0.2% C and 1% Mn. The most common specification is ASTM A-27, grade 70-40; other applicable specifications for carbon steel in turbines and pumps are A-216, grade WCB; A-516, grade 60; and A-283, grades A, B, C, and D. When higher strength is needed, cast martensitic stainless steels are specified. Many machines, especially high-head pumps, were constructed of 12 to 13% Cr steel (CA-15), air cooled from about 1000 °C (1832 °F) and tempered in the range of 590 to 815 °C (1095 to 1500 °F). In recent years, CA-15 has been largely supplanted by CA-6NM, a 12 to 14% Cr casting grade with 4% Ni, which has somewhat better weldability. The martensitic stainless steels have yield stresses about twice those of carbon-manganese steels, typically about 550 MPa (80 ksi) compared with 250 MPa (36 ksi), and resistance to cavitation erosion that is superior to A-27. Some smaller hydroturbines and pumps have been fabricated from cast austenitic stainless steels of the 18 to 20% Cr and 8 to 10% Ni class (CF-8).

Manifestations of Cavitation Erosion. Cavitation erosion typically begins with deformation rumpling and work hardening of metal surfaces, often with delineation of grain and phase boundaries, as shown in Fig. 1(a), 2(a), and 3(a) for AISI 1020 steel, CA-6NM, and type 304 stainless steel, respectively. Thereafter, material removal takes place first at surface upheavals, which usually occur at boundaries, in the ferrite phase at triple points and ferrite-pearlite interfaces in 1020 steel (Fig. 1b), at martensite-lath interfaces in CA-6NM (Fig. 2b), and at grain and twin boundaries in type 304 stainless steel (Fig. 3b). Fracture surfaces thus exposed may be striated, as in Fig. 4. Material loss then spreads over the entire region subject to cavitation collapse, resulting in extremely roughened surfaces with completely obscured microstructural features (Fig. 5).

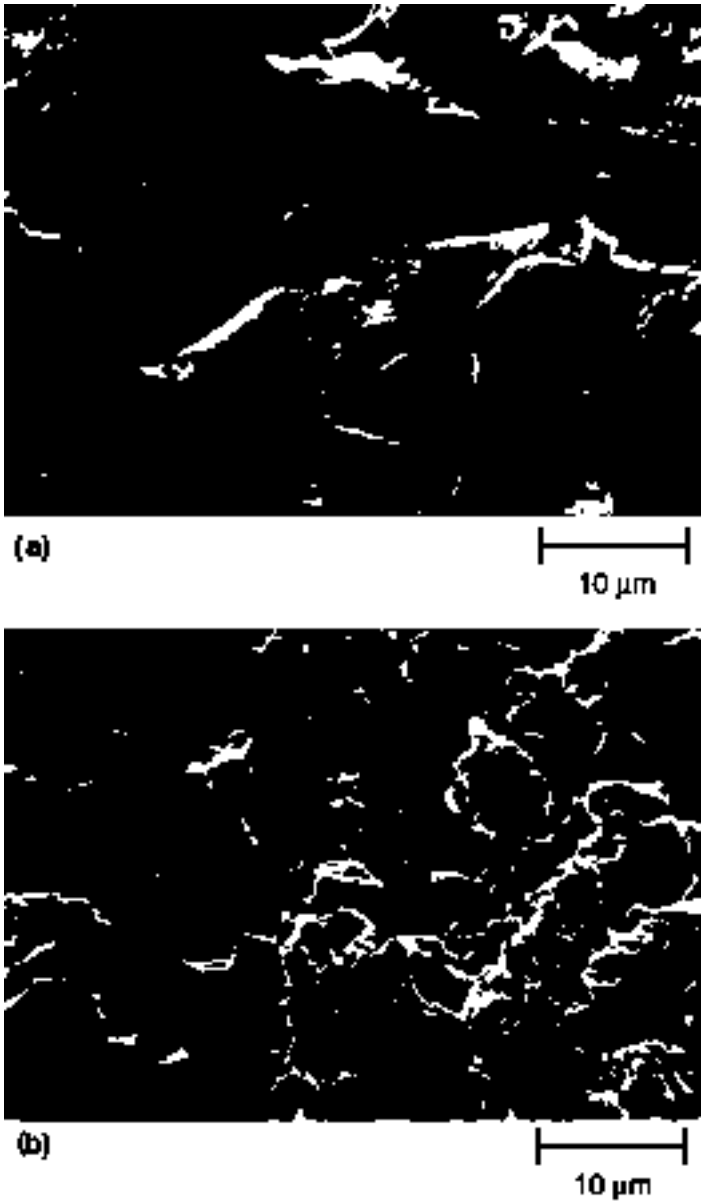


Fig. 1 Exposure to vibratory cavitation of normalized AISI 1020 steel. (a) Damage after 5 min. (b) Material removal after 10 min

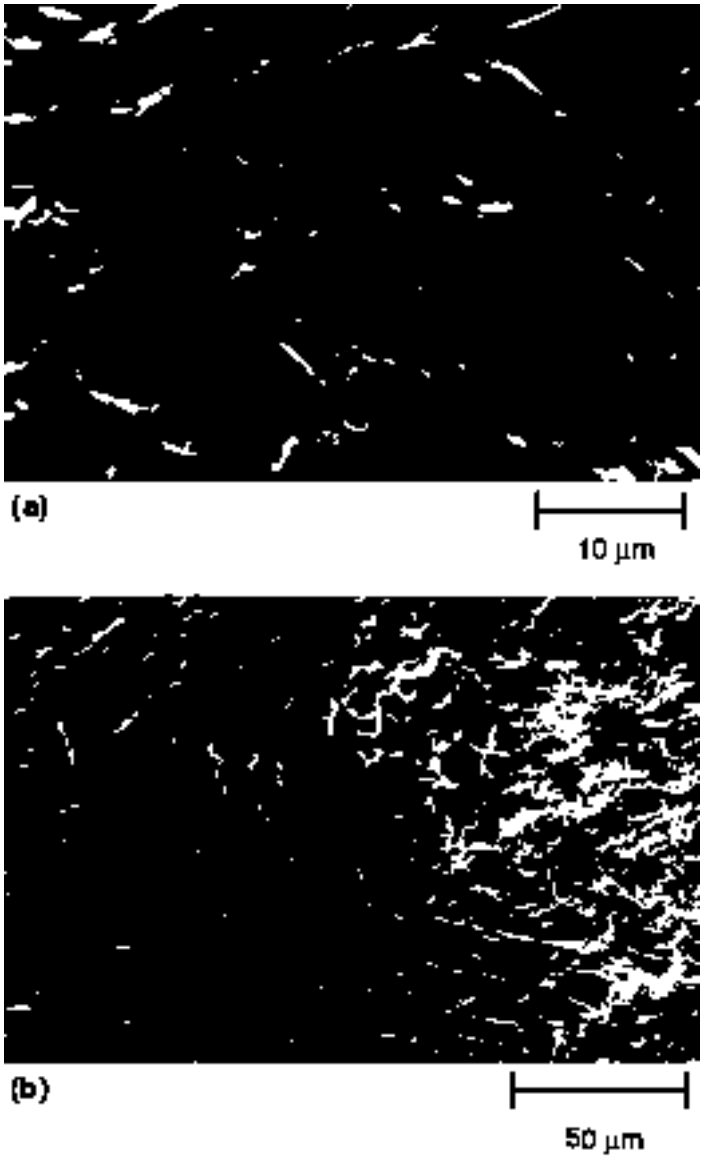


Fig. 2 Vibratory cavitation erosion of CA-6NM martensitic stainless steel. (a) Deformation rumpling and pitting at lath boundaries. (b) Early stage of material removal

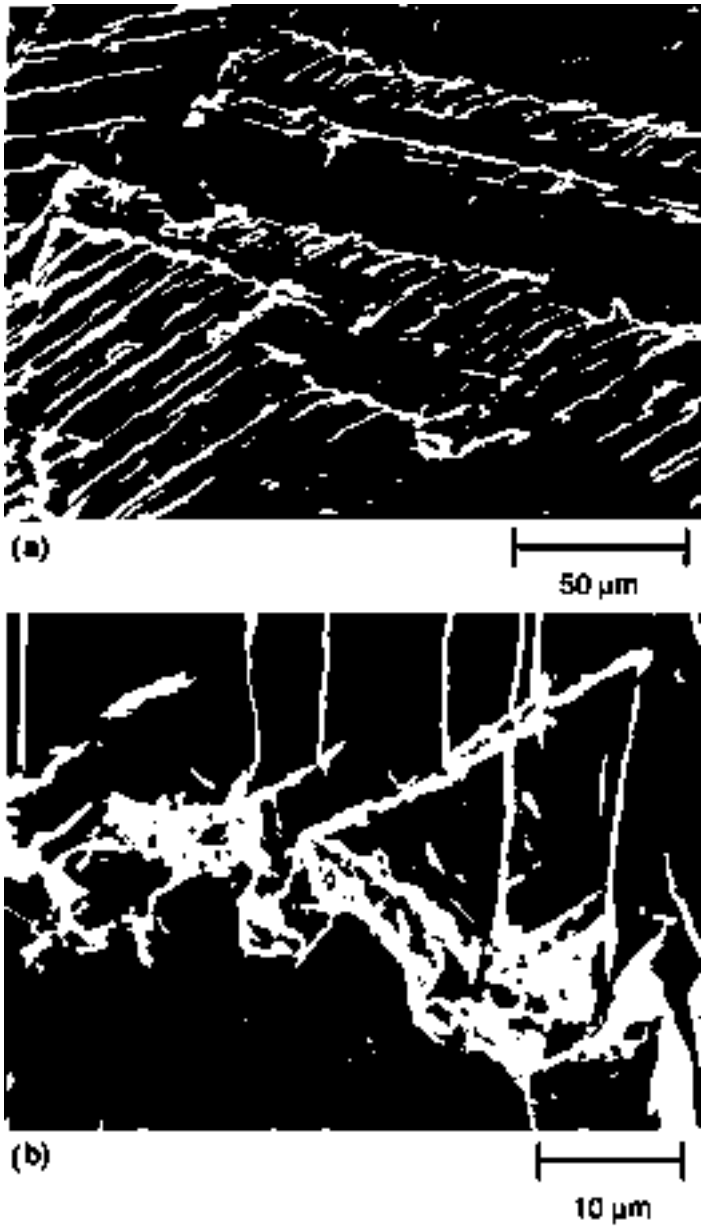


Fig. 3 Vibratory cavitation erosion of type 304 austenitic stainless steel. (a) Linear deformation features and boundary definition. (b) Material removal at upheaved grain boundary

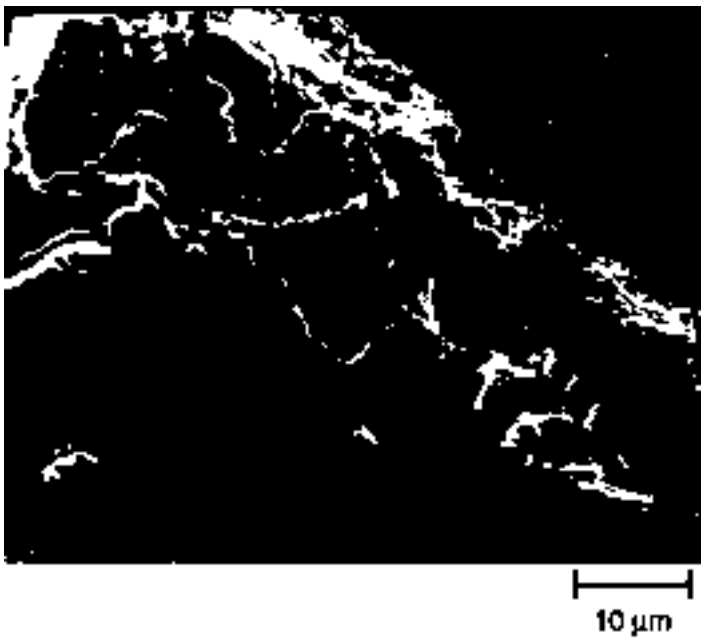


Fig. 4 Striated fracture surface in type 304 stainless steel early in material-removal stage of cavitation erosion

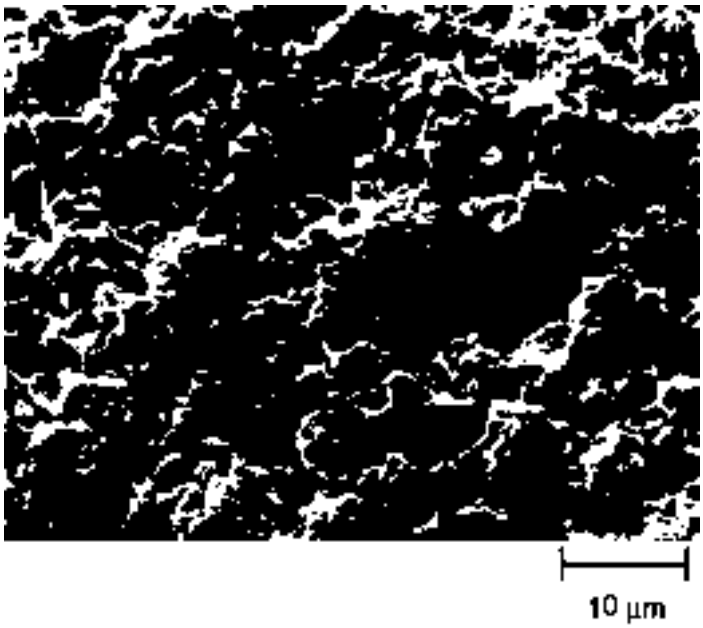


Fig. 5 Eroded surface of type 304 stainless steel with all microstructural features obliterated

The rate at which cavitation erosion proceeds through the various stages is difficult to predict, except in closely controlled laboratory tests (Ref 1). In full-scale hydraulic machinery, various factors make prediction uncertain or impossible: the intensity of cavitation, the properties of the construction material, the original surface condition, and the changing hydraulic profile as damage develops. Damage can accumulate in a relatively short time, as illustrated by deep erosion in a stainless steel overlay on a carbon steel blade in a Francis turbine after about 7000 h (Fig. 6), and erosion completely through a carbon steel impeller in a 5000 hp pump after about 20,000 h (Fig. 7). A previous repair with austenitic stainless steel at the toe of the vane can be seen in Fig. 7. Both examples of aggressive cavitation erosion are attributable to improper hydraulic profiles in the manufactured components. Similar machines, under what appear to be identical conditions except for better profiles, might operate for years longer without visible erosion damage.

Fig. 6 Deep cavitation erosion of austenitic stainless steel weld overlay on a carbon steel turbine blade. Courtesy of T.J. Spicher

Fig. 7 Erosion through a carbon steel impeller in a pump

References cited in this section

1. C.M. Hansson and I.L.H. Hansson, Cavitation Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 214–220

3. A. Karimi and J.L. Martin, Cavitation Erosion of Materials, *Int. Met. Rev.,* Vol 31(No. 1), 1986, p 1–26

4. T. Okada, Y. Iwai, and K. Awazu, A Study of Cavitation Bubble Collapse Pressures and Erosion, Part 1: A Method for Measurement of Collapse Pressures, *Wear,* Vol 133, 1989, p 219–232

5. F.G. Hammitt, *Cavitation and Multiphase Flow Phenomena,* McGraw-Hill, 1980

6. R.E.A. Arndt, Cavitation in Fluid Machinery and Hydraulic Structures, *Ann. Rev. Fluid Mech.,* Vol 13, 1981, p 273–328

Liquid-Impact Erosion

R.H. Richman, Daedalus Associates, Inc.

Liquid-Droplet Erosion

This form of liquid-impact erosion is caused by continued impingement of liquid droplets, usually water moving at high relative velocities (>50 m/s, or >164 ft/s). Because water droplets cannot move at high velocities without breaking up, the practical problem of liquid-droplet erosion arises when a solid body moves at high velocity through a field of slower-moving droplets. High contact ("water hammer") pressure is generated at each impact, in addition to which the liquid droplet then spreads laterally at a velocity substantially higher than the impact velocity (Ref 2). Surface regions of impacted solids are deformed in much the same way as they are under cavitation collapse, except that individual liquid-droplet impacts are usually more damaging

than the impacts from cavitation collapse, owing to the larger size of the impacting liquid jets and the high-velocity lateral outflows from the impact sites.

Affected Systems and Materials. Material degradation by liquid-droplet erosion is a significant problem in just a few kinds of engineering systems: blades in the low-pressure ends of steam turbines where the steam is “wet,” and aircraft, missiles, and helicopter rotors flying at high speeds through rainstorms. Much research has been devoted to liquid-droplet erosion, and good summaries can be found in Ref 2 and 7.

The standard blade material in steam turbines manufactured all over the world is martensitic steel containing 12 to 13% Cr (similar to types 403 and 410) and a variety of secondary alloying elements, depending on the manufacturer. High magnetomechanical damping is a key property of 12% Cr steel, which serves admirably as blading in high-purity steam. Some turbines have been fitted with precipitation-hardened stainless steel (17-4 PH) blades in the next-to-last row of the low-pressure stages; the intention was to forestall corrosion fatigue failures in the location where steam first becomes wet and impurities are concentrated. Titanium alloys, exemplified by Ti-6Al-4V, are widely used as blading in small industrial turbines in the United States and elsewhere, and as long blades in large steam turbines in Russia and Switzerland, owing to their high strength-to-weight ratio and better resistance to erosion and corrosion than 12% Cr steel. All of the commonly used blade alloys are protected from liquid-droplet impingement in the last blade row by local hardening (by flame or induction heating), brazed-on shields of Stellite 6B, or welded-on hard facing.

Helicopter rotors are made of titanium sheet laid over a structural framework and pressurized internally with nitrogen, or of fiber-reinforced composites with a polymeric foam core.

Manifestations of Liquid-Droplet Erosion. As with cavitation erosion, damage from liquid-droplet impacts starts with deformation and work hardening of metal surfaces, characterized by rumpling and emergence of boundaries and phases. Material loss occurs first at regions of weakness, as is seen for a Stellite 6B shield on a turbine blade in Fig. 8. Unlike the randomly roughened surfaces characteristic of cavitation erosion, surfaces eroded by liquid droplets tend to have a somewhat more regular columnar appearance after significant exposure; surface configurations are shown in Fig. 9 and 10. Figure 9 is a low-magnification view of liquid-droplet erosion damage on a Stellite 6B shield brazed to a 12% Cr blade; note that the braze layer has been eroded past the focus plane of the photograph. At higher magnification, the eroded surface of 12% Cr steel is seen in Fig. 10(a) and the Stellite 6B shield in Fig. 10(b). Figure 11 is a cross section of eroded Stellite 6B. The characteristic columnar topography of liquid-droplet erosion is probably a consequence of impulse pressures focused by pits that were formed earlier in the process (Ref 2). Such amplification effects are much less evident in cavitation erosion.

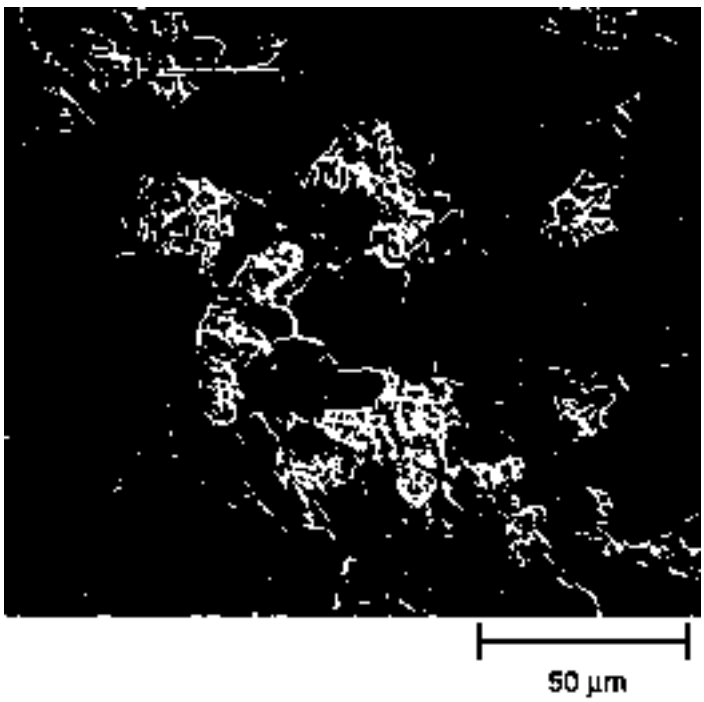


Fig. 8 Early stage of liquid-droplet erosion of Stellite 6B

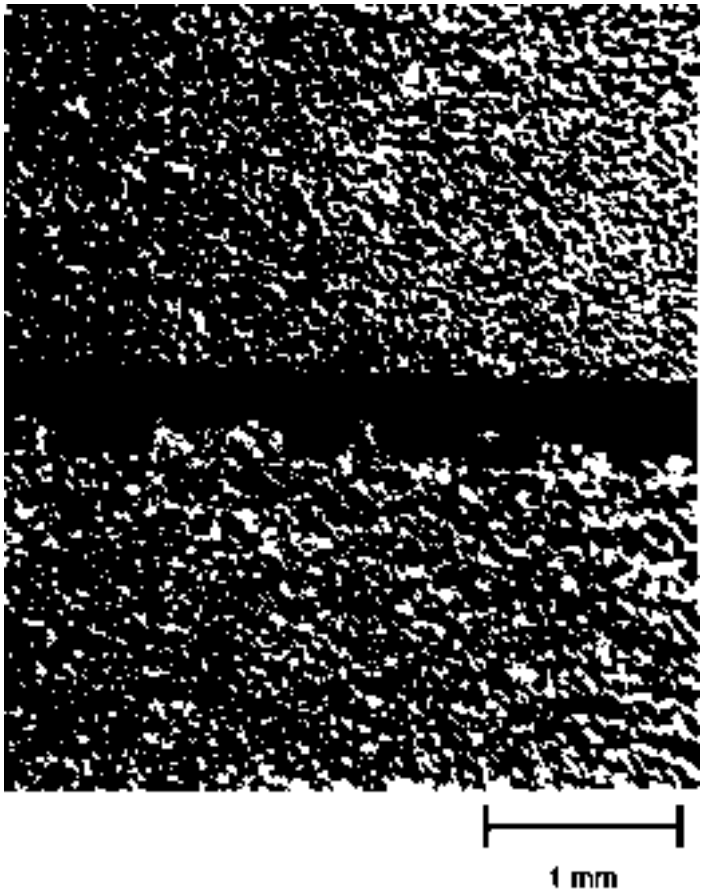


Fig. 9 Joint area between Stellite 6B (top) and 12% Cr steel (bottom) of a steam turbine blade eroded by water droplets

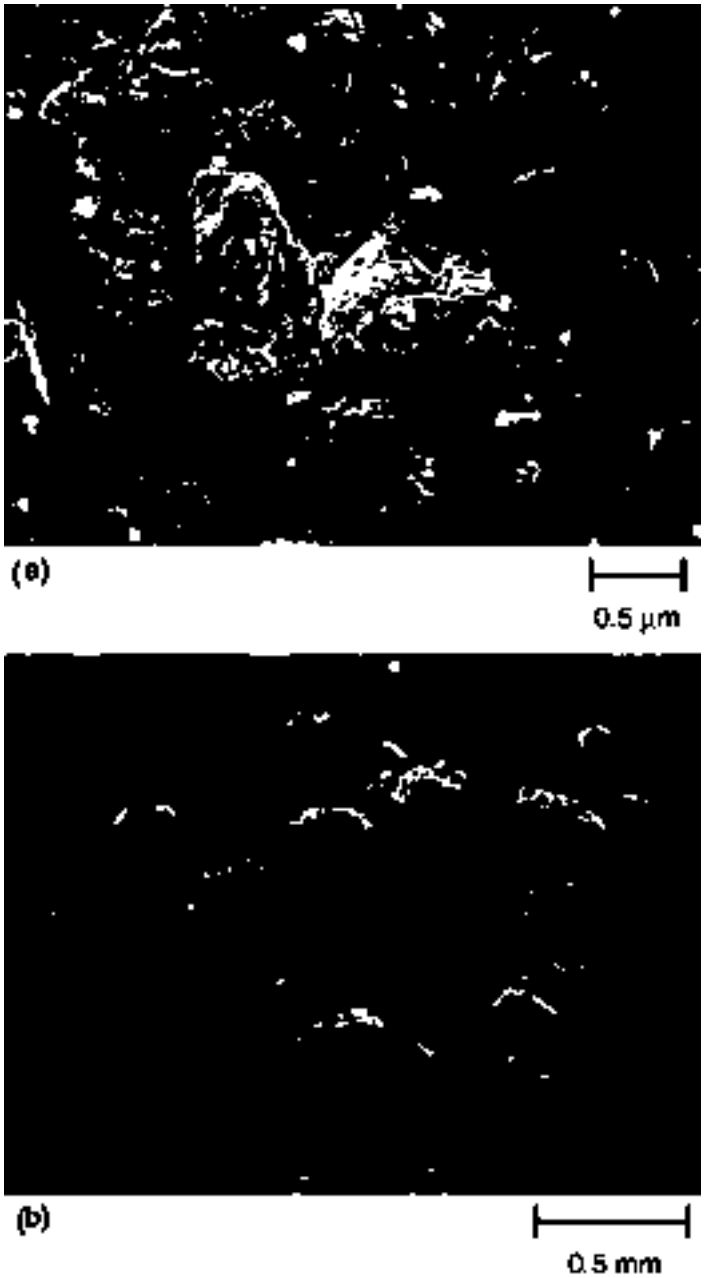


Fig. 10 Surface appearance at low magnification of a steam turbine blade eroded by water droplets. (a) 12% Cr steel blade material. (b) Stellite 6B shield

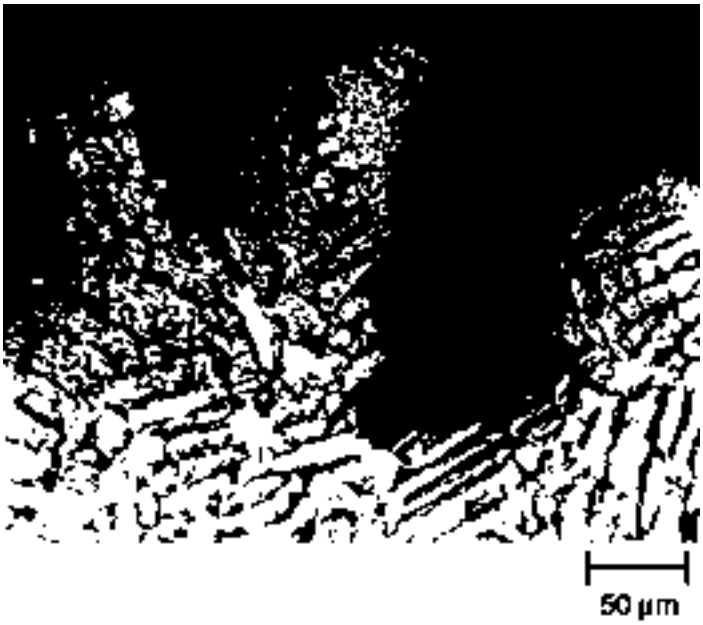


Fig. 11 Eroded Stellite 6B shield in cross section

References cited in this section

2. F.J. Heymann, Liquid Impingement Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 221–232

7. J.H. Brunton and M.C. Rochester, Erosion of Solid Surfaces by the Impact of Liquid Drops, *Treatise on Materials Science and Technology,* Vol 16, *Erosion,* C.M. Preece, Ed., Academic Press, 1979, p 185–248

Liquid-Impact Erosion

R.H. Richman, Daedalus Associates, Inc.

Materials Issues in Liquid Impact

Specifying materials for hydraulic machinery has traditionally been a matter of empiricism, because the damage mechanism in liquid-impact erosion was not known with certainty. Thus, it was not possible to specify the material properties that would ensure good erosion resistance. Experience had shown that, in general, hardness and ductility were beneficial, as was high work hardenability (Ref 3). Nevertheless, quantitative correlations between measured erosion rates and single properties or combinations of properties could not be obtained except within narrow classes of materials with similar structure (Ref 5). Therefore, equipment builders have relied on empirical comparisons of erosion resistance, such as that compiled by Heymann (Fig. 13 in Ref 2) for all types of liquid impact.

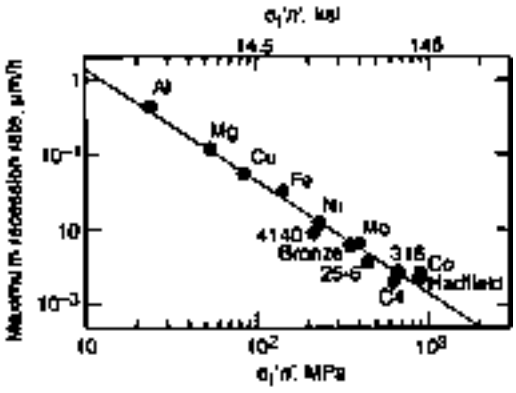


Fig. 12 Dependence of maximum erosion rate in cavitation erosion on the combined parameter $\sigma'_f n'$

Various investigators have attributed the failure mode in liquid-impact erosion to fatigue, and evidence of fatigue can be deduced from experiments (Ref 8 and references therein). The fact that, at best, only qualitative correlation between erosion rates and fatigue behavior could be obtained (Ref 9, for example) probably stems from the choice of relatively insensitive parameters (such as endurance limit) from stress- or load-controlled experiments for analysis. When relationships were sought between erosion rates and the parameters that characterize strain-based fatigue (Ref 10), damage rates in cavitation erosion and liquid-droplet erosion correlated very well with cyclic deformation properties (Ref 8). The main determinant of erosion resistance is the fatigue strength coefficient, σ'_f, which is a measure of cyclic *stress* resistance. Material removal rates correlate best with the product ($\sigma'_f n'$), in which *n′*, the cyclic strain-hardening exponent, reflects cyclic *strain* resistance. Figure 12 illustrates the correlation with cavitation-erosion rates in vibratory tests performed in conformance with ASTM G 32 (Ref 11). It will be appreciated that erosion behavior is not simply related to any monotonic property such as true fracture stress or ultimate tensile stress, because the monotonic stress-strain curve does not describe the cyclic behavior of metals and alloys.

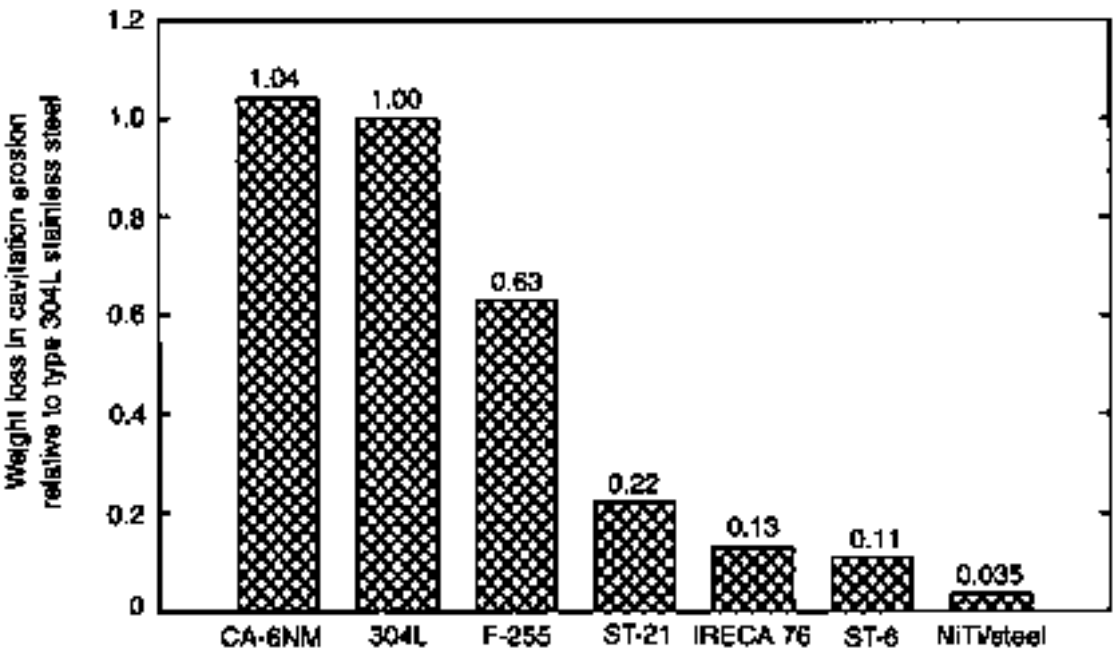


Fig. 13 Relative erosion rates in vibratory cavitation

The implication of the foregoing is that the metallurgical factors that promote resistance to low-cycle fatigue (Ref 12) should also confer resistance to liquid-impact erosion. This postulate was confirmed with near-equiatomic NiTi alloys; both the martensitic form (B19′ structure) and the "austenitic" form (B2 structure) are anomalously resistant to low-cycle fatigue (Ref 13) and, it turned out, to liquid-impact erosion (Ref 8, 14).

References cited in this section

2. F.J. Heymann, Liquid Impingement Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 221–232

3. A. Karimi and J.L. Martin, Cavitation Erosion of Materials, *Int. Met. Rev.,* Vol 31(No. 1), 1986, p 1–26

5. F.G. Hammitt, *Cavitation and Multiphase Flow Phenomena,* McGraw-Hill, 1980

8. R.H. Richman and W.P. McNaughton, A Metallurgical Approach to Improved Cavitation Erosion Resistance, *J. Mater. Eng. Perform.,* Vol 6(No. 5), 1997, p 633–641

9. G. Hoff, W. Herbert, and H. Rieger, "Rain and Sand Erosion, Phenomena of Material Destruction Caused by Repeated Loads," STP 474, *Characterization and Determination of Erosion Resistance,* ASTM, 1970, p 353–375

10. J.D. Morrow, "Cyclic Plastic Strain Energy and the Fatigue of Metals," STP 378, *Internal Friction, Damping, and Cyclic Plasticity,* ASTM, 1965, p 45–84

11. "Standard Method of Vibratory Cavitation Erosion Test," G 32, *Annual Book of ASTM Standards,* ASTM

12. C.E. Feltner and P. Beardmore, "Strengthening Mechanisms in Fatigue," STP 467, *Achievement of High Fatigue Resistance in Metals and Alloys,* ASTM, 1970, p 77–112

13. K.N. Melton and O. Mercier, Fatigue of NiTi and CuZnAl Shape Memory Alloys, *Proc. Fifth Int. Conf. on the Strength of Metals and Alloys,* P. Haasen, V. Gerold, and G. Kostorz, Ed., Pergamon, 1979, p 1243–1248

14. R.H. Richman, J.A. Maasberg, and A.S. Rao, Comparison of Actual and Simulated Liquid-Droplet Erosion, *Microstruct. Sci.,* Vol 25, 1998, p 183–190

R.H. Richman, Daedalus Associates, Inc.

Mitigation and Repair of Liquid- Impact Damage

Minimizing Damage. Although proper design is a necessary condition for long lives of equipment that could be exposed to liquid impact, it is often not a sufficient condition. For example, the design principles for avoiding cavitation are well known (Ref 5), and large hydroturbines can be configured to ensure that cavitation does not occur during operation at a specific load. However, generating units may be run at greater or lesser loads (i.e., in nonoptimal hydraulic regimes) for significant portions of their lifetimes, owing to generation-dispatch decisions that are remotely initiated. Likewise, liquid-droplet impingement in steam turbines can be minimized by interstage moisture removal to reduce the amount of condensed water that can impact downstream stages and by increasing the spacing between stator and rotor rows to break up the droplets. Nevertheless, threats from cavitation or liquid droplets cannot be eliminated completely from the operating environment of most hydraulic

machines; more erosion-resistant materials may be the only practical way to extend the life and decrease repair costs of affected equipment.

Using More Erosion-Resistant Materials. Metals and alloys that are far more erosion resistant than the standard constructional materials described earlier are well known. The best of these, exemplified by cobalt-base alloys and NiTi alloys, are characterized by energy-absorbing deformation modes: fine twinning, stress- or strain-induced phase transformation, and/or reversible motions of twinlike boundaries (Ref 15, 16, 17). Of course, building large machines out of such alloys is impractical for reasons of cost and fabricability. In principle, however, inexpensive base materials such as carbon-manganese steels can be clad with highly erosion-resistant alloys by fusion welding (Ref 18) or explosive bonding (Ref 19) to protect regions subject to liquid impact. In particular, the effectiveness of cobalt-containing weld overlays, such as Stellite 6 (ST-6), Stellite 21 (ST-21), and the IRECA series of stainless steels developed by Hydro Quebec (Ref 20), has been demonstrated in field trials; erosion rates relative to type 304 stainless steel are summarized in Fig. 13. In practice, it is rare to find protective weld overlays of anything except austenitic stainless steel in hydroturbines and pumps. This is partly attributable to the inadvisability of fusion cladding with specialty alloys in the field, where high-temperature postweld heat treatments for microstructural control and stress relief are usually not possible.

As already mentioned, last-stage blades in steam turbines are traditionally protected by strips of Stellite 6B attached to the leading edges by brazing. The weight penalty of Stellite, important in these long blades, could be avoided by substituting NiTi or an aluminide intermetallic (Ref 20) for Stellite.

Repairing Damage. This section is aimed primarily at field repair of cavitation damage in large hydraulic machines. Steam turbine blades damaged by liquid-droplet impingement are normally not repaired; they are replaced.

Conventional Weld Repairs. Cavitation damage in hydroturbines and pumps is usually repaired by air-arc gouging to sound metal, cleaning the gouged area with a wire wheel or wire brush, preheating the areas to be repaired, filling the cavities with weld metal, and grinding to final contour. This seems simple enough, but often, in practice it is not.

Improperly executed repairs result in cracking, accelerated erosion damage, or both. Appropriate welding procedures can be found, for example, in *Welding, Brazing, and Soldering,* Volume 6 of *ASM Handbook;* they need not be repeated here. Instead, three experiential guidelines for weld repairs of erosion damage (Ref 21) follow:

- Before starting a weld repair, look closely at the region immediately upstream of the damage to determine if a contour change occurs there. Grinding away high spots is often more effective than adding material to the erosion pit caused by the high spot. This approach is illustrated in Fig. 14, where metal-removal contours have been marked upstream of an eroded area on a hydroturbine blade.
- There is always a potential for distortion of a hydraulic profile from excessive heat buildup during weld repairs. A rule of thumb is not to allow an area 152 mm (6 in.) from the weld zone to get too hot to touch without severe discomfort.
- It may be tempting to repair damage in carbon steel with much more damage-resistant austenitic stainless steel. Problems with distortion, cracks from residual stresses, subsequent repairs, and postweld grinding can be minimized by making the repair with carbon steel and then ensuring a smooth contour upstream. If the erosion problem persists after perfecting the contour, then resort to stainless filler.

Fig. 14 Area marked for contour adjustment just upstream of cavitation damage on a turbine blade. Courtesy of T.J. Spicher

The difficulties associated with what some refer to as “mixed metallurgy” are worth emphasizing. Figure 15 is an example of accelerated erosion and cracking 27,000 h after weld repair on an impeller in a 80,000 hp pump. Cavitation damage in the pump, manufactured of CA-15 martensitic stainless steel, had been repaired on-site with E309 austenitic stainless steel filler after 10,700 h of operation. Cracking in and around the weld-overlaid areas was so extensive that whole sections of the impeller had to be cut out and replaced with fabricated pieces in the shop.

Fig. 15 Accelerated cavitation erosion and cracking associated with austenitic stainless steel weld deposits on a martensitic stainless steel (CA-15) impeller vane

Thermal Spray Repairs. Thermal spraying, with its great compositional flexibility and minimal thermal effects on substrates, would seem to be an obvious candidate for repairing erosion damage. Those benefits notwithstanding, as-sprayed deposits have not been very successful in resisting liquid impact, owing to the presence of brittle phases or porosity in the deposits or inadequate adhesion between deposits and substrates (Ref 22, 23). In contrast, Ni-Cr-Si-B alloys, plasma sprayed and then flame remelted, are used to repair cavitation damage in hydroturbines in China (Ref 24). Laboratory tests have confirmed that commercially available powder (e.g., Ni-15.5Cr-4Fe-4Si-3B, in weight percent), plasma sprayed on a variety of substrates and then flame remelted, form adherent, hard, erosion-resistant coatings (Ref 25). Performance in vibratory cavitation of as-sprayed and flame remelted Ni-Cr-Fe-Si-B coatings are compared with type 304L stainless steel in Fig. 16.

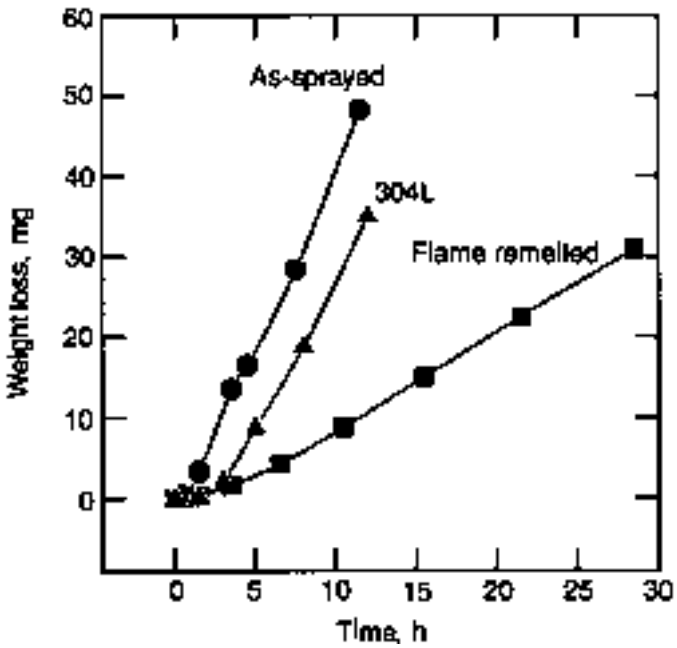


Fig. 16 Behavior comparison of plasma-sprayed Ni-Cr-Fe-Si-B powder with wrought type 304L stainless steel in vibratory cavitation

Repairs Made with Polymers. Over the years, elastomers and other polymers, either alone or mixed with inorganic substances, have been rumored to be the best way to repair cavitation damage. While it is true that elastomers can perform well in mild cavitation conditions, they are much less capable than metals of dissipating the heat generated by high-intensity liquid impacts, and they are difficult to bond strongly to metals. Intense cavitation invariably causes thermal decomposition and/or delamination of polymer repairs. Measures such as polyurethane paint definitely have their place for protecting low-head hydraulic machines, but repairs of erosion damage with polymers in high-head turbines and pumps should be regarded as a stopgap that allows a device to keep operating until the next scheduled outage, when proper repairs can be made.

References cited in this section

5. F.G. Hammitt, *Cavitation and Multiphase Flow Phenomena,* McGraw-Hill, 1980

15. D.A. Woodford, Cavitation-Erosion-Induced Phase Transformation in Alloys, *Metall. Trans.,* Vol 3, 1972, p 1137–1145

16. R.H. Richman, W.P. McNaughton, and A.S. Rao, Cyclic Deformation and Phase Transformation in Cavitation Erosion of Alloys, *Int. Conf. on Cavitation,* C453/051, Institution of Mechanical Engineers, 1992, p 87–94

17. R.H. Richman, Deformation-Induced Martensite and Resistance to Cavitation Erosion, *J. Phys. (France) IV,* Colloque 8, Vol 5, 1995, p 1193–1198

18. A. Akhtar, A.S. Rao, and D. Kung, Cavitation Erosion of Stainless Steel, Nickel, and Cobalt Alloy Weld Overlay Materials, *Coatings and Bimetallics for Aggressive Environments,* R.D. Sisson, Jr., Ed., American Society for Metals, 1985, p 125–142

19. C.A. Zimmerly, O.T. Inal, and R.H. Richman, Explosive Welding of a Near-Equiatomic Nickel-Titanium Alloy to Low-Carbon Steel, *Mater. Sci. Eng. A,* Vol 188, 1994, p 251–254

20. M. Johnson, D.E. Mikkola, P.A. March, and R.N. Wright, The Resistance of Nickel and Iron Aluminides in Cavitation Erosion and Abrasive Wear, *Wear,* Vol 140, 1990, p 279–289

21. T.J. Spicher, Alaska Power Authority, personal communications, 1985–1989

22. A. Akhtar, Plasma Sprayed Coatings for Cavitation Protection in Hydraulic Turbines, *Mater. Perform.,* Vol 21(No. 8), 1982, p 15–18

23. M. Scholl and P. Clayton, Abrasive and Erosive Wear of Some Hypervelocity Air Plasma Sprayed Coatings, *Thermal Spray Coatings: Properties, Processes, and Applications, Proc. Fourth Nat. Thermal Spray Conf.,* T. Bernecki, Ed., ASM International, 1992, p 39–52

24. K.-Z. Sang and Y.-G. Li, Cavitation Erosion of Flame Spray Weld Coating of Nickel-Base Powder, *Wear,* Vol 189, 1995, p 20–24

25. C.A. Zimmerly and O.T. Inal, Dept. of Materials and Metallurgical Engineering, and R.H. Richman, Daedalus Associates Inc., unpublished research, New Mexico Institute of Mining and Technology, Socorro, NM, 1996

Acknowledgments

The author thanks J.A. Maasberg for metallography, W.P. McNaughton and Dr. A.S. Rao for long-term collaboration, and T.J. Spicher and K. Kyle for guidance about practices in hydrogenerating and pumping stations. Most of the research in which the author participated was sponsored by the Electric Power Research Institute (EPRI); I am grateful for that support and for the encouragement of Dr. J. Stringer, the EPRI project manager.

Liquid-Impact Erosion

R.H. Richman, Daedalus Associates, Inc.

References

1. C.M. Hansson and I.L.H. Hansson, Cavitation Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 214–220

2. F.J. Heymann, Liquid Impingement Erosion, *Friction, Lubrication, and Wear Technology,* Vol 18, *ASM Handbook,* ASM International, 1992, p 221–232

3. A. Karimi and J.L. Martin, Cavitation Erosion of Materials, *Int. Met. Rev.,* Vol 31(No. 1), 1986, p 1–26

4. T. Okada, Y. Iwai, and K. Awazu, A Study of Cavitation Bubble Collapse Pressures and Erosion, Part 1: A Method for Measurement of Collapse Pressures, *Wear,* Vol 133, 1989, p 219–232

5. F.G. Hammitt, *Cavitation and Multiphase Flow Phenomena,* McGraw-Hill, 1980

6. R.E.A. Arndt, Cavitation in Fluid Machinery and Hydraulic Structures, *Ann. Rev. Fluid Mech.,* Vol 13, 1981, p 273–328

7. J.H. Brunton and M.C. Rochester, Erosion of Solid Surfaces by the Impact of Liquid Drops, *Treatise on Materials Science and Technology,* Vol 16, *Erosion,* C.M. Preece, Ed., Academic Press, 1979, p 185–248

8. R.H. Richman and W.P. McNaughton, A Metallurgical Approach to Improved Cavitation Erosion Resistance, *J. Mater. Eng. Perform.,* Vol 6(No. 5), 1997, p 633–641

9. G. Hoff, W. Herbert, and H. Rieger, "Rain and Sand Erosion, Phenomena of Material Destruction Caused by Repeated Loads," STP 474, *Characterization and Determination of Erosion Resistance,* ASTM, 1970, p 353–375

10. J.D. Morrow, "Cyclic Plastic Strain Energy and the Fatigue of Metals," STP 378, *Internal Friction, Damping, and Cyclic Plasticity,* ASTM, 1965, p 45–84

11. "Standard Method of Vibratory Cavitation Erosion Test," G 32, *Annual Book of ASTM Standards,* ASTM

12. C.E. Feltner and P. Beardmore, “Strengthening Mechanisms in Fatigue,” STP 467, *Achievement of High Fatigue Resistance in Metals and Alloys,* ASTM, 1970, p 77–112

13. K.N. Melton and O. Mercier, Fatigue of NiTi and CuZnAl Shape Memory Alloys, *Proc. Fifth Int. Conf. on the Strength of Metals and Alloys,* P. Haasen, V. Gerold, and G. Kostorz, Ed., Pergamon, 1979, p 1243–1248

14. R.H. Richman, J.A. Maasberg, and A.S. Rao, Comparison of Actual and Simulated Liquid-Droplet Erosion, *Microstruct. Sci.,* Vol 25, 1998, p 183–190

15. D.A. Woodford, Cavitation-Erosion-Induced Phase Transformation in Alloys, *Metall. Trans.,* Vol 3, 1972, p 1137–1145

16. R.H. Richman, W.P. McNaughton, and A.S. Rao, Cyclic Deformation and Phase Transformation in Cavitation Erosion of Alloys, *Int. Conf. on Cavitation,* C453/051, Institution of Mechanical Engineers, 1992, p 87–94

17. R.H. Richman, Deformation-Induced Martensite and Resistance to Cavitation Erosion, *J. Phys. (France) IV,* Colloque 8, Vol 5, 1995, p 1193–1198

18. A. Akhtar, A.S. Rao, and D. Kung, Cavitation Erosion of Stainless Steel, Nickel, and Cobalt Alloy Weld Overlay Materials, *Coatings and Bimetallics for Aggressive Environments,* R.D. Sisson, Jr., Ed., American Society for Metals, 1985, p 125–142

19. C.A. Zimmerly, O.T. Inal, and R.H. Richman, Explosive Welding of a Near-Equiatomic Nickel-Titanium Alloy to Low-Carbon Steel, *Mater. Sci. Eng. A,* Vol 188, 1994, p 251–254

20. M. Johnson, D.E. Mikkola, P.A. March, and R.N. Wright, The Resistance of Nickel and Iron Aluminides in Cavitation Erosion and Abrasive Wear, *Wear,* Vol 140, 1990, p 279–289

21. T.J. Spicher, Alaska Power Authority, personal communications, 1985–1989

22. A. Akhtar, Plasma Sprayed Coatings for Cavitation Protection in Hydraulic Turbines, *Mater. Perform.,* Vol 21(No. 8), 1982, p 15–18

23. M. Scholl and P. Clayton, Abrasive and Erosive Wear of Some Hypervelocity Air Plasma Sprayed Coatings, *Thermal Spray Coatings: Properties, Processes, and Applications, Proc. Fourth Nat. Thermal Spray Conf.,* T. Bernecki, Ed., ASM International, 1992, p 39–52

24. K.-Z. Sang and Y.-G. Li, Cavitation Erosion of Flame Spray Weld Coating of Nickel-Base Powder, *Wear,* Vol 189, 1995, p 20–24

25. C.A. Zimmerly and O.T. Inal, Dept. of Materials and Metallurgical Engineering, and R.H. Richman, Daedalus Associates Inc., unpublished research, New Mexico Institute of Mining and Technology, Socorro, NM, 1996

Wear Failures of Plastics

Sujeet K. Sinha, National University of Singapore

Introduction

PLASTICS (or polymers*) are used in a variety of engineering and nonengineering applications where they are subjected to surface damage and wear. Examples of the tribological (involving sliding between two surfaces) use of plastics include gears and cams of various machines, tires, break pads, conveyors, hoppers, automobile body parts, aircraft, spacecrafts, hip/knee joint replacement, roller-skating wheels, and household appliances (washing machine, tubs, etc.). Wear of material parts is a very common cause of failure or low working life of machines, leading to financial loss and life hazards. Therefore, it is important to understand how polymers and other materials wear. Similar to the wear of metal, polymer wear is affected by several factors that may be broadly divided into three groups: mechanical, environmental, and thermal. These three groups of factors largely decide the mechanism of wear of a polymer surface when it comes in contact with another surface.

Historically, polymer wear has been studied based on the prevailing wear mechanisms at the contact zone (between the polymer surface and a hard counterface), which led to several methods of classification. The classification of polymer wear mechanisms that has often been followed in the literature is based on three methodologies of defining types of wear (Ref 1). The first classification is based on the two-term model that divides wear mechanisms into two types—interfacial and bulk. The second classification is more phenomenological and is based on the perceived wear mechanism. This classification includes fatigue wear, chemical wear, delamination wear, fretting, erosion, abrasion, and transfer wear. The third classification is specific to polymers and draws the distinction based on mechanical properties of polymers. In the third classification, wear study is separated as "Elastomers," "Thermosets," "Glassy Thermoplastics," and "Semicrystalline Thermoplastics." These classifications provide a useful basis for understanding wear failures in polymers. More often than not, wear of a polymer is a complex phenomenon that involves several of the wear mechanisms listed previously in any one wear process. For the purpose of this article, details on several of the aforementioned classifications are expanded, using wear data and micrographs from published works. The primary goals are to present the mechanisms of polymer wear and to quantify wear in terms of wear rate (rate of removal of the material). This analysis is restricted mostly to base polymers (with no fillers). Normally, polymers used in tribological applications are subjected to sliding against hard surfaces such as metals. A polymer-polymer sliding pair, except in few instances, usually produces undesirable high friction and high wear conditions due to enhanced adhesion between the polymer. Also, poor conductivity of the polymers results in elevated temperature at the polymer/polymer interface, leading to melting and rapid wear. Therefore, the focus of this article is on the wear of polymers when slid against metallic surfaces.

Footnote

* *The terms "plastic" and "polymers" have some distinctions. However, in this article, the two terms mean engineering plastics. Engineering plastics are polymers that contain very small percentage of additives such as plasticizers and antioxidants in order to enhance their physical and mechanical properties.

Reference cited in this section

1. B.J. Briscoe and S.K. Sinha, Wear of Polymers, *Proc. Inst. Mech. Eng. J., J. Eng. Tribol.,* Vol 216, 2002

Wear Failures of Plastics

Sujeet K. Sinha, National University of Singapore

Interfacial Wear

The notion of interfacial wear arises from the popular two-term model of frictional energy dissipation (Ref 2). This model states that in any frictional phenomenon, where frictional energy is released at the contact points between two sliding surfaces, there can be two types of energy dissipation—interfacial and bulk. Though subjectively defined, the interface may be considered the region of the material very close (a few microns) to the contact point. This region of the material is almost instantly affected by the stress and thermal conditions

arising at the contact points due to sliding. The interfacial wear is defined as the removal of the material due to interfacial friction energy dissipation between asperities leading to events such as material softening, transfer wear, and chemical wear. A schematic of the processes involved in the interfacial wear is shown in Fig. 1.

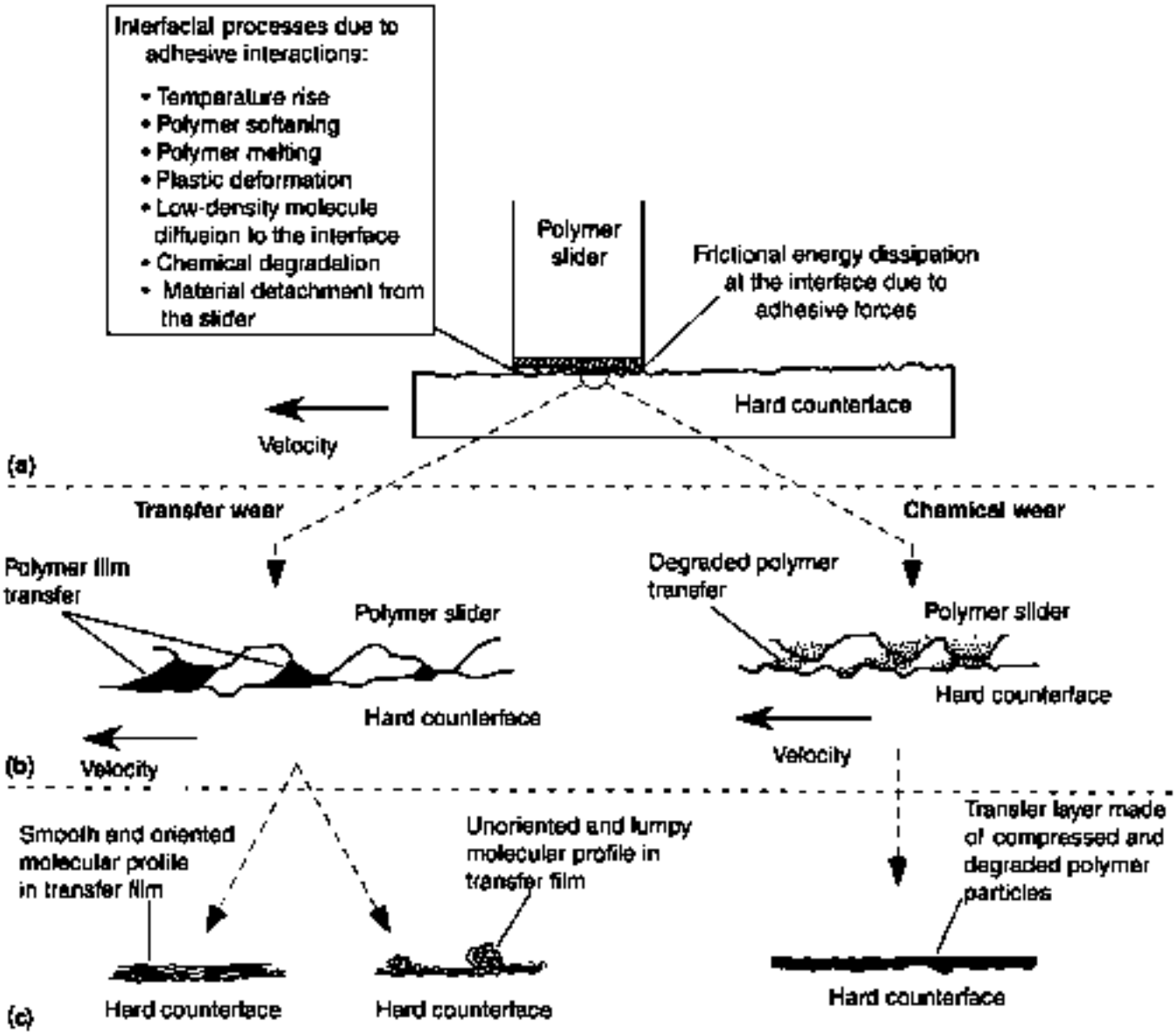


Fig. 1 Interfacial wear processes. (a) Initial contact of the two surfaces. (b) Running-in process where the soft polymer molecules are gradually transferred to the hard counterface as third-body. (c) Steady-state wear process where the wear and friction phenomena are influenced mainly by the shear and adhesive properties of the transferred film. Reprinted with permission from Ref 1

A distinction within the interfacial wear process may be made based on whether or not the frictional heat dissipation is isothermal or quasi-adiabatic. Isothermal heat dissipation can change the mechanical property of the interface zone as opposed to the quasi-adiabatic, which affects only the transfer layer normally present at the true interface. The chemical-wear mechanism is initiated if the frictional heat can chemically affect the polymer surface, resulting in the production of degraded polymer molecules. The other important parameter to consider in interfacial wear is the roughness of the counterface. For rough and hard counterfaces, the wear mode is generally that of bulk or cohesive wear. Interfacial wear is initiated only when the counterface is smooth enough to form interfacial junctions between the polymer and the counterface. An excellent example of interfacial wear with isothermal condition is that of polytetrafluoroethylene (PTFE) sliding against a metal surface. When PTFE is slid against a smooth metal surface, friction is high in the beginning but drops to a lower value after some sliding. Because of the presence of frictional stress and heat, the PTFE molecular chains are oriented in the direction of sliding and a transfer film is deposited onto the counterface. The molecular

orientation in PTFE is responsible for the drop in friction coefficient. Although the friction coefficient is low, for PTFE wear is generally high because of the thermal softening of the interface zone and easy removal of the material. This is one of the reasons why PTFE has not been used very widely for tribological applications. Figure 2 shows micrographs of oriented PTFE molecules deposited on the counterface after wear.

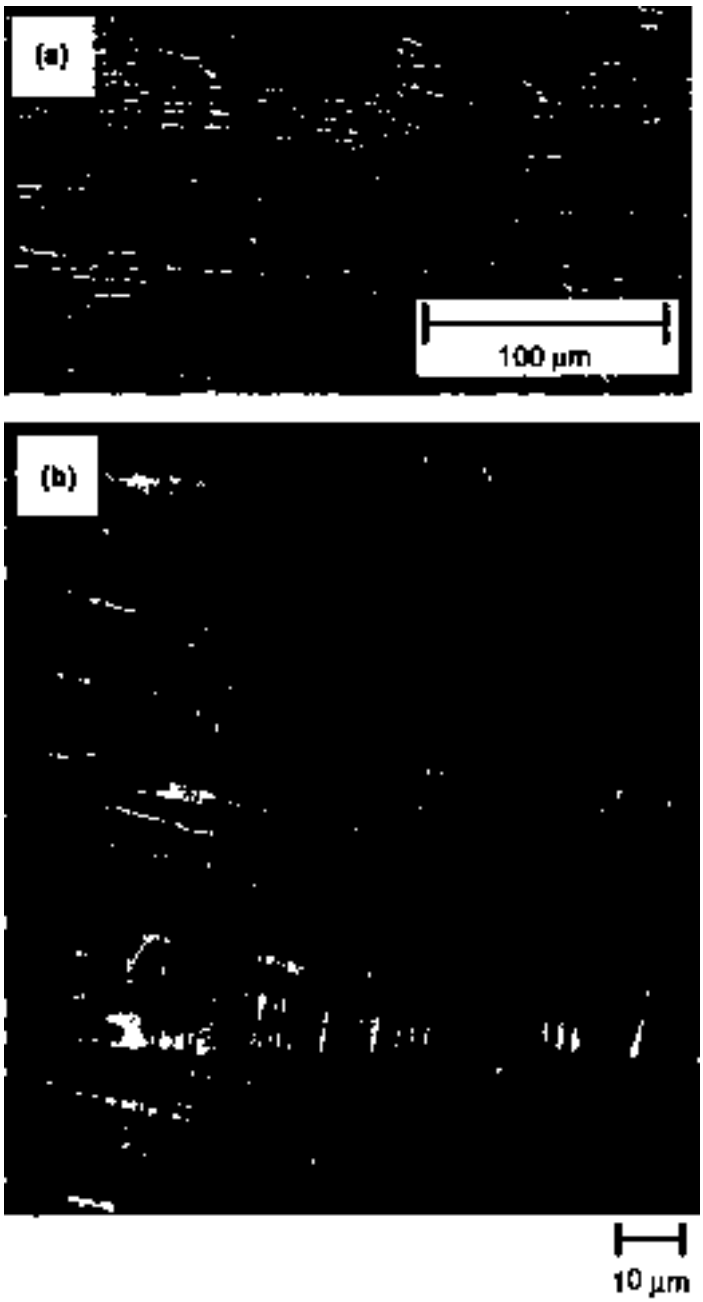


Fig. 2 Micrographs of oriented PTFE films on the counterface. (a) PTFE transfer film on a glass slide. The film thickness varies between 50 and 500 nm, and sometimes it can show a lumpy feature when the sliding test is carried out at high loads. The film is highly birefringent, indicating that the molecules are oriented parallel to the sliding direction. Reprinted with permission from Ref 3. (b) PTFE transfer film when a PTFE pin is slid over a metallic surface. PTFE covers the counterface, making fibers and layers over one another. The orientation of the fibers in the transfer film can easily change if the sliding direction is changed. Reprinted with permission from Ref 4

The quasi-adiabatic interfacial wear involves glassy thermoplastics (not cross linked) and cross-linked polymer systems such as elastomers and thermosets. These polymers show a range of wear behavior. For example, thermosets, which do not soften due to thermal energy, undergo chemical degradation at the interface. These degraded products detach themselves from the main body of the polymer and form transfer film and debris at the interface. The wear rate can be very high if the prevailing interface temperature is high. An important

application of thermosets in a tribological context is in brake pads where the base polymer is mixed with several additives for optimal friction, wear, and mechanical strength.

Although friction models are available for interfacial sliding, theoretical wear quantification is difficult. This is because wear depends on a number of parameters other than the mechanical and physical properties of the material. These parameters include temperature, sliding speed, normal pressure, counterface roughness, and the rheological properties of transfer film. The exact influence of each parameter on wear is rarely known. Few attempts have been made to obtain wear laws using empirical means. In one such example involving PTFE, Uchiyama and Tanaka (Ref 5) have rationalized the effects of temperature and normal pressure in relating linear wear (thickness removed per unit sliding distance) with sliding speed. According to their work, if linear wear, x (length per unit sliding distance), is assumed to be directly proportional to the sliding speed (v) at a constant temperature, T_o, and pressure, p_o, then linear wear can be expressed by:

$$x = \frac{k_0 (a_T v)(p / p_0)^n}{b_s} \qquad \text{(Eq 1)}$$

where n is a constant greater than unity and k_o is a proportionality constant. a_T and b_s are shift factors that depend on the temperature. a_T and b_s are obtained through experimentation by shifting the data on the speed axis and wear rate axis (on a wear rate/sliding speed plot), respectively, such that they coincide with similar data obtained at a temperature of 29 °C (84 °F). The authors claim that the relation can be applied to other polymer systems, too.

References cited in this section

1. B.J. Briscoe and S.K. Sinha, Wear of Polymers, *Proc. Inst. Mech. Eng. J., J. Eng. Tribol.,* Vol 216, 2002

2. B.J. Briscoe and D. Tabor, Friction and Wear of Polymers, Chapter 1, *Polymer Surfaces,* D.T. Clark and J. Feast, Ed., John Wiley & Sons, 1978

3. C.M. Pooley and D. Tabor, Friction and Molecular Structure: The Behaviour of Some Thermoplastics, *Proc. R. Soc. (London) A,* Vol 329, 1972, p 251–274

4. R.P. Steijn, The Sliding Surface of Polytetrafluoroethylene: An Investigation with the Electron Microscope, *Wear,* Vol 12, 1968, p 193–212

5. Y. Uchiyama and K. Tanaka, Wear Law for Polytetrafluoroethylene, *Wear,* Vol 58, 1980, p 223–235

Wear Failures of Plastics

Sujeet K. Sinha, National University of Singapore

Cohesive Wear

Cohesive wear is defined as subsurface or bulk wear when the interacting surfaces produce damage to the material far deeper into the material than only at the interface. This type of wear is also referred to in the literature as plowing or abrasive wear. Subsurface damage in material can be caused by surface sliding in two ways. First, if a polymer is sliding against a rough and hard surface, the asperities of the hard surface can plow into the bulk of the polymer removing debris. These debris materials generally get transferred to the counterface, forming a transfer film (also known as the “third body”), which eventually makes the counterface appear smoother. The formation of a stable film at the counterface leads to a change in the wear rate of the polymer. The second cause of subsurface damage is through subsurface fatigue cracks, which can lead to the

removal of material when these cracks grow to the surface of the polymer. Fatigue wear removes the material in chunks or flakes.

Considerable attention has been given by researchers to the creation of a model for cohesive or abrasive wear of polymers. The most notable model for wear involving bulk properties of the polymer was given by Ratner-Lancaster (Ref 6). The relation is given as:

$$V = \frac{K\mu W v}{HS\varepsilon} \qquad \text{(Eq 2)}$$

where V is the wear volume, K is a proportionality constant also termed wear rate, v is the sliding speed, μ is the coefficient of friction, H is the indentation hardness, S is the ultimate tensile strength, and ε is the elongation to break of the polymer. Some evidence of the usefulness of Ratner-Lancaster relation may be found in the work by Briscoe (Ref 7). In this work, the wear rate ($mm^3\ mm^{-1}\ kg^{-1}$) was plotted against the reciprocal of the product of S and ε, which furnished, as predicted by the Eq 2, a straight line. In contrast to Eq 2, Rhee has followed a different approach where wear is thought to be nonlinearly proportional to pressure, sliding velocity, and temperature (Ref 8). He proposed an empirical relation of the type:

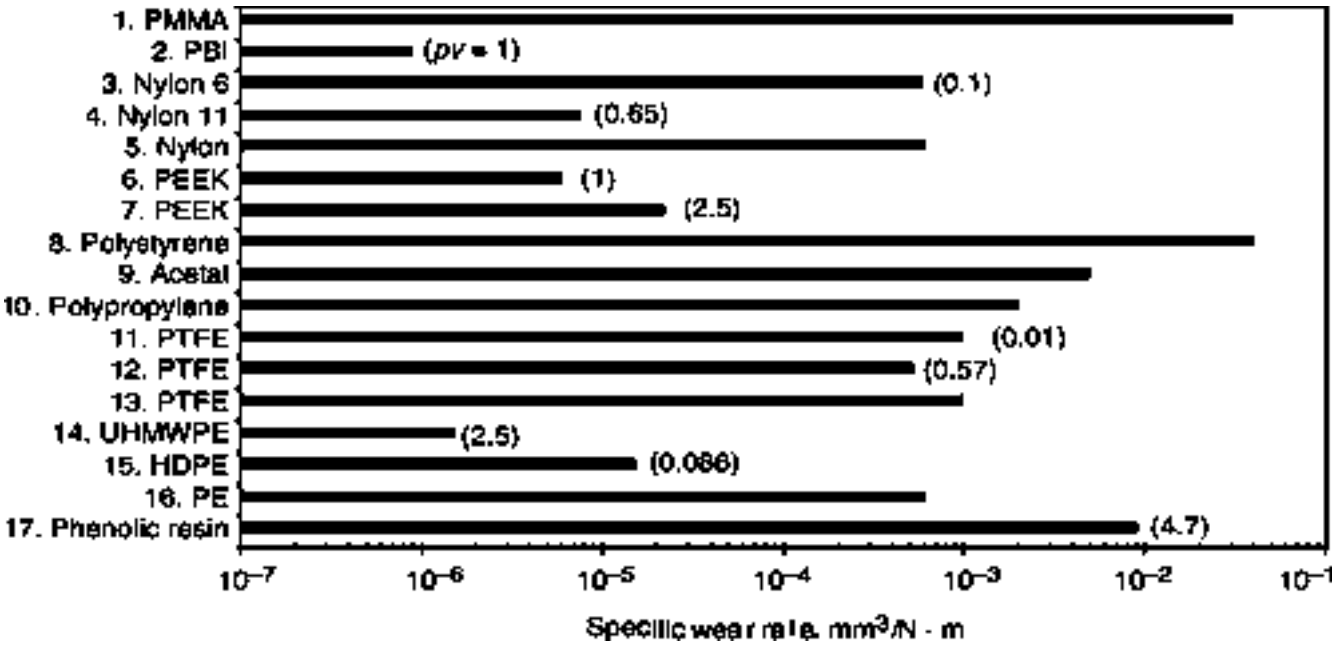
$$\Delta w = K p^a v^b T^c$$

where Δw is the weight loss of the polymer, and a, b, and c are material-dependent variables. Yet another wear model was proposed by Kar and Bahadur (Ref 9) in which they arrived at an empirical relation using the principles of dimensional analysis. Using data obtained for polyoxymethylene (POM) and PTFE-filled POM, Kar and Bahadur obtained a relation given as:

$$V = \frac{1.5\, K \gamma^{1.775} p^{1.47} Z^{1.25}}{E^{3.225}} \qquad \text{(Eq 3)}$$

where γ is the surface energy, Z is the sliding distance, and E the modulus of elasticity of the polymer. Another variation of Eq 3 may be found in the work by Viswanath and Bellow (Ref 10).

Figure 3 presents specific wear rate (wear volume per unit sliding distance per unit normal load) for a number of polymer systems under abrasive or nonabrasive sliding conditions (Ref 5, 11, 12, 13, 14, 15, 16, 17, 18). The data are shown for a variety of experimental conditions as reported in literature. Though the experimental conditions used in these tests were different, some trends may be noticed. Polybenzimidazole (PBI) and ultrahigh molecular weight polyethylene (UHMWPE) show, among all polymers, very high wear resistance. Extremely poor wear resistance is demonstrated by polymethyl methacrylate (PMMA), polystyrene (PS), and phenolic resin. Figure 4 shows worn surfaces of polyetheretherketone (PEEK) (Ref 19) and UHMWPE (Ref 20). These polymer surfaces show scars of wear by plowing and plastic deformation.



Specimen	Material	Counterface roughness (R_a),	Sliding speed (v),	$1/S\varepsilon^{(a)}$	Normal pressure (p)	Temperature

		µm	m/s		MPa	ksi	°C	°F	Ref	
1	PMMA	1.2	…	0.09	…	…	…	…		11
2	PBI	…	1	…	1	0.15	20	68	17	
3	Nylon 6	…	5×10^{-3}	…	20	2.9	…	…	15	
4	Nylon 11	0.11	1	…	0.65	0.09	…	…	13	
5	Nylon	1.2	…	0.1	…	…	…	…	11	
6	PEEK	…	1	…	1	0.15	20	68	17	
7	PEEK	0.05	0.5	…	5	0.73	…	…	12	
8	Polystyrene	1.2	…	5	…	…	…	…	11	
9	Acetal	1.2	…	0.5	…	…	…	…	11	
10	Polypropylene	1.2	…	0.1	…	…	…	…	11	
11	PTFE	…	0.2	…	0.05	0.007	…	…	16	
12	PTFE	…	0.1	…	5.66	0.82	29	84	5	
13	PTFE	1.2	…	0.2	…	…	…	…	11	
14	UHMWPE	0.05	0.5	…	5	0.73	…	…	12	
15	HDPE	0.9	0.03	…	2.8	0.41	…	…	14	
16	Polyethylene	1.2	…	0.09	…	…	…	…	11	
17	Phenolic resin	0.05	5.6	…	0.84	0.12	…	…	18	

PMMA, polymethyl methacrylate; PBI, polybenzimidazole; PEEK, polyetheretherketone; PTFE, polytetrafluoroethylene; UHMWPE, ultrahigh molecular weight polyethylene; HDPE, high-density polyethylene.
(a) *S*, tensile strength; ε, elongation to break

Fig. 3 Specific wear rate for a number of polymers as reported in the literature. The experimental conditions as reported in the literature are given in the table.

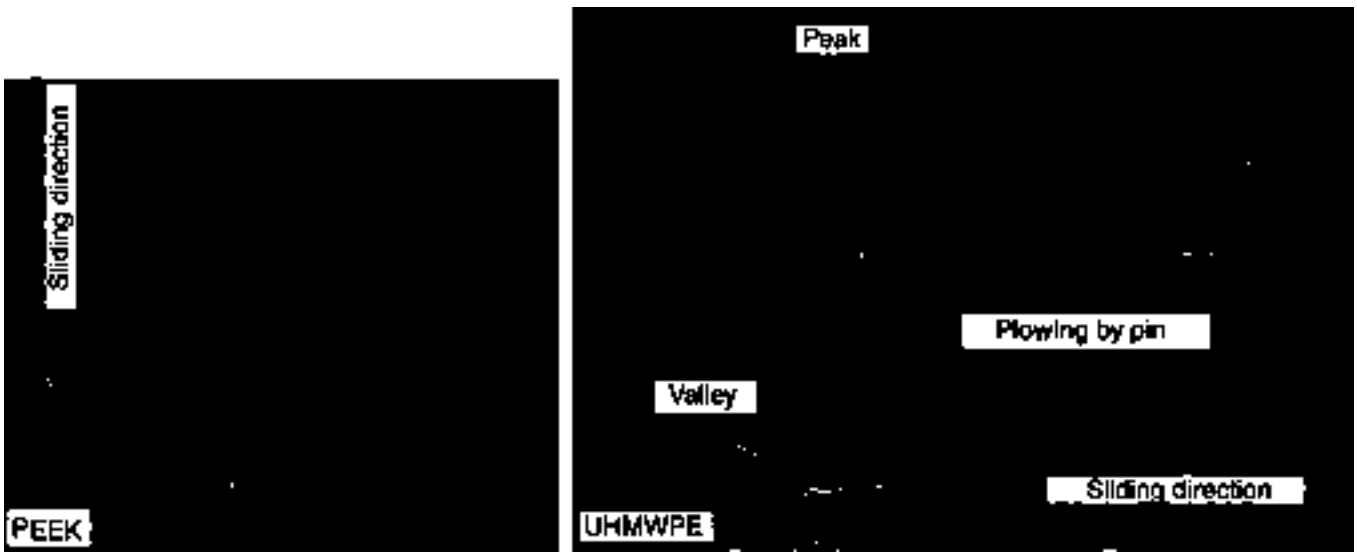


Fig. 4 Micrographs showing surfaces of worn polymers when they were slid against abrasive surfaces. PEEK (left) reprinted with permission from Ref 19. UHMWPE (right) (reprinted with permission from Ref 20) surfaces show scars of abrasive and plowing actions of hard counterfaces.

References cited in this section

5. Y. Uchiyama and K. Tanaka, Wear Law for Polytetrafluoroethylene, *Wear*, Vol 58, 1980, p 223–235

6. J.K. Lancaster, Friction and Wear, Chapter 14, *Polymer Science, A Materials Science Handbook*, A.D. Jenkins, Ed., North-Holland Publishing, 1972

7. B.J. Briscoe, Wear of Polymers: An Essay on Fundamental Aspects, *Tribol. Int.,* Aug 1981, p 231–243

8. S.K. Rhee, *Wear,* Vol 16, 1970, p 431

9. M.K. Kar and S. Bahadur, The Wear Equation for Unfilled and Filled Polyoxymethylene, *Wear,* Vol 30, 1974, p 337–348

10. N. Viswanath and D.G. Bellow, Development of an Equation for the Wear of Polymers, *Wear,* Vol 181–183, 1995, p 42–49

11. D.G. Evans and J.K. Lancaster, The Wear of Polymers, *Wear,* Vol 13, *Materials Science and Technology,* D. Scott, Ed., Academic Press, 1979, p 85–139

12. J.P. Davim and N. Marques, Evaluation of Tribological Behaviour of Polymeric Materials for Hip Protheses Applications, *Tribol. Lett.,* Vol 11 (No. 2), 2001, p 91–94

13. S. Bahadur and V.K. Polineni, Tribological Studies of Glass Fabric Reinforced Polyamide Composites Filled with CuO and PTFE, *Wear,* Vol 200, 1996, p 95–104

14. C.H. da Silva, D.K. Tanaka, and A. Sinatora, The Effect of Load and Relative Humidity on Friction Coefficient between High Density Polyethylene on Galvanized Steel—Preliminary Results, *Wear,* Vol 225–229, 1999, p 339–342

15. F. Van De Velde and P. De Baets, The Friction and Wear Behaviour of Polyamide 6 Sliding against Steel at Low Velocity under Very High Contact Pressure, *Wear,* Vol 209, 1997, p 106–114

16. S. Bahadur and D. Gong, The Action of Fillers in the Modification of the Tribological Behaviour of Polymers, *Wear,* Vol 158, 1992, p 41–59

17. K. Friedrich, Z. Lu, and A.M. Hager, Recent Advances in Polymer Composites' Tribology, *Wear,* Vol 190, 1995, p 139–144

18. T. Kato and A. Magario, The Wear of Aramid Fibre Reinforced Brake Pads: The Role of Aramid Fibres, *STLE Trib. Trans.,* Vol 37 (No. 3), 1994, p 559–565

19. J.V. Voort and S. Bahadur, The Growth and Bonding of Transfer Film and the Role of CuS and PTFE in the Tribological Behavior of PEEK, *Wear,* Vol 181–183, 1995, p 212–221

20. J. Song, P. Liu, M. Cremens, and P. Bonutti, Effects of Machining on Tribological Effects of Ultra High Molecular Weight Polyethylene (UHMWPE) under Dry Reciprocating Sliding, *Wear,* Vol 225–229, 1999, p 716–723

Sujeet K. Sinha, National University of Singapore

Elastomers

The study of wear of elastomers has evolved primarily from the interest in the friction and wear of automobile tires and industrial seals. Schallamach carried out extensive studies on relatively softer rubbers such as polyisoprene, butyl rubber, and natural rubber (Ref 21, 22). Through extensive experimentation on the sliding of rubber against hard surfaces, he found that the process of sliding for rubber takes place through series of

detachments at the contact points, giving it the look of a wave (Fig. 5). These waves initiate at the front edge of the slider due to excessive buckling of rubber in the front and runs to the rear of the slider. When a slider in contact with elastomer is pushed forward, the adhesive force (between the slider and the elastomer) generates compressive tensile stress at the front edge leading to buckling and folding of the elastomer in the form of a wave. The detached part further relaxes the material, thus facilitating the movement of the slider. Moore, in a later study of the wear of rubbers and tires (Ref 23), concluded that for elastomeric materials there are two ways in which the frictional energy is dissipated, leading to wear. The flow chart he produced is redrawn in Fig. 6. In order to model abrasive action of asperities on elastomers, several tests using sharp needles have also been carried out in the past. The process of wear by a sharp needle or an asperity is schematically shown in Fig. 7. Though there are a number of models available that quantify the frictional work done during sliding on rubber, a wear model for elastomers is still unavailable presently except for the abrasive case where the Ratner-Lancaster relation can be applied.

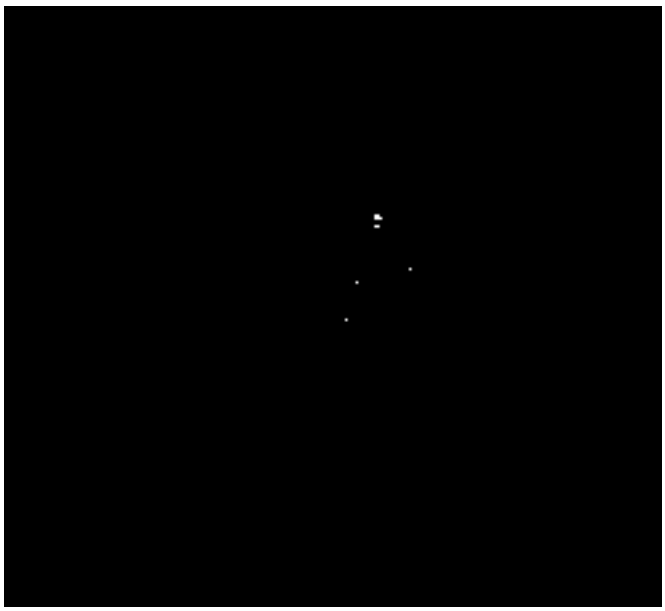


Fig. 5 Waves of detachment when an elastomer is slid against a hard and smooth surface. The rubber moves forward in the form of ripples of wave on its contact surface with a smooth and hard counterface. These so-called waves of detachment can produce wear in the form of rolls of detached material or the third body. Reprinted with permission from Ref 22

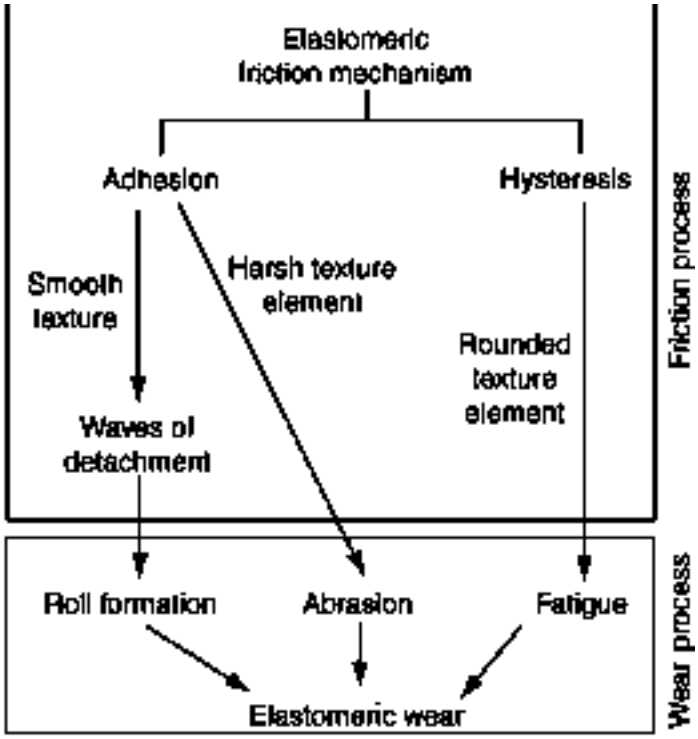


Fig. 6 Classification of the processes of friction leading to wear for elastomers (adapted after Moore, Ref 23). The diagram clarifies the role of friction in determining the wear mechanism for elastomeric polymers.

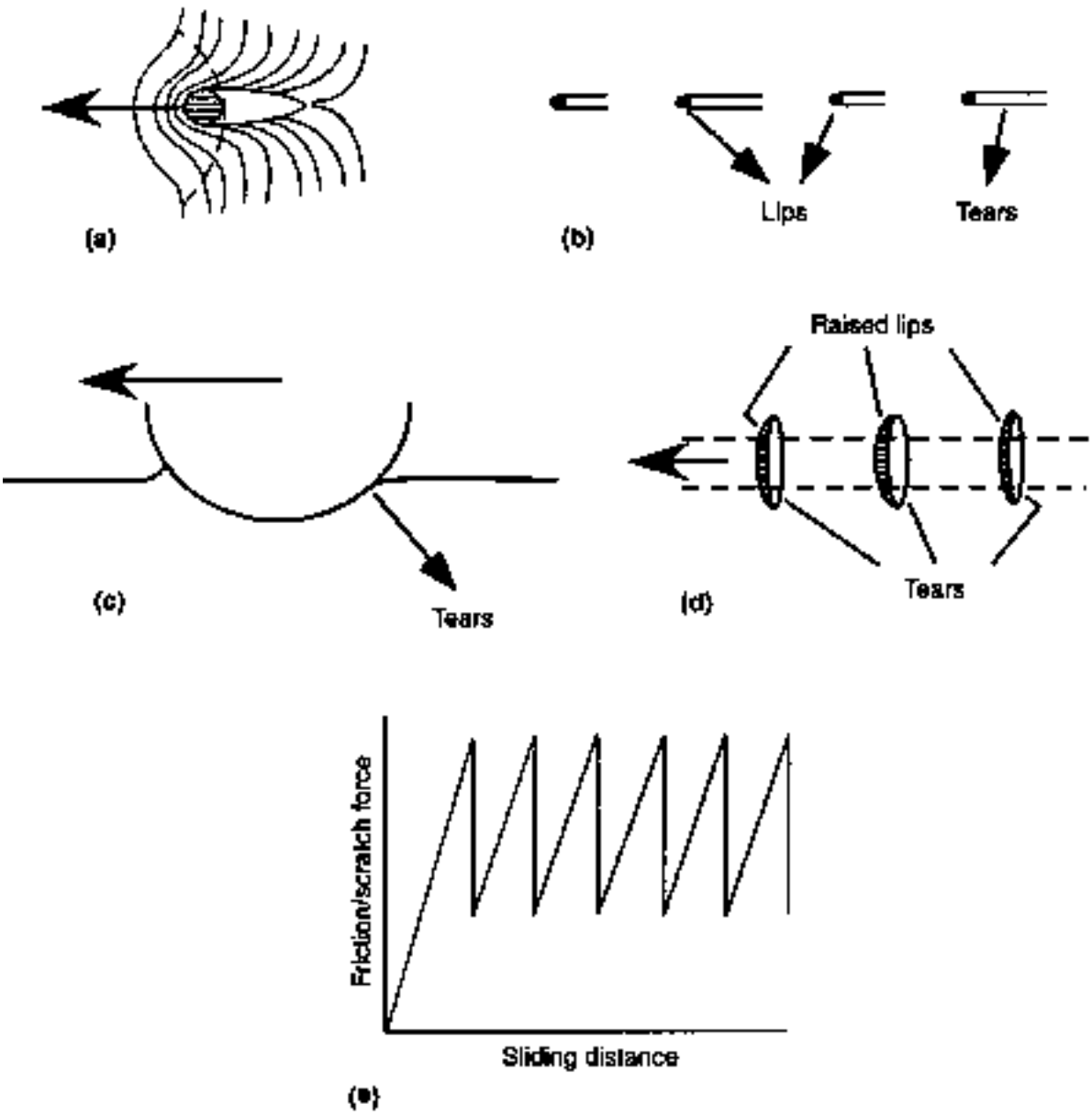


Fig. 7 Damage created on the surface of an elastomer by isolated stress concentration. (a) Surface deformation pattern when a sharp needle or conical indentor with acute angle is slid on the surface of an elastomer. The elastomer surface is pulled in the direction of motion and fails in tension behind the contact at $\pi/2$ to the tensile field. (b) After the needle jumps forward, the surface relaxes and tensile tears are evident on the surface but are now in the direction of motion. (c) Tearing of an elastomer due to tractive stress with a large unlubricated indentor. The tear is generated at the rear of the contact region and is almost at right angles to the sliding motion. (d) A raised lip of elastomer is formed, but no material is actually removed. (e) A typical friction/scratching force profile when a slider is passed over an elastomer. Reprinted with permission from Ref 1

References cited in this section

1. B.J. Briscoe and S.K. Sinha, Wear of Polymers, *Proc. Inst. Mech. Eng. J., J. Eng. Tribol.,* Vol 216, 2002

21. A. Schallamach, Abrasion of Rubber by a Needle, *J. Polymer Sci.,* Vol 9 (No. 5), 1952, p 385–404

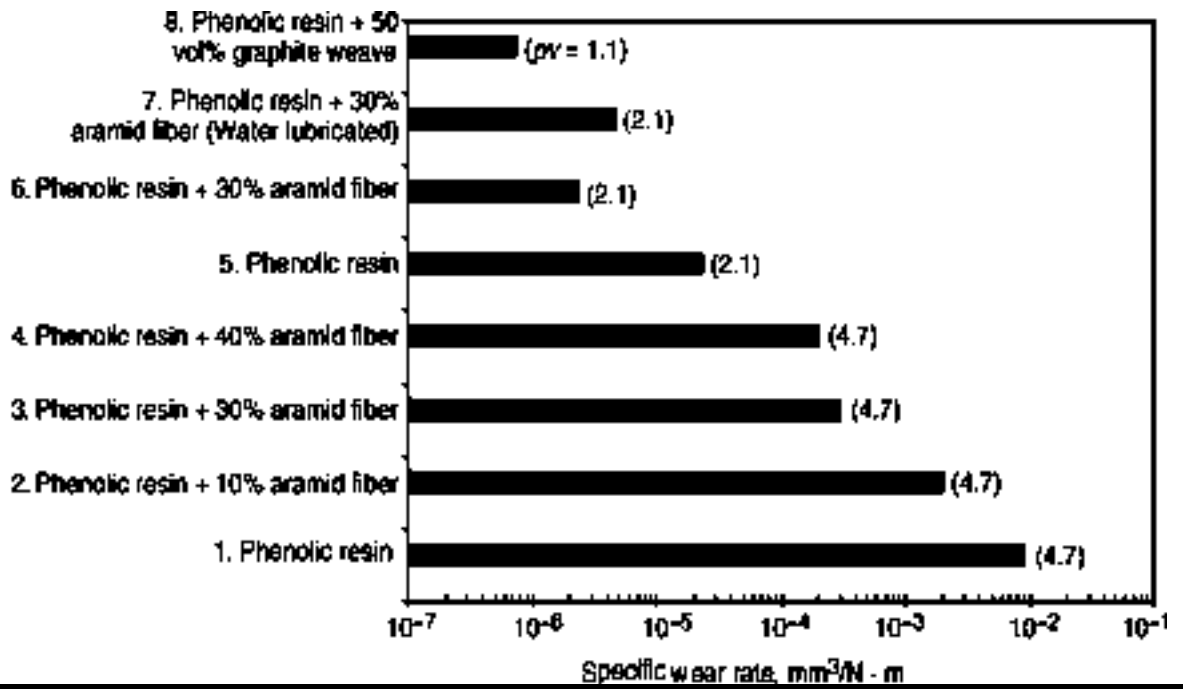
22. A. Schallamach, How Does Rubber Slide? *Wear,* Vol 17, 1971, p 301–312

23. D.F. Moore, Friction and Wear in Rubbers and Tyres, *Wear,* Vol 61, 1980, p 273–282

Thermosets

Thermosets have found applications mainly in automobile brakes, gears, cams, and clutch parts where they are subjected to sliding. Brake pads are one area where thermosets such as phenolic and epoxy resins have been used and studied extensively. These polymers do not soften when the temperature rises at the interface, and thus they prevent the component from yielding or failing in a catastrophic manner during service. However, thermal energy dissipated due to frictional work can induce chemical degradation and wear at the sliding surface. Thermosets have generally been filled with fibers and particles as additives in order to increase the strength and wear resistance of the material (Ref 24, 25, 26, 27). Fillers include glass fiber, aramid fiber, and metal oxide particles of various kinds.

The role of aramid fibers, in the context of brake pads, caught special attention from tribologists when there was an effort to replace asbestos used in brake pads with aramid fibers (Ref 18, 28, 29, 30). Figure 8 compares the specific wear rate of a few formulations of thermoset composites. The relevant micrograph is given in Fig. 9. It is seen from these results that the wear resistance of phenolic resin increases by almost two orders of magnitude when fillers such as carbon and aramid fibers are added to the phenolic resin matrix. The micrograph (Fig. 9) shows that a transfer layer is formed on the polymer surface in addition to the transfer layer found on the counterface. These strong and highly adhesive transfer layers help improve the wear resistance of the polymer composite.



Specimen	Sliding speed (v), m/s	Normal pressure (p)		Counterface roughness (R_a), µm	Ref
		MPa	ksi		
1–4	5.6	0.84	0.12	0.5	18
5–7	0.5	4.25	0.62	0.05–0.1	29
8[(a)]	1.6	0.69	0.10	0.05	27

(a) N_2 atmosphere at room temperature

Fig. 8 Specific wear rates for phenolic resin and its composites. The data are reported for various experimental conditions and *pv* (pressure × velocity) factors as reported in the literature.

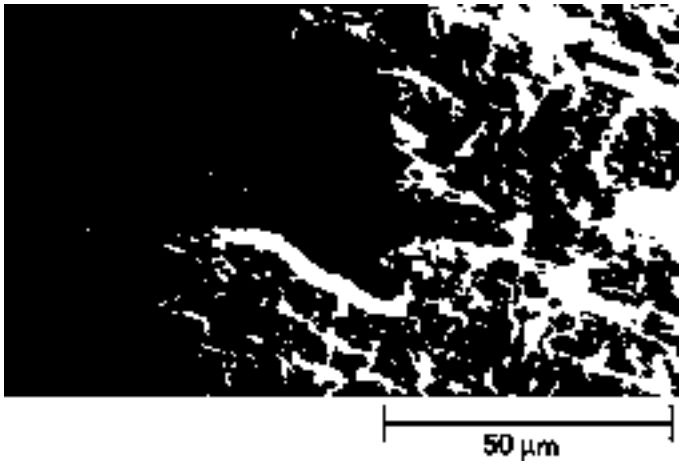


Fig. 9 Micrograph of the worn surface for a phenolic resin-aramid fiber composite (Ref 29) showing partial coverage of the polymer pin by transfer film

References cited in this section

18. T. Kato and A. Magario, The Wear of Aramid Fibre Reinforced Brake Pads: The Role of Aramid Fibres, *STLE Trib. Trans.*, Vol 37 (No. 3), 1994, p 559–565

24. T. Liu and S.K. Rhee, High Temperature Wear of Semi-Metallic Disk Brake Pads, *Wear of Materials 1977*, American Society of Mechanical Engineers, 1977, p 552–554

25. S.K. Rhee and P.A. Thesier, "Effects of Surface Roughness of Brake Drums on Coefficient of Friction and Lining Wear," paper 720449, Society of Automotive Engineers, 1971

26. J.K. Lancaster and J.P. Giltrow, The Role of the Counterface Roughness in the Friction and Wear of Carbon Fibre Reinforced Thermosetting Resins, *Wear*, Vol 16, 1970, p 357–374

27. B. Bhushan and D.F. Wilcock, Wear Behaviour of Polymeric Compositions in Dry Reciprocating Sliding, *Wear*, Vol 75, 1982, p 41–70

28. B.J. Briscoe, I. Ramirez, and P.J. Tweedle, *Proc. Int. Conf. Disc Brakes for Commercial Vehicles* (London), 1–2 Nov 1988, Mechanical Engineering Publications Ltd., 1988, p 5

29. S.K. Sinha and S.K. Biswas, Friction and Wear Behaviour of Continuous Fibre as Cast Kevlar-Phenolic Resin Composite, *J. Mater. Sci.*, Vol 27, 1992, p 3085–3091

30. S.K. Sinha and S.K. Biswas, Effect of Sliding Speed on Friction and Wear of Uni-Directional Aramid Fibre-Phenolic Resin Composite, *J. Mater. Sci.*, Vol 30, 1995, p 2430–2437

Sujeet K. Sinha, National University of Singapore

Glassy Thermoplastics

Traditionally, glassy thermoplastics have not been used as typical tribological materials. This is because they show mechanical instability at the glass-transition temperature. However, they are often subjected to sliding, scratching, or abrasion in various working environments. For example, a window pan or automoble body part made of glassy polymer may be subjected to water, dust, and occasional scratching, or a bathtub may have

water plasticization coupled with sliding and compression. Some glassy thermoplastics filled with fibers or particulate fillers have been used for tribological applications. The problem encountered with such polymers is their tendency to fail in a catastrophic manner when the glass-transition temperature is reached. Examples of this class of polymer are PMMA, PS, and polycarbonate (PC). Cross-linked polymer PEEK also behaves in a way similar to glassy polymers (Ref 31). See Fig. 10 for the changes in deformation behavior in sliding of PEEK when the operating temperature is close to the glass-transition temperature for PEEK.

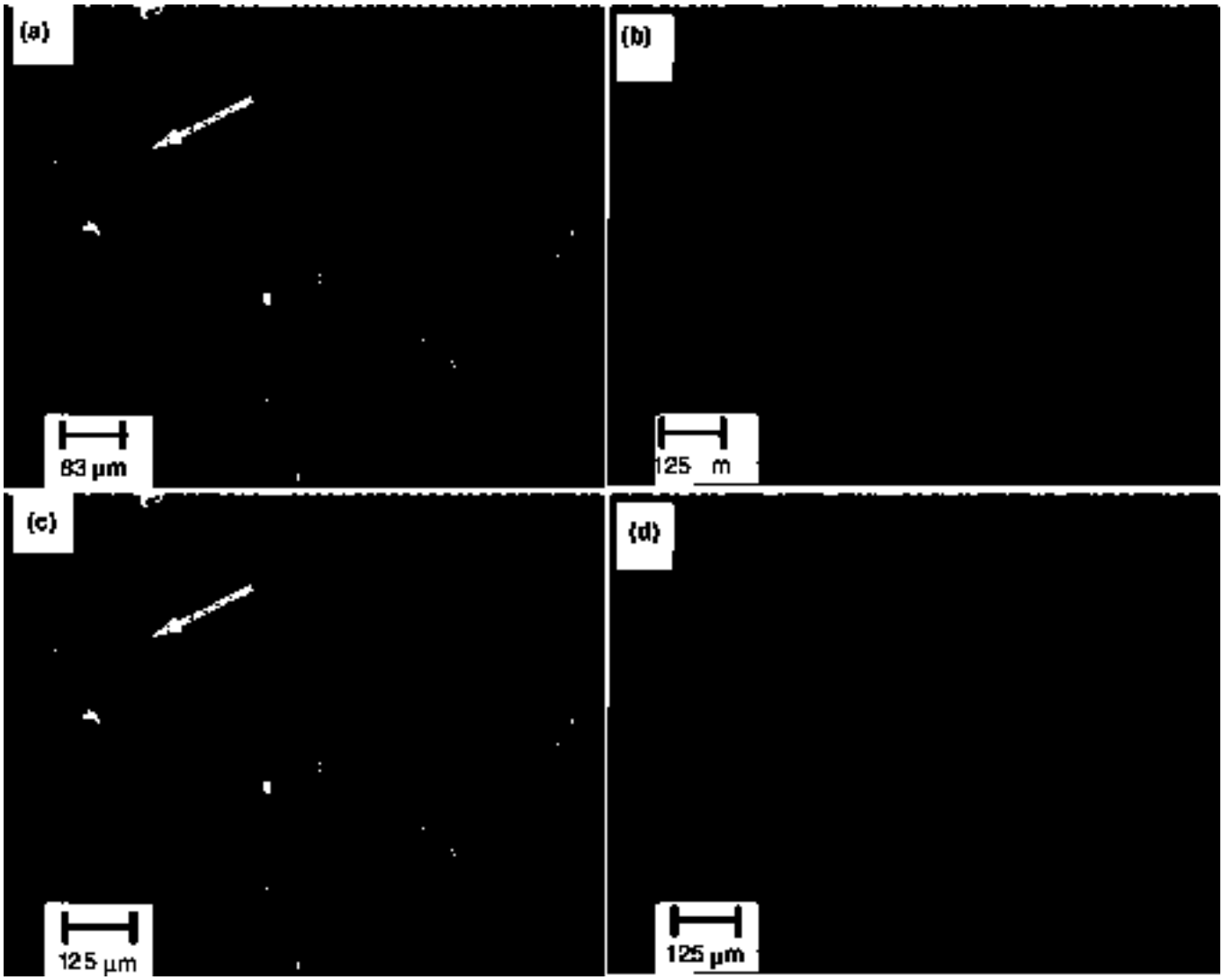


Fig. 10 Micrographs of worn PEEK surfaces at various operating temperatures. These pictures highlight the changes in the surface deformation behavior of the polymer with temperature. (a) 90 °C (194 °F). (b) 152 °C (306 °F). (c) 180 °C (356 °F). (d) 225 °C (437 °F). Arrows indicate the sliding direction. Glass-transition temperature for PEEK used in the experiment was 148 °C (300 °F). Reprinted with permission from Ref 29

The study of glassy thermoplastic surfaces has mainly focused on understanding the damage processes under a variety of experimental and ambient conditions (Ref 32, 33). For example, damage modes can be studied using the concept of wear maps. Figure 11 gives such a map of PMMA for different normal load and imposed strain conditions. A range of studies have been carried out by Briscoe, Chateauminois, and coworkers (Ref 34, 35) in order to understand the role of the third body in fretting wear of PMMA. Their study with PMMA concluded that the formation of the third body and the wear rate depend on the kinematics of sliding. In linear sliding, as opposed to torsional sliding, the wear rate is low. The worn area showed debris material in rolled and compacted forms. The authors concluded that the energy dissipation in the linear sliding case occurred mainly by the rolling and shearing actions on the rolled debris that reduced the frictional work required for sliding. Therefore, wear in the linear sliding case was low.

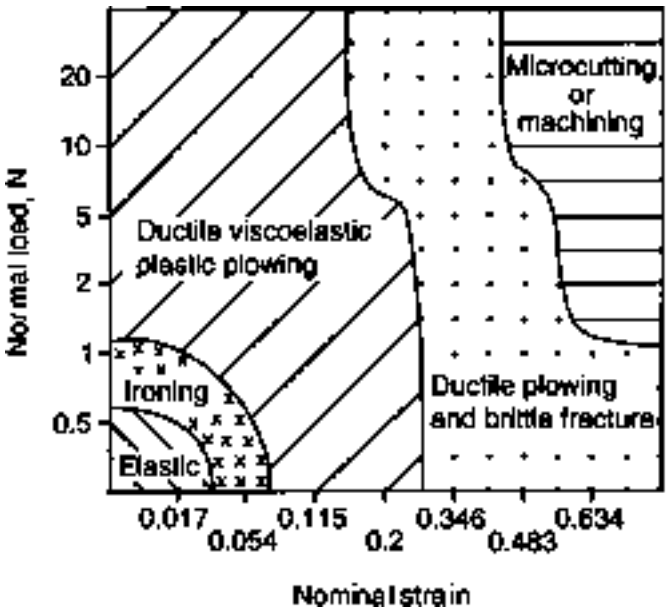


Fig. 11 Scratching damage maps for PMMA. Scratching velocity = 0.004 mm/s and nominal strain is defined as 0.2 × tan θ; 2θ being the included angle of the indenter.

References cited in this section

29. S.K. Sinha and S.K. Biswas, Friction and Wear Behaviour of Continuous Fibre as Cast Kevlar-Phenolic Resin Composite, *J. Mater. Sci.,* Vol 27, 1992, p 3085–3091

31. J. Hanchi and N.S. Eiss, Jr., Tribological Behaviour of Polyetheretherketone, a Thermotropic Liquid Crystalline Polymer and in situ Composites Based on Their Blends under Dry Sliding Conditions at Elevated Temperatures, *Wear,* Vol 200, 1996, p 105–121

32. B.J. Briscoe, P.D. Evans, E. Pelillo, and S.K. Sinha, Scratching Maps for Polymers, *Wear,* Vol 200, 1996, p137–147

33. B.J. Briscoe, E. Pelillo, and S.K. Sinha, Characterisation of the Scratch Deformation Mechanisms for Poly(methylmethacrylate) Using Surface Optical Reflectivity, *Polymer Int.,* Vol 43, 1997, p 359–367

34. B.J. Briscoe, A. Chateauminois, T.C. Lindley, and D. Parsonage, Fretting Wear Behaviour of Polymethylmethacrylate under Linear Motions and Torsional Contact Conditions, *Tribol. Int.,* Vol 31 (No. 11), 1998, p 701–711

35. B.J. Briscoe, A. Chateauminois, T.C. Lindley, and D. Parsonage, Contact Damage of Poly(methylmethacrylate) During Complex Microdisplacements, *Wear,* Vol 240, 2000, p 27–39

Sujeet K. Sinha, National University of Singapore

Semicrystalline Thermoplastics

The most versatile use of polymers in tribological application has been for the semicrystalline group of polymers. Semicrystalline thermoplastics include PTFE, polyethylene (PE), UHMWPE, and nylon. These polymers, in homogeneous or heterogeneous forms, have found applications in gears, bearings, automobile piston seals, knee/hip joint replacement, and so forth. Semicrystalline thermoplastics do soften in the presence

of thermal energy; however, the way thermal energy is transmitted from the interface to the bulk depends on the thermal properties of the individual polymer. Based on this behavior, the mode of wear for semicrystalline polymers can be divided into two groups: adiabatic and isothermal. Furthermore, the isothermal type, a common case, is subdivided into three categories based on the way polymer transfer film is deposited onto the hard counterface. Figure 12 delineates these groups of wear processes for semicrystalline thermoplastics in isothermal heat transfer conditions.

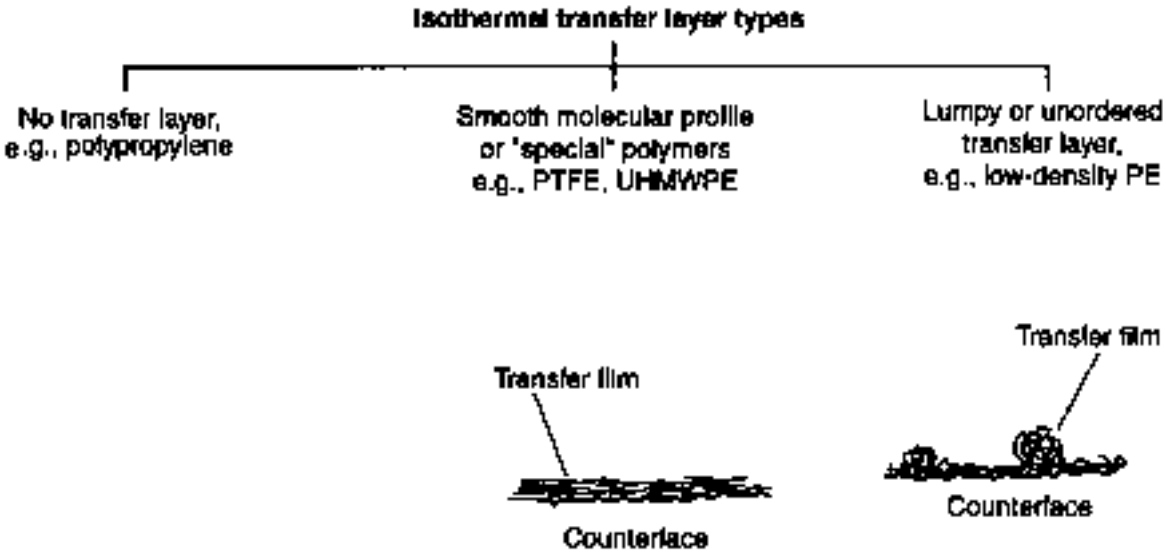


Fig. 12 Generic types of transfer wear behavior when semicrystalline polymers are slid on a hard smooth surface. In most of the cases there is a formation of transfer layer on the counterface though the shear, and adhesive properties of the transfer films will vary depending upon the mechanical properties of the polymer and the surface topography of the counterface. Reprinted with permission from Ref 1

Early studies on the friction and wear of thermoplastics was motivated by the prospect of finding an ultralow friction polymer material (Ref 3, 36). Polytetrafluoroethylene provided very low friction coefficient (~0.06), though the corresponding wear rate was high. The reason for low friction was found to be highly oriented PTFE molecules that were transferred to the counterface during sliding (Ref 3). The interface of the polymer also showed highly oriented molecules that extended out of the samples showing fibers. In order to reduce the wear rate and utilize the excellent low friction property of PTFE, this polymer has often been used with fillers to form composites. Polytetrafluoroethylene itself has also been used as filler for other polymeric systems such as PE. Figure 13 gives the wear rate of PTFE, and some of its composites when slid against hard metallic surfaces. For surface-treated PTFE (such as γ-irradiation), the situation may be different. Evidence shows that for such a system there may be an increase in the crystallinity of the polymer at the surface and consequently a decrease in the wear rate (Ref 37).

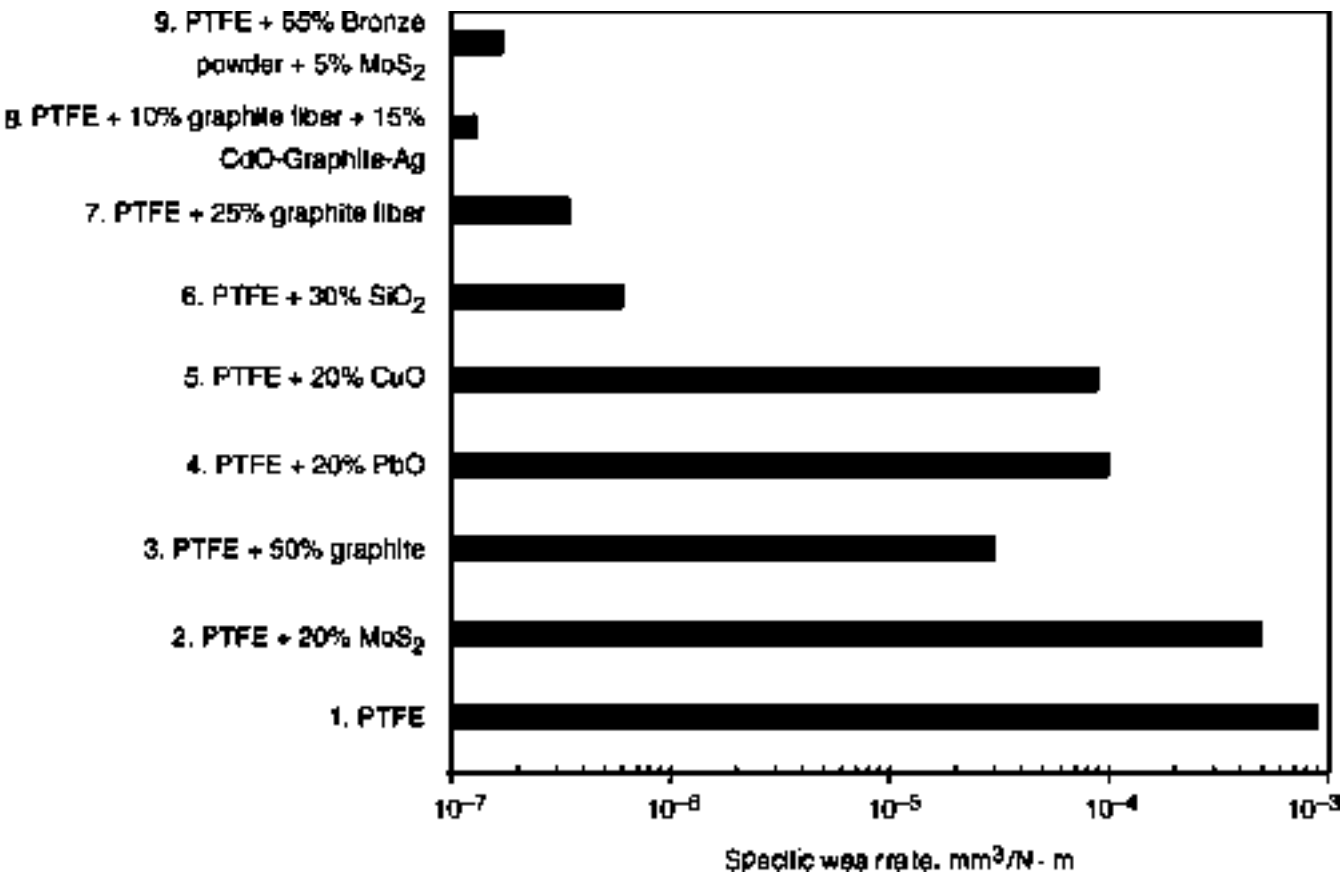


Fig. 13 Wear rate of PTFE and its composites under different experimental conditions. For specimens 1 to 4: sliding speed (*v*) = 0.2 m/s; normal pressure (*p*) = 0.05 MPa (0.007 ksi). Source: Ref 16. For specimens 7 to 9: sliding speed (*v*) = 1.6 m/s; normal pressure (*p*) = 0.69 MPa (0.10 ksi); counterface roughness (R_a) = 0.025 µm. Source: Ref 27

The wear process for semicrystalline thermoplastic polymer may seem to depend very much on the transfer film and its rheological properties, though evidence is also available where it shows that the loading condition can also change the wear mechanism. For UHMWPE, Wang et al. (Ref 38) found that the microscopic surface wear depends on the tensile and elongation properties of the polymer. However, under intense and nonconformal loading conditions the wear mechanism could change to macroscopic subsurface wear due to fatigue. Thus, the wear mechanism can change if the loading condition is changed. Wang et al. provided a model for the wear of semicrystalline thermoplastics that resembles the Ratner-Lancaster model for abrasive wear of polymers: for microscopic surface wear

$$V \propto \frac{L^{3/2} R_a^{3/2}}{S^{3/2} \varepsilon}$$

for macroscopic sub-surface wear,

$$V \propto \frac{1}{N} = (\Delta\varepsilon_p / \varepsilon)^{1/\alpha}$$

where V is the wear volume, L is the normal load, R_a is the counterface roughness, S is the ultimate tensile strength of the polymer, ε is the elongation at break, N is the cyclic fatigue life of the polymer, $\Delta\varepsilon_p$ is the inelastic strain amplitude, and α is a material constant obtained from low-cycle fatigue test using the Coffin-Manson equation (Ref 39).

References cited in this section

1. B.J. Briscoe and S.K. Sinha, Wear of Polymers, *Proc. Inst. Mech. Eng. J., J. Eng. Tribol.,* Vol 216, 2002

3. C.M. Pooley and D. Tabor, Friction and Molecular Structure: The Behaviour of Some Thermoplastics, *Proc. R. Soc. (London) A,* Vol 329, 1972, p 251–274

16. S. Bahadur and D. Gong, The Action of Fillers in the Modification of the Tribological Behaviour of Polymers, *Wear,* Vol 158, 1992, p 41–59

27. B. Bhushan and D.F. Wilcock, Wear Behaviour of Polymeric Compositions in Dry Reciprocating Sliding, *Wear,* Vol 75, 1982, p 41–70

36. K. Tanaka and T. Miyata, Studies on the Friction and Transfer of Semi-Crystalline Polymers, *Wear,* Vol 41, 1977, p 383–398

37. B.J. Briscoe, P.D. Evans, and J.K. Lancaster, Single Point Deformation and Abrasion of γ-Irradiated Poly(tetrafluoroethylene), *J. Phys. D, Appl. Phys.,* Vol 20, 1987, p 346–353

38. A. Wang, D.C. Sun, C. Stark, and J.H. Dumbleton, Wear Mechanism of UHMWPE in Total Joint Replacement, *Wear,* Vol 181–183, 1995, p 241–249

39. R.W. Hertzberg and J.A. Manson, *Fatigue Engineering Plastics,* Academic Press, 1980, p 54–61

Sujeet K. Sinha, National University of Singapore

Environmental and Lubricant Effects on the Wear Failures of Polymers

Except for elastomers, polymers in general are not used in lubricated conditions. However, polymers are often subjected to environmental conditions that affect their friction and wear performances. For example, polymers used in marine applications get exposed to sea water, and a machine component such as gear or brake pad may come in direct contact with leaking oil or water. For elastomers, their applications in seal rings and automobile tires regularly expose the material to lubricants, chemicals, and water. For industrial seals, the presence of lubricant protects it from dry contact with metal parts and the consequent severe wear. This kind of wear not only lowers the life of the seal, but also affects the metal part. It has been observed that soft elastomer can wear the metal part it comes in contact with (Ref 40). In an effort to increase the life of seals, a number of studies have been carried out to estimate the film thickness of the lubricant for elastomer pressed against a metal (Ref 41, 42, 43). The other example of the use of polymers in a lubricating environment is that of the knee/hip joint replacement using UHMWPE (Ref 44, 45). UHMWPE is widely used in making acetabular sockets for hip joints that normally slide against a ceramic ball. The presence of synovial body fluid ensures low friction by lubricating the surfaces. This fluid does not seem to chemically affect the polymer, though it does affect the way transfer film is formed at the counterface. The main problems in the application of UHMWPE for knee/hip joint replacement are the production of wear particles, which tend to become points of bacterial infection growth for the patient, and the wear of the metallic or ceramic counterface leading to increased wear of the polymer.

Environmental fluids and humidity have been found to affect many polymers in two ways. The first is the change in the adhesive and flow properties of the transfer film, and the second effect is that of changing the mechanical properties of the bulk of the polymer due to plasticization. In the presence of a liquid the adhesion of the transfer film is normally decreased, leading to high wear of the polymer (see Fig. 8 for wear data on water-lubricated sliding case). This is because the deposited polymer on the counterface is constantly removed during sliding, requiring further wear of the bulk of the polymer. The effect of liquid on the mechanical properties of the bulk polymer largely depends on the polarities of the polymer and the liquid, as well as on the surface tension of the liquid (thus the surface energy of the polymer) (Ref 46). Many polymers plasticize in the presence of water and some chemical liquids because liquid molecules can easily migrate into the bulk of the

polymer. Plasticization of a polymer drastically reduces its mechanical strength and hardness, which gives rise to a substantial reduction in the wear resistance. Briscoe et al. (Ref 47) studied the effect of lubricants on PE. They found that when oleamide and stearamide are applied to the surface of PE, the lubricants interact with polymer molecules and form a chemically bonded monolayer on the outer surface of the polymer. This can drastically reduce the coefficient of friction when the polymer slides against a hard surface.

References cited in this section

40. A.N. Gent and C.T.R. Pulford, Wear of Steel by Rubber, *Wear*, Vol 49, 1978, p 135–139

41. D. Dowson and P.D. Swales, An Elastohydrodynamic Approach to the Problem of the Reciprocating Seal, paper F3, *Proc. Third Int. Conf. Fluid Sealing*, British Hydromechanics Research Association

42. G.J. Field and B.S. Nau, Theoretical Study of EHL of Reciprocating Rubber Seals, *ASLE Trans.*, Vol 18, 1974, p 48–54

43. S.C. Richards and A.D. Roberts, Boundary Lubrication of Rubber by Aqueous Surfactant, *J. Phys. D, Appl. Phys.*, Vol 25, 1992, p A76–A80

44. A. Wang, A. Essner, V.K. Polineni, D.C. Sun, C. Stark, and J.H. Dumbleton, Lubrication and Wear of Ultra-High Molecular Weight Polyethylene in Total Joint Replacements, *New Directions in Tribology*, I. Hutchings, Ed., Mechanical Engineering Publications Ltd., London, 1997, p 443–458

45. A. Unsworth, The Effects of Lubrication in Hip Joint Prostheses, *Phys. Med. Biol.*, Vol 23, 1978, p 253–268

46. J.M. Senior and G.H. West, Interaction between Lubricants and Plastic Bearing Surfaces, *Wear*, Vol 18, 1971, p 311–323

47. B.J. Briscoe, V. Mustafaev, and D. Tabor, Lubrication of Polythlene by Oleamide and Stearamide, *Wear*, Vol 19, 1972, p 399–414

Wear Failures of Plastics

Sujeet K. Sinha, National University of Singapore

Summary and Case Study

Wear of polymers is an important aspect of their failure analysis and lifetime prediction. Wear failure of polymers is controlled by a number of factors, which include mechanical properties of polymers, such as ultimate tensile strength, elongation to break and hardness, sliding speed, normal load, coefficient of friction, counterface roughness, rheology and adhesive property of the transfer film, and thermal properties of polymers. The adhesive strength of the transfer layer to the counterface has strong influence upon the wear rate. Strong adherent transfer film normally gives low wear rate. Abrasive action of the asperities, adhesive force, thermal softening, chemical degradation and subsurface fatigue are some of the factors that initiate material removal during the process of polymer wear. Effect of lubricants depends upon how lubricant molecules attach themselves to the polymer molecules making bonds between the two molecular entities. Many polymers, in the presence of water or lubricant molecules, plasticize, which reduces friction, but wear can be high because of the decrease in the mechanical strength of the polymer due to plasticization. Lubricants in general reduce the adhesion of the transfer layer to the counterface leading to easy removal of the transfer layer and a high wear situation for the polymer.

A Case Study: Nylon as a Tribological Material. First synthesized in 1935 by Carothers (Ref 48), nylon is among few very important semicrystalline industrial thermoplastics. Nylon is the commercial name for those aliphatic polyamides that are made exclusively from ω-amino acids (Ref 49). There are several forms of nylon, generally denoted by nylon-*n* or nylon-*m,n*, where *m* and *n* stand for the number of main chain carbon atoms in constituent monomer(s). Among all varieties of nylons, nylon 6 and nylon 6/6 are the most widely produced and used materials because of their excellent mechanical properties and low cost. Nylon 11 and nylon 12, which show better performance in terms of low moisture absorption when compared to other nylons, are also used extensively; however, they are expensive.

Historically, nylons have been very popular materials for many tribological applications such as sliding fittings, bearings, and gears. Possibly the greatest advantage of using nylon as tribological material over metals is that no external lubricant is needed and the vibration noise is far less for nylon than for metals. Nylon parts can be extrusion molded with superior strength properties and low overall production cost.

Nylon sliding against nylon is a poor tribological pair due to high friction and high thermal effects (Ref 50). Even pure nylon sliding against metal surfaces does not perform well. However, nylon is an excellent low-friction and wear-resistant material if used in the form of plastic composite sliding against metal surfaces. This can be observed from the few studies that are available in the literature on nylon. Table 1 provides friction and wear results on a few types of nylon and its composites.

Table 1 Friction and Wear for Nylons

Nylon type	**Friction coefficient**	**Specific wear rate, $\times 10^{-6}$ $mm^3/N \cdot m$**	**Test conditions**	**Ref**
Nylon 11	0.31	7.48	Normal pressure = 0.65 MPa; sliding speed = 1 m/s; quench-hardened AISI steel counterface (R_a = 0.11 μm)	13, 51
Nylon 11 + 35% CuS	0.42	1.8	Normal pressure = 0.65 MPa; sliding speed = 1 m/s; quench-hardened AISI steel counterface (R_a = 0.11 μm)	13, 51
Nylon 11 + 5.6% glass fiber	0.38–0.5	2.97	Normal pressure = 0.65 MPa; sliding speed = 1 m/s; quench-hardened AISI steel counterface (R_a = 0.11 μm)	13
Nylon 11 + 20.7% glass fiber	0.38–0.5	1.66	Normal pressure = 0.65 MPa; sliding speed = 1 m/s; quench-hardened AISI steel counterface (R_a = 0.11 μm)	13
Nylon 6/6	0.62	…	Normal load = 200 N; sliding against nylon 6/6	50
Nylon 6/6 + 30% glass fiber	0.1–0.3	…	Normal load = 200 N; sliding against nylon 6/6	50
Nylon 6/6 + 30% glass fiber + 15% PTFE	0.05–0.1	…	Normal load = 200 N; sliding against nylon 6/6	…
Nylon 6	0.3	589	Normal load = 825 kN (pressure = 20 MPa); sliding velocity = 5 mm/s; steel counterface ($R_a \approx 5$ μm), extremely high pressure	15

Similar to the case of many other plastics, the tribological performance of nylon greatly depends on its ability to form adherent and stable transfer film on the hard metal counterface. Several studies have shown that if pure nylon is used in sliding, the transfer film is weak and patchy. This kind of transfer film can be easily removed from the counterface due to dynamic actions of sliding. Interfacial temperature also plays its role in making the transfer layer soft and weak. With certain types of fillers in nylon it has been found that the composite makes a very thin but adherent transfer layer. This transfer layer protects the bulk of the polymer from further wear. Common fillers with advantageous effects on the wear resistance of nylon are glass fiber (Ref 50), CuS, CuO, CuF_2 (Ref 51), and PTFE (Ref 13). In the study by Hooke et al. (Ref 50), aramid and carbon fibers were also used as fillers for nylon. However, the authors found high friction for these two fillers and concluded that

interfacial heating due to high friction could damage the nylon matrix leading to accelerated wear, especially in the high load and speed conditions.

The main disadvantage with the use of nylons is their water-absorbent characteristics. Mechanical properties such as elastic modulus and hardness as well as physical properties such as glass-transition temperature of nylon drastically reduce with the increase in the absorbed water content in nylon. In this respect nylon 11 and nylon 12 are superior to nylon 6 and nylon 6/6. The percentage water absorption at saturation and 20 °C temperature for nylon 11 and nylon 12 is 1.6% each (Ref 48), while this value for nylon 6 is 10.9% (Ref 48). The percentage of water absorption depends on the amount of crystallinity in the polymer—the higher the crystallinity, the lower the water absorption. A loss of mechanical strength for nylon results in increased wear rate.

One can conclude from this case study that for nylon, the wear resistance characteristics can be enhanced if low water absorbing forms (such as nylon 11 or nylon 12) of nylon reinforced with fillers such as glass fiber, CuS, CuO, or PTFE, are used. To author's knowledge, so far there is no available published work on the friction and wear characteristics of nylon 12.

References cited in this section

13. S. Bahadur and V.K. Polineni, Tribological Studies of Glass Fabric Reinforced Polyamide Composites Filled with CuO and PTFE, *Wear,* Vol 200, 1996, p 95–104

15. F. Van De Velde and P. De Baets, The Friction and Wear Behaviour of Polyamide 6 Sliding against Steel at Low Velocity under Very High Contact Pressure, *Wear,* Vol 209, 1997, p 106–114

48. W.H. Carothers, U.S. Patents 2130947 and 2130948

49. S.M. Aharoni, *n-Nylons: Their Synthesis, Structure, and Properties,* John Wiley & Sons, 1997

50. C.J. Hooke, S.N. Kukureka, P. Liao, M. Rao, and Y.K. Chen, The Friction and Wear of Polymers in Non-Conformal Contacts, *Wear,* Vol 200, 1996, p 83–94

51. S. Bahadur, D. Gong, and J.W. Anderegg, The Role of Copper Compounds as Fillers in Transfer Film Formation and Wear of Nylon, *Wear,* Vol 154, 1992, p 207–223

Wear Failures of Plastics

Sujeet K. Sinha, National University of Singapore

Failure Examples (Ref 52)

Example 1: Wear Failure of an Antifriction Bearing. Shown in Fig. 14 is the worn surface of an antifriction bearing made from a nylon/polyethylene blend. The bearing was worn in contact with a steel shaft. Movement of the shaft against the bearing caused abrasive marks (Fig. 14). Fine iron oxide particles acted as an abrasive, producing the failure mechanism observed.

Fig. 14 Wear marks on the surface of a nylon/polyethylene antifriction bearing. The bearing was in contact with a rotating steel shaft. 417×. Source: Ref 53

Example 2: Failure of a Nylon Driving Gear. Figure 15 shows pitting on the tooth flank of a nylon oil-lubricated driving gear. The pitting produced numerous surface microcracks in association with large-scale fragmentation (frictional wear). The stress-cracking effect of the lubricating oil is believed to have played a role in initiating the observed microcracks.

Fig. 15 Pitting and surface microcracks on the tooth flank of an oil-lubricated nylon driving gear. 37×. Source: Ref 53

Example 3: Failure of a Polyoxymethylene Gear Wheel. A polyoxymethylene gear wheel (Fig. 16) exhibits a different failure mechanism. This component had been in operation in a boiler room and is believed to have failed because of considerable shrinkage. The oriented crystalline superstructures and the microporosity are reported to be due to postcrystallization. The porosity is attributed to the difference in densities between the amorphous (1.05 g/cm^3, or 0.04 lb/in.^{3}) and the semicrystalline (1.45 g/cm^3, or 0.05 lb/in.^{3}) states (Ref 53). Breakdown along the crystalline superstructure started mainly at the mechanically stressed tooth flanks. In addition, oil vapors, humidity, and other degradative agents could also have contributed to the observed failure.

Fig. 16 Failed polyoxymethylene gear wheel that had been in operation in a boiler-room environment. 305×. Source: Ref 53

References cited in this section

52. A. Moet, Failure Analysis of Polymers, *Failure Analysis and Prevention,* Vol 11, *ASM Handbook,* ASM International, 1986

53. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage,* Prentice Hall, 1981

Wear Failures of Plastics

Sujeet K. Sinha, National University of Singapore

References

1. B.J. Briscoe and S.K. Sinha, Wear of Polymers, *Proc. Inst. Mech. Eng. J., J. Eng. Tribol.,* Vol 216, 2002

2. B.J. Briscoe and D. Tabor, Friction and Wear of Polymers, Chapter 1, *Polymer Surfaces,* D.T. Clark and J. Feast, Ed., John Wiley & Sons, 1978

3. C.M. Pooley and D. Tabor, Friction and Molecular Structure: The Behaviour of Some Thermoplastics, *Proc. R. Soc. (London) A,* Vol 329, 1972, p 251–274

4. R.P. Steijn, The Sliding Surface of Polytetrafluoroethylene: An Investigation with the Electron Microscope, *Wear,* Vol 12, 1968, p 193–212

5. Y. Uchiyama and K. Tanaka, Wear Law for Polytetrafluoroethylene, *Wear,* Vol 58, 1980, p 223–235

6. J.K. Lancaster, Friction and Wear, Chapter 14, *Polymer Science, A Materials Science Handbook,* A.D. Jenkins, Ed., North-Holland Publishing, 1972

7. B.J. Briscoe, Wear of Polymers: An Essay on Fundamental Aspects, *Tribol. Int.,* Aug 1981, p 231–243

8. S.K. Rhee, *Wear,* Vol 16, 1970, p 431

9. M.K. Kar and S. Bahadur, The Wear Equation for Unfilled and Filled Polyoxymethylene, *Wear,* Vol 30, 1974, p 337–348

10. N. Viswanath and D.G. Bellow, Development of an Equation for the Wear of Polymers, *Wear,* Vol 181–183, 1995, p 42–49

11. D.G. Evans and J.K. Lancaster, The Wear of Polymers, *Wear,* Vol 13, *Materials Science and Technology,* D. Scott, Ed., Academic Press, 1979, p 85–139

12. J.P. Davim and N. Marques, Evaluation of Tribological Behaviour of Polymeric Materials for Hip Protheses Applications, *Tribol. Lett.,* Vol 11 (No. 2), 2001, p 91–94

13. S. Bahadur and V.K. Polineni, Tribological Studies of Glass Fabric Reinforced Polyamide Composites Filled with CuO and PTFE, *Wear,* Vol 200, 1996, p 95–104

14. C.H. da Silva, D.K. Tanaka, and A. Sinatora, The Effect of Load and Relative Humidity on Friction Coefficient between High Density Polyethylene on Galvanized Steel—Preliminary Results, *Wear,* Vol 225–229, 1999, p 339–342

15. F. Van De Velde and P. De Baets, The Friction and Wear Behaviour of Polyamide 6 Sliding against Steel at Low Velocity under Very High Contact Pressure, *Wear,* Vol 209, 1997, p 106–114

16. S. Bahadur and D. Gong, The Action of Fillers in the Modification of the Tribological Behaviour of Polymers, *Wear,* Vol 158, 1992, p 41–59

17. K. Friedrich, Z. Lu, and A.M. Hager, Recent Advances in Polymer Composites' Tribology, *Wear,* Vol 190, 1995, p 139–144

18. T. Kato and A. Magario, The Wear of Aramid Fibre Reinforced Brake Pads: The Role of Aramid Fibres, *STLE Trib. Trans.,* Vol 37 (No. 3), 1994, p 559–565

19. J.V. Voort and S. Bahadur, The Growth and Bonding of Transfer Film and the Role of CuS and PTFE in the Tribological Behavior of PEEK, *Wear,* Vol 181–183, 1995, p 212–221

20. J. Song, P. Liu, M. Cremens, and P. Bonutti, Effects of Machining on Tribological Effects of Ultra High Molecular Weight Polyethylene (UHMWPE) under Dry Reciprocating Sliding, *Wear,* Vol 225–229, 1999, p 716–723

21. A. Schallamach, Abrasion of Rubber by a Needle, *J. Polymer Sci.,* Vol 9 (No. 5), 1952, p 385–404

22. A. Schallamach, How Does Rubber Slide? *Wear,* Vol 17, 1971, p 301–312

23. D.F. Moore, Friction and Wear in Rubbers and Tyres, *Wear,* Vol 61, 1980, p 273–282

24. T. Liu and S.K. Rhee, High Temperature Wear of Semi-Metallic Disk Brake Pads, *Wear of Materials 1977,* American Society of Mechanical Engineers, 1977, p 552–554

25. S.K. Rhee and P.A. Thesier, "Effects of Surface Roughness of Brake Drums on Coefficient of Friction and Lining Wear," paper 720449, Society of Automotive Engineers, 1971

26. J.K. Lancaster and J.P. Giltrow, The Role of the Counterface Roughness in the Friction and Wear of Carbon Fibre Reinforced Thermosetting Resins, *Wear,* Vol 16, 1970, p 357–374

27. B. Bhushan and D.F. Wilcock, Wear Behaviour of Polymeric Compositions in Dry Reciprocating Sliding, *Wear,* Vol 75, 1982, p 41–70

28. B.J. Briscoe, I. Ramirez, and P.J. Tweedle, *Proc. Int. Conf. Disc Brakes for Commercial Vehicles* (London), 1–2 Nov 1988, Mechanical Engineering Publications Ltd., 1988, p 5

29. S.K. Sinha and S.K. Biswas, Friction and Wear Behaviour of Continuous Fibre as Cast Kevlar-Phenolic Resin Composite, *J. Mater. Sci.,* Vol 27, 1992, p 3085–3091

30. S.K. Sinha and S.K. Biswas, Effect of Sliding Speed on Friction and Wear of Uni-Directional Aramid Fibre-Phenolic Resin Composite, *J. Mater. Sci.,* Vol 30, 1995, p 2430–2437

31. J. Hanchi and N.S. Eiss, Jr., Tribological Behaviour of Polyetheretherketone, a Thermotropic Liquid Crystalline Polymer and in situ Composites Based on Their Blends under Dry Sliding Conditions at Elevated Temperatures, *Wear,* Vol 200, 1996, p 105–121

32. B.J. Briscoe, P.D. Evans, E. Pelillo, and S.K. Sinha, Scratching Maps for Polymers, *Wear,* Vol 200, 1996, p137–147

33. B.J. Briscoe, E. Pelillo, and S.K. Sinha, Characterisation of the Scratch Deformation Mechanisms for Poly(methylmethacrylate) Using Surface Optical Reflectivity, *Polymer Int.,* Vol 43, 1997, p 359–367

34. B.J. Briscoe, A. Chateauminois, T.C. Lindley, and D. Parsonage, Fretting Wear Behaviour of Polymethylmethacrylate under Linear Motions and Torsional Contact Conditions, *Tribol. Int.,* Vol 31 (No. 11), 1998, p 701–711

35. B.J. Briscoe, A. Chateauminois, T.C. Lindley, and D. Parsonage, Contact Damage of Poly(methylmethacrylate) During Complex Microdisplacements, *Wear,* Vol 240, 2000, p 27–39

36. K. Tanaka and T. Miyata, Studies on the Friction and Transfer of Semi-Crystalline Polymers, *Wear,* Vol 41, 1977, p 383–398

37. B.J. Briscoe, P.D. Evans, and J.K. Lancaster, Single Point Deformation and Abrasion of γ-Irradiated Poly(tetrafluoroethylene), *J. Phys. D, Appl. Phys.,* Vol 20, 1987, p 346–353

38. A. Wang, D.C. Sun, C. Stark, and J.H. Dumbleton, Wear Mechanism of UHMWPE in Total Joint Replacement, *Wear,* Vol 181–183, 1995, p 241–249

39. R.W. Hertzberg and J.A. Manson, *Fatigue Engineering Plastics,* Academic Press, 1980, p 54–61

40. A.N. Gent and C.T.R. Pulford, Wear of Steel by Rubber, *Wear,* Vol 49, 1978, p 135–139

41. D. Dowson and P.D. Swales, An Elastohydrodynamic Approach to the Problem of the Reciprocating Seal, paper F3, *Proc. Third Int. Conf. Fluid Sealing,* British Hydromechanics Research Association

42. G.J. Field and B.S. Nau, Theoretical Study of EHL of Reciprocating Rubber Seals, *ASLE Trans.,* Vol 18, 1974, p 48–54

43. S.C. Richards and A.D. Roberts, Boundary Lubrication of Rubber by Aqueous Surfactant, *J. Phys. D, Appl. Phys.,* Vol 25, 1992, p A76–A80

44. A. Wang, A. Essner, V.K. Polineni, D.C. Sun, C. Stark, and J.H. Dumbleton, Lubrication and Wear of Ultra-High Molecular Weight Polyethylene in Total Joint Replacements, *New Directions in Tribology,* I. Hutchings, Ed., Mechanical Engineering Publications Ltd., London, 1997, p 443–458

45. A. Unsworth, The Effects of Lubrication in Hip Joint Prostheses, *Phys. Med. Biol.,* Vol 23, 1978, p 253–268

46. J.M. Senior and G.H. West, Interaction between Lubricants and Plastic Bearing Surfaces, *Wear,* Vol 18, 1971, p 311–323

47. B.J. Briscoe, V. Mustafaev, and D. Tabor, Lubrication of Polythlene by Oleamide and Stearamide, *Wear,* Vol 19, 1972, p 399–414

48. W.H. Carothers, U.S. Patents 2130947 and 2130948

49. S.M. Aharoni, *n-Nylons: Their Synthesis, Structure, and Properties,* John Wiley & Sons, 1997

50. C.J. Hooke, S.N. Kukureka, P. Liao, M. Rao, and Y.K. Chen, The Friction and Wear of Polymers in Non-Conformal Contacts, *Wear,* Vol 200, 1996, p 83–94

51. S. Bahadur, D. Gong, and J.W. Anderegg, The Role of Copper Compounds as Fillers in Transfer Film Formation and Wear of Nylon, *Wear,* Vol 154, 1992, p 207–223

52. A. Moet, Failure Analysis of Polymers, *Failure Analysis and Prevention,* Vol 11, *ASM Handbook,* ASM International, 1986

53. L. Engel, H. Klingele, G.W. Ehrenstein, and H. Schaper, *An Atlas of Polymer Damage,* Prentice Hall, 1981

Wear Failures of Reinforced Polymers

J. Bijwe, Indian Institute of Technology, Delhi

Introduction

REINFORCED POLYMERS are used extensively in applications where resistance to adhesive and abrasive wear failure is important for materials selection. Polymers form a special class of materials because of their self-lubricity, which allows them to function without external conventional liquid lubrication. However, polymers also have some inherent tribological limitations, such as significantly low thermal conductivity, dissipativity, and diffusivity as compared with metals. Frictional heat generated at the sliding contacts cannot be dissipated properly, and hence flash temperatures at sliding contacts remain high. Their poor thermal stability also makes them more vulnerable due to loss of mechanical strength with an increase in the surface

temperature. The thermal expansion coefficients of polymers are ten times greater than those of metals, posing problems related to dimensional clearances. In addition to the creeping tendency, polymers have low dimensional stability and rigidity. They have poor compressive strengths (approximately 30 times less) compared with those of other classes of tribomaterials. These inherent limitations restrict the utility of the polymers under severe operating conditions such as high loads, speeds, and temperatures. Therefore, reinforcements (fibrous or particulate) generally are used to increase the load carrying capacity, strength, resistance to creep, and wear. Limitations of strength and thermal conductivity can be overcome efficiently by the right selection of reinforcements and fillers in the appropriate amount, combination, and processing technology.

The tribological performance of reinforced polymers is governed by the type of base matrix, the nature of the filler (type, amount, size, shape, aspect ratio, distribution, orientation, combination with fillers, and the quality of bonding with the matrix), and operating conditions. Fibers are far more wear resistant than the matrix and hence control the wear of composite. Continuous fiber-reinforced composites with a thermoset-polymer matrix (such as phenolics, epoxy, etc.) may have low wear rates and higher strength than those with thermoplastics because a higher ratio of fiber is achievable with a thermoset matrix. Incorporation of fillers also can modify wear resistance of polymers up to the order of four. Solid lubricants mostly reduce wear and, rather essentially, friction coefficient. Similarly, reinforcement generally reduces wear but not always the coefficient of friction.

Generally, reinforced polymers have multiple or multifunctional fillers that can have synergistic and/or antagonistic interactions with respect to wear performance. Therefore, the tribological behavior of reinforced polymers is an empirical evaluation depending on specific material conditions, operating parameters, and environment. It is also ill advised to compare tribological properties of materials evaluated in different laboratories, configurations, environments, and so on. Their ranges or performance rankings are important for materials selection in a particular wear situation. Table 1 indicates tribological regimes for various kinds of reinforced polymers. Several recent review articles also provide background on the tribology of polymers and composites (Ref 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11).

Table 1 Tribo-potential of polymers and composites for variety of applications

Composite material	**Tribological applications**	**Maximum tribological regime**
Neat and short fiber reinforced composites (SFRP)	Seals, gears, slideways bearings, and abrasive wear applications	$PV < 15$ MPa · m/s $V < 5$ m/s, $\mu > 0.03$ $T < 250$ °C, $W_s > 10^{-16}$ m^3/Nm
Continuous fiber reinforced composites (UD)	Under-water or high temperature applications, aerospace seals and bearings	$PV < 100$ MPa · m/s $V < 5$ m/s $T < 320$ °C, $\mu > 0.09$ $W_s > 10^{-17}$ m^3/Nm
Thin layer composites with metallic supports	Pivot bearings, high pressure applications	$PV < 300$ MPa · m/s $V < 1$ m/s $T < 320$ °C, $\mu > 0.06$ $W_s > 10^{-18}$m^3/Nm

P, pressure; *V*, sliding speed; W_s, specific wear rate; *T*, temperature. Source: Ref 1

This article briefly reviews the abrasive and adhesive wear failure of:

- Particulate-filled polymers
- Short fiber reinforced polymers (SFRP)
- Polymers with continuous fibers
- Mixed reinforcements (either with fiber and filler, or with two types of fillers known as *hybrid composites*) and fabrics

Each section discusses various aspects such as friction and wear performance of the composites, correlation of performance with various materials properties, and studies on wear-of failure mechanisms by scanning electron microscopy (SEM).

Apart from the anisotropy of reinforced polymers (especially fiber-reinforced polymers, or FRP), synergistic and/or antagonistic effects in the case of a combination of two fibers or fillers is one of the most important aspects of composite tribology. This topic is briefly discussed with some emphasis on various mathematical models starting with the simple rule of mixtures. More complex methods developed for describing the wear performance of each type of composite are also available, but such models, along with the data on wear mechanisms and friction and wear performance, serve more for tailoring future composites. This subject is mentioned only briefly because it is beyond the main purpose of characterizing wear failures from the perspective of failure analysis and prevention.

References cited in this section

1. K. Friedrich and R. Walter, Microstructure and Tribological Properties of Short Fiber/Thermoplastic Composites, *Tribology of Composite Materials Conf. Proc.*, P.K. Rohatgi, P.J. Blau, and C.S. Yust, Ed., ASM International, 1990, p 217–226

2. K-H. Zum Gahr, *Microstructure and Wear of Materials, Tribology Series 10,* Elsevier, Amsterdam, 1987

3. J. Bijwe and M. Fahim, Tribology of High Performance Polymers–State of Art, *Handbook of Advanced Functional Molecules and Polymers,* H.S. Nalwa, Ed., Gordon and Breach Publishers, Amsterdam, 2001, p 265–321

4. K. Friedrich, Z. Lu and A.M. Hager, Recent Advances in Polymer Composite Tribology, *Wear,* Vol 190, 1995, p 139–144

5. K. Friedrich, Wear Models for Multiphase Materials and Synergistic Effects in Polymeric Hybrid Composites, *Advances in Composite Tribology, Composite Materials Series 8,* K. Friedrich, Ed., Elsevier, 1993, p 209–273

6. N.P. Suh, *Tribophysics,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1986

7. K. Friedrich, Ed., *Friction and Wear of Polymer Composites, Composite Materials Series 1,* Elsevier, 1986

8. U.S. Tewari and J. Bijwe, Tribological Behavior of Polyimides, *Polyimides: Fundamentals and Applications,* M.K. Ghosh and K.L. Mittal, Ed., Marcel Dekker Inc., 1996, p 533–586

9. S.W. Zhang, State-of-the-Art of Polymer Tribology, *Tribol. Int.,* Vol 31 (No. 1–3), 1998, p 46–60

10. H. Czichos, *Tribology: A System Approach to the Science and Technology of Friction, Lubrication and Wear, Tribology Series 1,* Elsevier, 1978

11. E. Santner and H. Czichos, Tribology of Polymers, *Tribol. Int.,* Vol 22, 1989, p 103–109

J. Bijwe, Indian Institute of Technology, Delhi

Abrasive Wear Failure of Reinforced Polymers

Polymer composites are extensively used for sliding components in earth-moving equipment, rock- and ore crushers, dies in powder metallurgy, extruders and chutes, and so on, where the major wear failure mechanism is either two-body or third-body abrasion. The abrasion is a net result of microscopic interactions of the surface and the abradant as shown in Fig. 1 (Ref 12). Microplowing and microcutting are the dominant processes in the abrasion of ductile materials, while microcracking is important in brittle materials.

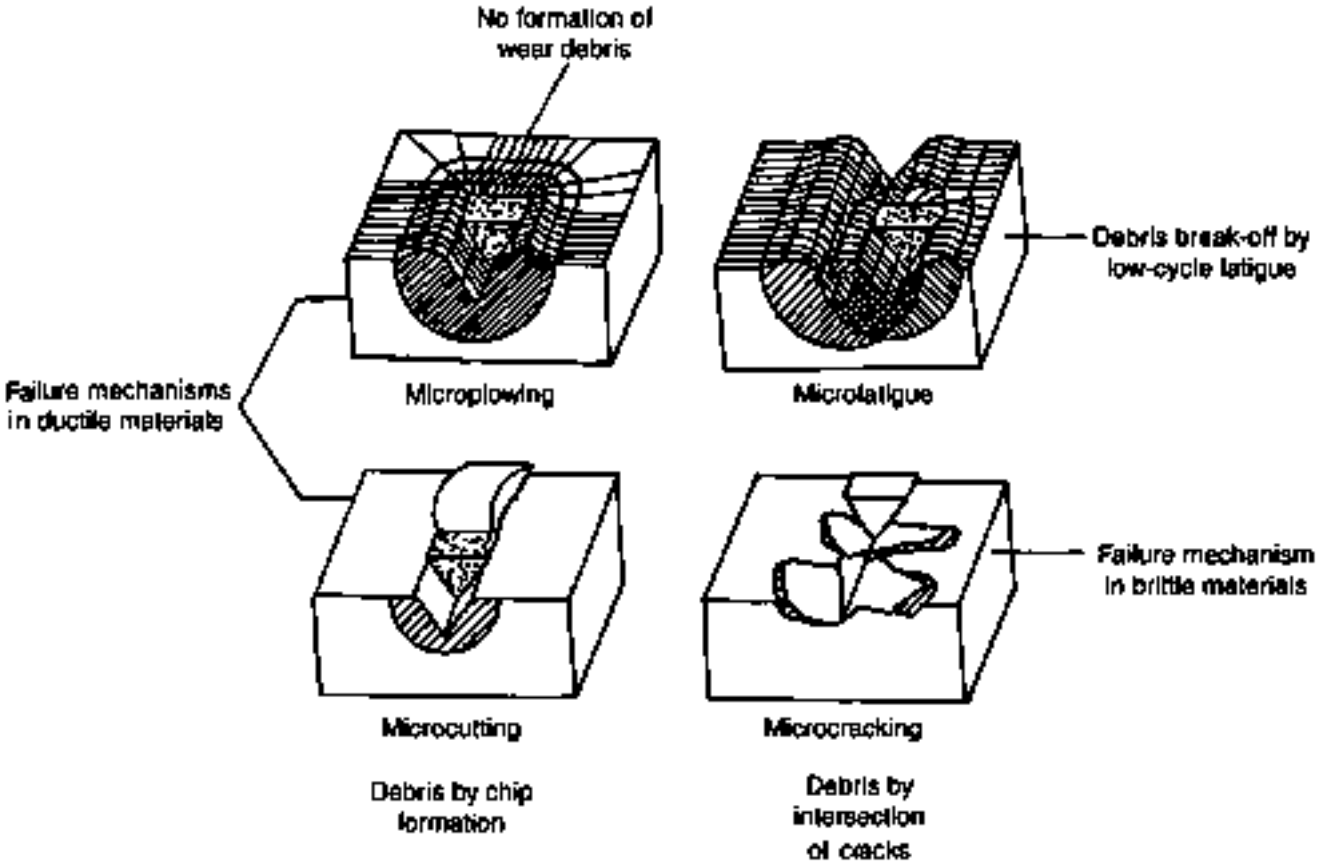


Fig. 1 Schematic of different interactions during sliding of abrasive particles against the surface of material. Source: Ref 12

As previously noted, the wear of reinforced polymers is influenced by the properties of the matrix, the filler, and their bonding; the ratio of grooving depth to the filler size; the shape and size; orientation distribution; hardness of the grit and filler; and the operating parameters. In the case of abrasive wear of reinforced polymers, the role of filler is very much different from the role of filler in adhesive wear mode. Abrasive wear behavior generally is evaluated by abrading it under load against hard rough surfaces such as paper or wheel impregnated with silicon carbide (SiC), flint, alumina, and so on, or a rough metallic disc.

Abrasive wear failure of FRPs occurs generally because of matrix shearing and fiber debonding followed by fiber cracking and cutting. Generally, fibers in the normal (N) direction are more wear resistant than those in the parallel (P) direction, which in turn, are better than those in the antiparallel (AP) direction. If the size of filler is smaller than the abrading grit, then the filler particle is easily dislodged and dug out. Fillers of medium size are most effective in this case. With increase in volume fraction (V_f) of the filler or fiber, wear may increase, decrease, or show minima at some concentration, generally around 20 to 30%. Abrasive wear may increase with a decrease in modulus of elasticity of fiber or matrix as a result of higher debonding of fibers. If

the filler hardness is higher than the abrading grit, wear decreases. If the filler or matrix is brittle, wear increases due to cracking and flaking (Ref 2).

Abrasive Wear of Particulate-Reinforced Polymers. Extensive work has been done in this area. (Ref 13) Figure 2 shows relative abrasive wear loss of quartz and glass-filled polymethylmethacrylate (PMMA) slid against SiC, SiO_2, and $CaCO_3$ abrasives as a function of filler volume fraction. The performance clearly depends on the ratio of hardness of the abrasives to the filler, the interfacial adhesion of the filler with the matrix PMMA, and volume fraction.

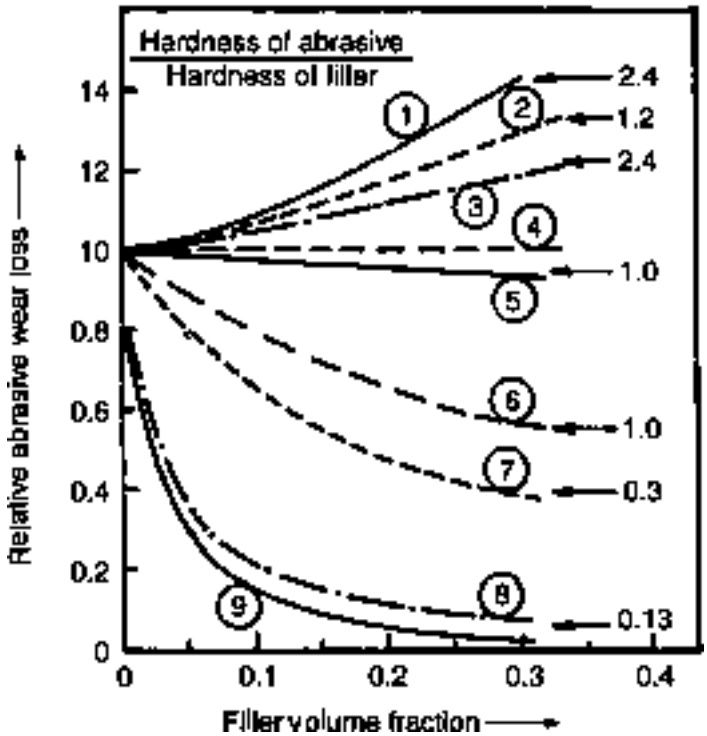


Fig. 2 Relative abrasive wear loss of polymethylmethacrylate (PMMA) and composites filled with quartz and glass against abrasives SiC (45 µm), WIB, SiO_2 (10 µm) and $CaCO_3$ (3 µm) as a function of filler volume fraction, V_f. WIB, weak interfacial bond; SIB, strong interfacial bond: 1, WIB-quartz, filler against SiC; 2, WIB-glass filler against SiO_2; 3, SIB quartz filler against SiC; 4, unfilled PMMA; 5, WIB-quartz filler against SiO_2; 6, SIB-quartz filler against SiO_2; 7, WIB-glass filler against $CaCO_3$; 8, WIB-quartz filler against $CaCO_3$; 9, SIB-quartz filler against $CaCO_3$. LROM, linear rule of mixtures. IROM, inverse rule of mixtures. Source: Ref 13

An upper bound on abrasive wear resistance is modeled by the following equation, known as the *inverse rule of mixtures* (IROM), where overall wear behavior is assumed to be a function of the individual contribution from each phase. Wear resistance (W^{-1}) (as modeled by Khruschov) (Ref 14) is:

$$W^{-1} = \sum V_i W_i^{-1}$$

where W is the total wear volume, as a linear function of volume fraction of the phase present, V_i is the volume fraction of the *i*th phase, and W_i is the wear volume due to *i*th phase. Equation 1 provides an upper bound to the wear resistance. However, the basic assumption that each phase shows a wear rate proportional to the applied load is not always correct, especially when hard phases and ceramic fillers are involved. IROM cannot be applied in such a case, because of nonlinearity in wear load relation. Hence, another model known as the *linear rule of mixtures* (LROM) (Eq 2), was developed (Ref 15) and is suitable for composites containing a combination of similar phases,

$$W = \sum V_i W_i$$

These models based on volume fraction of the reinforcing phases showed considerable deviation from the experimental values, especially when the size of the abrasives was large and the load was high. As seen in Fig. 3, (Ref 16) the deviation clearly increased with the size of the filler and applied load. In reality, the abrasive wear of a multiphase system is the macroscopic sum of all the microscopic events generated by the abradants and hence depends on the size of the grain. When grain size equals or exceeds the microstructure, filler pullout

is inevitable, leading to large wear to the extent depending on the quality of the interface. The deviation depends on the size of grain, microstructure, load, and volume fraction. The rules of mixture are not useful in the quantitative prediction of industrial wear rates but for the development of wear-resistant materials in the laboratory.

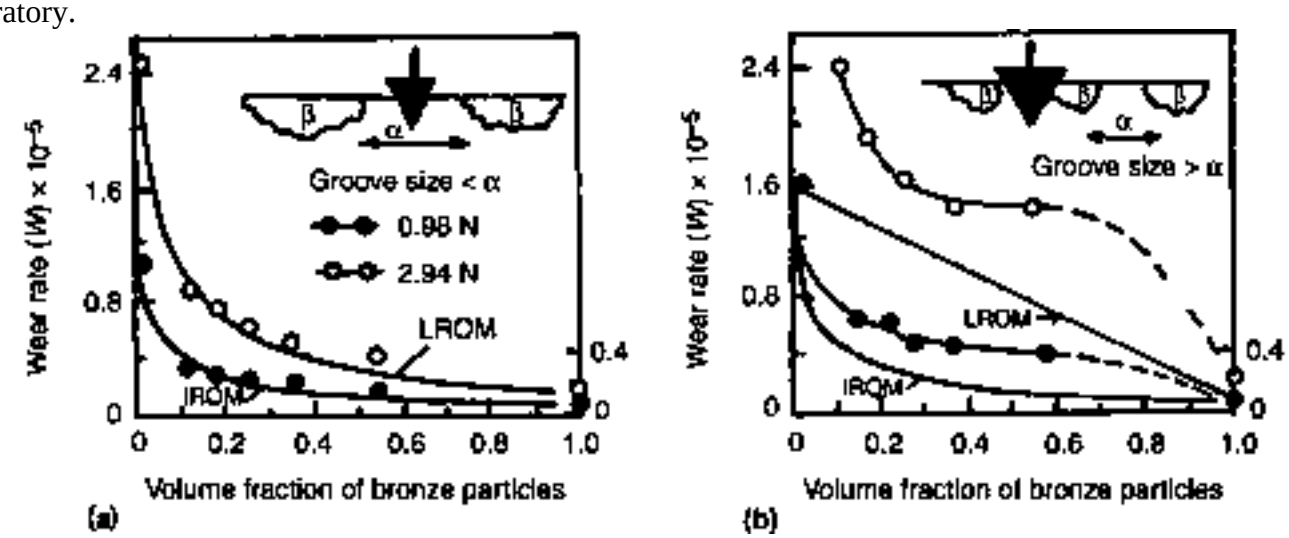


Fig. 3 Dimensionless wear rate, *W*, at two loads as a function of volume fraction of bronze particles in epoxy-Cu-Al system (Cu-Al particle diameter ≃100 μm). Abrading surface, silicon carbide. (a) Fine grade; (b) coarse grade. α-space between two particles, β. LROM; linear rule of mixtures; IROM, inverse rule of mixtures. Source: Ref 16

Bahadur and Gong (Ref 17) developed a theoretical model for the optimal filler loading based on the random packing model of particles in a polymer. Using the computer program based on this model and the data on filler particle size distribution with the density of filler and polymer and the distance between the particles of the filler in the critical packing state, the filler proportion in the polyamide (PA) 11 was calculated. The maximum mechanical strength and highest abrasive resistance of the composites with optimal filler contents calculated with the equation agreed well with the experimental data.

Abrasive Wear of Short-Fiber Reinforced Polymers (SFRPs). Figure 4indicates effect of size, orientation, hardness, modulus, and brittleness of the fiber on the abrasive wear performance of the composites (Ref 2). The performance of the composites generally deteriorates in the case of short-fiber reinforced polymers (SFRPs). An increase in wear performance was reported for seven composites and deterioration for six in the case of thirteen polymers reinforced with 30% short carbon fibers (CF) (Ref 18). Generally, a reduction in *Se* factor or *HSe* factor (where *S* is ultimate tensile strength, *e* is the elongation to break, and *H* is hardness) is observed to be responsible for the deterioration in performance, and Ratner-Lancaster plots show good correlation (Ref 8, 18, 19). Table 2 shows several properties correlated with the wear behavior of such composites by various researchers.

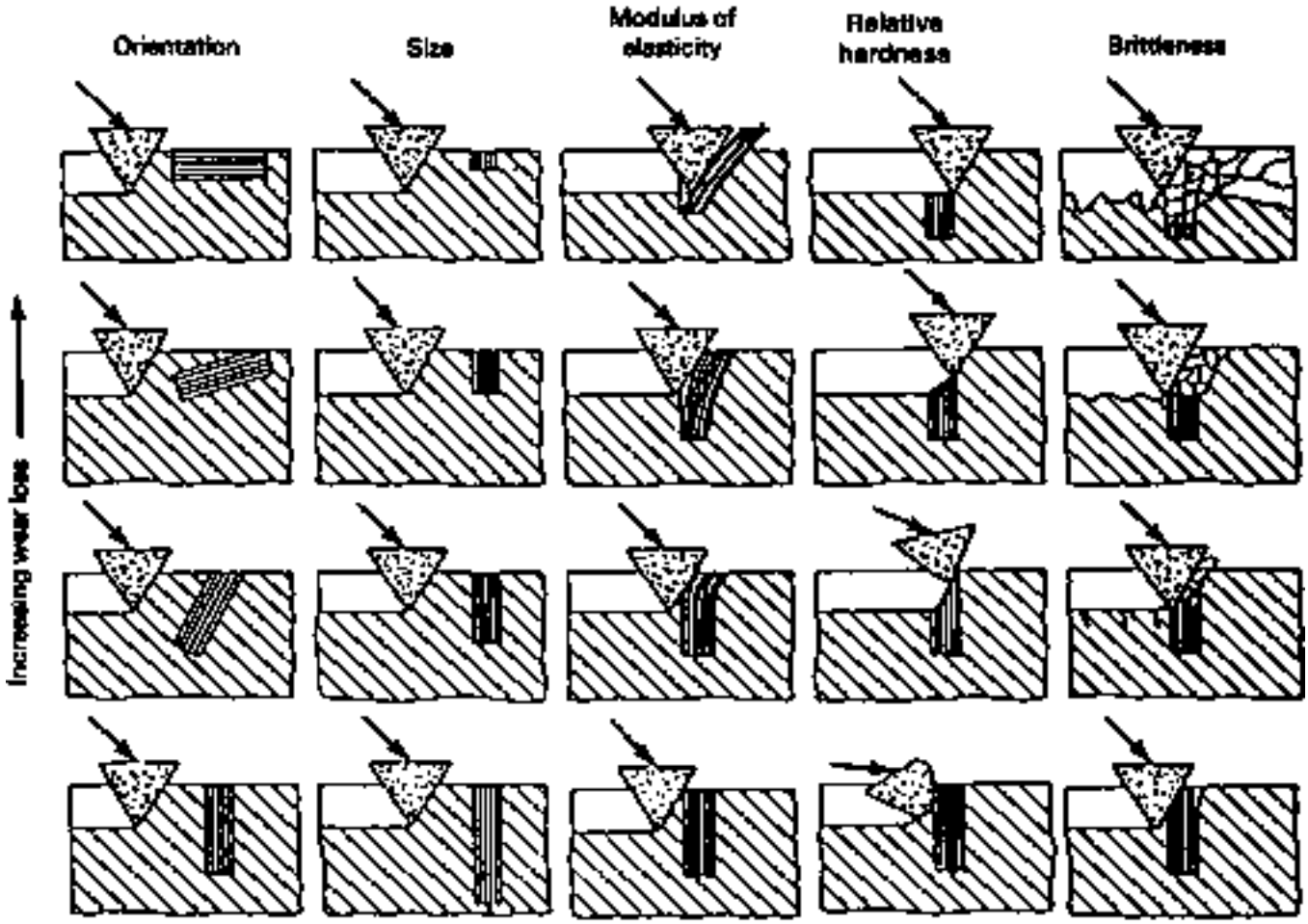


Fig. 4 Influence of various properties of reinforcing phase on abrasive wear of composite. Source: Ref 2

Table 2 Details of the literature on (abrasive) wear property correlation of polymers and composites

Resin	**Fiber/filler, wt%**	**Wear rate due to fillers**	**Correlation of abrasive wear with**	**Ref**
PA 6, PA 66, PVC, PTFE, PP, epoxy, PMMA, polyester, PC, phenolic, PE, acetal copolymer	CF/30	Increased for six polymers and decreased for seven polymers	$(Se)^{-1}$	18
PA 66	CF/10, 20, 30, and 40 PTFE/15	Increased	$(Se)^{-1}$	19
PA 6, PMMA, PE, POM, PA 66, PP, PTFE, PC, PTFCE, polyphenylene oxide, and PVC	…	…	Cohesive energies	20
PA 6, PTFE, PP, POM, HDPE, PVC, and PMMA	…	…	Plowing component of friction, hardness, tensile strength, and elongation to failure of polymers	21
PET	GF/30 Glass/30 Sphere	Increased	The hardness, macrofracture energy, and the probability factor for microcracking	22
PA 6	GF/15 and 20	Increased	$(HSe)^{-1}$, fracture toughness,	23

Resin	Fiber/filler, wt%	Wear rate due to fillers	Correlation of abrasive wear with	Ref
	PTEE/3 Bronze and copper powder/6		fracture energy, and ductility factor	
PTFE	…	…	$(S^2e)^{-1}$	24
PTFE	…	…	(S_{max}/S_y) and asperity strain (r/R)	25
PA 66, PPS, PEEK, PC, PES, and PEI	CF/30 GF/30 and 40	…	Roughness of abrasive paper, $(Se)^{-1}$ and (ESe^{-1})	26
EP and PEEK	CF, GF and AF/60	…	Asperity size	27
ABS, PA, PE, PP, PS, and POM	…	…	(Depth of wear groove × f_{ab} value)$^{-1}$	2
PES and PMMA	…	…	K_{Ic}	28
PEI	GF/16, 20, and 25 PTFE/15 Graphite and MoS_2/15	Increased	$(Se)^{-1}$	8
PI	Graphite/15 and 40 PTFE/15 MoS_2/15	Increased	$(Se)^{-1}$	8

PP, polypropylene; POM, polyoxymethylene; PC, polycarbonate; PTFCE, polytrifluorochloroethylene; HDPE, high-density polyethylene; PET, polyethylene terephthalate; S, ultimate tensile strength; e, elongation to break; E, elastic modulus; K_{Ic}, fracture toughness; f_{ab}, the ratio of volume of material removed as wear debris to the volume of the wear groove

The maxima in the specific wear rate-volume fraction relation also depends on the size of abrasives. As seen in Fig. 5, (Ref 29) the 10% glass fiber (GF) loading in PEI showed maximum wear for the abrasive grain size of ≅118 μm while for 175 μm size grits, loading was at 20%. The wear minima in both cases, however, was at 30% loading of short GF. The difference in the severity of the fiber damage in the 10% polyetherimide (PEI) composite due to these different grit sizes can be seen in Fig. 6(a) and (b). The various stages of fiber cracking, cutting, and pulverization in 30% GF composite are shown in Fig. 6(c). Generally, the wear rate of SFRPs shows increase with fiber volume fraction. The performance, however, deteriorates disproportionately if the combination of filler, solid lubricants, and fibers is included in the composite (Ref 8).

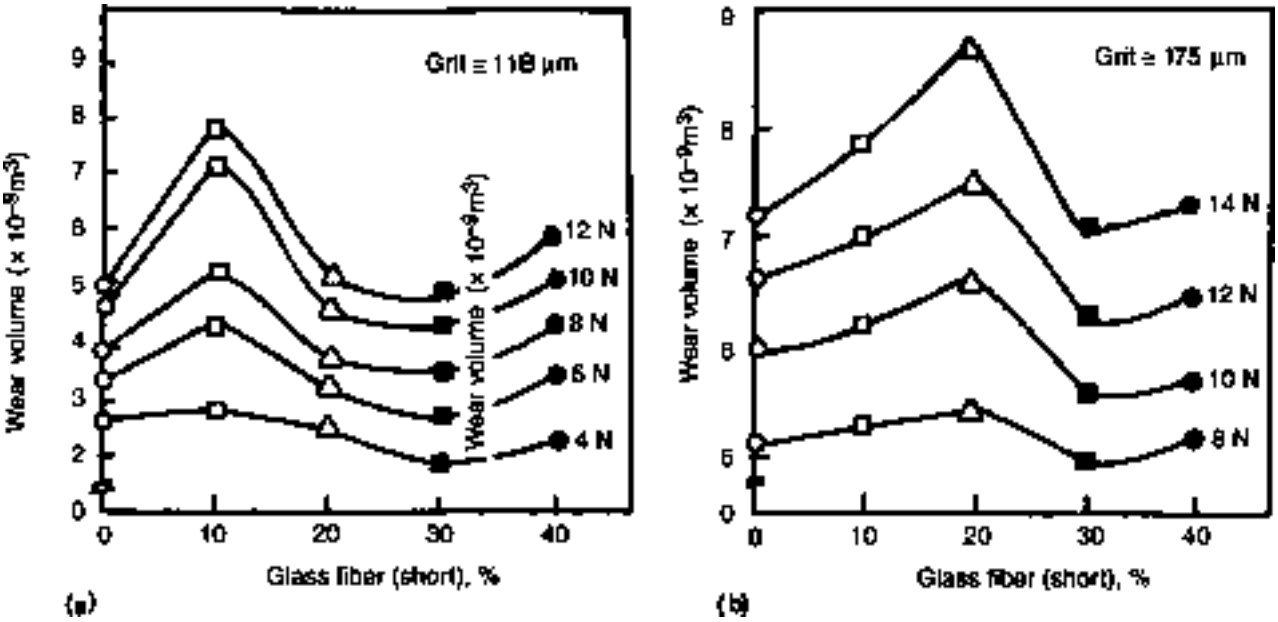


Fig. 5 Abrasive wear volume at various loads and SiC abrasive papers as a function of volume fraction of short glass fibers (GF) in PEI. Speed 5 cm/s in single pass condition; distance slid 3.26 m. (a) 120 grade, grit size ≅118 µm. (b) 80 grade, grit size ≅175 µm. Source: Ref 29

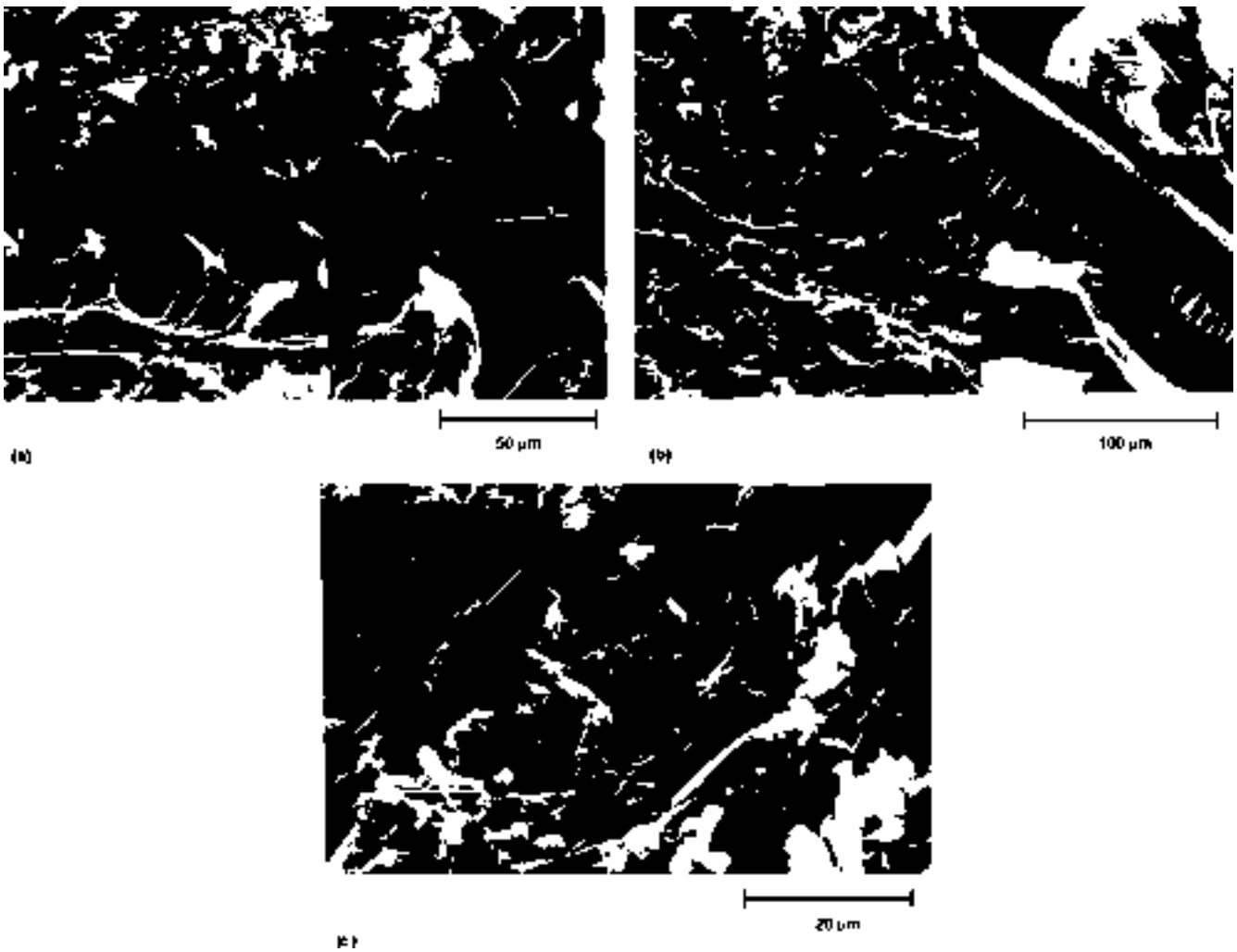


Fig. 6 Scanning electron micrographs of abraded surfaces of composites against 80 grade SiC paper and under 14 N load. (a) Polyetherimide (PEI) + 10% glass fiber (GF) showing extensive damage to matrix and fiber; cavities left after fiber consumption. PEI + 30% GF. (b) Fiber on the stage of microcracking. (c) Initiation of fiber pulverization. Source: Ref 29

A wear equation was developed for SFRPs based on crack propagation theory for describing the effect of load (F_N) on specific wear rate (W_S) (Ref 30):

$$W_S = K_6 \frac{V_S V_C}{E H \varepsilon_f \mu_\alpha F_N^\delta}$$

where V_S is sliding speed, E is elastic modulus, H is hardness, and ε_f is wear failure strain. The equation is valid only when thermal activation is insignificant, as in the case of low sliding speed. μ_α is the number between unity and zero; V_C, the crack growth velocity, is related to fiber aspect ratio, elastic modulus of fiber, shear modulus of the matrix, stress of friction on damaged matrix-fiber interface, fiber fraction, and many other complex characteristic properties. If F_N is large, thermal effect becomes significant, and the exponential term in the following equation controls the wear rate:

$$W_S = K_9 \frac{V_S F_N^{2/(\beta - 5)}}{E H \varepsilon_f \mu_\alpha} \exp \frac{-\gamma_2}{F_N}$$

where K_9 is a parametric constant and γ_2 is a dimensionless constant. When V_S is high, material softens, causing crack propagation to be easier than in the case of low V_S. Hence, depending on the magnitude of V_S, wear rate increases, decreases, or remains constant. W_S decreases with F_N when $\delta > 2/\beta$ since $V_C \propto F_N^{2/\beta}$. When V_C becomes thermally activated then, at a certain point, the effect of V_C can override the F_N^δ term and, thus, W_S increases with F_N.

A wear model was developed for correlating wear behavior of various types of composites with the materials properties (Ref 31). As seen in Fig. 7, surface deformation and wear mechanisms are functions of hardness and fracture energy of the matrix. When pressure, P, increases a certain critical value, P_{crit}, the contribution due to brittle fracture of the material to the wear increases during the transition in wear mechanism at certain V_f. The fracture toughness of a material (K_{Ic}) can be correlated to H as:

$$P_{crit} \propto \frac{K_{Ic}^2}{H}$$

If load is high, P_{crit} is approached earlier. For a given apparent contact area (A), effective pressure (P_{eff}) is related to effective contact area (A_{eff}) as:

$$P_{eff} \propto \frac{PA}{A_{eff}}$$

If the material has high hardness, low fracture toughness, sharp grains, or rough counterface, the probability of P_{eff} reaching P_{crit} is high (Fig. 7a). Wear rate (W) as a function of hardness shows a change in wear mechanism as shown in Fig. 7(b). Transition II shows a disproportionately higher wear rate with an increase in P (for a material of hardness H). Transition I represents reduction in P_{crit} ($P < P_{crit}$ to $P > P_{crit}$) due to increase in H at constant pressure P leading to additional microcracking events. When load increases and H also increases due to higher V_f, the same roughness of the abradant becomes more detrimental, and wear dominated by microcracking becomes more prominent. Taking into account various factors such as probability factor (Ω^*) of microcracking, hardness, and size of abradant, particle density/area, specific wear rate W_S was correlated with H of the composite, modified wear coefficient (Ω), and fracture energy G_{Ic} (Ref 31).

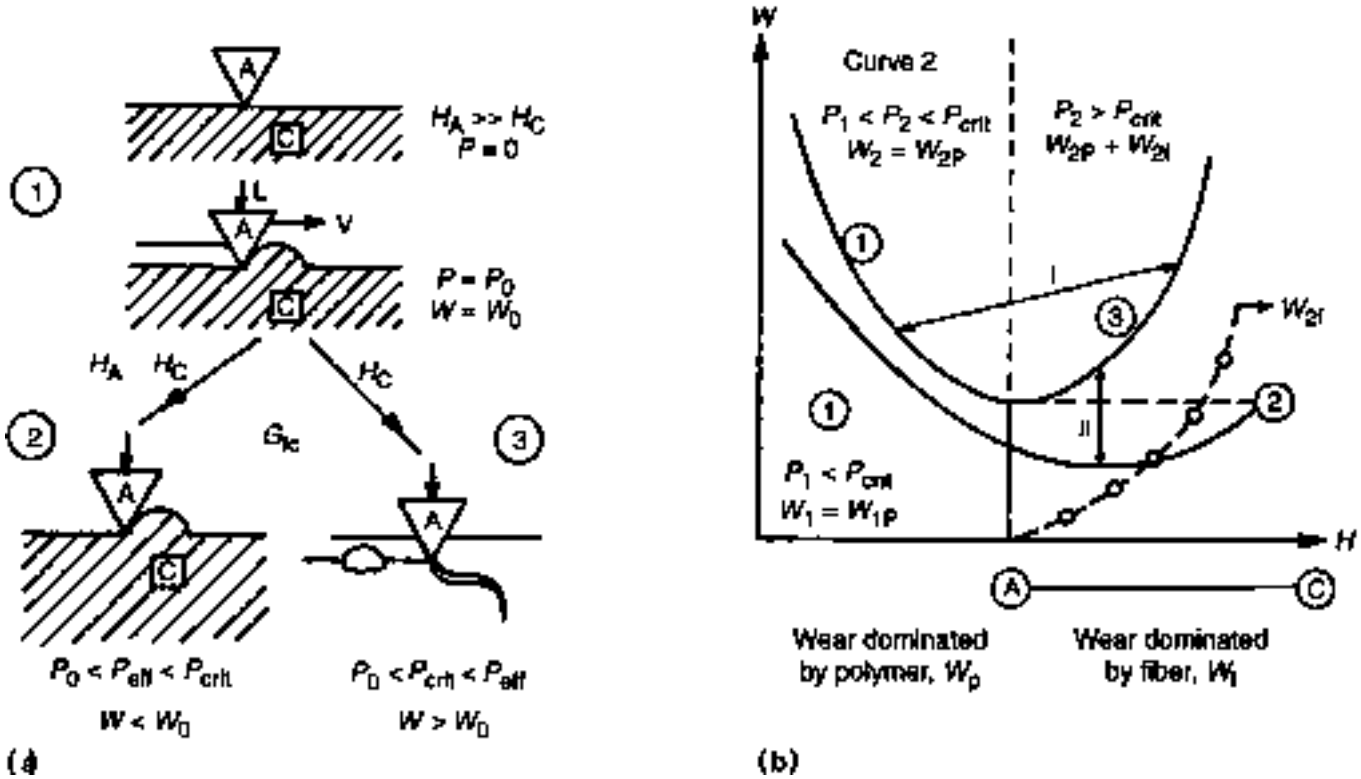


Fig. 7 (a) Abrasive wear mechanisms and surface deformation as a function of pressure, *P*; material hardness, *H*; and fracture energy, G_{Ic}. (b) Curves 1 to 3 correspond to the schematic in (a), possible schematic of the wear rate, *W* as a function of hardness, *H* of wearing material. Curve 1 is a normal curve showing a reduction in wear with increase in hardness while curve 2 reflects changes in trends when microcracking. W_f plays a role at higher pressure. Source: Ref 31

Abrasive Wear of Continuous Unidirectional Fiber Reinforced Polymers (FRP). In-depth studies on continuous steel fiber reinforced, epoxy polymer (EP), and polymethylmethacrylate (PMMA) (Ref 2) focused on various aspects such as area fraction, volume fraction, mean free path between the fibers (λ), and the ratio λ/*D*, where *D* was the size of abrasives. Based on these studies, the general trends in abrasive wear of FRP composites and the properties and alignment of fibers were summarized (Fig. 4) (Ref 2). Studies were done on abrasive wear performance of FRP composites of epoxy (thermoset polymer) and polyetheretherketone (PEEK) (thermoplastic polymer) reinforced with various fibers such as AS4 carbon fiber (CF) (62 vol%), glass fiber (GF) (58 vol%) and K49, aramid fibers (AF) in epoxy and AS4CF (55 vol%) and K49, AF (60 vol%) in PEEK (Refs 32, 33).

Summary of the wear behavior (Table 3) indicates clearly that unidirectional (UD) fiber reinforcement in the antiparallel (AP) orientation was detrimental in all the cases, while normal orientation (O_N) was always beneficial. Aramid fiber was most effective in improving the resistance when it was in normal orientation. Interestingly, parallel orientation of AF was worse than the AP orientation in the case of PEEK. The various wear mechanisms suggested in different orientations are schematically shown in Fig. 8(a), while Fig. 8b shows the ideal composite with very good abrasive wear resistance based on these studies. It should contain thermoplastic polymer such as PEEK and continuous AFs in the (O_N) and CFs in the parallel (O_P) orientation (Ref 32). A cyclic wear model based on volume fraction of fiber (V_f), wear resistance, and elastic moduli of its constituents shows a very good matching of experimental data with the calculated one (Ref 34). The wear model developed for FRPs (Ref 31) also shows good correlation for CF-reinforced UD composite of epoxy matrix.

Table 3 Influence of fiber orientation on the abrasive wear behavior of continuous fiber reinforced polymer composite

Polymer	UD reinforcement and wear resistance (normalized) $(W_s^{-1})_{composite}/(W_s^{-1})_{epoxy}$			
	CF	AF	GF	Neat polymer

EP	…	…	…	1.0
N	1.8	7.9	2.3	…
P	1.7	1.1	1.4	…
AP	0.9	1.1	0.8	…
PEEK	…	…	…	1.8
N	1.7	15.2	2.4	…
P	1.6	0.9	1.8	…
AP	1.0	1.3	1.2	…

UD, unidirectional; W_s, wear rate; CF, carbon fiber; AF, aramid fiber; GF, glass fiber; EP, epoxy; N, normal; P, parallel; AP, antiparallel; PEEK, polyetheretherketone. Source: Ref 33

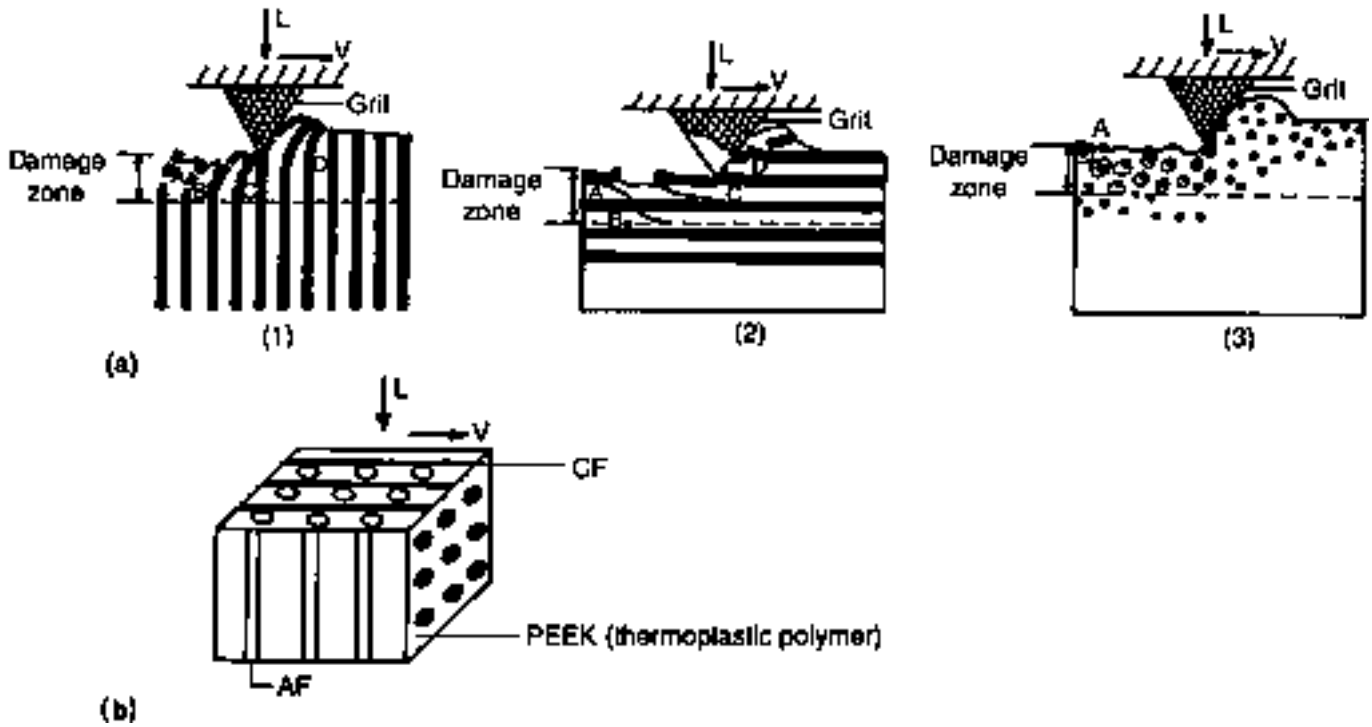


Fig. 8 (a) Schematic of basic wear failure mechanisms observed in (a1) (a2) parallel, P and (a3) antiparallel AP orientations. (a1) A, fiber slicing, B, fiber matrix debonding; C, fiber cracking, j and D, fiber bending (especially in the case of aramid fiber or carbon fiber). (a2) A, interlaminar crack propagation; B, fiber cracking; C, fiber matrix debonding; and D, fiber fracturing. (a3) A, fiber fracturing; B, fiber matrix debonding. (b) Ideal composite for high abrasive wear resistance. *L*, normal load; *V*, velocity Source: Ref 32

Abrasive Wear of Fabric-Reinforced Polymer Composites. Very limited research work is done on the abrasive wear behavior of bidirectionally (BD) reinforced composites. Influence of various fabrics and their orientations on the abrasive wear behavior of composites of thermoplastic PEI is covered in Ref 35. The weave of glass fabric and load were also influencing factors. As compared with short glass fibers (Ref 29), fabric proved to be significantly beneficial for enhancing the abrasive wear resistance of polyetherimide (PEI). Worn surfaces of PEI_{AF} (O_P); PEI_{AF} (O_N); PEI_{HY} (O_N), or polyetherimide hybrid, O_N; PEI_{CF} (O_P); and PEI_{CF} (O_N) are shown in Fig. 9 10 (Ref 36). The extensive softening of AF due to high-contact pressure especially in O_N followed by fibrillation was the most distinct feature observed on PEI_{AF} and PEI_{HY} surfaces. The fiber cracking breakage, pulverization, removal, and plastic deformation of the matrix were minimal in the case of AF composites. The presence of AF hindered the removal of PEI matrix also. In fact, the smooth topography looked more like

adhesive wear case rather than the abrasive wear against SiC paper of 175 μm size. In other cases, excessive breakage and removal of fibers, which led to higher wear were observed. The influence of amount and type of fiber fraction in BD composites and fiber fraction orientation on abrasive wear behavior by sliding the epoxy composites against 70 μm Al_2O_3 paper has been reported (Ref 31).

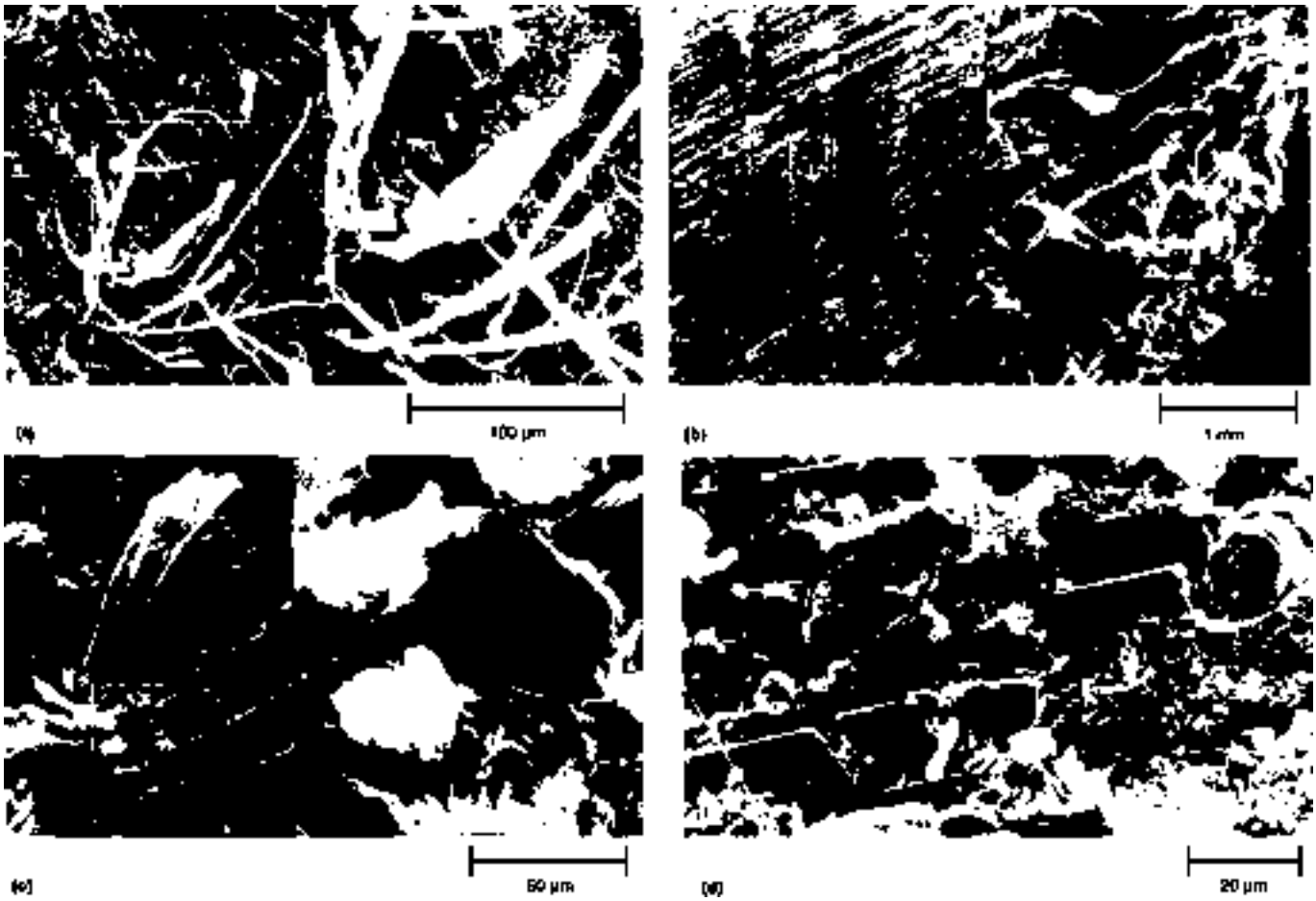


Fig. 9 Scanning electron microscope micrographs of abraded PEI composites reinforced by various fabrics; *L* 12 N, SiC paper 80 grade (grit size 175 μm); distance slid 10 m (33 ft). O_P, fabric parallel to the sliding plane; O_N, fabric normal to the sliding plane. (a) PEI_{AF} (O_P) showing extensive elongation and fibrillation of ductile and soft aramid fibers (AF) during abrasion (b and c) for PEI_{AF} (O_N). (b) Smooth surface topography due to molten AF (high contact pressure, half the fibers being in normal direction). (c) Enlarged view of AF tip indicating extensive elongation and melting. (d and e) $PEI_{CF+AF\ (HY)}$. (d) PEI_{HY} (O_P) abraded from CF side showing multiple microcutting in CF. (e) PEI_{HY} (O_N) showing excessive melting of AF and third body abrasion due to loose grit (middle portion) on the softened matrix. (f) PEI_{CF} (O_P) excessive breakage of an array of CF in both the directions resulting in high wear. (g) PEI_{CF} (O_N) CF tips showing less fiber damage (and hence less wear); cavities due to fiber pull out. (h) PEI_{G3} (O_P) excessive damage to GF in both the directions due to microcutting. Source: Ref 36

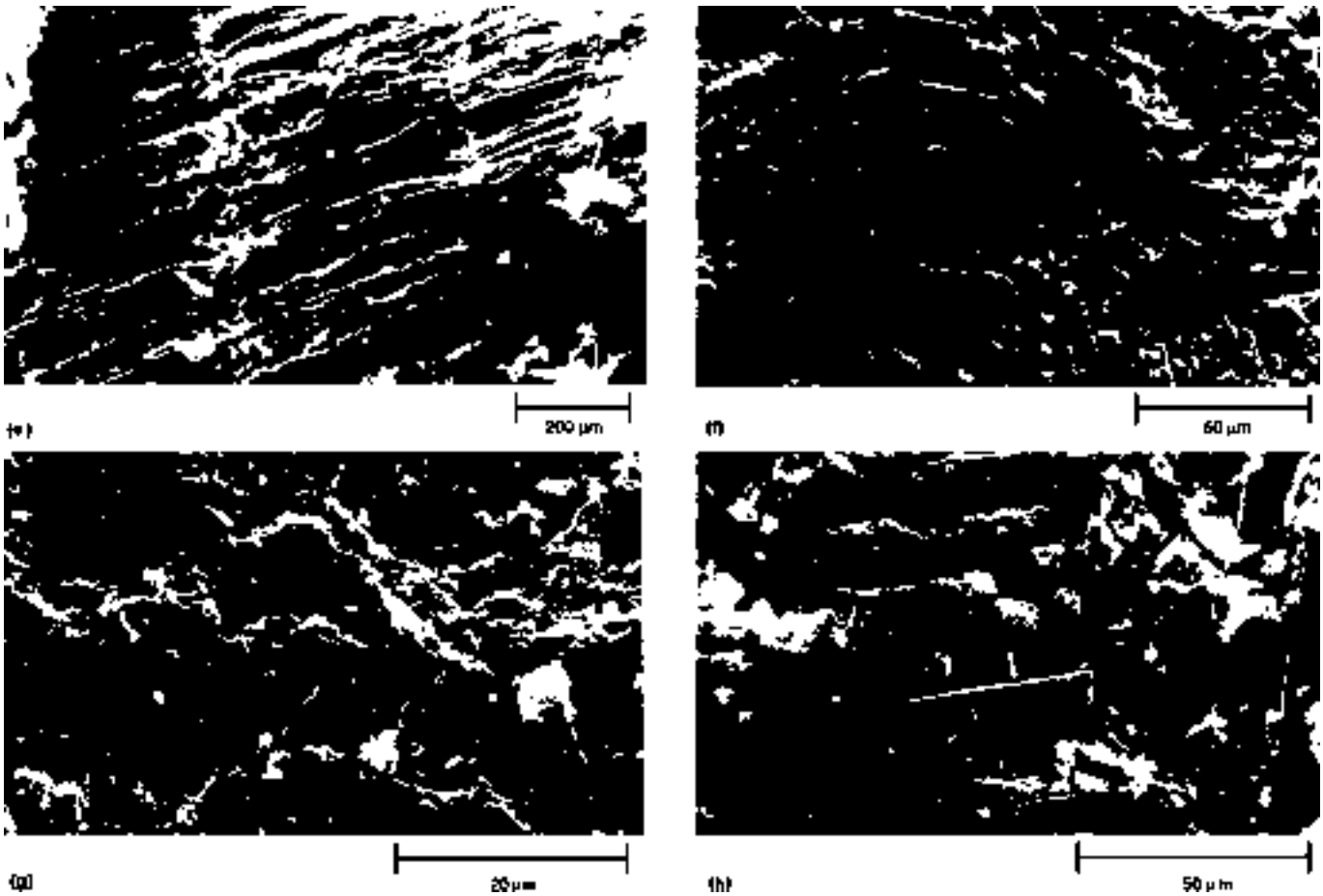


Fig. 10 (continued) (e) PEI_{HY} (O_N) showing excessive melting of AF and third body abrasion due to loose grit (middle portion) on the softened matrix. (f) PEI_{CF} (O_P) excessive breakage of an array of CF in both the directions resulting in high wear. (g) PEI_{CF} (O_N) CF tips showing less fiber damage (and hence less wear); cavities due to fiber pull out. (h) PEI_{G3} (O_P) excessive damage to GF in both the directions due to microcutting. Source: Ref 36

References cited in this section

2. K-H. Zum Gahr, *Microstructure and Wear of Materials, Tribology Series 10,* Elsevier, Amsterdam, 1987

8. U.S. Tewari and J. Bijwe, Tribological Behavior of Polyimides, *Polyimides: Fundamentals and Applications,* M.K. Ghosh and K.L. Mittal, Ed., Marcel Dekker Inc., 1996, p 533–586

12. K-H Zum Gahr, Wear by Hard Particles, *Tribol. Int.,* Vol 31, 1998, p 587–596

13. S.V. Prasad and P.D. Calvert, Abrasive Wear of Particle Filled Polymers, *J. Mater. Sci.,* Vol 15, 1980, p 1746–1754

14. M.M. Khruschov, Principles of Abrasive Wear, *Wear,* Vol 28, 1974, p 69–88

15. K-H Zum Gahr, *Proc. International Conf. on Wear of Materials,* Vancouver, American Society of Mechanical Engineers, 14–18 April 1985, p 45–58

16. W. Simm and S. Freti, Abrasive Wear of Multiphase Materials, *Wear,* Vol 129, 1989, p 105–121

17. S. Bahadur and D. Gong, Formulation of the Model for Optimal Proportion of Filler in Polymer for Abrasive Wear Resistance, *Wear of Materials,* American Society of Mechanical Engineers, 1991, p 177–185

18. J.K. Lancaster, Friction and Wear, *Polymer Science: A Material Science Handbook,* A.D. Jenkins, Ed., North Holland, 1972

19. U.S. Tewari, J. Bijwe, J.N. Mathur, and I. Sharma, Studies on Abrasive Wear of Carbon Fiber (Short) Reinforced Polyamide Composites, *Tribol. Int.,* Vol 25, 1992, p 53–60

20. J.P. Giltrow, A Relationship Between Abrasive Wear and the Cohesive Energy of Materials, *Wear,* Vol 15, 1970, p 71–78

21. M. Vaziri, R.T. Spurr, and F.H. Stott, An Investigation of the Wear of Polymeric Materials, *Wear,* Vol 122, 1988, p 329–342

23. J.J. Rajesh, J. Bijwe and U.S. Tewari, Influence of Fillers on Abrasive Wear of Short Glass Fiber Reinforced Polyamide Composites, *J. Mater. Sci.,* Vol. 36, 2001, p 351–356

24. B.J. Briscoe, P.D. Evans, and J.K. Lancaster, The Influence of Debris Inclusion on Abrasive Wear Relationships of PTFE, *Wear,* Vol 124, 1988, p 177–194

25. B.J. Briscoe, and P.D. Evans, The Influence of Asperity Deformation Condition on the Abrasive Wear of γ-Irradiated Polytetrafluoroethylenes, *Wear,* Vol 133, 1989, p 47–64

26. C. Lhymn, K.E. Templemeyer, and P.K. Davis, Abrasive Wear of Short Fiber Composites, *Composites,* Vol 16 (No. 2), 1985, p 127–136

27. M. Cirino, K. Friedrich, and R.B. Pipes, Evaluation of Polymer Composites for Sliding and Abrasive Wear Applications, *Composites,* Vol 19 (No. 5), 1988, p 383–392

28. M.K. Omar, A.G. Atkins, and J.K. Lancaster, The Role of Crack Resistance Parameters in Polymer Wear, *Physica D: Appl. Phys.,* Vol 19, 1986, p 177–195

29. J. Bijwe, J. Indumathi and A.K. Ghosh, Evaluation of Engineering Polymeric Composites for Abrasive Wear Performance, *J. Reinf. Plast. Compos.,* Vol 18 (No. 17), 1999, p 1573–1591

30. C. Lhymn, Effect of Normal Load on Specific Wear Rate of Fibrous Composites, *Wear,* Vol 120, 1987, p 1–27

31. K. Friedrich, Wear of Reinforced Polymers by Different Abrasive Counterparts; *Friction and Wear of Polymeric Composites, Composite Materials Series 1,* K. Friedrich, Ed., Elsevier, 1986, p 233–287

32. M. Cirino, R.B. Pipes, and K. Friedrich, The Abrasive Wear of Continuous Fiber Reinforced Polymer Composites, *J. Mater. Sci.,* Vol 22, 1987, p 2481–2492

33. M. Cirino, K. Friedrich, and R.B. Pipes, The Effect of Fiber Orientation on the Abrasive Wear Behavior of Polymer Composite Materials, *Wear,* Vol 121, 1988, p 127–141

34. B.K. Yen and C.K.H. Dharan, A Model for Abrasive Wear of Fiber Reinforced Polymer Composites, *Wear,* Vol 195, 1996, p 123–127

35. J. Bijwe, J. Indumathi, and A.K. Ghosh, On the Abrasive Wear Behavior of Fabric Reinforced Polyetherimide Composites, *Wear,* in press

36. J. Indumathi, “Friction and Wear Studies on Polyetherimide and Composites”, Ph.D. dissertation, Indian Institute of Technology Delhi, New Delhi, 2000

J. Bijwe, Indian Institute of Technology, Delhi

Sliding (Adhesive) Wear Failure of Polymer Composites

The polymer composites that are used for sliding wear applications such as bush bearings, bearing cages, slides, gear seals, and so forth in industries such as textile, food, paper, pharmaceutical and such are particulate-filled, fiber reinforced, or mixed composites. In the case of particulate fillers, the particle size is very important for achieving desired performance. The fillers such as ZrO_2, SiC, and so forth are known for their hardness and beneficial effects on abrasive wear resistance of a composite. However, they have a detrimental effect on adhesive wear of polymer composites. Interestingly, the same fillers have proved to be very beneficial when the size is in nanometers (Ref 37, 38, 39, 40).

The performance of FRP composites depends on the type of fiber and matrix, volume fraction, distribution, aspect ratio, alignment, and adhesion to the matrix. As per Eq 7, the higher the aspect ratio (l/r, where l and r are the length and radius of fiber, respectively), the greater is the contact load transferred from the matrix to the fiber and the greater the wear resistance, W_R (inverse of wear rate) (Ref 41):

$$\sigma_f = 2\tau l r^{-1} + \sigma_m$$

where σ_f is the contact stress, σ_m is the compressive stress of the matrix in the composite loaded against counterface under a load L, and τ is the tangential stress produced because of the difference in the moduli of matrix and fiber. It is, however, not true that with increase in the concentration of fibers, wear resistance increases continuously. In fact, either it deteriorates or becomes constant beyond a typical optimum concentration in the case of short fibers.

Generally, short fibers (0.1 mm to 3 mm, or 0.004 to 0.118 in.) in approximately the 20 to 30% concentration range are used for reinforcement and reducing wear in thermoplastics.

Figure 10 (Ref 2) highlights some trends generally observed in the wearing of composites against a smooth metal. Increase in load and speed results in higher wear of FRP through different mechanisms. High load results in more fiber cracking and pulverization leading to deterioration in load carrying capacity, while high speed accelerates the debonding of fibers/fillers. This results in easy peeling off or pulling out of the reinforcing phase. High-modulus fibers are more effective in wear reduction than the high-strength fibers. Moreover, the higher the modulus of fiber or composite, the less is the wear. The higher the aspect ratio of the reinforcement, the lower is the wear. For a typical volume fraction (V_f) of a filler, there exists an optimum value of the mean free path for the minimum wear; the same is true for the filler size. High-strength and high elastic modulus of the matrix enhance the support to the fillers. The higher the brittleness of the matrix is, the higher the crack propagation tendency and the higher the wear of a composite.

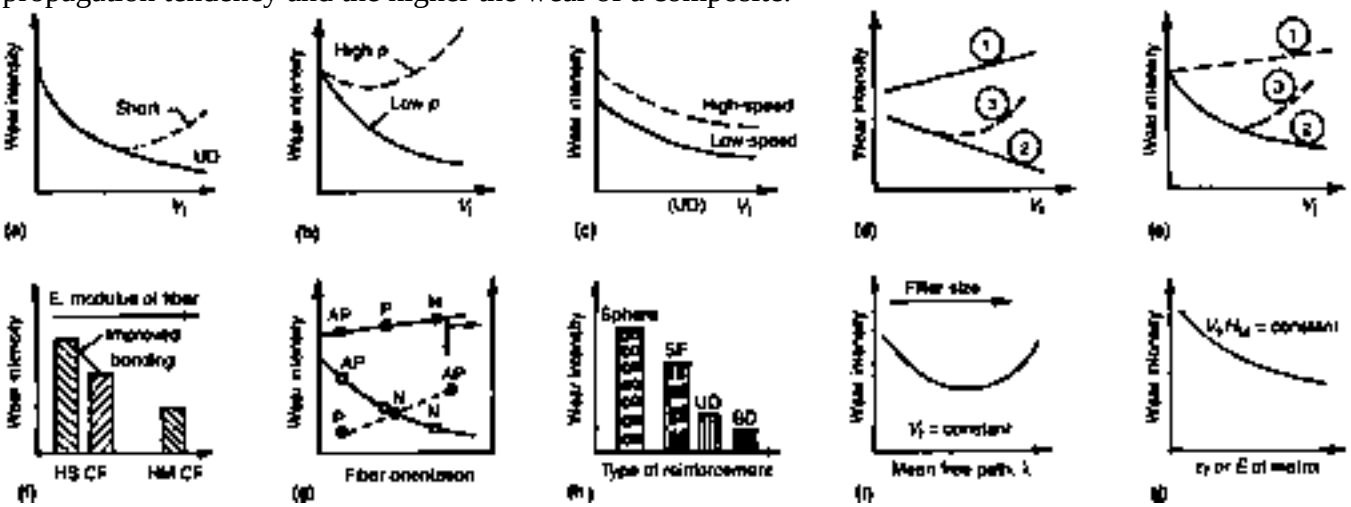


Fig. 10 General trends indicating effect of microstructure of a composite and the properties of fillers on adhesive wear of composites. *p*, applied pressure; H_M, hardness of matrix. AP, P, and N refer to orientations of fibers with respect to sliding direction: AP, antiparallel; P, parallel; and N, normal. HS CF and HM CF, (high-strength and high-modulus carbon fibers); ε_f, fracture strain; and *E*, Young's modulus. 1, abrasive of larger size; 2, nonabrasive filler/solid lubricant/abrasive filler in nanometer size/long fibers or fabric; 3, short fibers. Source: Ref 2

Apart from these factors, the most important wear-controlling factor in the case of a polymer composite is the efficiency of interaction of the polymer and filler with the counterface. If the interaction reinforces the thin-film transfer efficiency of the polymer, the friction coefficient (μ) and the extent of frictional heating reduce, leading to less damage to the polymer, filler, and their adhesion. This generally results in less wear. The nature of transferred film on the counterface plays a key role in controlling wear performance of a composite. If this adheres to the counterface firmly, wear of the composite decreases. If the fillers are capable of enhancing this adhesion by forming chemical bonds through chemical or physical interaction with the counterface during sliding, the film is more firmly attached, resulting in significant reduction in wear.

Adhesive Wear of Particulate-Filled Composites. The influence of nanometer-sized particulate fillers such as SiC, ZrO_2 and Si_3N_4 has been studied in PEEK (Ref 37, 38, 39, 40). At particular concentrations and sizes of the filler, minimum wear rate (K_0) and minimum μ were observed. However, minima of K_0 and μ did not always match. For example, these matched in the case of ZrO_2 at 7.5% (K_0 decreased by 1.8 times and μ by 1.3 times). For Si_3N_4, K_0 decreased continuously with increased filler amount. The μ, however, was minimal at 8 vol% loading (1.5 times decrease in μ and 7 times decrease in K_0). In a further work on simultaneous addition of SiC filler (3.3 vol%, which proved to be the best for maximum reduction in the wear rate and μ) and solid lubricant polytetrafluoroethylene (PTFE) in increasing amounts up to 40 vol%, a synergistic effect was observed.

Figure 11 (Ref 40) shows the influence of PTFE on μ and K_0 of PEEK and PEEK + SiC (3.3% vol%) composite. It was interesting to observe that at up to 5% PTFE loading in a PEEK-SiC composite, μ rose to a high value showing a negative contribution of PTFE towards μ. Without PTFE, μ was quite low Fig 11b. Beyond 10% loading, a combination of PTFE and SiC showed synergism, and μ lower than that with the individual fillers was observed. Thus this combination showed antagonism and synergism in particular ranges. K_0, however, was lowest without PTFE in the PEEK-SiC composite. Combination of PTFE with SiC thus showed a negative effect for wear behavior. Interestingly, inclusion of the single filler PTFE proved beneficial. These studies brought out a very important aspect of the influence of solid lubricant. It worsened the performance of a composite when combined with other potential fillers. This detrimental effect was due to the formulation of SiF_x confirmed during x-ray photoelectron spectroscopic (XPS) analysis, due to chemical reaction between SiC and PTFE. This SiF_x was responsible for raising the μ of a composite. When PTFE contents exceeded the amount required for the chemical reaction with 3.3 vol% SiC, the excess unreacted PTFE started showing a beneficial effect and μ started decreasing. The film transfer efficiency of the filler was poor when PTFE and SiC were in combination, and this led to deterioration in wear performance.

Fig. 11 Influence of fillers on friction and wear behavior of PEEK composites; *L*, Normal load, 196 N; speed 0.445 m/s; counterface plain carbon steel ring. (a) nanometer-sized SiC in PEEK; (b), and (c) PTFE in PEEK and PEEK + SiC (3.3 vol% constant) composites. Source: Ref 40

The influence of increasing the amount of PTFE in PEEK (Ref 42, 43) (Fig. 12a) showed that the PTFE significantly benefited both μ and the specific wear rate of PEEK. However, with an increase in PTFE though μ decreased continuously, K_0 showed excessive increase beyond 80% loading of PTFE. K_0 was minimal at 5% PTFE contents while μ was minimal for 100% PTFE. A 12 to 18% loading range of PTFE was found to be optimal for the friction and wear combination.

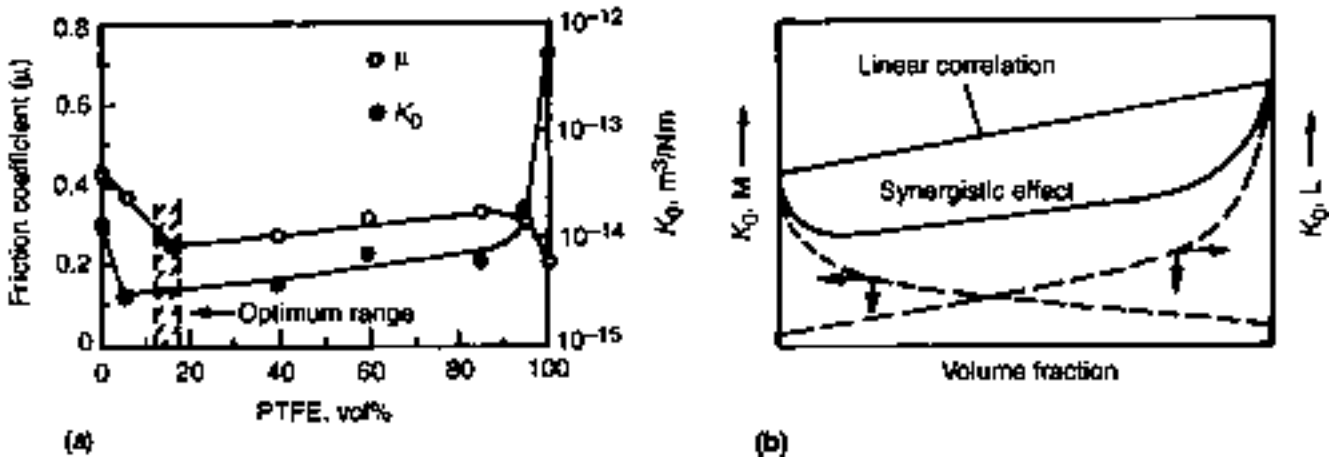


Fig. 12 (a) Influence of PTFE on friction and wear performance of PEEK composites and the optimum range of PTFE amount for best combination of μ and K_0. (b) Linear correlation and synergistic effect as a result of two opposite trends. $K_{0,M}$ and $K_{0,L}$ represent specific wear rates of matrix and lubricant (PTFE). Source: Ref 42, 43

Wear rates (W_C) of PEEK-PTFE composite have been described (Ref 42, 43) as follows:

$$W_C = (1 - V_{fL})W_M + V_{fL}W_L$$

where V_{fL} is the volume fraction of lubricant PTFE, and W_M and W_L are the wear rates of the PEEK matrix and lubricant PTFE. Figure 12(b) shows the synergistic effect of the lubricant on wear. This could not be explained with the help of linear correlation. Such synergism could be described as:

$$\frac{1}{W_C} = (1 - V_{fL})\frac{1}{W_M^*} + V_{fL}\frac{1}{W_L^*}$$

where W_M^* indicates wear rate in the presence of PTFE.

$$W_M^* = W_M \cdot f$$

where f is the lubricating efficiency factor and W_M^* is the effective wear rate of the matrix (Ref 4).

Adhesive Wear of SFRP or Mixed (SFRP + Particulate Filled) Composites. Short fibers are very effective in modifying the wear performance and friction behavior (except in the case of glass fibers) of the composite. The combination of fibers and lubricants usually shows synergism. Inclusion of short glass fibers benefited the polyphenylene sulfide (PPS) maximum and the PEEK minimum (Ref 4). Beyond typical fiber loading, extent of improvement slowed down for polyethernitrile (PEN) and polyether sulfone (PES) but not for polyphenylene sulfide (PPS). Similarly, the type of carbon fiber and its volume fraction in polyethernitrile (PEN) influenced the wear behavior significantly. Pitch-based CF proved more beneficial than the polyacrylonitrile (PAN)-based. Beyond 15% loading, the extent of improvement, however, was marginal (Ref 4). Results of a study on the short-fiber and solid-lubricated composites are shown in Fig. 13(a), while Fig. 13(b) shows the influence on pressure × velocity (PV) factor and high temperature on the wear rate of composites. Figure 13(b) indicates that unreinforced polybenzimidozole (PBI) is far superior to the reinforced and lubricated PEEK. Systematic and step-by-step inclusion of the lubricant and reinforcement (short glass fibers) in PEI could improve friction and wear behavior very significantly as seen in Fig. 14 (Ref 8). Polyetherimide is a hard and ductile polymer. It did not transfer any film on the counterface but did transfer a molten material in the severe *PV* condition. The thin coherent film of PTFE transferred on the mild steel counterface (Fig. 15a) in the case of $PEI_{PTFE\ 15\%}$ composite was responsible for the lowest μ in the series of composites. For the composite containing three lubricants and short glass fibers, the film transfer was not as coherent and thin (Fig. 15b), as in the earlier case, and μ was little

higher. Various wear failure mechanisms observed on the worn surfaces of composites during scanning electron microscopic (SEM) studies are shown in Fig. 16 16 (Ref 44, 45, 46, 47, 48).

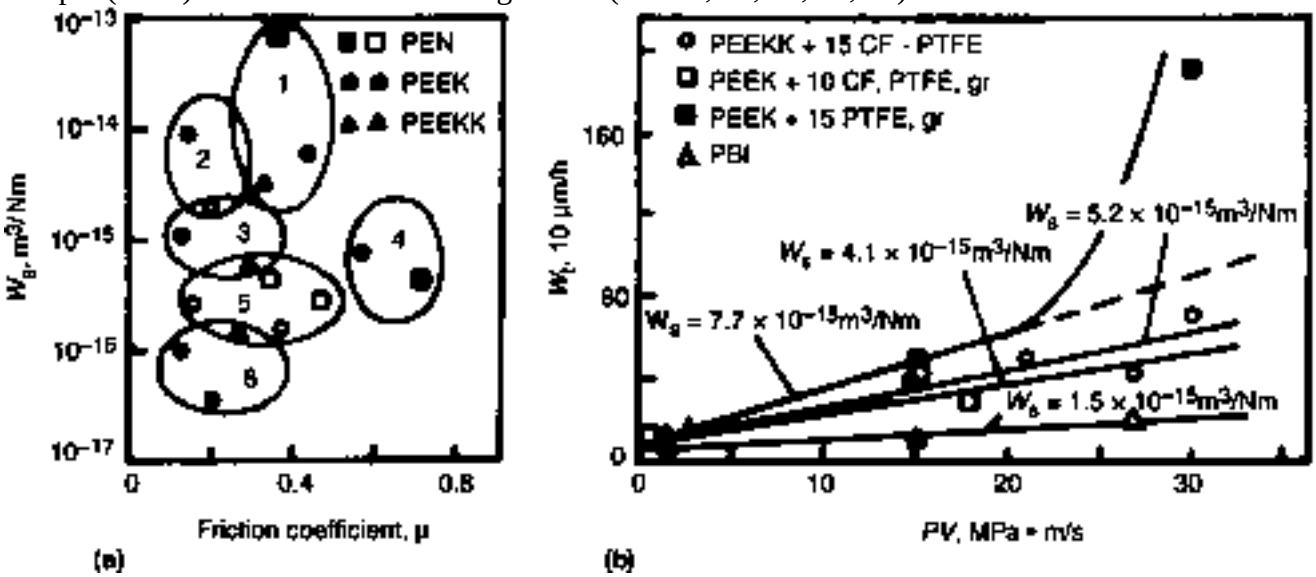


Fig. 13 (a) Indicative trends in influence of reinforcement and solid lubrication on friction and wear of high-temperature polymers. *P* = 1 MPa; *V* = 1 m/s. PEN, polyethernitrile; PEEK, polyetheretherketone; PEEKK, polyetheretherketoneketone; gr, graphite; TF, Teflon. 1, neat polymers; 2, polymers + PTFE; 3, polymers + graphite/PTFE; 4, polymers + glass fibers (GF); 5, polymers + carbon fibers (CF); and 6, polymers + CF/GF + PTFE. (b) Influence of pressure × velocity (PV) factor on wear rate of fiber-reinforced plastics (*T*, 220 °C; V, 3 m/s). Source: Ref 4

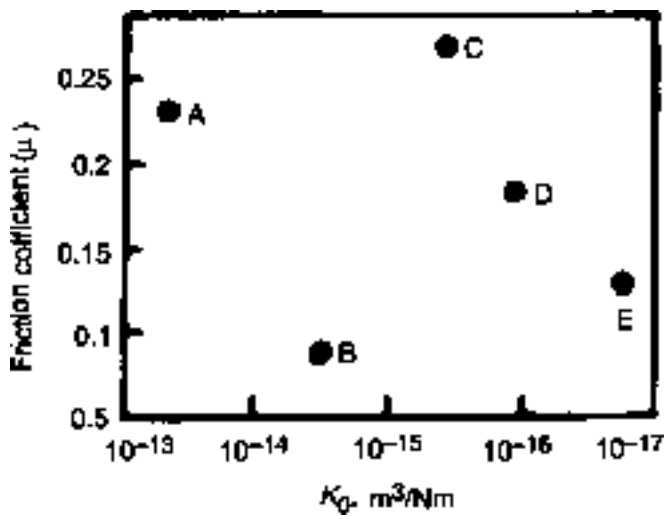


Fig. 14 Influence of inclusion of fillers (individually and simultaneously) on friction and wear performance of PEI composites against mild steel (L 43 N, speed 2.1 m/s, for A 25 N counterface mild steel); A, polyetherimide (PEI); B, ($PEI_{PTFE15\%}$); C, ($PEI_{GF20\%}$); D, ($PEI_{GF\ 16\%+graphite\ 20\%}$); and E, ($PEI_{CF25\%+PTFE15\%+(MoS}$)

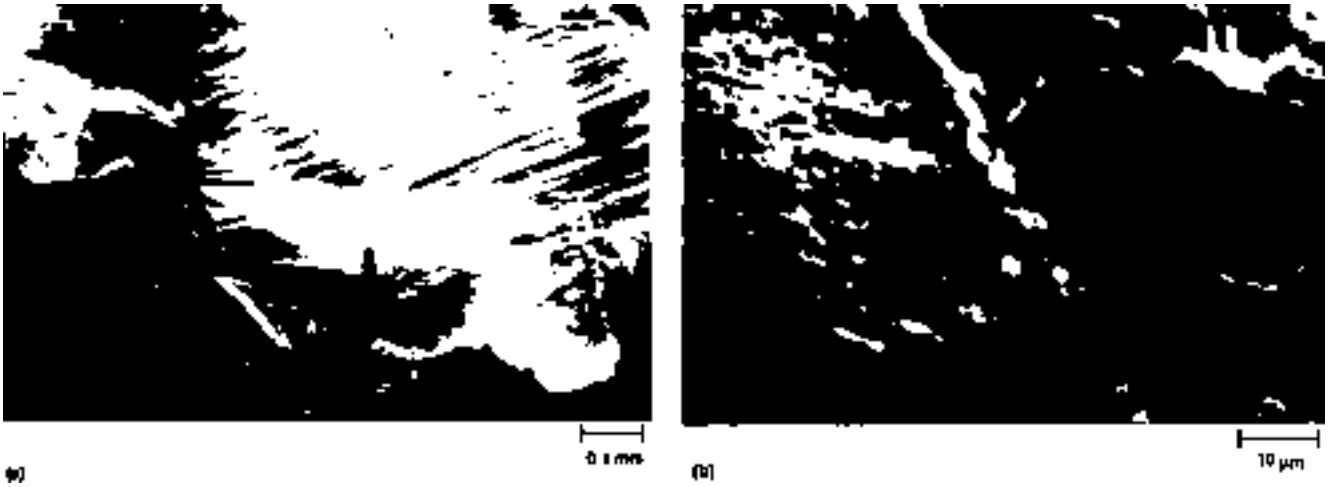


Fig. 15 Scanning electron micrographs of worn surfaces of PEI composites indicating (a) transfer of thin and coherent film of PTFE on the steel disc responsible for lowest μ exhibited by ($PEI_{PTFE15\%}$). (b) Film transfer (less coherent and thin) in the case of ($PEI_{CF25\%+PTFE15\%+MoS_2}$) responsible for slightly higher μ than the $PEI_{PTFE15\%}$. Source Fig. 15(a): Ref 44. Source Fig. 15(b): Ref 45

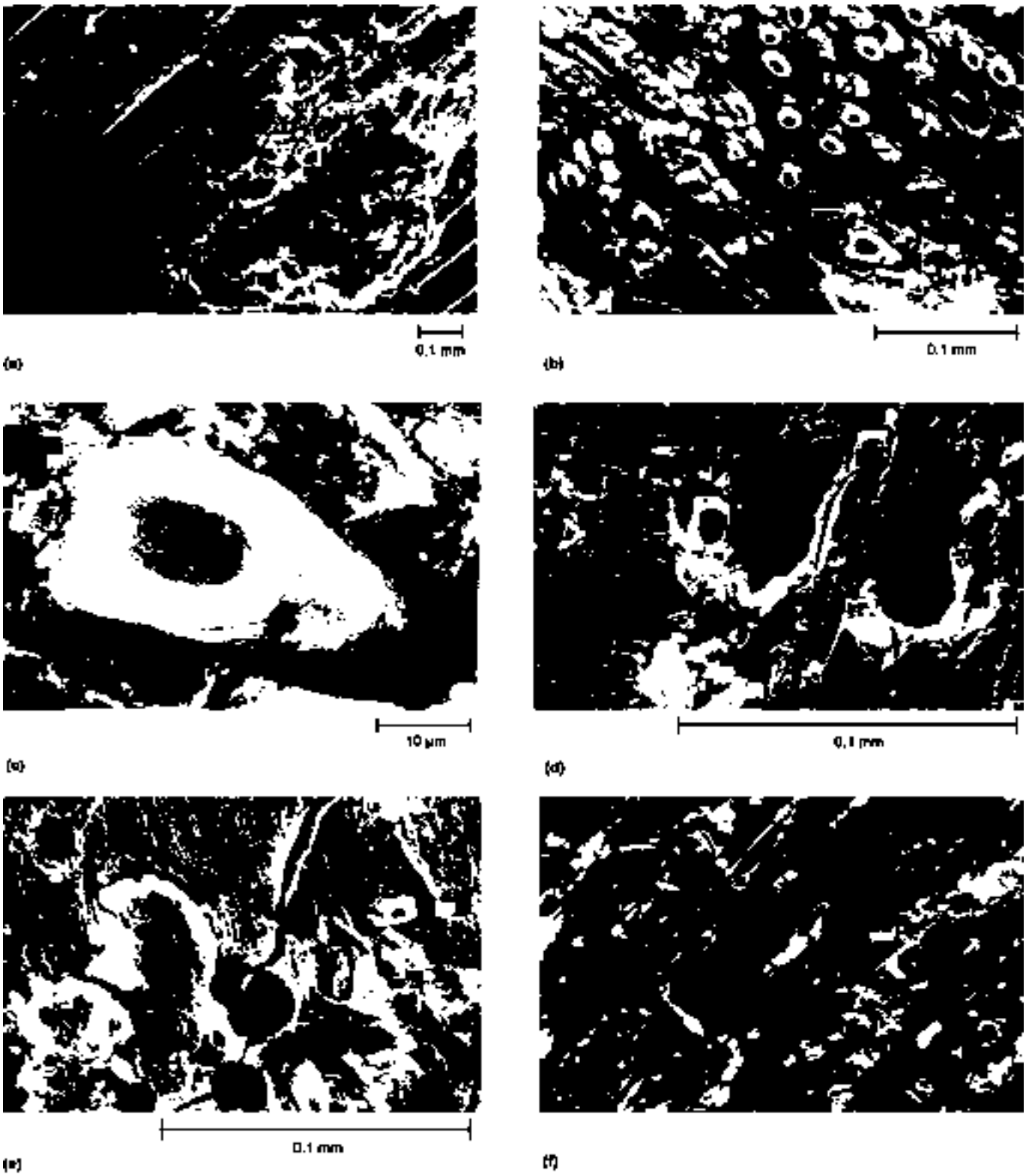


Fig. 16 Wear failure of PEI and composites (a) Failed surface of PEI while sliding against very smooth (R_a 0.06 µm) aluminum surface resulting in high µ (L 28 N; v 2.1 m/s) Left part shows severe melt flow of PEI; middle portion shows crater with chipped-off molten material (Ref 46). (b–e) Worn surface of PEI_{GF+gr} (*L*, 112 N; *V*, 2.1 m/s). (b) Severe melt flow of polymer in sliding direction with maximum fibers normal to the surface, cracks generated in sliding direction, and a pulled out fiber, (c) Magnified view of pulled-out fiber from the matrix with worn elliptical and polished tip with excessive fiber-matrix debonding aggravating wear of composite. (d) Multiple parallel microcracks perpendicular to the sliding direction indicating fatigue with cavities due to fiber consumption, deterioration in fiber matrix adhesion, and wear thinning of longitudinal fiber. (e) Deep cracks initiating and propagating from fiber to fiberith pits formed due to graphite extraction and fiber consumption, back transfer of molten polymer from the disc to the pin surface (patches in the left portion of the micrographs) (*L*, 132 N; V, 1

m/s) (f–h) Worn surfaces of PEI GF+gf+PTFE+MoS (with fibers parallel to the sliding surface), worn under *L*, 72 N and *V*, 2.1 m/s, showing (f) microcracking of fibers, (g) deterioration in the fiber-matrix adhesion and peeled off fiber, and (h) wear thinning of fibers with still more deterioration in fiber-matrix adhesion (*V*, 2.1 m/s; *L*, 132 N). Source: Ref 44, 45, 46, 47, 48

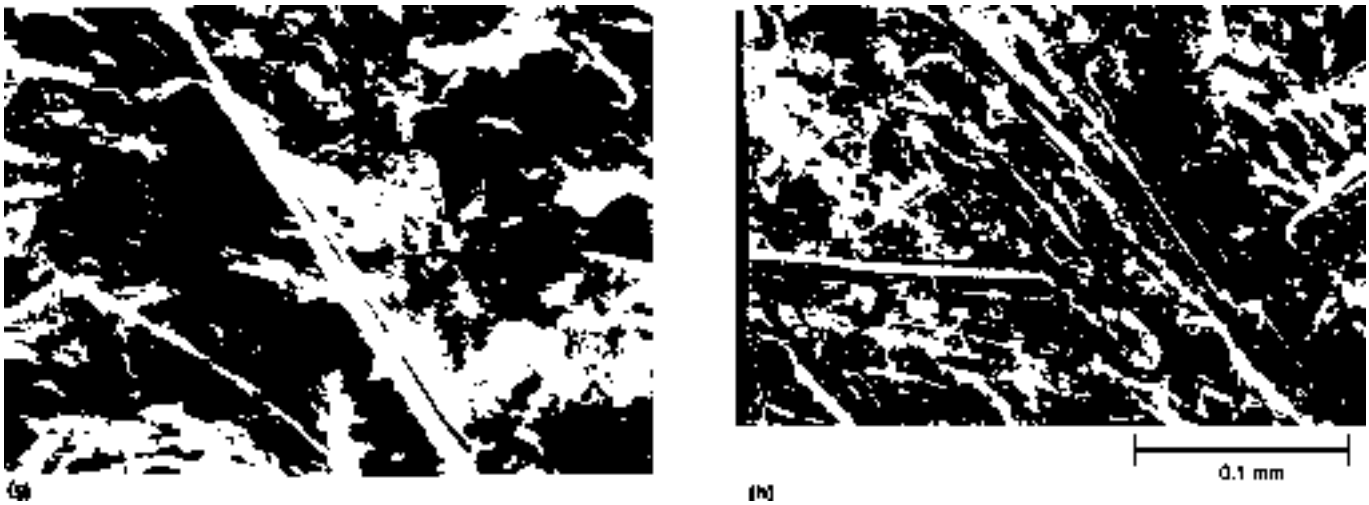


Fig. 16 (continued) (g) deterioration in the fiber-matrix adhesion and peeled off fiber, and (h) wear thinning of fibers with still more deterioration in fiber-matrix adhesion (*V*, 2.1 m/s; *L*, 132 N). Source: Ref 44, 45, 46, 47, 48

A wear model to describe the adhesive wear behavior of SFRPs based on the microscopic observations on the worn surfaces of the FRPs has been developed (Ref 49). Since fiber cracking and fiber-matrix debonding occur sequentially, a combined process can be considered. The fiber debris removed from the matrix can act as third body abrasives and also needs to be included in models. Hence, the sliding wear rate of composite ($W_{s,c}$) is the sum of wear rates that account for the sliding process ($W_{s,s}$) and others, which accounts for the additional wear mechanism ($W_{s.fci}$), the post-sliding wear process; that is, wear due to fiber fracture, fiber-matrix interfacial debonding pulverization, fibrillation, and so forth. Correlation was observed between the experimental and calculated data (Ref 49).

Adhesive Wear of Unidirectional FRP Composites. Pioneering research in this area included investigation of various factors influencing the wear performance of FRP (Ref 18, 41, 50, 51, 52) while others developed the wear model based on in-depth studies on UDFRPs (Ref 53, 54, 55, 56). The tribo-studies were later extended to various composites containing reinforcement with short and continuous fibers and fabrics, (Ref 1, 4, 5, 22, 27, 31, 32, 33, 42, 49, 57, 58). Figure 17 4 summarizes some results of selected UD composites (Ref 54).

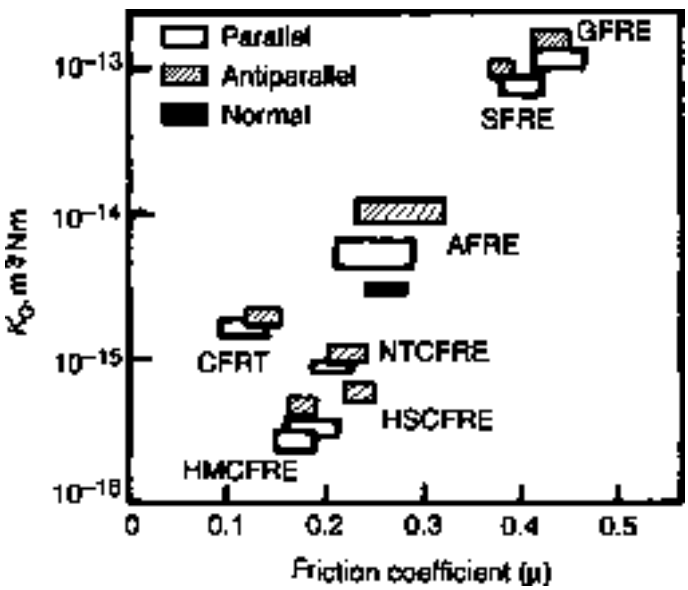


Fig. 17 Specific wear rate and friction coefficient of unidirectional composites (see Table 4) in three orientations (*P*, 1.5 N/mm^2; *V*, 0.83 m/s; distance slid, 16 km).

Table 4 Details of UD composites studied in adhesive wear mode

No.	Reinforcement (UD)	Resin and V_f		
		Epoxy	**Polyester**	**PTFE/Teflon**
1	High-strength carbon fiber (HS-CFR)	HS-CFR-E (42, 52, 59, 65)	HS-CFR-EST (42, 50, 57, 65)	HS-CFRT (42, 67)
2	High-modulus carbon fiber (HM-CFR)	HM-CFR-E (65)	…	…
3	High-strength carbon fiber NT-CFR	NT-CFR-E (65)	…	…
4	E-glass fiber (GFR)	GFR-E (60, 68, 76)	GFR-EST (52, 58, 70)	…
5	Stainless steel fiber (SFR)	SFR-E (56, 62, 69, 75)	SFR-EST (48, 54, 70, 76)	SFRT (42, 67)
6	Aramid fiber (Kevlar fiber) (AFR)	AFR-E (40, 50, 60, 70)	AFR-EST (70)	AFRT (42, 67)

NT, nontreated. Source: Ref 54

Wear behavior of FRP depends on the properties of fibers, their orientations, and bonding with the matrix and the counterface material, including operating conditions. In the case of a brittle matrix with epoxy, generated cracks propagate right through the fiber if the bonding between fiber and matrix is strong. In the case of highly ductile matrix such as polyurethane (PU), the cracks cannot propagate through the matrix and fiber. The fiber bends with the matrix under the asperity contact, and the wear rate is controlled by the wear rate of the fiber (Ref 6). Analysis of worn surfaces of UD graphite-fiber reinforced PU indicated that the fiber tips (fibers perpendicular to sliding plane), which were originally circular, became elliptical and bent during sliding.

The following mechanisms for wear failure of FRPs sliding against smooth metals under pressure, *p*, have been proposed (Ref 53):

1. Wear thinning of the fiber due to continuous sliding for a distance *D* under load *L*
2. Subsequent breakdown of the fiber due to strain $\mu p/E$ (*E*, modulus of elasticity) of FRP caused by the frictional force, load, and sliding distance
3. Peeling off of the fibers from the matrix because of strain $\mu p/E$ exceeding interlaminar shear strength (I_S)

In-depth studies on UD composites (Ref 5) focused on investigating the influence of type and orientation of reinforcements in the selected matrices on friction and wear behavior. The various failure mechanisms operative in wearing FRP are shown in Fig. 18 (Ref 57). Various finite elemental micromodels have been developed for explaining the failure mechanisms in different fiber orientations based on the evaluation of contact and stress conditions produced by a sliding of hemispherical steel asperity. Various deformed shapes of microstuctures in three orientations have been discussed based on deformation of fibers mainly by compression and bending/shear type loadings (Ref 57).

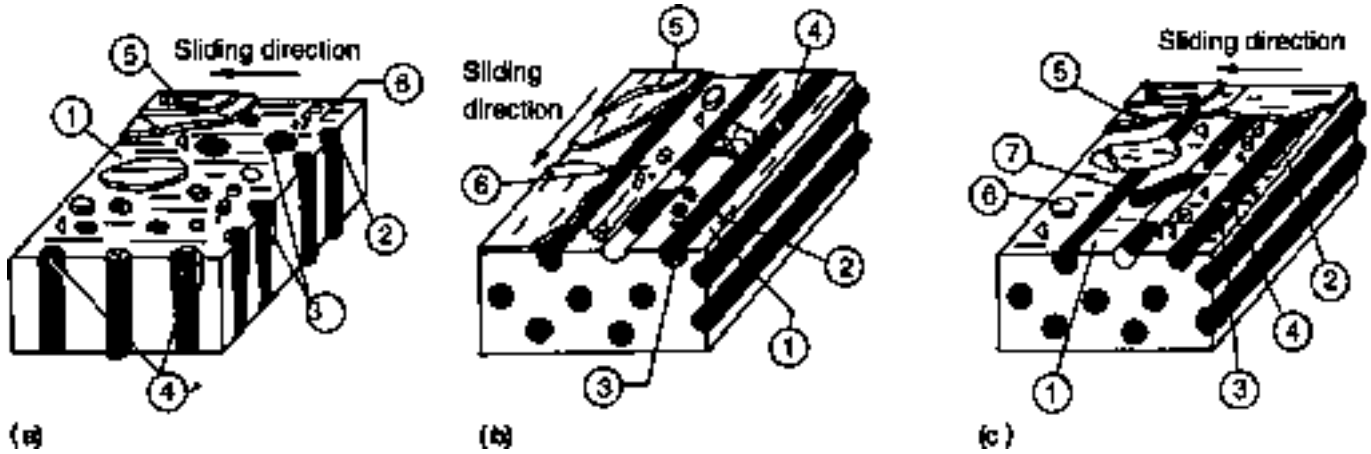


Fig. 18 Failure wear mechanisms in fiber-reinforced polymers sliding with fibers in different orientations. (a) N orientation; (b) parallel orientation; (c) antiparallel orientation. 1, wear failure of matrix by microplowing, microcracking, and microcutting; microplowing; 2, sliding and wear thinning of fibers; 3, interfacial separation of fiber and matrix; 4, fiber cracking; 5, back-transferred polymer or organic fibers (film and layered wear debris) showing delamination and cracking; 6, metallic and wear debris transferred from the counterface; 7, pulled-out or peeled-off fiber pieces

Various wear failure mechanisms evident from SEM studies on the worn surfaces of UD composites of PA66 are shown in Fig. 19 (Ref 5). CF/PA66 (P) composite shows wear failure of CFs parallel to the sliding direction by various mechanisms leading to fiber thinning, cracking, pulverization, and debonding from the matrix. Wearing of AF in O_N led to a smooth surface with little fiber-matrix debonding (Fig. 19b). The surface also showed microcracking (middle portion parallel to the width of the micrograph), delamination and microcracking of the fibers at the edge. When the AF was in the O_P direction, it showed a tendency to be peeled from the surface due to poor wetting to the matrix (Fig. 19c).

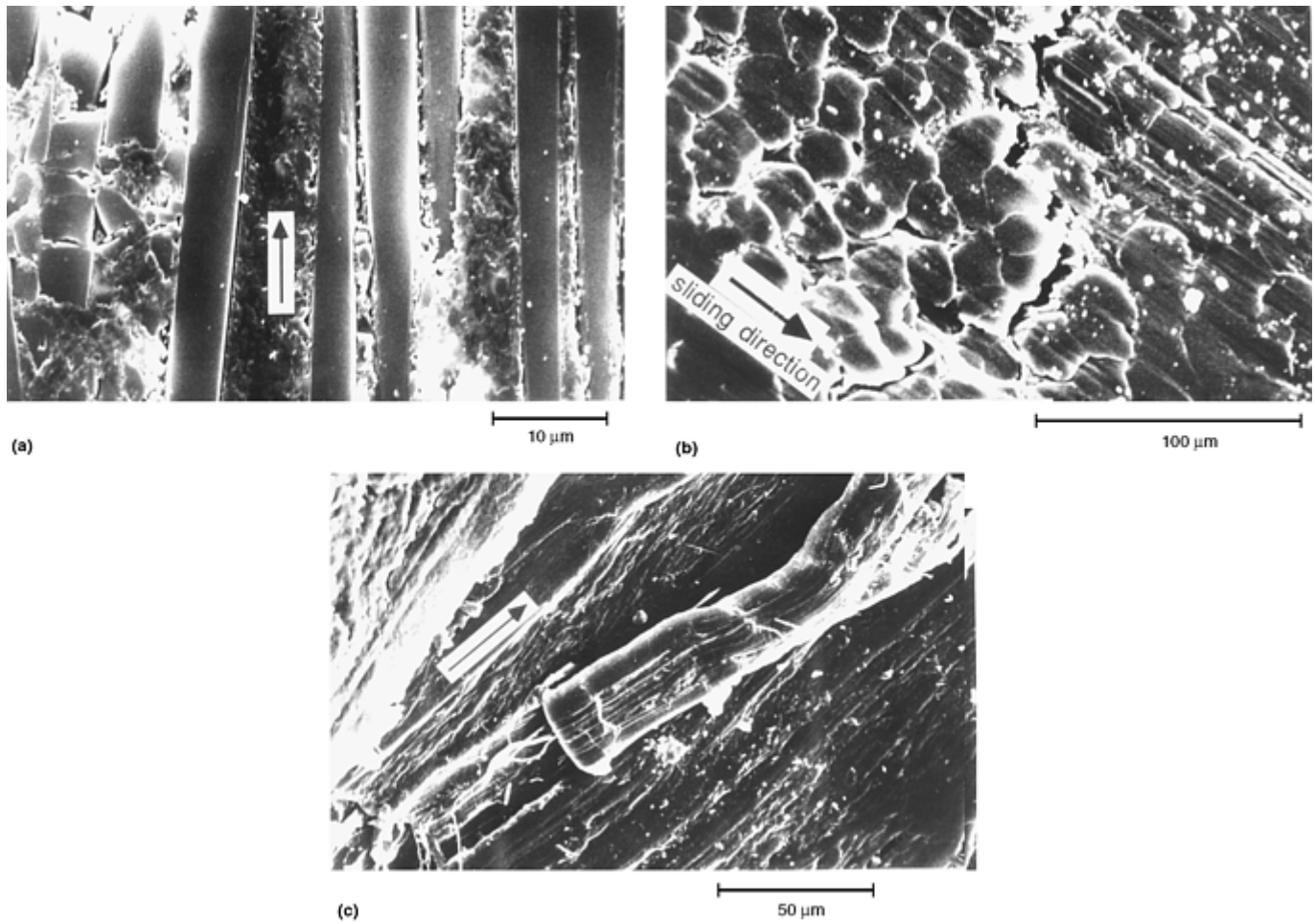


Fig. 19 Scanning electron microscope micrographs of worn surfaces of PA66 unidirectional composites. (a) Carbon fiber (parallel, P) showing fiber thinning, fiber fracture, fiber pulverization (left portion) and fiber matrix debonding (middle portion). (b) Aramid fiber (AF) in the normal orientation showing fiber cracking (edge); (c) AF(P) showing pull-out of aramid fiber. Source: Ref 5

Adhesive Wear of Fabric Reinforced Composites. Modi and others (Ref 58) compared the friction and wear performance of GF-PEEK (UD) composite and graphite fabric (five-harness satin weave) PEEK (BD) composite and concluded that the wear rate of BD composite was lower than that of the UD composite by an order of one. Friction behavior was also better for the BD rather than the UD composite. The temperature sensitivity of the former was remarkably lower than that of the latter.

Adhesive Wear of Hybrid Composites. The tribology of composites reinforced with continuous fibers of two types in different proportions in epoxy matrix was investigated (Ref 53). The wear behavior of hybrid UD composites containing fibers of glass and carbon in epoxy composite was also studied (Ref 59). Figure 20 shows that the wear rates of these composites were lower than the values expected from the LROM equation but higher than the minimum values indicated by the IROM. The dotted curve shown in the Fig. 20 is for the calculated values as per an equation (developed in Ref 59), which fit reasonably well. However, the model did not consider the possibility of mutual interaction of the constituents causing a deviation in the wear resistance of a hybrid composite based on the rule of mixture calculation. The practical application of such composite was justified on the basis of performance to cost ratio.

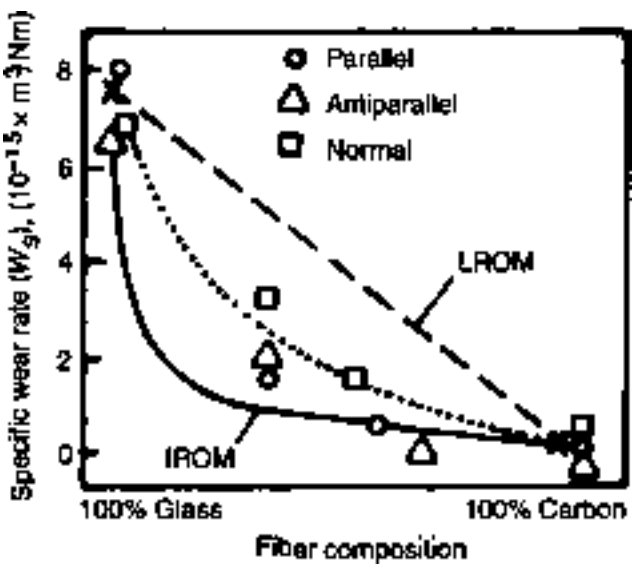


Fig. 20 Specific wear rate as a function of fiber composition in hybrid composite (*L* 93 N, velocity *V*) 0.5 m/s, nominal V_f 0.57 with dotted curve for calculated values as per equation in Ref 59. IROM, inverse rule of mixture; LROM, linear rule of mixture. Source: Ref 59

Various hybrid composites based on three matrices (namely amorphous polyamide (PA), PA 66, and epoxy) containing three reinforcements as shown in Table 5 were tailored. (Ref 5). Among various investigations on these composites, behavior of just one composite (AF-CF-PA66) is shown in Fig. 21 (Ref 5). The stacking sequence for the sandwich hybrids (namely the composite with CF placed in the surface layer and AF in the core) was an important influencing factor and proved to be superior to CF in the core and AF in the surface layer. Thus the positive hybrid effect was found in the former case. In the latter, wear behavior was in accordance with a linear correlation between two limits. Fig. 22 highlights schematics of various wear failure mechanisms operative in the wearing of such BD and UD composites (Fig. 22 a, b). Studies by SEM on the worn surfaces of BD hybrid composites (sandwich structure) are shown in Fig. 23. The following favorable interactions could be seen on the surfaces of the hybrid composites:

- The probability of termination of cracks by the strong barrier of CFs was high when cracks were generated in the core of AF due to poor wetting.
- Stiffer CFs were better bonded to the matrix and prevented edge delamination and fibrillation of AFs.

- Third-body formation due to transferred material consisting of fiber debris separates the counterface and contributes to the load carrying capacity. Thus, debris of AF O_N in the core, especially in the vicinity of the CF/AF interfacial region, stayed there temporarily. This third body decelerated the wear process further.

Table 5 Details of hybrid composite

Matrix	No.	Designation	Total V_f $V_{f1} + V_{f2}$	f_1/f_2, %	Hybrid type	Sliding direction	Hybrid efficiency factor	W_s reduction at f_1/f_2 = 50/50[(a)]
Am PA	1	CF(0)-AF(90)-CF(0)	59–65	V	S	P-N-P	>0	21%
	2	AF(0)-CF(90)-AF(0)	61	50/50	S	N-P-N	0	…
	3[(a)]		61	50/50	S	P-P-P	<0	22%
	3(b)	CF(0)-AF(0)-CF(0)				N-N-N	>0	27%
PA 66	4	CF(0)-AF(90)-CF(0)	20–40	V	S	P-N-P	>0	45%
	5	CF(0)-AF(90)-CF(0)	20–40	V	L	P-N-P	>0	64%
	6[(a)]		25	50/50	S	N-P-N	0	…
	7[(a)]	CF(0)-GF(90)-CF(0)	35	50/50	S	P-N-P	<0	135%
	7(b)	GF(0)-AF(90)-GF(0)				AP-N-AP		127%
EP_2	8[(a)]	CF(0)-AF(0)-CF(0)	62	50/50	S	P-P-P	<0	63%
	8(b)					N-N-N	>0	36%

V, variable; S, sandwich; L, layer structure.
(a) $W_{S\ red}$ expressed in % of rule of mixtures value $W_{S\ f1/f2}$.
Source: Ref 5

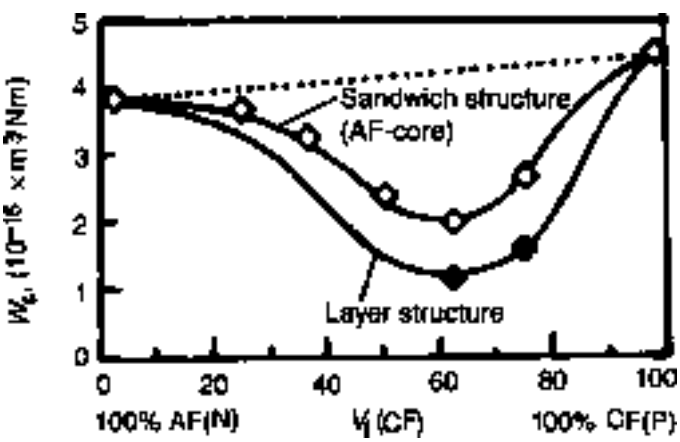


Fig. 21 Specific wear rates of hybrid composites formulated by two structures, sandwich and layer, (composite aramid fiber-carbon fiber polyamide am). Source: Ref 5

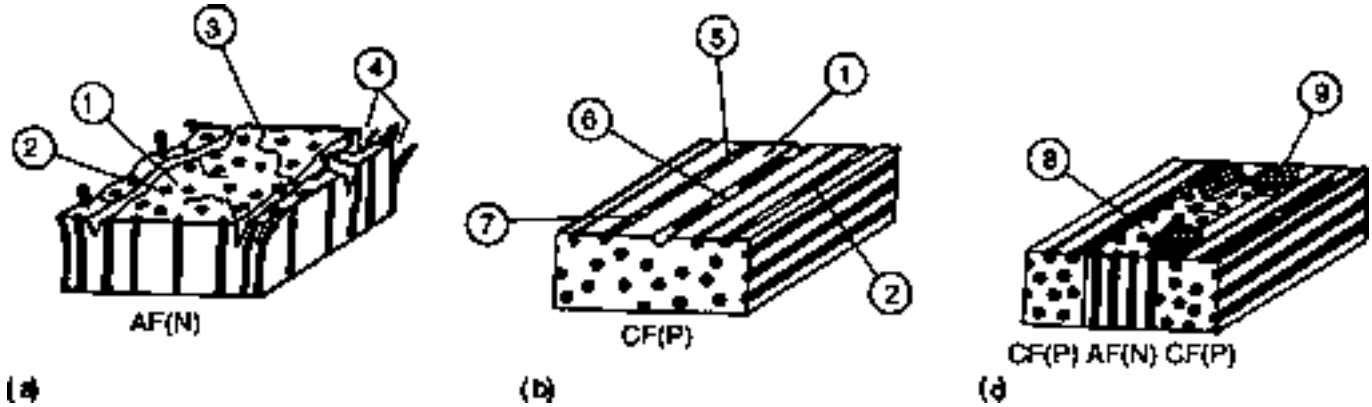


Fig. 22 Failure wear mechanisms of unidirectional fiber reinforced polymer composites with different orientations of fibers with respect to sliding direction against a smooth metal surface. (a) Normal aramid fibers. (b) Parallel carbon fibers. (c) Wear reduction mechanism due to hybridization. 1, wear of matrix by plowing, cracking, cutting as a result of plastic deformation; 2, wear thinning of fiber or tip resulting in elliptical, well-polished tip; 3, matrix cracking; 4, edge delamination and fiber fibrillation; 5, fiber cracking; 6, pulverized fiber wear debris; 7, deterioration in fiber matrix adhesion, fiber pulverization, pullout, and peeling off followed by removal; 8, inhibition to matrix cracking; and 9, layer of back-transferred film or wear debris. Source: Ref 2

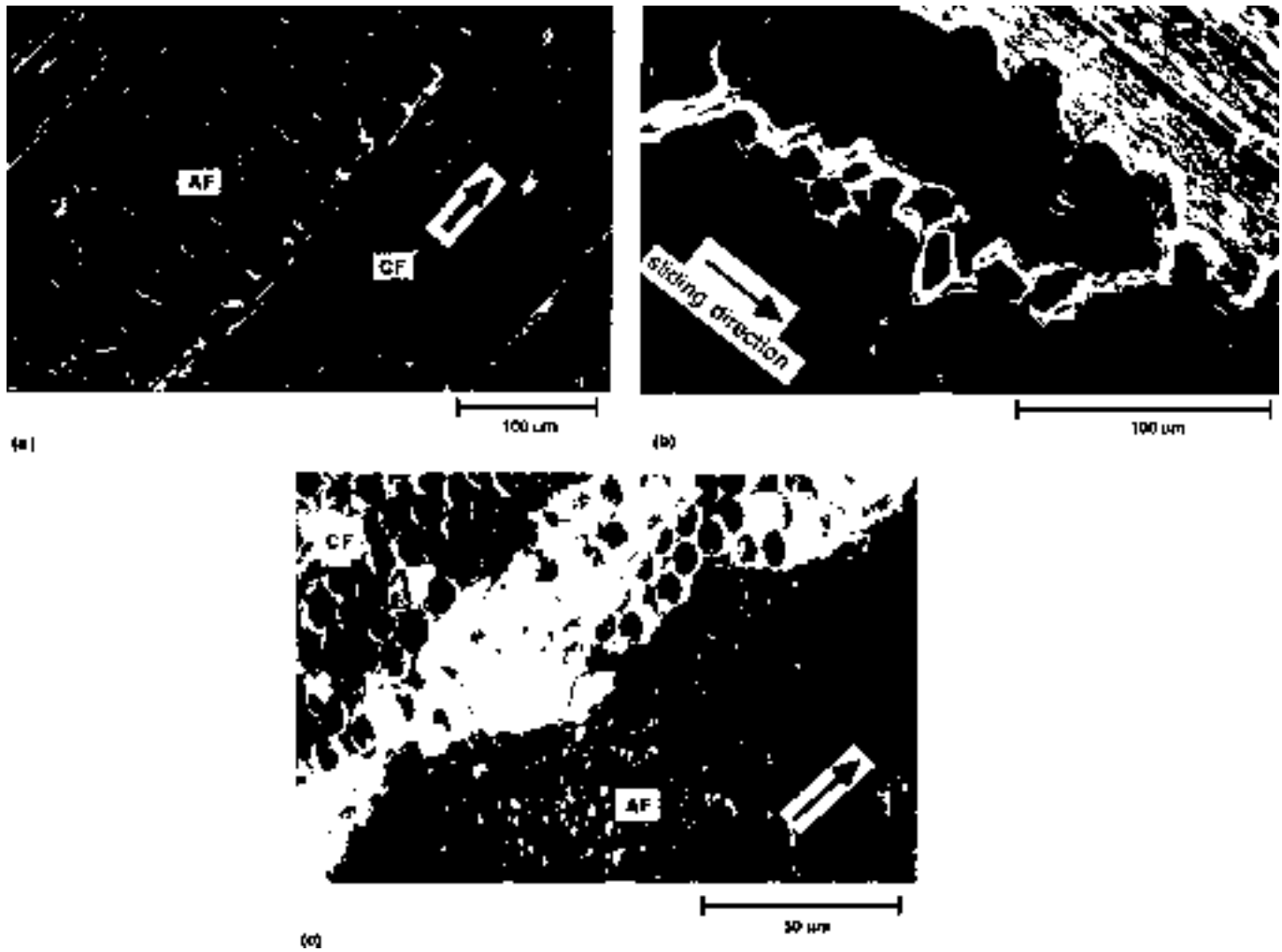


Fig. 23 Scanning electron microscope micrographs of worn surfaces of PA66 hybrid composites. (a) Hybrid (layer) composite-AF(N)/CF(P). (b) Hybrid (sandwich) composite-AF(N)/CF(P) stopping crack responsible for less wear. (c) AF(N)/CF(P) composite-accumulation of protective patch work (back transferred layer) (middle portion). Source: Ref 5

References cited in this section

1. K. Friedrich and R. Walter, Microstructure and Tribological Properties of Short Fiber/Thermoplastic Composites, *Tribology of Composite Materials Conf. Proc.,* P.K. Rohatgi, P.J. Blau, and C.S. Yust, Ed., ASM International, 1990, p 217–226

2. K-H. Zum Gahr, *Microstructure and Wear of Materials, Tribology Series 10,* Elsevier, Amsterdam, 1987

4. K. Friedrich, Z. Lu and A.M. Hager, Recent Advances in Polymer Composite Tribology, *Wear,* Vol 190, 1995, p 139–144

5. K. Friedrich, Wear Models for Multiphase Materials and Synergistic Effects in Polymeric Hybrid Composites, *Advances in Composite Tribology, Composite Materials Series 8,* K. Friedrich, Ed., Elsevier, 1993, p 209–273

6. N.P. Suh, *Tribophysics,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1986

8. U.S. Tewari and J. Bijwe, Tribological Behavior of Polyimides, *Polyimides: Fundamentals and Applications,* M.K. Ghosh and K.L. Mittal, Ed., Marcel Dekker Inc., 1996, p 533–586

18. J.K. Lancaster, Friction and Wear, *Polymer Science: A Material Science Handbook,* A.D. Jenkins, Ed., North Holland, 1972

22. K. Friedrich and M. Cyffka, On the Wear of Reinforced Thermoplastics by Different Abrasive Papers, *Wear,* Vol 103, 1985, p 333–344

27. M. Cirino, K. Friedrich, and R.B. Pipes, Evaluation of Polymer Composites for Sliding and Abrasive Wear Applications, *Composites,* Vol 19 (No. 5), 1988, p 383–392

31. K. Friedrich, Wear of Reinforced Polymers by Different Abrasive Counterparts; *Friction and Wear of Polymeric Composites, Composite Materials Series 1,* K. Friedrich, Ed., Elsevier, 1986, p 233–287

32. M. Cirino, R.B. Pipes, and K. Friedrich, The Abrasive Wear of Continuous Fiber Reinforced Polymer Composites, *J. Mater. Sci.,* Vol 22, 1987, p 2481–2492

33. M. Cirino, K. Friedrich, and R.B. Pipes, The Effect of Fiber Orientation on the Abrasive Wear Behavior of Polymer Composite Materials, *Wear,* Vol 121, 1988, p 127–141

37. Q.H. Wang, Q.J. Xue, H. Liu, and J. Xu, The Effect of Particle Size on Nanometer ZrO_2 on Tribo-behavior of PEEK, *Wear,* Vol 198, 1996, p 216–219

38. Q.H. Wang, J. Xu, W.C. Shen, and Q. Liu, An Investigation of the Friction and Wear Properties of Nanometer Si_3N_4 Filled PEEK, *Wear,* Vol 196, 1996, p 82–86

39. Q.H. Wang, W.C. Shen, J.F. Xu, and Q.J. Xue, The Effect of Nanometer SiC Filler on the Tribological Behavior of PEEK, *Wear,* Vol 209, 1997, p 316–321

40. Q.H. Wang, J. Xue, W.M. Liu, and J.M. Chen, Friction and Wear Characteristics of Nanometer SiC and Poly(tetraflouroethylene) Filled Poly(etheretherketone), *Wear,* Vol 243, 2000, p 140–146

41. J.K. Lancaster, Composites for Aerospace Dry Bearing Application, *Friction and Wear of Polymer Composites, Composite Material Series 1,* K. Friedrich, Ed., Elsevier, 1986

42. P. Lu and K. Friedrich, On Sliding of Friction and Wear of PEEK and its Composite, *Wear,* Vol 181–183, 1995, p 624–631

43. A.M. Hager and M. Davis, Short Fiber Reinforced, High Temperature Resistant Polymer for a Wide Field of Tribological Applications, *Advances in Composite Tribology, Composite Materials Series 8,* K. Friedrich, Ed., Elsevier, 1993, p 107–157

44. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of Internally Lubricated Polyetherimide Composite, *J. Synth. Lub.,* Vol 6, 1989, p 179–202

45. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of Polyetherimide Composite, *Wear,* Vol 138, 1990, p 61–76

46. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of Bulk Polyetherimide, *J. Mater. Sci.,* Vol 25, 1990, p 548–556

47. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of a Short Glass Fiber Reinforced Polyetherimide Composite, *Wear,* Vol 132, 1989, p 247–264

48. U.S. Tewari and J. Bijwe, Tribological Investigations of Polyetherimide Composite, *J. Mater. Sci.,* Vol 27, 1992, p 328–334

49. H. Voss and K. Friedrich, Wear Performance of a Bulk Liquid Crystal Polymer and its Short Fiber Composites, *Tribol. Int.,* Vol 19, 1986, p 145–156

50. J.P. Giltrow and J.K. Lancaster, Carbon Fiber Reinforced Polymer as Self Lubricating Materials, *Proc. Institute of Mechanical Engineering,* Vol 182 (3N), 1967–1968, p 147

51. J.P. Giltrow and J.K. Lancaster, The Role of the Counterface in the Friction and Wear of Carbon Fiber Reinforced Thermosetting Resin, *Wear,* Vol 16, 1970, p 359–374

52. J.K. Lancaster, Estimation of the Limiting PV Relationship for Thermoplastic Bearing Materials, *Tribology* Vol. 4, 1971, p 82–86

53. T. Tsukizoe and N. Ohmae, Friction and Wear Performance of Unidirectionally Oriented Glass, Carbon, Aramid and Stainless Steel Fiber-Reinforced Plastics, *Friction and Wear of Polymer Composites,* K. Friedrich, Ed., Elsevier, 1986, p 203–231

54. T. Tsukizoe and N. Ohmae, Wear Mechanism of Unidirectionally Oriented Fiber Reinforced Plastics, *ASME Trans. J. Lubr. Technol.* Vol 99 F(4), 1977, p 401–407

55. T. Tsukizoe and N. Ohmae, Friction and Wear of Partially Oriented Fiber Reinforced Plastics-Tribological Assessment for CFRP, GFRP and SFRP, *J. Jpn. Soc. Lubr. Eng.,* Vol 2 (No. 5), 1976, p 330 (in Japanese)

56. T. Tsukizoe and N. Ohmae, Friction Properties of Composite Materials, *Trans. Jpn. Soc. Mech. Eng.,* Vol 43 (No. 367), 1977, p 115 (in Japanese)

57. T. Goda, K. Varadi, and K. Friedrich, Finite Element Micro-models to Study Contact States, Stress and Failure Mechanisms in a Polymer Composites Subjected to a Sliding Steel Asperity, *Research News,* Budapest University of Technology and Economics, (3), 2000, p 8–14

58. P.B. Mody, T.W. Chou, and K. Friedrich, Effect of Testing Conditions and Microstructure on the Sliding Wear of Graphite Fiber/PEEK Matrix Composites, *J. Mater. Sci.,* Vol 23, 1988, p 4319–4330

59. H.M. Hawthorne, Wear in Hybrid Carbon/Glass Fiber Epoxy Composite Materials, *Proc. International Conf. on Wear of Materials,* Reston, VA, K.C. Ludema, Ed., American Society of Mechanical Engineers, 11–14 April 1983, p 576–582

Acknowledgments

The author would like to acknowledge Professor K. Friedrich, Research Director, Institute for Composite Materials Ltd., University of Kaiserslautern, Germany, for the kind consent to present his work and the provision of SEMs describing failure mechanisms of some important composites. The author would also like to thank J. John Rajesh and Nidhi Dhingra for help in research and typing.

Wear Failures of Reinforced Polymers

J. Bijwe, Indian Institute of Technology, Delhi

References

1. K. Friedrich and R. Walter, Microstructure and Tribological Properties of Short Fiber/Thermoplastic Composites, *Tribology of Composite Materials Conf. Proc.,* P.K. Rohatgi, P.J. Blau, and C.S. Yust, Ed., ASM International, 1990, p 217–226

2. K-H. Zum Gahr, *Microstructure and Wear of Materials, Tribology Series 10,* Elsevier, Amsterdam, 1987

3. J. Bijwe and M. Fahim, Tribology of High Performance Polymers–State of Art, *Handbook of Advanced Functional Molecules and Polymers,* H.S. Nalwa, Ed., Gordon and Breach Publishers, Amsterdam, 2001, p 265–321

4. K. Friedrich, Z. Lu and A.M. Hager, Recent Advances in Polymer Composite Tribology, *Wear,* Vol 190, 1995, p 139–144

5. K. Friedrich, Wear Models for Multiphase Materials and Synergistic Effects in Polymeric Hybrid Composites, *Advances in Composite Tribology, Composite Materials Series 8,* K. Friedrich, Ed., Elsevier, 1993, p 209–273

6. N.P. Suh, *Tribophysics,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1986

7. K. Friedrich, Ed., *Friction and Wear of Polymer Composites, Composite Materials Series 1,* Elsevier, 1986

8. U.S. Tewari and J. Bijwe, Tribological Behavior of Polyimides, *Polyimides: Fundamentals and Applications,* M.K. Ghosh and K.L. Mittal, Ed., Marcel Dekker Inc., 1996, p 533–586

9. S.W. Zhang, State-of-the-Art of Polymer Tribology, *Tribol. Int.,* Vol 31 (No. 1–3), 1998, p 46–60

10. H. Czichos, *Tribology: A System Approach to the Science and Technology of Friction, Lubrication and Wear, Tribology Series 1,* Elsevier, 1978

11. E. Santner and H. Czichos, Tribology of Polymers, *Tribol. Int.,* Vol 22, 1989, p 103–109

12. K-H Zum Gahr, Wear by Hard Particles, *Tribol. Int.,* Vol 31, 1998, p 587–596

13. S.V. Prasad and P.D. Calvert, Abrasive Wear of Particle Filled Polymers, *J. Mater. Sci.,* Vol 15, 1980, p 1746–1754

14. M.M. Khruschov, Principles of Abrasive Wear, *Wear,* Vol 28, 1974, p 69–88

15. K-H Zum Gahr, *Proc. International Conf. on Wear of Materials,* Vancouver, American Society of Mechanical Engineers, 14–18 April 1985, p 45–58

16. W. Simm and S. Freti, Abrasive Wear of Multiphase Materials, *Wear,* Vol 129, 1989, p 105–121

17. S. Bahadur and D. Gong, Formulation of the Model for Optimal Proportion of Filler in Polymer for Abrasive Wear Resistance, *Wear of Materials,* American Society of Mechanical Engineers, 1991, p 177–185

18. J.K. Lancaster, Friction and Wear, *Polymer Science: A Material Science Handbook,* A.D. Jenkins, Ed., North Holland, 1972

19. U.S. Tewari, J. Bijwe, J.N. Mathur, and I. Sharma, Studies on Abrasive Wear of Carbon Fiber (Short) Reinforced Polyamide Composites, *Tribol. Int.,* Vol 25, 1992, p 53–60

20. J.P. Giltrow, A Relationship Between Abrasive Wear and the Cohesive Energy of Materials, *Wear,* Vol 15, 1970, p 71–78

21. M. Vaziri, R.T. Spurr, and F.H. Stott, An Investigation of the Wear of Polymeric Materials, *Wear,* Vol 122, 1988, p 329–342

22. K. Friedrich and M. Cyffka, On the Wear of Reinforced Thermoplastics by Different Abrasive Papers, *Wear,* Vol 103, 1985, p 333–344

23. J.J. Rajesh, J. Bijwe and U.S. Tewari, Influence of Fillers on Abrasive Wear of Short Glass Fiber Reinforced Polyamide Composites, *J. Mater. Sci.,* Vol. 36, 2001, p 351–356

24. B.J. Briscoe, P.D. Evans, and J.K. Lancaster, The Influence of Debris Inclusion on Abrasive Wear Relationships of PTFE, *Wear,* Vol 124, 1988, p 177–194

25. B.J. Briscoe, and P.D. Evans, The Influence of Asperity Deformation Condition on the Abrasive Wear of γ-Irradiated Polytetrafluoroethylenes, *Wear,* Vol 133, 1989, p 47–64

26. C. Lhymn, K.E. Templemeyer, and P.K. Davis, Abrasive Wear of Short Fiber Composites, *Composites,* Vol 16 (No. 2), 1985, p 127–136

27. M. Cirino, K. Friedrich, and R.B. Pipes, Evaluation of Polymer Composites for Sliding and Abrasive Wear Applications, *Composites,* Vol 19 (No. 5), 1988, p 383–392

28. M.K. Omar, A.G. Atkins, and J.K. Lancaster, The Role of Crack Resistance Parameters in Polymer Wear, *Physica D: Appl. Phys.,* Vol 19, 1986, p 177–195

29. J. Bijwe, J. Indumathi and A.K. Ghosh, Evaluation of Engineering Polymeric Composites for Abrasive Wear Performance, *J. Reinf. Plast. Compos.,* Vol 18 (No. 17), 1999, p 1573–1591

30. C. Lhymn, Effect of Normal Load on Specific Wear Rate of Fibrous Composites, *Wear,* Vol 120, 1987, p 1–27

31. K. Friedrich, Wear of Reinforced Polymers by Different Abrasive Counterparts; *Friction and Wear of Polymeric Composites, Composite Materials Series 1,* K. Friedrich, Ed., Elsevier, 1986, p 233–287

32. M. Cirino, R.B. Pipes, and K. Friedrich, The Abrasive Wear of Continuous Fiber Reinforced Polymer Composites, *J. Mater. Sci.,* Vol 22, 1987, p 2481–2492

33. M. Cirino, K. Friedrich, and R.B. Pipes, The Effect of Fiber Orientation on the Abrasive Wear Behavior of Polymer Composite Materials, *Wear,* Vol 121, 1988, p 127–141

34. B.K. Yen and C.K.H. Dharan, A Model for Abrasive Wear of Fiber Reinforced Polymer Composites, *Wear,* Vol 195, 1996, p 123–127

35. J. Bijwe, J. Indumathi, and A.K. Ghosh, On the Abrasive Wear Behavior of Fabric Reinforced Polyetherimide Composites, *Wear,* in press

36. J. Indumathi, "Friction and Wear Studies on Polyetherimide and Composites", Ph.D. dissertation, Indian Institute of Technology Delhi, New Delhi, 2000

37. Q.H. Wang, Q.J. Xue, H. Liu, and J. Xu, The Effect of Particle Size on Nanometer ZrO_2 on Tribo-behavior of PEEK, *Wear,* Vol 198, 1996, p 216–219

38. Q.H. Wang, J. Xu, W.C. Shen, and Q. Liu, An Investigation of the Friction and Wear Properties of Nanometer Si_3N_4 Filled PEEK, *Wear,* Vol 196, 1996, p 82–86

39. Q.H. Wang, W.C. Shen, J.F. Xu, and Q.J. Xue, The Effect of Nanometer SiC Filler on the Tribological Behavior of PEEK, *Wear,* Vol 209, 1997, p 316–321

40. Q.H. Wang, J. Xue, W.M. Liu, and J.M. Chen, Friction and Wear Characteristics of Nanometer SiC and Poly(tetraflouroethylene) Filled Poly(etheretherketone), *Wear,* Vol 243, 2000, p 140–146

41. J.K. Lancaster, Composites for Aerospace Dry Bearing Application, *Friction and Wear of Polymer Composites, Composite Material Series 1,* K. Friedrich, Ed., Elsevier, 1986

42. P. Lu and K. Friedrich, On Sliding of Friction and Wear of PEEK and its Composite, *Wear,* Vol 181–183, 1995, p 624–631

43. A.M. Hager and M. Davis, Short Fiber Reinforced, High Temperature Resistant Polymer for a Wide Field of Tribological Applications, *Advances in Composite Tribology, Composite Materials Series 8,* K. Friedrich, Ed., Elsevier, 1993, p 107–157

44. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of Internally Lubricated Polyetherimide Composite, *J. Synth. Lub.,* Vol 6, 1989, p 179–202

45. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of Polyetherimide Composite, *Wear,* Vol 138, 1990, p 61–76

46. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of Bulk Polyetherimide, *J. Mater. Sci.,* Vol 25, 1990, p 548–556

47. J. Bijwe, U.S. Tewari, and P. Vasudevan, Friction and Wear Studies of a Short Glass Fiber Reinforced Polyetherimide Composite, *Wear,* Vol 132, 1989, p 247–264

48. U.S. Tewari and J. Bijwe, Tribological Investigations of Polyetherimide Composite, *J. Mater. Sci.,* Vol 27, 1992, p 328–334

49. H. Voss and K. Friedrich, Wear Performance of a Bulk Liquid Crystal Polymer and its Short Fiber Composites, *Tribol. Int.,* Vol 19, 1986, p 145–156

50. J.P. Giltrow and J.K. Lancaster, Carbon Fiber Reinforced Polymer as Self Lubricating Materials, *Proc. Institute of Mechanical Engineering,* Vol 182 (3N), 1967–1968, p 147

51. J.P. Giltrow and J.K. Lancaster, The Role of the Counterface in the Friction and Wear of Carbon Fiber Reinforced Thermosetting Resin, *Wear,* Vol 16, 1970, p 359–374

52. J.K. Lancaster, Estimation of the Limiting PV Relationship for Thermoplastic Bearing Materials, *Tribology* Vol. 4, 1971, p 82–86

53. T. Tsukizoe and N. Ohmae, Friction and Wear Performance of Unidirectionally Oriented Glass, Carbon, Aramid and Stainless Steel Fiber-Reinforced Plastics, *Friction and Wear of Polymer Composites,* K. Friedrich, Ed., Elsevier, 1986, p 203–231

54. T. Tsukizoe and N. Ohmae, Wear Mechanism of Unidirectionally Oriented Fiber Reinforced Plastics, *ASME Trans. J. Lubr. Technol.* Vol 99 F(4), 1977, p 401–407

55. T. Tsukizoe and N. Ohmae, Friction and Wear of Partially Oriented Fiber Reinforced Plastics-Tribological Assessment for CFRP, GFRP and SFRP, *J. Jpn. Soc. Lubr. Eng.,* Vol 2 (No. 5), 1976, p 330 (in Japanese)

56. T. Tsukizoe and N. Ohmae, Friction Properties of Composite Materials, *Trans. Jpn. Soc. Mech. Eng.,* Vol 43 (No. 367), 1977, p 115 (in Japanese)

57. T. Goda, K. Varadi, and K. Friedrich, Finite Element Micro-models to Study Contact States, Stress and Failure Mechanisms in a Polymer Composites Subjected to a Sliding Steel Asperity, *Research News,* Budapest University of Technology and Economics, (3), 2000, p 8–14

58. P.B. Mody, T.W. Chou, and K. Friedrich, Effect of Testing Conditions and Microstructure on the Sliding Wear of Graphite Fiber/PEEK Matrix Composites, *J. Mater. Sci.,* Vol 23, 1988, p 4319–4330

59. H.M. Hawthorne, Wear in Hybrid Carbon/Glass Fiber Epoxy Composite Materials, *Proc. International Conf. on Wear of Materials,* Reston, VA, K.C. Ludema, Ed., American Society of Mechanical Engineers, 11–14 April 1983, p 576–582

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Introduction

THIS HANDBOOK is organized according to four general categories of failure: fracture, corrosion, wear, and the subject of this article, distortion. One reason metals are so widely used as engineering materials is that they have high strength but also generally have the capability to respond to load (stress) by deforming. In fact, much of metallurgical engineering is concerned with balancing strength and ductility. Thus, distortion often is observed in analysis of other types of failures, and consideration of the distortion can be an important part of the analysis. Energy is absorbed during deformation, and in some situations, the amount of energy absorbed may also be an important factor. Furthermore, it should be noted that not all distortion necessarily constitutes a "failure."

This article first considers true distortion failures, that is, situations in which distortion occurs when it should not have occurred and in which the distortion is associated with a functional failure. Then, a more general consideration of distortion in failure analysis is introduced. As used here, distortion will refer to a condition in which the shape of a component has changed without loss of material. Deformation will refer to the process that results in the distortion.

Distortion failure occurs when a structure or component is deformed so that it can no longer support the load it was intended to carry, is incapable of performing its intended function, or interferes with the operation of another component. Distortion failures can be plastic or elastic and may or may not be accompanied by fracture. There are two main types of distortion: size distortion, which refers to a change in volume (growth or shrinkage), and shape distortion (bending or warping), which refers to a change in geometric form. Most of the examples in this article deal with metals, but the concepts also apply to nonmetals. Materials as diverse as metals, polymers, and wood are all susceptible to distortion, although the mechanisms may differ somewhat among the different classes of material.

Distortion failures are ordinarily considered to be self-evident, for example, damage of a car body in a collision or bending of a nail being driven into hard wood. However, the failure analyst is often faced with more subtle situations. For example, the immediate cause of distortion (bending) of an automobile-engine valve stem is contact of the valve head with the piston, but the failure analyst must go beyond this immediate cause in order to recommend proper corrective measures. The valve may have stuck open because of faulty lubrication; the valve spring may have broken because corrosion had weakened it. The spring may have had insufficient strength and taken a set, allowing the valve to drop into the path of the piston, or the engine may have been raced beyond its revolutions per minute limit many times, causing coil clash and subsequent fatigue fracture of the spring. Without careful consideration of all the evidence, the failure analyst may overlook the true cause of a distortion failure. Several common aspects of failure by distortion are discussed in this article, and suitable examples of distortion failures are presented for illustration.

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Overloading

Every structure has a load limit beyond which it is considered unsafe or unreliable. Applied loads that exceed this limit are known as overloads and sometimes result (depending on the factor of safety used in design) in distortion or fracture of

one or more structural members. Estimation of load limits is one of the most important aspects of design and is commonly computed by one of two methods—classical design or limit analysis.

Classical Design. The conservative, classical method of design (assuming monotonic or static loading) assumes that failure occurs whenever the stress at any point in a structure exceeds the yield strength of the material. Except for members that are loaded in pure tension, the fact that yielding occurs at some point in a structure has little influence on the ability of the structure to support the load. However, yielding has long been considered a prelude to structural collapse or fracture and is therefore a reasonable basis for limiting applied loads. This classical approach inherently assumes that the stress to cause fracture is greater than the stress to cause yield. As fracture mechanics analysis clearly shows, this may not be the case. Fracture may occur at loads less than that required to cause yield if a sufficiently large imperfection is present in the material.

Classical design keeps allowable stresses entirely within the elastic region and is used routinely in the design of parts. Allowable stresses for static service are generally set at one-half the yield strength for ductile materials and one-sixth for brittle materials, although other fractions may be more suitable for specific applications. For very brittle materials, there may be little difference between the “yield” and ultimate strength, and the latter is used in design computations. The reason for using such low fractions of yield (or ultimate) strength is to allow for such factors as possible errors in computational assumptions, accidental overload, introduction of residual stress during processing, temperature effects, variations in material quality (including imperfections), degradation (for example, from corrosion), and inadvertent local increases in applied stress resulting from notch effects.

Classical design is also used for setting allowable stresses in other applications, for example, where fracture can occur by fatigue or stress rupture. In these instances, fatigue strength or stress-rupture strength is substituted for yield strength as a point of reference, typically with different factors of safety.

Limit Analysis. The upper limit in design is defined as the load at which a structure will break or collapse under a single application of force. This load can be calculated by a method known as limit analysis (Ref 1, 2). With limit analysis, it is unnecessary to estimate stress distributions, which makes stress analysis much simpler by this method than by classical design. However, limit analysis is based on the concept of tolerance to yielding in the most highly stressed regions of the structure and therefore cannot be used in designing for resistance to fatigue or elastic buckling or in designing flaw-tolerant structures.

Limit analysis assumes an idealized material—one that behaves elastically up to a certain yield strength, then does not work harden but undergoes an indefinite amount of plastic deformation with no change in stress. The inherent safety of a structure is more realistically estimated by limit analysis in those instances when the structure will tolerate some plastic deformation before it collapses. Because low-carbon steel, one of the most common materials used in structural members, behaves somewhat like the idealized material, limit analysis is very useful to the designer, especially in the analysis of statically indeterminate structures.

Figure 1 illustrates the relative stress-strain behavior of a low-carbon steel, a strain-hardening material, and an idealized material—all with the same yield strength (the upper yield point for the low-carbon steel and the stress at 0.2% offset for the strain-hardening material). Load limits for parts made of materials that strain harden significantly when stressed in the plastic region can be estimated by limit analysis, as can those for parts made of other materials whose stress-strain behavior differs from that of the idealized material. In these situations, the designer bases his design calculations on an assumed strength that may actually lie well within the plastic region for the material.

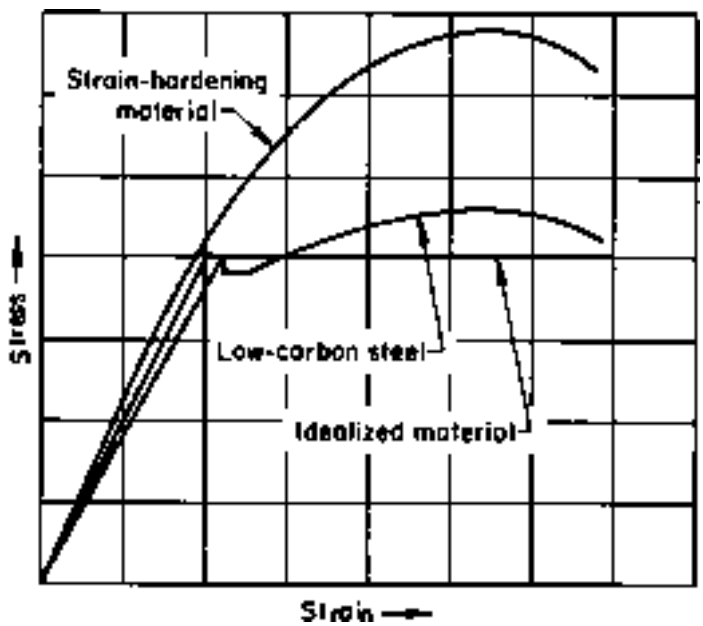


Fig. 1 Comparison of the conventional stress-strain behavior of a low-carbon steel, a strain-hardening material, and the idealized material assumed in limit analysis. All have the same yield strength.

Buckling. Collapse due to instability under compressive stress, or buckling, may or may not be permanent deformation, depending on whether or not the yield strength was exceeded. Long, slender, straight bars, tubes, or columns under axial compressive forces will buckle when the buckling load is exceeded. Buckling failure may also be encountered on the compressive sides of tubes, I-beams, channels, and angles under bending forces. Tubes may also buckle due to torsional forces, causing waves, or folds, generally perpendicular to the direction of the compressive-stress component. Parts under bending load are also subject to buckling failures on the compressive (concave) side (Fig. 2).

Fig. 2 Buckled flange (lower arrow) of an extruded aluminum section deliberately loaded with a lateral force (upper arrow). Source: Ref 3.

The buckling load depends only on the dimensions of the part and the modulus of elasticity of the material. Therefore, buckling cannot be prevented by changing the strength or hardness of the metal. The modulus of elasticity of a given metal is affected only by temperature, increasing at lower temperature and decreasing at higher temperature. Buckling can be prevented only by changing the size or shape of the part with respect to the load imposed on it (Ref 3).

The failure analyst should be sensitive to situations in which buckling has occurred but may not be immediately apparent. A beam in bending will be more susceptible to buckling on the compression side if it is relatively deep and narrow. A thin, circular shaft in torsion may buckle into a helical configuration when a critical moment is exceeded. Creep or distortion from other causes may change the dimensions of a structure so that it becomes susceptible to buckling. Further details can be found in references such as Ref 4.

Safety Factors. In both classical design and limit analysis, yielding is assumed to be the criterion for calculating safe loads on statically loaded structures. For a given design and applied load, the two methods differ in that the safety factor (the ratio of the theoretical capacity of a structural member to the maximum allowable load) is generally higher when calculated by limit analysis. For example, classical design limits the capacity of a rectangular beam to the bending

moment that will produce tensile yielding in the regions farthest from the neutral axis; limit analysis predicts that complete collapse will occur at a bending moment 1.5 times the limiting bending moment determined by classical design. It is important that the designer be able to relate the actual behavior of a structure to its assumed behavior because, for a given applied load and selected safety factor, a structure designed by limit analysis will usually have thinner sections than a structure designed by classical methods.

Safety factors are important design considerations because they allow for factors that cannot be computed in advance. Overload failure can occur either when the applied loads increase the stress above the design value or when the material strength is degraded. If either situation is a characteristic of the fabricated structure, the design must be changed to allow for these factors more realistically.

Example 1: Collapse of Extension Ladders by Overloading of Side Rails. Several aluminum alloy extension ladders of the same size and type collapsed in service in the same manner; the extruded aluminum alloy 6063-T6 side rails buckled, but the rungs and hardware remained firmly in place. The ladders had a maximum extended length of 6.4 m (21 ft), and the recommended maximum angle of inclination to the vertical was 15°.

Investigation. Visual examination disclosed that the side-rail extrusions, which had the I-beam shape shown in Fig. 3(a), had failed by plastic buckling, with only slight surface cracking in the most severely deformed areas. There were no visible defects in materials or workmanship, and all dimensions of the side rails were within specified tolerances.

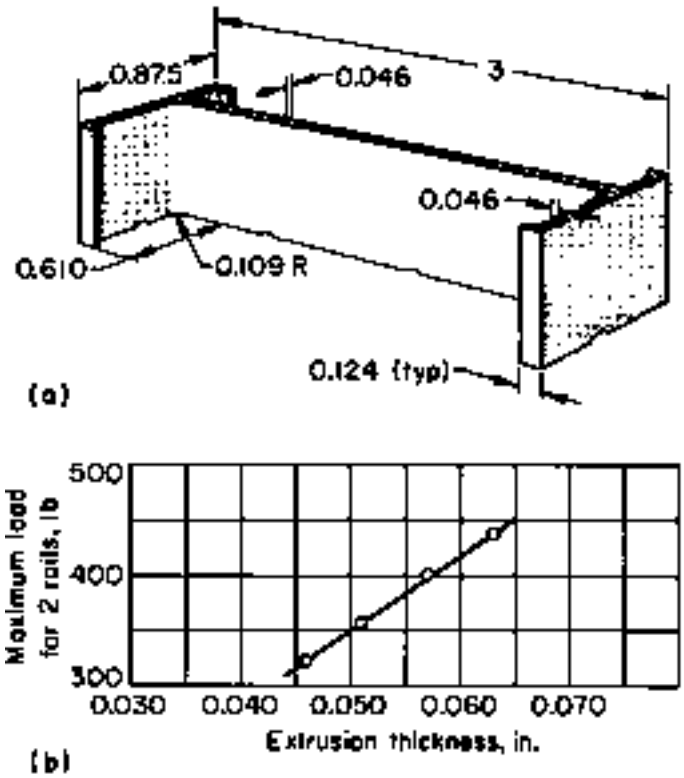


Fig. 3 Aluminum alloy 6063-T6 extension-ladder side-rail extrusion that failed by plastic deformation and subsequent buckling. (a) Configuration and dimensions (given in inches). (b) Relation of maximum applied load to the section thickness of the flanges and web of the side-rail extrusion.

Hardness tests using a portable hardness tester, metallographic examination, and tensile tests of specimens from the buckled side rails were conducted. All results agreed with the typical properties reported for aluminum alloy 6063-T6 extrusions.

Stress analysis of the design of the ladder, using actual dimensions, indicated that the side-rail extrusions had been designed with a thickness that would provide a safety factor of 1.2 at ideal loading conditions (15° maximum inclination) and that use of the ladders under other conditions could subject the side rails to stresses beyond the yield strength of the material. Once yielding occurred, buckling would continue until the ladder collapsed.

The stress analysis was extended to include an evaluation of the relation of maximum applied load to section thickness with the ladder at its maximum extension of 6.4 m (21 ft) and at an inclination of 15°. This relation is shown in Fig. 3(b).

Conclusions. The side rails of the ladders buckled when subjected to loads that produced stresses beyond the yield strength of the alloy. Failure was by plastic deformation, with only slight tearing in the most severely deformed regions.

Corrective Measure. The flange and web of the side-rail extrusion were increased in thickness from 1.2 to 1.4 mm (0.046 to 0.057 in.). This increased the safety factor from 1.2 to 1.56. After this change, no further failures were reported. Note that this example is presented for illustrative purposes only, and this safety factor may not be appropriate in other applications.

Amount of Distortion. When designing structures using limit analysis, the designer does not always consider the amount of distortion that will be encountered. A rough illustration of the distortion that resulted from overloading of small cantilever beams is given in Fig. 4. Known loads were applied to rectangular-section beams of low-carbon steel and of stainless steel, and the permanent deflection at the loading point was measured. Maximum fiber stresses were calculated from the applied load and original specimen dimensions.

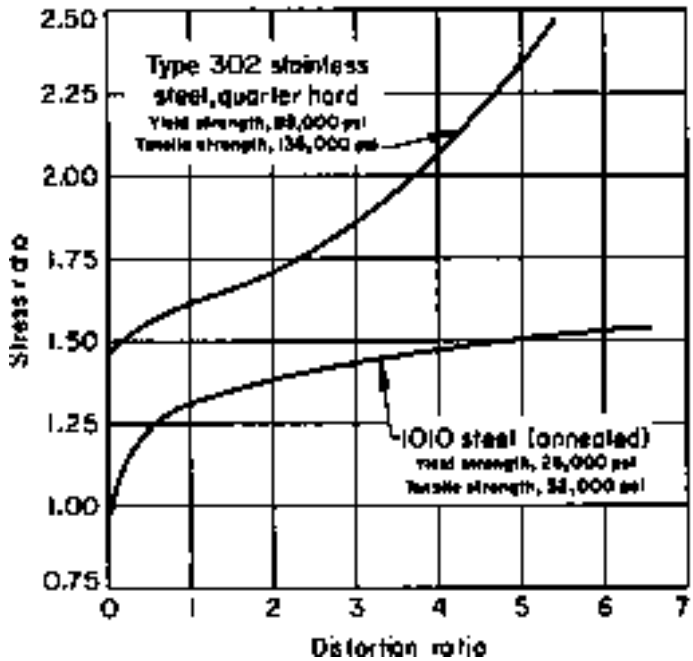


Fig. 4 Relation of distortion ratio to stress ratio for two steel cantilever beams of rectangular cross section. Distortion ratio is permanent deflection, measured at a distance from the support ten times the beam thickness, divided by beam thickness. Stress ratio is maximum stress, calculated from applied load and original beam dimensions, divided by yield strength.

This type of test provides a simplistic but useful concept of distortion by showing how much distortion occurs at strains beyond the yield point. As shown in Fig. 4, the beam of low-carbon steel, which strain hardens only slightly, exhibited no distortion when the calculated maximum fiber stress was equal to the yield strength (at a stress ratio of 1.00). However, this beam collapsed at a load equivalent to a fiber stress just above the tensile strength, as shown in Fig. 4 where the lower curve became essentially horizontal. This collapse load agrees with the limit-analysis collapse load of 1.5 times the load at yield. The beam made of stainless steel, which strain hardens at a rather high rate, showed no distortion at fiber stresses up to 1.47 times the yield strength. When the calculated stress equaled the tensile strength (at a stress ratio of 1.59), distortion was 0.7 times the beam thickness, and the beam supported a calculated stress of 1.5 times the tensile strength without collapse.

It should be noted that the preceding example is intended simply to illustrate the differences in deformation behavior between two different materials. As a practical matter, one would not substitute stainless steel for low-carbon steel to increase load capacity. One would use a heavier section, or perhaps, a higher-strength alloy.

When loads increase gradually, distortion is gradual, and design can be based on knowledge of the amount of distortion that can be tolerated. Thus, simple bench tests of full-size or scaled-down models can often be used in estimating the loads required to produce various amounts of distortion.

Effect of Impact/Very High Strain Rates. When rapid or impulse loads are applied, as in impact, shock loading, or high-frequency vibration, the amount of distortion that can occur without fracture is considerably less predictable. The crystallographic processes involved in deformation and fracture are influenced by strain rate as well as temperature. For most structural materials, measured values of strength are higher under impulse loading, and values of ductility are lower, than the values measured under static loading. Tensile and yield strengths as much as 20% higher than the slow-tension-test values have been measured under very high rates of loading. Strain-rate sensitivity data have been compiled for many

materials and is important in analysis of manufacturing operations involving deformation (see, for example, *Forming and Forging*, Volume 14, *ASM Handbook*). The amount of distortion that can occur at high rates of loading is difficult to analyze or predict precisely because:

- The variation, or scatter, among replicate tests of mechanical properties is greater when strain rates are high than it is when strain rates are low.
- Impulse or impact loading involves the propagation of high-velocity stress waves through the structure, which may not be well quantified, with the most extreme example being ballistic impact with strain rates on the order of 10^4/min.
- Impulse loading creates an adiabatic condition that causes a local increase in temperature.

Effect of Temperature. Distortion failures caused by overload can occur at any temperature at which the flow strength of the material is less than the fracture strength. In this discussion, flow strength is defined as the average true stress required to produce detectable plastic deformation caused by a relatively slow, continuously increasing application of load; fracture strength is the average true stress at fracture caused by a relatively slow, continuously increasing application of load. The flow strength and fracture strength of a material are temperature dependent, as is the elastic modulus (Young's modulus, bulk modulus, or shear modulus).

Figure 5 illustrates this temperature dependence schematically for polycrystalline materials that do not undergo a solid-state transformation. Two flow strengths are shown: one for a material that does not have a ductile-to-brittle transition in fracture behavior, such as metal with a face-centered cubic (fcc) crystal structure, and one for a body-centered cubic (bcc) material that exhibits a ductile-to-brittle transition.

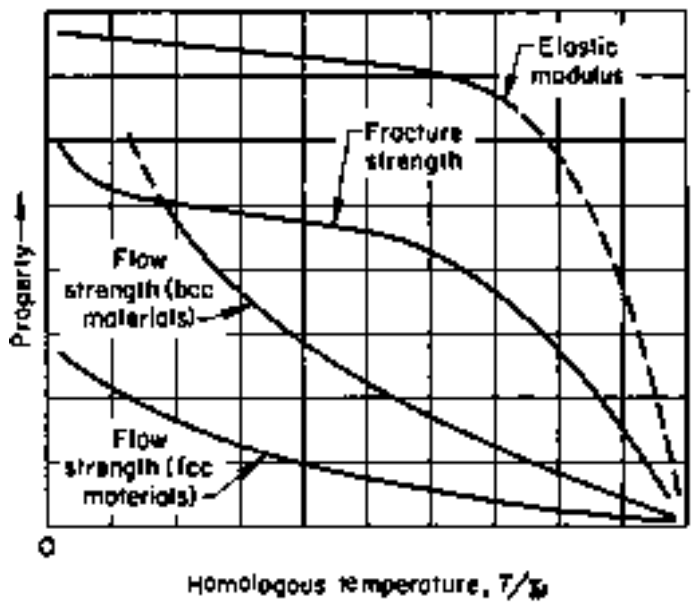


Fig. 5 Diagram of the temperature dependence of elastic, plastic, and fracture behavior of polycrystalline materials that do not exhibit a solid-state transformation. bcc, body-centered cubic; fcc, face-centered cubic; *T*, instantaneous absolute temperature; T_M, absolute melting temperature of the material

As shown in Fig. 5, the flow strength, fracture strength, and elastic modulus of a material generally decrease as temperature increases. If a structure can carry a certain load at 20 °C (70 °F), it can carry the same load without deforming at lower temperatures. Stressed members made of materials having a ductile-to-brittle fracture transition will sometimes fracture spontaneously if the temperature is lowered to a value below the transition temperature (e.g., see the article “Overload Failures” in this Volume).

If the temperature is increased so that the flow strength becomes lower than the applied stress, a structure may deform spontaneously with no increase in load. A change in temperature may also cause an elastic-distortion failure because of a change in modulus, as might occur in a control device where accuracy depends on a predictable elastic deflection of a control element or a sensing element. For most structural materials, the curve defining the temperature dependence of elastic and plastic properties is relatively flat at temperatures near 20 °C (70 °F). For steels, the modulus is slightly decreasing until temperatures of approximately 320 to 370 °C (600 to 700 °F), are reached, at which point modulus starts to decrease more rapidly.

In fcc materials, and in bcc materials at temperatures above the transition temperature but at lower homologous temperatures (<0.3 T_M), distortion (net section yielding) always accompanies overload fracture in sections that do not contain a severe stress raiser. At higher homologous temperature, an increase in temperature may cause a change from transgranular (TG) to intergranular (IG) fracture with a concurrent decrease in ductility. That is, there is a minimum ductility at elevated temperature (within the "creep" range) where the fracture mechanism changes from TG to IG fracture with a concurrent loss in ductility.

Creep, or time-dependent strain, is a relatively long-term phenomenon and can be distinguished from overload distortion by relating the length of time at temperature to the amount of distortion, as discussed in the article "Creep and Stress Rupture Failures" in this Volume. The specific mechanisms and associated kinetics of creep are temperature dependent. While creep is sometimes considered to be limited to temperatures above one-half of the absolute melting point and is usually associated with an intergranular mechanism at those temperatures, it is important for the failure analyst to know that long-term creep deformation and even fracture can also occur at lower temperatures and via other mechanisms.

Changes in operating temperature can affect the properties of a structure in other ways. At temperatures higher than about 30 to 40% of T_M depending on the alloy, microstructural changes may occur over time and degrade properties, allowing distortion and even fracture to occur. For example, if a martensitic steel is tempered at a given temperature and then encounters a higher temperature in service, yield strength and tensile strength will decrease because of overtempering. Long-time exposure to moderately elevated temperatures may cause overaging in a precipitation-hardening alloy, with a corresponding loss in strength. It is well known to metallurgists that exposure to cryogenic temperatures may cause cracking in a martensitic steel due to the volume change accompanying the transformation of retained austenite. What may not be as well appreciated is that even if cracking does not occur, the transformation may create a distortion failure due to dimensional growth or warpage in a close-tolerance assembly, such as a precision bearing.

When the temperature is changed, different coefficients of thermal expansion for different materials in a heterogeneous structure can cause interference between structural members (or can produce permanent distortion because of thermally induced stresses if the members are joined together). The failure analyst must understand the effect of temperature on properties of the specific materials involved when analyzing failures that have occurred at temperatures substantially above or below the design or fabrication temperature.

Effect of Stress Raisers and Complex Stress States. In sections that do contain stress raisers, net section yielding can still occur if the state of stress is plane stress; that is, one normal stress component is 0. If the stress raiser results in sufficient constraint to produce plane strain, then gross yielding and distortion will not be observed. However, localized distortion may accompany crack extension as the inherent ductility of the material manifests itself. At the microscale, the fracture may or may not show evidence of ductility. That is, the material may be microscale ductile or brittle. If it is microscale ductile, there may still be no evidence of ductility at the macroscale. In inherently brittle materials, where the fracture stress is equal to the flow stress, no gross or localized distortion accompanies fracture, as discussed further in the article "Fractures Appearance and Mechanisms of Deformation and Fracture" in this Volume.

The yield stress is generally taken as the critical value at which plastic deformation initiates. In uniaxial tension, it is clear when the applied stress reaches the yield point. However, for more complex multiaxial stress states, the point at which yield is anticipated may not be as clear. Theories such as maximum shear stress, maximum distortion energy, and others are detailed in Ref 4 and may be applied by the failure analyst if there is uncertainty as to whether the stresses known to be applied were sufficient to cause the distortion observed.

References cited in this section

1. J.W. Jones, Limit Analysis, *Mach. Des.,* Vol 45 (No. 23), 20 Sept 1973, p 146–151
2. D. Goldner, Plastic Bending in Tubular Beams, *Mach. Des.,* Vol 45 (No. 24), 4 Oct 1973, p 152–155
3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999, p 16–19
4. J.A. Collins, *Failure of Materials in Mechanical Design,* 2nd ed., Wiley and Sons, 1993

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Inappropriate Specifications

Errors in specification of material or method of processing for a part can lead to distortion failures. These errors are often the result of faulty or incomplete information being available to the designer. In such instances, the designer has to make assumptions concerning the conditions of service.

Example 2: Distortion Failure of an Automotive Valve Spring. The engine of an automobile lost power and compression and emitted an uneven exhaust sound after several thousand miles of operation. When the engine was dismantled, it was found that the outer spring on one of the exhaust valves was too short to function properly. The short steel spring and an outer spring taken from another cylinder in the same engine (both shown in Fig. 6) were examined in the laboratory to determine why one had distorted and the other had not.

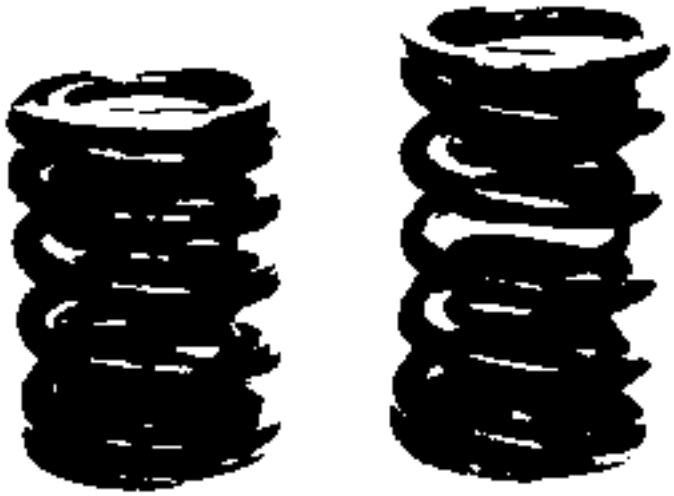


Fig. 6 Valve springs made from patented and drawn high-carbon steel wire. Distorted outer spring (left) exhibited about 25% set because of proeutectoid ferrite in the microstructure and high operating temperature. Outer spring (right) is satisfactory.

Investigation. The failed outer spring (at left, Fig. 6) had decreased in length to about the same free length as that of its companion inner spring. Most of the distortion had occurred in the first active coil (at top, Fig. 6), and a surface residue of baked-on oil present on this end of the spring indicated that a temperature of 175 to 205 °C (350 to 400 °F) had been reached. Temperatures lower than 120 °C (250 °F) usually do not cause relaxation (or set) in high-carbon steel springs.

The load required to compress each outer spring to a length of 2.5 cm (1 in.) was measured. The distorted spring needed only 30 kg (67 lb), whereas the longer spring needed 41 kg (90 lb). The distorted spring had suffered 25% set, which was the immediate cause of the engine malfunction.

The microstructure of both springs was primarily heavily cold-drawn fine pearlite, but the microstructure of the distorted spring contained small amounts of proeutectoid ferrite. Although the composition of the spring alloy was unknown, the microstructure indicated that the material was patented and cold-drawn high-carbon steel wire. The distorted spring had a hardness of 43 HRC, and the longer spring had a hardness of 46 HRC. Both hardness and microstructure indicated that the material in the deformed spring had 10% lower yield strength than material in the undeformed spring. The estimates of yield strength were considered valid because of two factors: the accuracy of the hardness testing and characteristically consistent ratios of yield strength to tensile strength for the grades of steel commonly used in spring wire.

Conclusions. The engine malfunctioned because one of the exhaust-valve springs had taken a 25% set in service. Relaxation in the spring material occurred because of the combined effect of improper microstructure (proeutectoid ferrite) plus a relatively high operating temperature. The undeformed spring exhibited little or no set because the tensile strength and corresponding yield strength of the material (estimated from hardness measurements) were about 10% higher than those of the material in the deformed spring.

Recommendations. A higher yield strength and a higher ratio of yield strength to tensile strength can be achieved in the springs by using quenched-and-tempered steel instead of patented and cold-drawn steel. An alternative would be to use a more expensive chromium-vanadium alloy steel instead of plain carbon steel; the chromium-vanadium steel should be quenched and tempered. Regardless of material or processing specifications, if springs are stressed close to the yield point of the material, close control of material and processing plus stringent inspection are needed to ensure satisfactory performance.

Service conditions are sometimes changed, invalidating certain assumptions that were made when the part was originally designed. Such changes include an increase in operating temperature to one at which the material no longer has the required strength, an increase in the load rating of an associated component, which the user may interpret as an increase in the allowable load on the structure as a whole, and an arbitrary increase in applied load by the user on the assumption that the component has a high enough safety factor to accommodate the added load.

Example 3: Bulging of a Shotgun Barrel Caused by a Change from Lead Shot to Iron Shot. A standard commercial shotgun barrel fabricated from 1138 steel deformed during a test that was made with a new type of ammunition. Use of the new ammunition, which contained soft iron shot with a hardness of about 72 HB, was intended to reduce toxicity; the old ammunition had contained traditional lead shot with a hardness of 30 to 40 HB.

Investigation. The shotgun barrel was of uniform inside diameter from the breech to a point 7.5 cm (3 in.) from the muzzle; at this point, the inside diameter began to decrease ("Before test" curve, Fig. 7a). This taper, or integral choke, which is intended to concentrate the shot pattern, ended about 3.8 cm ($1\frac{1}{2}$ in.) from the muzzle, and the final portion of the barrel had a relatively uniform inside diameter.

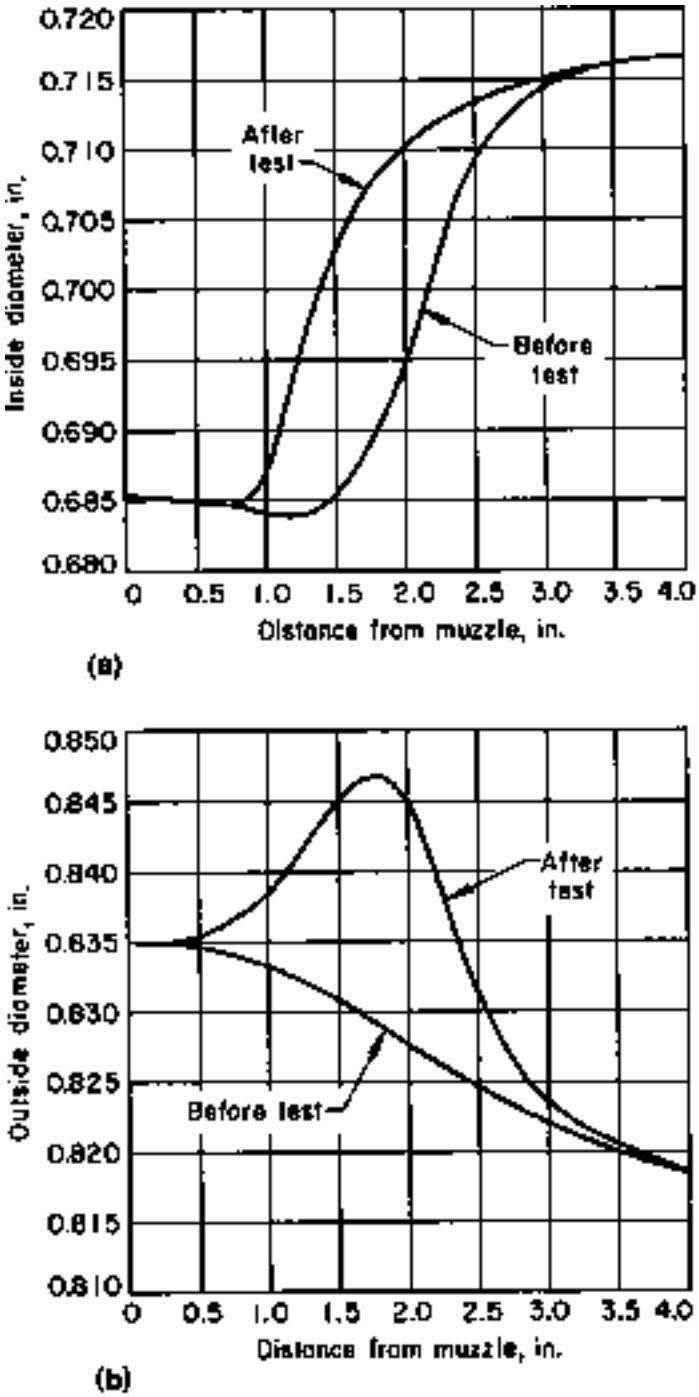


Fig. 7 Comparison of longitudinal profiles of an 1138 steel shotgun barrel before and after testing. 1000 rounds of a new type of ammunition were fired in the test. (a) Inside diameter. (b) Outside diameter

After a test in which 1000 rounds of ammunition containing soft iron shot were fired, the shotgun barrel had a longitudinal profile of inside diameter as shown in the "After test" curve in Fig. 7(a). Comparison of this curve with the

profile before the test shows that the effect of firing soft iron shot was to deform the gun barrel so that the choke taper was shifted toward the muzzle. After the test, there was a bulge on the outside surface of the barrel, shown in comparison to the longitudinal profile of the outside diameter before the test in Fig. 7(b). Deformation of the barrel had been detected after the first 100 rounds of iron-shot ammunition had been fired, and the bulge grew progressively larger as the test continued.

Apparently, the bore of the failed barrel was not concentric with the outside surface, because the wall thickness at a given distance from the breech varied widely among different points around the circumference. For example, at a distance of 5 mm (0.2 in.) from the muzzle, the wall thickness varied from 1.3 to 2 mm (0.051 to 0.080 in.).

The microstructure of the barrel material was a mixture of ferrite and coarse pearlite. The alloy had a hardness of 163 to 198 HB (converted from Vickers hardness measurements).

Based on previous tests, in which the hoop stress in shotgun barrels had been measured when lead-shot ammunition was fired, the safety factor had been estimated at 2.0. In this instance, it was concluded that wall thickness variations had reduced the safety factor to about 1.3 for lead-shot ammunition. Previous tests had also shown that lead shot was deformed extensively by impact with the bore in the choke zone of this type of gun barrel.

Analysis. The major stresses in the choke zone are produced by impact of shot pellets against the bore. When lead shot is used, the lead absorbs a considerable amount of the impact energy as it deforms. Soft iron shot, on the other hand, is much harder than lead and does not deform significantly. More of the impact energy is absorbed by the barrel when iron shot is used, producing higher stresses.

In this instance, had the gun barrel been of more uniform wall thickness around its circumference, it might not have deformed. However, it was believed that conversion to iron-shot ammunition would increase stresses in the barrel enough to warrant an increase in the strength of this type of barrel.

Conclusions. The shotgun barrel deformed because a change to iron-shot ammunition increased stresses in the choke zone of the barrel. Bulging was enhanced by a lack of uniformity in wall thickness.

Recommendations. Three alternative solutions to this problem were proposed, all involving changes in specifications:

- The barrel could be made of steel with a higher yield strength.
- The barrel could be made with a greater and more uniform wall thickness.
- An alternative nontoxic metal shot with a hardness of about 30 to 40 HB could be developed for use in the ammunition.

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Failure to Meet Specifications

Parts sometimes do not perform to expectations because the material or processing does not conform to requirements, leaving the part with insufficient strength. For instance, a part can be damaged by decarburization, as discussed here for a spiral power spring.

Figure 8 shows two spiral power springs that were designed to counterbalance a textile-machine beam. The spring at left in Fig. 8 was satisfactory and took a normal set when loaded to solid deflection in a presetting operation. The spring at right in Fig. 8, after having been intentionally overstressed in the same manner as the satisfactory spring, exhibited 15% less reaction force than was required at 180° angular deflection because it had taken a set that was 30° in excess of the normal set.

Fig. 8 Two spiral power springs from a textile machine. Spring at left is an acceptable part, whereas spring at right took an excessive set (the inner end of the spiral is 30° out of position) because of insufficient yield strength and a decarburized surface layer.

The material in the satisfactory spring had a hardness of 45 HRC, while the material in the spring that failed had a hardness of 41.5 HRC. This represents a 10% disparity in tensile and yield strengths between the two springs. The spring that failed had a 0.08 mm (0.003 in.) thick surface layer of partial decarburization that further weakened the region of the cross section where maximum stresses are developed in the spring under load. The decarburized layer, which had a lower yield strength than the bulk material, yielded excessively during presetting; therefore the spring did not attain its specified shape in the free state following this operation.

Another material deficiency that can lead to deformation failure is variability in response to heat treatment among parts in a given production lot. Certain alloys, particularly hardenable low-alloy steels and some precipitation-hardening alloys, can vary in their response to a specified heat treatment because of slight compositional variations from lot to lot or within a given lot. This can result in some parts having too low a strength for the application even though they were properly heat treated according to specification.

Remedies for variability in response to heat treatment usually involve changes in the heat treating process, ranging in complexity from tailoring the heat treating conditions for each lot or sublot to making a small adjustment in the heat treatment specification. Material composition and microstructural specifications can also be tightened to provide more uniform heat treatment response. Experiments on each lot or sublot are almost always needed to establish parameters when heat treating conditions are tailored.

An adjustment in heat treating conditions was successful in avoiding variation in properties among sublots of heat treated AISI type 631 (17-7 PH) stainless steel Belleville washers. Two of these washers—one of which was from an acceptable sublot and the other from a deficient sublot—were subjected to examination. The washer from the acceptable sublot had developed the required hardness upon solution heat treating at 955 °C (1750 °F), followed by refrigeration at -75 °C (-100 °F) and aging. The other washer was soft after an identical heat treatment and yielded under load (flattened). The microstructure of the acceptable washer was a mixture of austenite and martensite; the structure of the washer that flattened consisted almost entirely of austenite.

Previous experience with 17-7 PH stainless steel indicated that some alloy segregation was not unusual and that relatively minor variations in composition could affect response to heat treatment, perhaps by depressing the range of martensite-transformation temperatures to a variable degree. As noted in Ref 5, the solution-treating temperature has a marked effect on the martensite start (M_s) temperature in the precipitation-hardening stainless steels that are austenitic as solution annealed and martensitic as aged (17-7 PH, AM-350, AM-355 and PH15-7Mo). Consequently, although it never was clearly established whether temperature variations inside the solution-treating furnace or minor variations in composition were responsible for the observed variability in properties of the 17-7 PH Belleville washers, all sublots attained the required strength when the solution-treating temperature was lowered to 870 °C (1600 °F).

Faulty Heat Treatment. Mistakes made in heat treating hardenable alloys are among the most common causes of premature failure. Temperatures that are either too high or too low can result in the development of inadequate or undesirable mechanical properties. Quenching a steel part too fast can crack it; quenching too slowly can fail to produce the required strength or toughness. If parts are shielded from a heating or cooling medium, they can respond poorly to heat treatment, as discussed in this section.

Two hold-down clamps, both from the same lot, are shown in Fig. 9. Both clamps were bowed to the same degree after fabrication, as intended, but the clamp at the bottom flattened when it was installed. A small percentage of the clamps, all of which were made from hardened-and-tempered 1070 steel, deformed when a bolt was inserted through the hole and tightened. The clamp at top in Fig. 9 was acceptable, with a microstructure of tempered martensite and a hardness of 46 HRC; the clamp at bottom, which deformed, had a mixed structure of ferrite, coarse pearlite, and tempered martensite and a hardness of only 28 HRC.

Fig. 9 Two hardened-and-tempered 1070 steel hold-down clamps. The clamp at top was acceptable. The clamp at bottom was slack quenched because of faulty loading practice

(stacking), and it failed by distortion (flattening) because of the resultant mixed microstructure.

Analysis of the heat treating process revealed that the parts were stacked so that occasional groupings were slack quenched as a result of shielding; this promoted the formation of softer *high-temperature* transformation products. When the loading practice was changed to ensure more uniform quenching (so that transformation to 100% martensite was accomplished on all parts), the problem was solved.

Example 4: Bending of an Aircraft-Wing Slat Track. A curved member called a slat track (Fig. 10), which supported the extendable portion of the leading edge of the wing on a military aircraft, failed by bending at one end after very short service. It was estimated that the slat track, fabricated from heat treated 4140 steel, had undergone only one high-load cycle.

Fig. 10 4140 steel slat track from a military aircraft wing. The track bent because one end did not become fully austenitic during heat treatment, producing a low-strength structure of ferrite and tempered martensite.

Investigation. Hardness measurements were taken at various points along the length of the track. The end that bent (at right, Fig. 10) had a hardness of 30 HRC, compared with a hardness of 41 HRC for the remainder of the part.

Metallographic examination of specimens from both ends of the slat track revealed that the microstructure of the end that bent contained a large number of ferrite islands in a matrix of tempered martensite. The microstructure of the opposite end contained no ferrite.

Conclusions. Bending had occurred in a portion of the slat track because service stresses had exceeded the strength of the material in a region of mixed martensite and ferrite. It was determined that the most likely cause of the mixed structure was nonuniform austenitization during heat treatment. The end that bent never became fully austenitic, because the furnace temperature was locally too low or because the soaking time was too short or both. It was decided that material or design changes were not warranted, considering the nature of the failure and the probable cause of the mixed microstructure in one end of the slat track.

Corrective Measures. Steps were taken to improve control of temperature of parts during austenitization.

Warpage and Residual Stress Effects. Warpage during heat treatment or during stress-relief annealing is also a common type of distortion failure. Warpage can result from changes in residual stress or from thermal or phase transformation induced stresses and strains that are introduced during heating or cooling. When relief of residual stress causes distortion, the amount of distortion is proportional to the decrease of the residual stress. When distortion is caused by thermal or transformational stress, the extent of the distortion is greater for parts that have complex configuration or large differences in section thickness and for faster heating or cooling rates. Warpage can also be caused by inadequate support in the heat treating furnace, leading to sagging due to the weight of the part.

Warpage can be the result of plastic deformation that occurred in some region of the part at elevated temperature or during a change in temperature. Dimensional changes accompanying stress relief are the result of relief of those stresses involving both elastic and plastic strains. Distortion that occurs during other types of heat treatment involves mainly plastic strain and generally results in high levels of residual stress in the warped part. The magnitude and distribution of residual stresses are determined by the composition, shape, size, and heat treating conditions of a given part.

Warping is most often severe in heat treatments that involve quenching. In hardenable steels, the principal cause of warping on quenching is nonuniform rates of transformation. The effect of transformational stresses may be intensified if a nonuniform composition exists; this nonuniformity may be the result of segregation, or it may be the result of processing, as in a carburized part. Such inhomogeneities may produce a variation in transformation temperature at locations that are geometrically equivalent and that cool at the same rate. Nonuniform transformational stresses that result from inhomogeneity can also occur during tempering.

Warpage can often be minimized by modifying the heat treating conditions. For example, slow heating and cooling rates are less likely to cause warping because local variances in temperature and in rates of temperature change are minimized. Preheating before austenitizing is often used as a means of minimizing warping of some tool steels and heavy sections because preheating reduces the temperature gradient between the surface and the interior of the part. Induction hardening and nitriding have been used to minimize warping when surface hardness is of primary importance to the performance of

a part. Warping is of particular concern with respect to tool steels, and preventative measures have been published (e.g., see “Control of Distortion in Tool Steels” in *Heat Treating*, Volume 4 of *ASM Handbook*).

In heat treatments requiring rapid cooling or quenching, excessive warping can usually be reduced by changing the quenching conditions. In many instances, the orientation of a part as it enters the quenchant will influence the amount of distortion that occurs. Quenching in special fixtures (quench presses) is widely used in certain industries to minimize distortion by providing different, controlled cooling rates at different locations in a given part. Martempering has also been used to minimize distortion, because transformation rates upon cooling to the M_s temperature are equalized throughout the part.

While heat treatment operations are a common cause of distortion and residual stress, it should be noted that other manufacturing operations also cause deformation directly or through introduction of residual stresses. Residual stresses are created by strain gradients in a component. The strain gradients may be created mechanically or thermally. Very few manufacturing operations create uniform strain gradients, and uniform strain from mechanical working almost never occurs. Parts that are cold worked, including both cold-formed parts and stampings, will have a residual stress pattern characteristic of the process. Machining and other finishing processes introduce surface residual stresses that are often tensile and may be significant, especially in parts with thin sections. Grinding processes can be controlled to minimize residual stresses, but less-controlled grinding processes can produce tensile surface stresses and even cracks in severe cases.

Castings, at least until they are stress relieved, may have residual stresses arising from the sequence in which the solidification occurred. The liquid material contracts as it solidifies. Regions that are the last to solidify are constrained by the surrounding material. Similarly, welding will produce residual stresses and possible distortion. Electrodischarge machining also involves localized melting and resolidification, and associated residual stresses, although on a smaller scale.

The residual stresses existing in a component will always be balanced so that the sum of the forces is 0. Thus, a compressive residual stress cannot exist without tensile residual stresses somewhere else in the component. If a machining process removes material that is under a residual stress, the stress distribution in the remaining material will change and the component may warp unacceptably during the machining process.

Shot peening is another common process in the manufacture of some types of parts and is intended to introduce a surface compressive residual stress to improve fatigue resistance. However, shot peening may also cause warping, again more likely in parts with thin sections. Furthermore, if the peening is excessive, the high surface compressive stresses will be balanced by high subsurface tensile stresses, which, in turn, can cause further problems. Carburizing usually results in compressive residual stress at the surface.

Regardless of the origin of the residual stresses, they will add to the service-induced stresses and, if a critical point is exceeded, local yielding and distortion will be observed at a nominal stress at which yielding would not be anticipated. Residual stress effects can be difficult to verify in a failure analysis because the residual stresses existing in the deformed or fractured part are different than they were prior to the failure and the presence or absence of residual stress is not obvious in any case. Methods for measurement and study of residual stress include hole drilling, x-ray diffraction, and neutron diffraction techniques (e.g., see the articles “Residual Stress Measurements” in *Mechanical Testing and Evaluation*, Volume 8 of *ASM Handbook*, and “X-Ray Diffraction Residual Stress Measurements in Failure Analysis” in this Volume). More detailed discussion of residual stress is also available in Ref 6.

For the purposes of this article, without going into the details on the many causes and effects of residual stress, the main point is that residual stress is closely tied to distortion. The failure analyst must be aware of the potential contribution of residual stress to failures involving both distortion and fracture. Various types of residual stresses can be introduced in complex ways by manufacturing operations. For example, it is commonly understood that quenching operations create surface longitudinal and hoop stresses. This is true for lower-hardenability materials, but not for high-hardenability steels (Ref 7). Another example of the complexity of the effect of manufacturing on residual stresses is shot peening. If peening intensity is too high, large tensile stresses can occur below the surface, which can cause problems in bending and torsion.

Faulty Case Hardening. Carburizing, which both increases the surface hardness of a part and provides resistance to wear and indentation, can, if improperly controlled, produce a case that has too low or too high a carbon content. With too low a carbon content, the surface may not be hard enough to withstand normal service loads. This condition may be accompanied by shallow case depth, which aggravates the problem. With an excessively high carbon content, which generally is the result of a high carbon potential or improper diffusion during the carburizing cycle, large amounts of retained austenite may be present in the carburized zone after heat treatment, depending on the composition of the steel. Retained austenite softens the surface and, under certain conditions, may transform to martensite in service. When transformation in service occurs, the resulting untempered martensite may crack and thus promote early failure by surface fatigue, or a distortion failure in a close-fitting assembly may occur because of the volume change that accompanies transformation.

In addition to the carbon level and surface hardness, a carburized case is also characterized by its thickness or depth below the surface. Improper case depth can lead to some of the same types of problems associated previously with improper carbon content. The case depth is also important when maximum stresses occur at the surface of a part, as they

do in bending or torsion. To avoid failures due to distortion, as well as other failures, a designer must define the desired case characteristics and then work with the heat treater to ensure that the process is capable and controlled to produce that case. This communication must also include the necessary quality control. As should be clear from this discussion, surface hardness testing cannot completely characterize a case.

Example 5: Seizing of a Spool-Type Hydraulic Valve. Occasional failures were experienced in spool-type valves used in a hydraulic system. When a valve would fail, the close-fitting rotary valve would seize, causing loss of flow control of the hydraulic oil. The rotating spool in the valve was made of 8620 steel and was gas carburized. The cylinder in which the spool fitted was made of 1117 steel, also gas carburized.

Investigation. Low-magnification visual examination of the spool and cylinder from a failed valve revealed some burnishing, apparently the result of contact between the spool and the inside cylinder wall. Measurement of the surface profile of the spool showed no evidence of wear or galling. When this profile was compared with that of a spool from a valve that operated satisfactorily, no significant difference was found.

Metallographic sections were made of the cylinder from the failed valve and from the one that operated satisfactorily. The microstructure of the carburized case on the cylinder from the satisfactory valve was well-defined martensite interspersed with some austenite (light-etching) areas, Fig. 11a. On the other hand, the microstructure of the case from the failed valve contained a much greater amount of retained austenite, especially near the surface (Fig. 11b). In addition, some patches of untempered martensite were found close to the surface of the failed valve cylinder with use of a more aggressive etching procedure.

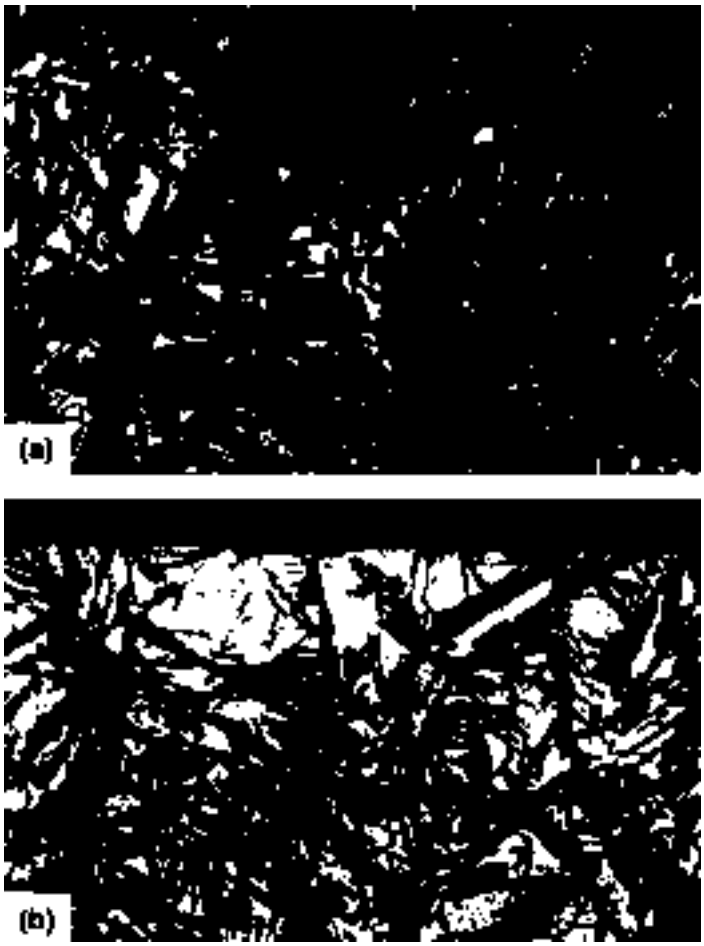


Fig. 11 Cross sections through the carburized 1117 steel cylinders from two spool-type hydraulic valves. The cylinder of the valve that operated satisfactorily (a) had little retained austenite in the case, whereas the cylinder of the seized valve (b) had much retained austenite that transformed to martensite in service, resulting in size distortion (growth). 500×

Microhardness traverses of the two sections revealed that, in general, the hardness of the case from the failed part was about 100 Knoop points lower than the case hardness of the unfailed part. At the surface, the case hardness of the failed part was 300 Knoop points lower.

Conclusions. Momentary sliding contact between the spool and the cylinder wall (probably during valve opening) caused unstable retained austenite in the failed cylinder to transform to martensite. The increase in volume resulted in sufficient

size distortion (growth) to cause interference between the cylinder and the spool, seizing, and loss of flow control. The failed parts had been carburized in a process in which the carbon potential was too high, which resulted in a microstructure having excessive retained austenite after heat treatment.

Corrective Measure. The composition of the carburizing atmosphere was modified to yield carburized parts that did not retain significant amounts of austenite when they were heat treated. After this change was made, occasional valve failure by seizing ceased.

Faulty Repairs. Products are often repaired to correct deficiencies that are found in new parts during quality-control inspections or in used parts after they have deteriorated in service. Repair welding and brazing are generally recognized as potential sources of unwanted alterations of the properties of heat treatable alloys. Parts can be made softer or more brittle by careless repair, depending on the alloy and the conditions under which the repair was made.

Substitution of a part, particularly a fastener, whose properties do not match the properties of the part it replaces can lead to failure of the substitute part, failure of another part, or both, as in the following instance.

A carrying handle of complex configuration was secured to a heavy portable device by means of a ring clamp at one end of the handle and a special $\frac{1}{4}$-20 hardened bolt through a small flange at the other end. For some unknown reason, the special slotted hexagonal-head bolt was replaced by a standard commercial hexagonal-head cap screw whose hardness was 93 HRB instead of the specified hardness of 28 HRC.

The commercial cap screw was distorted in service (Fig. 12a), causing the handle to become loose at one end. This in turn caused eccentric and excessive loading on the small flange of the handle with the result that the flange bent and then broke. Figure 12(b) shows the broken flange; the bright area next to the hole is where the loose flange chafed against the distorted bolt and is an indication of the eccentric load distribution on the flange. A fastener of the correct hardness and length is shown in Fig. 12(c) for comparison with the distorted commercial cap screw.

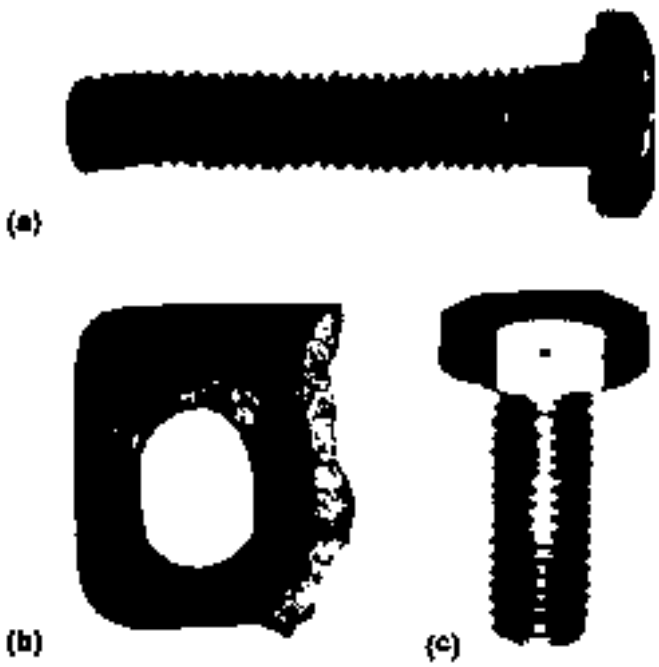


Fig. 12 Part substitution that resulted in a distortion failure. (a) Distorted commercial cap screw that was used as a replacement for a hardened bolt. (b) Carry-handle flange that broke because the cap screw bent. (c) Correct replacement part

References cited in this section

5. A.J. Lena, Precipitation Reactions in Iron-Base Alloys, in *Precipitation from Solid Solution,* American Society for Metals, 1959. p 224–327

6. G. Totten, M. Howes, and T. Inoue, Ed., *Handbook of Residual Stress and Deformation of Steel,* ASM International, 2002

7. F. McClintock and A. Argon, Mechanical Behavior of Material, Chapter 12, *Residual Stress*, Addison-Wesley, 1966, p 420–434

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Analyzing Distortion Failures

Distortion failures are often considered to be instances of relatively simple phenomena that are easy to analyze because deformation can occur only when the applied stress exceeds the flow strength of the material. On the contrary, distortion is not always the result of simple overload or use of an improperly processed part. Analysis of a distortion failure often must be exceptionally thorough and rigorous to determine the root cause of failure and, more importantly, to specify proper corrective action. The analyst must consider factors that may not have been anticipated in design of the part, such as material substitutions or process changes during manufacture and the misuse, abuse, or occurrence of complex stress fields in service.

A seemingly innocent substitution of material resulted in distortion of small volute springs made of cold-worked spring-tempered Inconel. Normally, the material was purchased as cold-flattened wire, but one lot of springs was formed from cold-rolled and slit strip because flattened wire could not be obtained in time to meet the delivery schedule. After a presetting operation, the strip springs were consistently out of tolerance; they had taken an excessive set. The Inconel strip had a hardness of 360 HV, compared to 390 HV for the wire that normally was used. This represents a difference of about 10% in strength and accounts for the observed distortion. Had this problem gone undetected during manufacture, it might well have resulted in distortion failures in service.

Analytical Procedure. The general process of materials failure analysis, adapted from the typical procedure for analysis of a distortion failure, is briefly summarized by the following ten steps:

1. Define the effect of the failure on the structure or assembly, and define the desired results of corrective action.
2. Obtain all available design and service information.
3. Examine the distorted part visually, making a record of observations, including a sketch or a photograph of the distorted part along with an undistorted part for comparison. Note all pertinent measurements of dimensions. It is usually helpful to enter these measurements, which should be made with at least the same precision as in a quality-control inspection, alongside the design dimensions on a blueprint of the part.
4. Perform laboratory tests as necessary to confirm the composition, structure, and other chemical or metallurgical characteristics of the distorted part.
5. Trace the failed part through all manufacturing processes to discover whether process deviations occurred during production.
6. Compare the actual conditions of service with design assumptions.
7. Compare the actual material properties with design specifications.
8. Determine whether any differences found in the preceding two steps fully account for the distortion observed in the failed structure. If the differences do not fully account for the observed distortion, the information obtained in the second or fourth step is incorrect or incomplete.
9. Prepare alternative courses of action to correct the variant factors that caused the observed distortion, and select the course that seems most likely to produce the desired result, which was defined in the first step.
10. Test the selected course of corrective action to verify its effectiveness. Evaluate side effects of the corrective action, such as its effects on cost or on ease of implementation.

Example 6: Deformation of a Gas-Nitrided Drive-Gear Assembly. Slipping of components in the left-side final drive train of a tracked military vehicle was detected after the vehicle had been driven 13,700 km (8500 miles) in combined highway and rough-terrain service. No abnormal service conditions were reported in the history of the vehicle. The slipping was traced to the mating surfaces of the final drive gear and the adjacent splined coupling sleeve (Fig. 13). Failure analysis was conducted to determine the cause of the malfunction and to recommend corrective measures that would prevent similar failures in other vehicles.

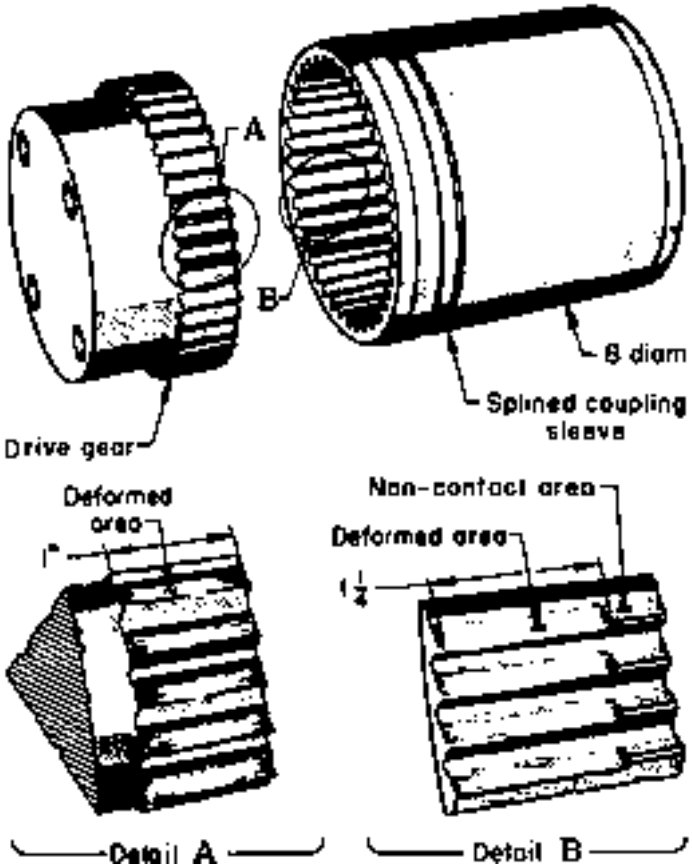


Fig. 13 Gas-nitrided 4140 steel (27–31 HRC) drive-gear assembly in which gear teeth deformed because of faulty design and low core hardness. Details A and B show deformed areas on drive-gear teeth and mating internal splines. Dimensions given in inches

Material and Fabrication. Specifications required that the gear and coupling be made from 4140 steel bar oil quenched and tempered to a hardness of 265 to 290 HB (equivalent to 27 to 31 HRC) and that the finish-machined parts be single-stage gas nitrided to produce a total case depth of 0.5 mm (0.020 in.) and a minimum surface hardness equivalent to 58 HRC.

Visual Examination. Low-magnification examination of the drive gear and coupling showed that the teeth of the gear and the mating contact surfaces of the internal splines on the coupling were almost completely worn away. The spline surfaces had been damaged mainly by severe indentation and plastic deformation. No spalling, cracking, or other indications of fatigue damage were visible on the splines nor was there any indication that abrasive wear had been the mechanism of failure.

Metal on the drive side of the gear teeth had been plastically deformed and subsequently removed due to wear (details A and B, Fig. 13). The damaged areas on the splines were wider (axial dimension) than the gear teeth, indicating that there had been excessive lateral play between the components.

Examination of the surfaces of the internal splines that contacted a gear on the opposite end of the coupling (not illustrated) revealed a much smaller amount of deterioration than on the failed end. For example, the machined flats on the crests of the splines showed no damage, the follower sides of the splines had very shallow, almost imperceptible wear areas that were only about 0.1 mm (0.004 in.) wide, and the drive side of each spline had a gently radiused, concave area of damage approximately 6.4 × 25 mm (0.25 × 1 in.) and 0.13 mm (0.005 in.) deep at its center. However, the amount of damage on the drive side of these splines was considered to be excessive for only 13,700 km (8500 miles) of service.

Conformance to Material Specifications. Spectrographic analysis of the two components confirmed that composition was within the range specified for 4140 steel. Core hardness was in the required range of 27 to 31 HRC (264 to 294 HB). Total case depth, as determined by the microscopic method on polished specimens etched in 2% nital, was satisfactory. Surface hardness, as measured using a Knoop indenter, was equivalent to 50 HRC, substantially lower than the required value of 58 HRC.

Microstructure of the cases and cores of the two components was examined at 500× on polished sections etched in 2% nital. There was a white layer (nitrogen-rich iron nitride, Fe_2N) about 0.025 mm (0.001 in.) thick on the surfaces of the gear teeth and the splines, and grain-boundary networks of iron nitride were present to a slight degree near the surface. The microstructure of the core consisted mainly of tempered martensite, but contained large amounts of blocky ferrite.

Gear and Spline Configuration. Measurements of the gear teeth and splines in areas showing little damage established that the parts had been manufactured in conformance with the engineering drawings. The splines were straight axially and convex radially. However, the gear teeth that failed were convex in both directions in the vicinity of the pitch line. This design, which was intended to facilitate alignment and adjustment, provided extremely small contact areas between tooth and spline surfaces and hence very heavy localized loading. The gear that engaged the other end of the internally splined coupling (not shown in Fig. 13) was designed to provide larger contact areas with the splines in the vicinity of the pitch line, providing lighter and more uniform local loading.

Conclusions. The premature failure occurred by crushing, or cracking, of the case as a result of several factors, which are listed below in order of importance:

- Design that produced excessively high localized stresses at the pitch line of the mating components
- Specification of a core hardness (27 to 31 HRC) too low to provide adequate support for a 0.5 mm (0.020 in.) thick case or to permit attainment of the specified surface hardness of 58 HRC after nitriding; actual surface hardness was 50 HRC
- The presence of large amounts of blocky ferrite in the core (a microstructure conducive to case crushing) as a result of faulty heat treatment before nitriding
- The presence of a white nitride layer about 0.025 mm (0.001 in.) thick at the surface and of nitride networks in the case

Recommendations. Measures to correct the first two deficiencies listed above were recommended as the most important steps in obtaining adequate drive-train performance. First, the excessively high local stresses at the pitch line should be reduced to an acceptable level by modifying the gear-tooth contour in the vicinity of the pitch line to provide a wider and longer initial contact area than in the original design. Second, a core hardness of 35 to 40 HRC should be specified to provide adequate support for the case and to permit attainment of the specified surface hardness of 58 HRC. Closer control of heat treating, which would be necessary to produce the recommended higher core hardness consistently, would also eliminate the presence of blocky ferrite in the core.

For maximum service life, consideration should also be given to controlling single-stage gas nitriding to minimize the thickness of the white layer and the extent of nitride networks in the case or to use double-stage gas nitriding to provide a diffused nitride layer and specifying final lapping or honing to remove the white layer.

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Special Types of Distortion Failure

Analysis of distortion failures can be particularly difficult when there is no apparent permanent deformation of the part or when complex stress fields are involved. In this section, three types of distortion failure are discussed, which may provide useful insights into the problems of analyzing unusual mechanisms of distortion.

Elastic Distortion. A distortion failure does not necessarily involve yielding under a single application of load. Unacceptable distortion can occur within the elastic regime. All parts will deflect elastically under load. But if a part ordinarily made of a high-modulus alloy is made of a low-modulus alloy, it will deflect more under a given load than if it were made of the high-modulus alloy. If this greater amount of deflection places the part in the path of another part in an assembly, it could be said to have failed by elastic distortion. As mentioned earlier, a change in the modulus of a material because of a change in temperature can cause an elastic-distortion failure. Elastic buckling of a long, slender column is another type of distortion failure in which the yield strength of the material is not exceeded (unless the structure collapses).

Ratcheting. Cyclic strain accumulation, or ratcheting, requires that a part be stressed by steady-state loading, either uniaxial or multiaxial, and that a cyclically varying strain in a direction other than the direction of principal stress be superimposed on the part. In ratcheting, an oscillating load or a cyclic variation of temperature strains the material beyond the yield point on alternate sides of a single member, or on alternate members of a structure, during each half-cycle. With succeeding cycles, plastic strain accumulates, with the result that one or more of the overall dimensions of the member or the structure change relatively uniformly along the direction of steady-state stress. Deformation produced by a cyclic variation in load is known as isothermal ratcheting (even though a temperature change may occur simultaneously with the

load variation). Progressive growth due to plastic strain incurred during a change in temperature is called thermal ratcheting. Ratcheting may ultimately result in ductile fracture or in failure by low-cycle fatigue.

At elevated temperatures, ratcheting must be distinguished from creep or stress relaxation. Ratcheting is solely a strain-dependent phenomenon, whereas creep and stress relaxation are time-dependent phenomena. Exposure to elevated temperature for an extended length of time, typically hours to thousands of hours, is necessary for creep or stress relaxation to occur, but extensive deformation by ratcheting can occur in short periods of time—sometimes only minutes. Ratcheting can appear to be time dependent when the cyclic strains are imposed at regular intervals. However, the factor that distinguishes ratcheting is the occurrence of plastic strain during both halves of the cyclic variation.

In general, the proper corrective action for failures by ratcheting involves changing the design of the part or the conditions of service to reduce the magnitude of the service stresses or specifying a material with a higher yield strength for the application.

Inelastic Cyclic Buckling. Cyclic strain softening is a continuous decrease of elastic limit or tangent modulus that occurs with imposition of alternating stresses below the net section yield strength. Columns made from materials that exhibit this behavior can fail by lateral displacement at the midspan (buckling) under stresses much lower than those predicted by classical design.

Table 1 presents the results of a test in which cylindrical specimens of cold-worked 1020 steel, resembling tensile specimens with threaded ends, were stressed by alternating tensile and compressive loads of equal magnitude (Ref 8). Buckling occurred with stresses approximately 60 to 90% of the 0.2% offset yield strength, and the cycles ranged from $\frac{1}{4}$ at the highest stresses to 1900 at the lowest. When an aluminum alloy was tested in the same manner, buckling did not occur in a range of stresses below the 0.2% offset yield strength. Aluminum alloys are among those that do not exhibit cyclic strain softening.

Table 1 Inelastic cycle buckling of cylindrical specimens of cold-worked 1020 steel

Peak stress		Stress cycles		
MPa	**ksi**	**At buckling**	**At failure**	**Failure mode**
600–648	87–94	$\frac{1}{4}$	$\frac{1}{4}$	(a)
552	80	14	20	(a)
490	71	130	143	(a)
441	64	380	691	(b)
421	61	1900	2377	(b)
400	58	…	4730	(c)

Tensile strength of the steel: 690 MPa (100 ksi); yield strength, 620 MPa (90 ksi).
(a) Buckling, with or without fracture.
(b) Buckling followed by buckling-induced fracture.
(c) Low-cycle fatigue fracture.
Source: Ref 8

Reference cited in this section

8. C.R. Preschmann and R.I. Stephens, Inelastic Cyclic Buckling, *Exp. Mech.*, Vol 12 (No. 9), Sept 1972, p 426–428

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Deformation Related to Other Types of Failure

Previous discussion and examples have shown how distortion of a critical component can impede the functionality of the component itself or a larger assembly. In some cases, the root cause of the distortion is a problem elsewhere in a

mechanism. The distortion results from overload (stress too high), material deficiency (material too weak), or a manufacturing process variation (residual stress related). Sometimes the root cause of the distortion is obvious, other times more subtle. Distortion is also often observed in connection with other types of failures. The characteristics of the distortion observed, or the absence of distortion, can provide important and sometimes critical insights into the sequence of events leading to an ultimate failure.

To restate the obvious, deformation occurs when stresses exceed the critical point for yield of the material. Fracture may or may not be the end result in the sequence, but if there is distortion, it most certainly occurred prior to the fracture. Corrosion or wear may contribute by removing material and decreasing the load-carrying capability of a cross section. While no two situations are exactly the same, a few composite examples illustrate the concept and the related thought processes of the failure analyst:

- A motor vehicle runs off the road, and the driver claims there was a problem with control immediately prior. Inspection of the damaged vehicle after the accident cannot locate all the parts associated with the steering and suspension, so the first investigator concludes that these parts were loose, finally fell off, and led to the loss of control. Subsequent engineering failure analysis and more careful examination of the parts that are available show clear distortion on the parts that were adjacent to the missing items. This deformation could not have occurred if the missing parts had simply fallen off. The forces of the collision caused deformation followed by fracture and separation of some parts from the remainder of the vehicle. Figure 14 shows one instance of the deformation observed in this type of situation. The threads of a missing bolt left impressions on the inside surface of a hole as the members sheared with respect to each other.
- In another motor vehicle accident, fractured parts are found after the accident. Again, the driver claims a problem with control. The fracture surfaces are so damaged that microscopic examination is inconclusive regarding the fracture mechanism. (In typical accidents, this damage occurs by post-accident smashing of the fracture, or exposure to heat, fire, or corrosive conditions.) One investigator concludes that a critical component failed in fatigue and led to the loss of control. However, careful examination of the component and comparison with an exemplar of the same exact part number shows the deformation that occurred prior to separation. Furthermore, the directionality of this deformation was such that it could not have occurred during vehicle operation, but it was consistent with the forces of the collision. Thus, the collision caused deformation followed by fracture. Macroscopic suggestions of a fatigue crack could be explained by geometric stress concentrators.
- A component of a piping system is erroneously fabricated from the wrong material and is in the process of failing prematurely by a combination of corrosion, erosion, and wear. The plant in which this component is installed experiences an explosion and fire, and the deteriorating component is found deformed and fractured afterward. Careful analysis of the material properties and component dimensions, including deformation patterns, is undertaken to determine if the deformation and fracture is consistent with normal system operating pressures and temperatures acting on a deteriorated component, causing a leak, and subsequent explosion. Analysis is also undertaken to consider the alternative possibility that the forces of an explosion originating elsewhere may have caused the deformation and fracture of this component.
- Some gas appliance flexible connectors manufactured and sold during the 1960s and 1970s are known to be susceptible to separation of the end fitting through long-term and progressive deterioration. Unfortunately, a small number of such connectors still remain in service. A home is destroyed by a gas explosion and fire that occur under suspicious circumstances. However, investigators also find one of the suspect connectors and question whether it separated and leaked natural gas before the explosion or whether it separated as a result of the forces of the explosion. Careful inspection of the post-explosion appearance of the connector can provide information to answer this question.

Fig. 14 Deformation from threads that left impressions on the inside surface of a hole as the members sheared with respect to each other.

These few examples illustrate that understanding and interpretation of distortion may be important in a wide range of failure investigations. These examples are by no means unique or unusual, and many readers may no doubt have experienced similar situations. These examples also show the need for a thorough investigation to ensure that the true root cause is identified. For example, when a component is found that does not meet specification, it is most tempting and may seem most expedient for the investigator to simply blame the deficient component. Similarly, if the cause of a failure is not immediately apparent, it may be easiest to blame the operator, driver, or pilot. However, if the true root cause is overlooked due to a superficial investigation, then it is possible that the accident, fire, and so forth will occur again, with unnecessary loss of life or property.

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

Deformation by Design

Distortion and deformation are not always indications of failure. It is hopefully well recognized that any component experiencing a stress will also exhibit an associated elastic distortion. In most designs, stress levels are such that this distortion is minimal and may not even be noticed. However, some components are expected to experience significant elastic and sometimes even plastic distortion during their service lives.

Springs of all types are an obvious example of a cyclic elastic distortion by design. Many of the examples earlier in this article have dealt with failures of springs through development of unwanted plastic distortion. Springs can also be susceptible to fatigue failures. (See the article “Fatigue Failures” in this Volume.)

Threaded fasteners are another type of component in which understanding of the elastic distortion is important. In the design of bolted connections, the elastic stretch of the threaded fastener is large relative to the compression of the joint members. With the compliance of the fastener being significantly greater than the joint, the cyclic stress amplitude on the fastener is minimized. It is generally desired to maintain as high a clamp load as possible without introducing plastic deformation into either the fastener or the joint components. Failure analysis of bolted connections requires consideration of the initial clamp load as well as factors that may cause a reduction in the clamp load in service. Reference 9 is a useful resource for those analyzing the failure of a bolted connection.

With respect to plastic distortion, a major consideration in the design of any automobile is how it will perform under various accident scenarios. The structure of the vehicle is sacrificed to protect the occupants. The vehicle components must possess a combination of strength and the capability to deform and absorb the energy of the collision, thereby reducing the forces and energy experienced by the occupants. Sometimes, and not surprisingly, the deformation process leads to fracture. Some observers may decry the deformed appearance of the automobile after an accident, saying “they don't make them like they used to.” More sophisticated analysis will consider the vehicle dynamics to calculate the forces of the accident. This, in turn, allows consideration of the stresses experienced by the vehicle components, and whether or not the deformation observed is that which would be expected based on design analysis and actual crash testing. For further details, see the article “Modeling and Accident Reconstruction” in this Volume as well as SAE Standards and Federal Motor Vehicle Safety Standards.

Armor provides another example of a type of engineered product that deforms in service without being considered “failed.” The armor succeeds when it is able to resist the bullet or projectile. While some armors are high strength and distortion will not be observed, others absorb the energy of the projectile by deformation. Again, success or failure is determined by comparison with design expectations.

Reference cited in this section

9. J.H. Bickford, *Introduction to the Design and Behaviour of Bolted Joints*, 3rd ed., Marcel Dekker, 1995

Acknowledgment

This article was revised from previous editions in *Failure Analysis and Prevention,* Volume 10, *Metals Handbook* 8th edition (1975) and Volume 11, *ASM Handbook,* 9th edition (1986).

Analysis of Distortion and Deformation

Revised by Roch J. Shipley and David A. Moore, Packer Engineering and William Dobson, Binary Engineering Associates, Inc.

References

1. J.W. Jones, Limit Analysis, *Mach. Des.,* Vol 45 (No. 23), 20 Sept 1973, p 146–151
2. D. Goldner, Plastic Bending in Tubular Beams, *Mach. Des.,* Vol 45 (No. 24), 4 Oct 1973, p 152–155
3. D.J. Wulpi, *Understanding How Components Fail,* 2nd ed., ASM International, 1999, p 16–19
4. J.A. Collins, *Failure of Materials in Mechanical Design,* 2nd ed., Wiley and Sons, 1993
5. A.J. Lena, Precipitation Reactions in Iron-Base Alloys, in *Precipitation from Solid Solution,* American Society for Metals, 1959. p 224–327
6. G. Totten, M. Howes, and T. Inoue, Ed., *Handbook of Residual Stress and Deformation of Steel,* ASM International, 2002
7. F. McClintock and A. Argon, Mechanical Behavior of Material, Chapter 12, *Residual Stress,* Addison-Wesley, 1966, p 420–434
8. C.R. Preschmann and R.I. Stephens, Inelastic Cyclic Buckling, *Exp. Mech.,* Vol 12 (No. 9), Sept 1972, p 426–428
9. J.H. Bickford, *Introduction to the Design and Behaviour of Bolted Joints,* 3rd ed., Marcel Dekker, 1995

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Introduction

A SHAFT is a metal bar—usually cylindrical in shape and solid, but sometimes hollow—that is used to support rotating components or to transmit power or motion by rotary or axial movement. Even fasteners, such as bolts or studs, can be considered to be stationary shafts, usually with tensile forces, but sometimes combined with bending and/or torsional forces. In addition to failures in shafts, this article will discuss failures in connecting rods, which translate rotary motion to linear motion (and conversely) and in piston rods, which translate the action of fluid power to linear motion.

Shafts operate under a broad range of service conditions, including dust-laden or corrosive atmospheres and temperatures that vary from extremely low, as in arctic or cryogenic environments, to extremely high, as in gas turbines. In addition, shafts may be subjected to a variety of loads—in general, tension, torsion, compression, bending, or combinations of these. Shafts are also sometimes subjected to vibratory stresses.

Apart from wear by bearings, which can be a major contributor to shaft failure (see the section “Wear” in this article), the most common cause of shaft failure is metal fatigue. Fatigue is a weakest link phenomenon; hence, failures start at the most vulnerable point in a dynamically stressed area—typically a stress raiser, which may be mechanical, metallurgical, or sometimes a combination of the two. Mechanical stress raisers include such features as small fillets, sharp corners, grooves, splines, keyways, nicks, and press or shrink fits. Shafts often break at edges of press-fitted or shrink-fitted members, where high degrees of stress concentration exist. Such stress concentration effectively reduces fatigue resistance, especially when coupled with fretting. Metallurgical stress raisers may be quench cracks, corrosion pits, gross nonmetallic inclusions, brittle second-phase particles, weld defects, or arc strikes.

Occasionally, brittle fractures are encountered, particularly in low-temperature environments or as a result of impact or a rapidly applied overload. Brittle fracture may thus be attributable to inappropriate choice of material because of incomplete knowledge of operating conditions and environment or failure to recognize their significance, but it may also be the result of abuse or misuse of the product under service conditions for which it was not intended.

Surface treatments can cause hydrogen to be dissolved in high-strength steels and may cause shafts to become embrittled even at room temperature. Electroplating, for instance, has caused failures of high-strength steel shafts. Baking treatments applied immediately after plating are used to ensure removal of hydrogen.

Ductile fracture of shafts is usually caused by accidental overload and is relatively rare in normal operation. Creep, a form of distortion at elevated temperatures, can lead to stress rupture and can also cause shafts having close tolerances to fail because of excessive changes in critical dimensions.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Fracture Origins

Fractures of shafts originate at points of stress concentration either inherent in design or introduced during fabrication or operation. Design features that concentrate stress include ends of keyways, edges of press-fitted members, fillets at shoulders, and edges of oil holes. Stress concentrators produced during fabrication include grinding damage, machining marks or nicks, and quench cracks resulting from heat treating operations.

Frequently, stress concentrators are introduced during hot or cold forming of shafts; these include surface discontinuities, such as laps, seams, pits and forging laps, and internal imperfections, such as bursts. Internal

stress concentrators can also be introduced during solidification of ingots from which forged shafts are made. Generally, these stress concentrators are internal discontinuities, such as pipe, segregation, porosity, shrinkage, and nonmetallic inclusions.

Fractures also result from bearing misalignment that is either introduced at assembly or caused by deflection of supporting members in service, from mismatch of mating parts, and from careless handling in which the shaft is nicked, gouged, or scratched.

To a lesser degree, shafts can fracture from misapplication of material. Such fractures result from use of materials having high ductile-to-brittle transition temperatures, low resistance to hydrogen embrittlement, temper embrittlement, or caustic embrittlement, or chemical compositions or mechanical properties other than those specified. In some instances, fractures may originate in regions of partial or total decarburization or excessive carburization, where mechanical properties are different because of variations in chemical composition.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Examination of Failed Shafts

As with any failure, examination of a failed shaft should include gathering as much background information as possible about the shaft. This information should include design parameters, operating environment, manufacturing procedures, and service history. Detailed knowledge of these factors can often be helpful in guiding the direction of failure investigation and corrective action.

Design Parameters. The failure analyst should have copies of the detail and assembly drawings as well as the material and testing specifications that involve the shaft. Potential stress raisers or points of stress concentration, such as splines, keyways, cross holes, and changes in shaft diameter, should be noted. The type of material, mechanical properties, heat treatment, test locations, nondestructive examination used, and other processing requirements should also be noted. Special processing or finishing treatments, such as shot peening, fillet rolling, burnishing, plating, metal spraying, and painting, can influence performance, and the analyst should be aware of such treatments.

Mechanical Conditions. How a shaft is supported or assembled in its working mechanism and the relationship between the failed part and its associated members can be valuable information. The number and location of bearings or supports, as well as how their alignment may be affected by deflections or distortions that can occur as a result of mechanical loads, shock, vibrations, or thermal gradients, should be considered.

The method of connecting the driving or driven member to the shaft, such as press fitting, welding, or use of a threaded connection, a set screw, or a keyway, can influence failure. It is also important if power is transmitted to or taken from the shaft by gears, splines, belts, chains, or torque converters.

Manufacturing records may indicate when the part was made and the material supplier, heat, or lot number. Inspection records may provide the inspection history of the part and indicate any questionable areas.

Service History. Checking the service records of an assembly should reveal when the parts were installed, serviced, overhauled, and inspected. These records should also show whether service or maintenance operations were conducted in accordance with the manufacturer's recommendations. Often, talking with operators or maintenance personnel may reveal pertinent unrecorded information.

Initial Examination. Ideally, the entire assembly that was involved in the failure should be made available for examination. However, in the real world this is not always possible. Samples of oil, grease, and loose debris should be carefully removed from all components, identified, and stored for future reference. The components can then be cleaned with a solvent that will not remove or obliterate any rust, oxidation, burnishing marks, or other pertinent evidence.

Surfaces of all areas that may have been involved in or may have contributed to the failure should be examined, noting scuff marks, burnished areas, abnormal surface blemishes, and wear. These marks should be associated with some abnormal service condition, if possible. In addition, the failed part should be examined to establish

the general area or location of failure, noting the proximity to any possible stress concentrations found when examining the part drawings.

Fracture surfaces should be examined visually to determine if there are indications of one or more fracture mechanisms and if there is an apparent crack origin. Surfaces of the component adjacent to the fracture surface should be examined for secondary cracks, pits, or imperfections. Photographs should be taken to record the condition of the pertinent parts before physical evidence is destroyed by subsequent examinations.

Nondestructive methods of inspection, such as ultrasonic inspection, can sometimes provide useful information. Such methods may reveal other cracks that have not progressed to rupture; these cracks may have other cracks or fracture surfaces that are not as badly damaged as the primary fracture and that can be diagnosed more readily.

Also, some machines and structures may have other shafts similar to the one that failed; these shafts may have the same service history as the failed shaft, and examination may reveal cracks that can provide useful information. Whenever possible, the analyst should also compare shafts that are from machines in service but have not failed.

Macroscopic Examination (from less than actual size to approximately 50×). Many characteristic marks on fracture surfaces, though visually identifiable, can be better distinguished with macroscopic examination. A magnifying glass or binocular microscope can be used to study the unique vestigial marks left on tensile-fracture surfaces in the form of fibrous, radial, and shear-lip zones from which the relative amounts of ductility and toughness possessed by the metal can be appraised. In the study of fatigue fractures, macroscopic examination of such features as beach marks, ratchet marks, fast-fracture zones, and crack-initiation sites, if present, may yield information relative to the kinds and magnitude of the stresses that caused failure.

Microscopic Examination (50 to 2000×.) Metallographic sections taken through fractures are used to classify fracture paths (transgranular or intergranular), to establish the mode of fracture (shear or cleavage), and to locate and identify crack-initiation sites. Plating before mounting can be used to preserve the edge or edges of the fracture surface. In addition, metallographic examination can reveal microstructure near the fracture surface, the grain size of the material, and the presence of undue segregation, inclusions, alloy concentrations, brittle grain-boundary phases, decarburization, and fabricating imperfections.

Scanning Electron Microscopy (SEM) and Transmission Electron Microscopy (TEM) Examination (10 to More Than 20,000×). The optical metallograph is limited for fracture studies by its restricted depth of field (the scanning electron microscope has a depth of field about 300 times that of the optical microscope). Thus, the electron microscope is better suited for fractographic work. With it, fractures may be classified by fracture path, fracture mechanism, and fracture features. There are two fracture paths: transgranular (or transcrystalline) and intergranular (or intercrystalline).

The fracture mechanisms and related microscopic features of transgranular fracture are microvoid coalescence (dimples, associated with ductile fracture), tearing (tear ridges), cleavage (river patterns, feather marks, Wallner lines, and cleavage tongues, all associated with brittle fracture), and fatigue (striations and tire tracks).

The fracture mechanism of intergranular fracture is grain-boundary separation. The causes may be the presence of grain-boundary phases, alloy-depleted boundaries, and environmental or mechanical factors such as stress-corrosion cracking, hydrogen damage, heat damage, and triaxial stress states.

Mechanical testing and chemical analysis occasionally pinpoint the cause of a failure as wrong material, improper heat treatment, or in-service changes in properties. Hardness testing and spectroscopic analysis should be performed as a matter of course. Impact tests, tensile tests, and other special mechanical tests may be performed if manufacturing procedures are in doubt or if other paths of investigation are not fruitful.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Stress Systems Acting on Shafts

The stress systems acting on a shaft must be clearly understood before the cause of a fracture in that shaft can be determined. Also, both ductile and brittle behavior under static loading or single overload, as well as the characteristic fracture surfaces produced by these types of behavior, must be clearly understood for proper analysis of shaft fractures.
Figure 1 shows simplified, two-dimensional, free-body diagrams illustrating the orientations of the normal-stress and shear-stress systems at any internal point in a shaft loaded in pure tension, torsion, and compression. Also, the single-overload fracture behavior of both ductile and brittle materials is illustrated for each type of load.

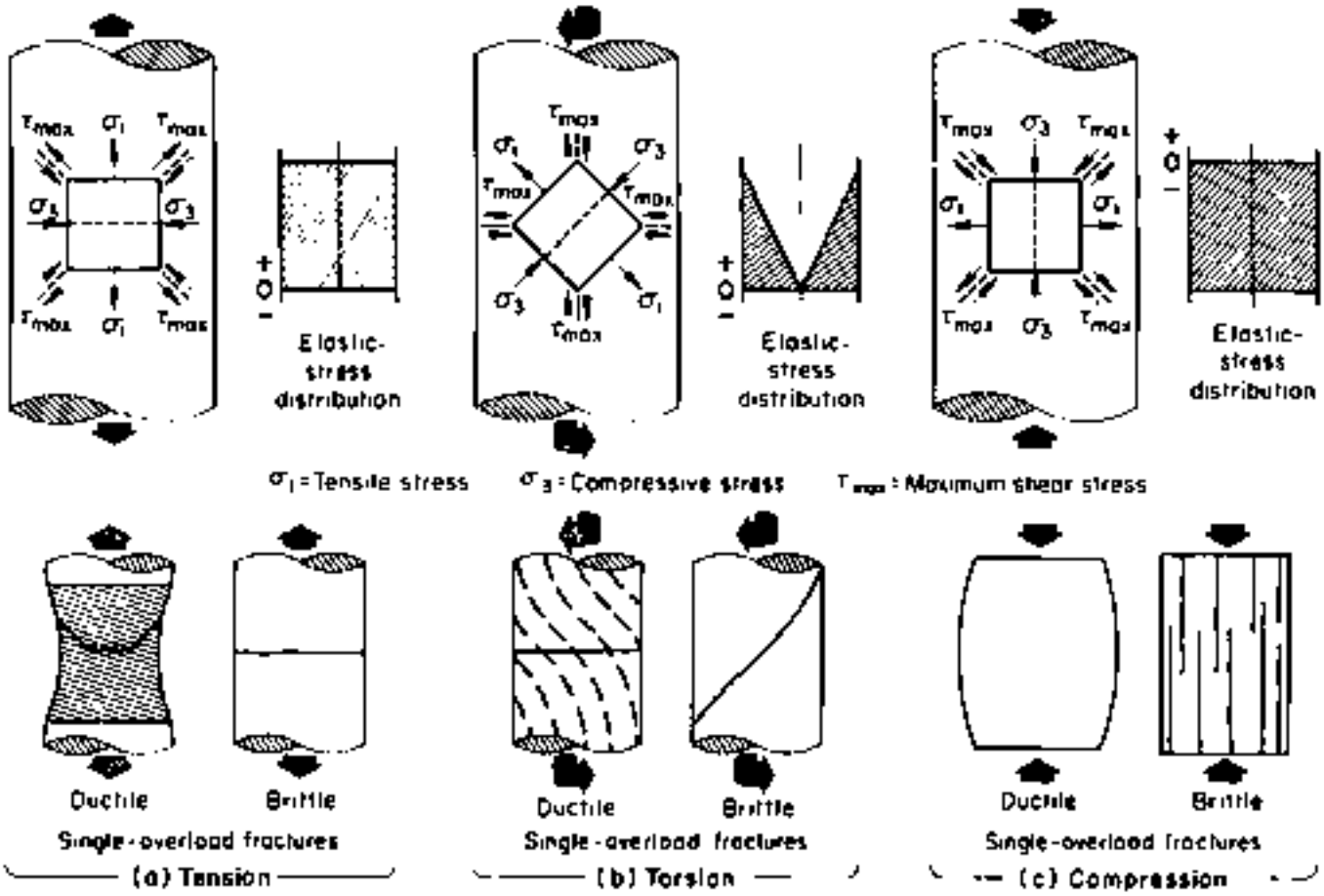


Fig. 1 Free-body diagrams showing orientation of normal stresses and shear stresses in a shaft and the single-overload fracture behavior of ductile and brittle materials. (a) Under simple tension. (b) Under torsion. (c) Under compression loading. See text for discussion.

A free-body stress system may be considered to be a square of infinitely small dimensions. Tensile and compressive stresses act perpendicular to each other and to the sides of the square to stretch and squeeze the sides, respectively. The shear, or sliding, stresses act on the diagonals of the square, 45° to the normal stresses. The third-dimension radial stresses are ignored in this description. The effects of the shear and normal stresses on ductile and brittle materials under the three types of loads illustrated in Fig. 1 and under bending load are discussed below.
Tension. Under tension loading, the tensile stresses, σ_1, are longitudinal, whereas the compressive-stress components, σ_3, are transverse to the shaft axis. The maximum-shear-stress components, τ_{max}, are at 45° to the shaft axis (Fig. 1a).
In a ductile material, shear stresses developed by tensile loading cause considerable deformation (elongation and necking) before fracture, which originates near the center of the shaft and propagates toward the surface, ending with a conical shear lip usually about 45° to the shaft axis. However, in a brittle material, a fracture from a single tensile overload is roughly perpendicular to the direction of tensile stress, but involves little or no permanent deformation. The fracture surface is usually rough and crystalline in appearance.
The elastic-stress distribution in pure tension loading, in the absence of a stress concentration, is uniform across the section. Thus, fracture can originate at any point within the highly stressed volume.
Torsion. The stress system rotates 45° when a shaft is loaded in torsion (Fig. 1b). Both the tensile and compressive stresses are 45° to the shaft axis and remain mutually perpendicular. One shear-stress component is parallel with the shaft axis; the other is perpendicular to the shaft axis.
In a ductile material loaded to failure in torsion, shear stresses cause considerable deformation before fracture. However, this deformation is usually not obvious, because the shape of the shaft has not been changed. The distortion will be

obvious if there were axial grooves or lines on the shaft before twisting, or if the metal is hot etched to reveal grain-flow twisting. If a shaft loaded in torsion is assumed to consist of an infinite number of infinitely thin disks that slip slightly with respect to each other under the torsional stress, visualization of deformation is simplified. Torsional single-overload fracture of a ductile material usually occurs on the transverse plane, perpendicular to the axis of the shaft. In pure torsion, the final-fracture region is at the center of the shaft; the presence of slight bending will cause it to be off-center.

A brittle material in pure torsion will again fracture perpendicular to the tensile-stress component, which is now 45° to the shaft axis. The resulting fracture surfaces usually have the shape of a spiral.

The elastic-stress distribution in pure torsion is maximum at the surface and zero at the center of the shaft. Thus, in pure torsion, fracture normally originates at the surface, which is the region of highest stress.

Compression. When a shaft is loaded in axial compression (Fig. 1c), the stress system rotates so that the compressive stress, σ_3, is axial and the tensile stress, σ_1, is transverse. The shear stresses, τ_{max}, are 45° to the shaft axis, as they are during axial tension loading.

In a ductile material overloaded in compression, shear stresses cause considerable deformation but usually do not result in fracture. The shaft is shortened and bulges laterally under the influence of shear stress. A brittle material loaded in pure compression, if it does not buckle, again will fracture perpendicular to the maximum tensile-stress component. Because the tensile stress is transverse, the direction of brittle fracture is parallel to the shaft axis.

The elastic-stress distribution in pure compression loading, in the absence of a stress concentration, is uniform across the section. If fracture occurs, it will likely be in the longitudinal direction, because compression loading increases the shaft diameter and stretches the metal at the circumference.

Bending. When a shaft is stressed in bending, the convex surface is stressed in tension and has an elastic-stress distribution similar to that shown in Fig. 1(a). The concave surface is stressed in compression and has an elastic-stress distribution similar to that shown in Fig. 1(c). Approximately midway between the convex and concave surfaces is a neutral axis, where all stresses are zero.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Fatigue Failures

Fatigue in shafts can generally be classified into three basic subdivisions: bending fatigue, torsional fatigue, and axial fatigue. Bending fatigue can result from these types of bending loads: unidirectional (one-way), reversed (two-way), and rotating. In unidirectional bending, the stress at any point fluctuates. Fluctuating stress refers to a change in magnitude without changing algebraic sign. In reversed bending and rotating bending, the stress at any point alternates. Alternating stress refers to cycling between two stresses of opposite algebraic sign, that is, tension (+) to compression (-) or compression to tension. Torsional fatigue can result from application of a fluctuating or an alternating twisting moment (torque). Axial fatigue can result from application of alternating (tension-and-compression) loading or fluctuating (tension-tension) loading. More complete information is available in the article “Fatigue Failures” in this Volume.

Unidirectional-Bending Fatigue. The axial location of the origin of a fatigue crack in a stationary cylindrical bar or shaft subjected to a fluctuating unidirectional-bending moment evenly distributed along the length will be determined by some minor stress raiser, such as a surface discontinuity. Beach marks (also called clamshell, conchoidal, and crack-arrest marks) of the form shown in Fig. 2(a) and (b) are indicative of a fatigue crack having a single origin at the point indicated by the arrow. The crack front, which formed the beach marks, is symmetrical relative to the origin and retains a concave form throughout. Both the single origin and the smallness of the final-fracture zone in Fig. 2(a) suggest that the nominal stress was low. The larger final-fracture zone in Fig. 2(b) suggests a higher nominal stress.

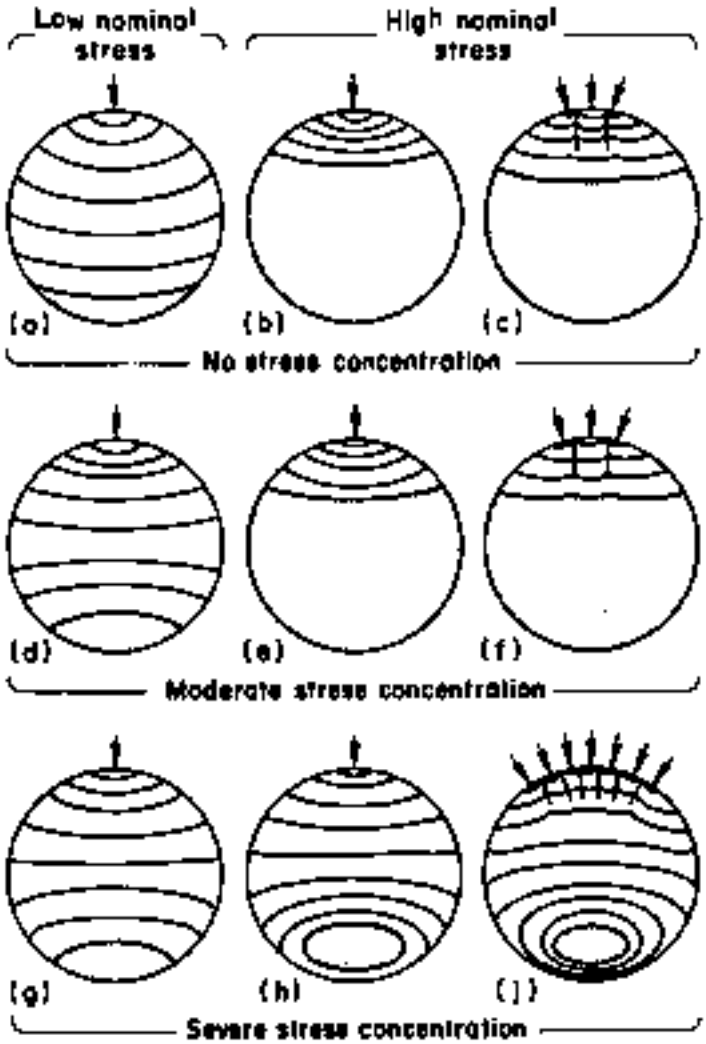


Fig. 2 Fatigue marks produced from single origins at low and high nominal stresses and from multiple origins at high nominal stresses. Fatigue marks are typical for a uniformly loaded shaft subjected to unidirectional bending. Arrows indicate crack origins; final-fracture zones are shaded.

Figure 2(c) shows a typical fatigue crack originating as several individual cracks that ultimately merged to form a single crack front. Such multiple origins are usually indicative of high nominal stress. Radial steps (ratchet marks) are present between crack origins.

Figures 2(d), (e), and (f) show typical fatigue beach marks that result when a change in section in a uniformly loaded shaft provides a moderate stress concentration. With a low nominal stress, the crack front changes from concave to convex before rapture (Fig. 2d). At higher nominal stresses, the crack front flattens and may not become convex before final fracture (Fig. 2e and f).

A change in section in a uniformly loaded shaft that produces a severe stress concentration will lead to a pattern of beach marks such as that shown in Fig. 2(g), (h), or (j). An example of a severe stress concentration is a small-radius fillet at the junction of a shoulder and a smaller-diameter portion of a shaft or at the bottom of a keyway. Such a fillet usually results in the contour of the fracture surface being convex with respect to the smaller-section side. The crack-front pattern shown in Fig. 2(g) was produced by a low nominal stress. The crack front in Fig. 2(h) developed more rapidly because of a higher stress in the peripheral zone. Multiple crack origins, high nominal stress, and unidirectional bending usually produce the beach-mark pattern shown in Fig. 2(j).

Example 1: Unidirectional-Bending Fatigue Failure of an A6 Tool Steel Shaft. The shaft shown in Fig. 3 was part of a clamping device on a tooling assembly used for bending 5.7-cm (2.25-in.) outer-diameter tubing on an 8.6-cm (3.375-in.) radius. The assembly contained two of these shafts, both of which failed simultaneously and were sent to the laboratory for examination. The maximum clamping force on the assembly was 54,430 kg (120,000 lb). The material specified for the shafts was a free-machining grade of A6 tool steel.

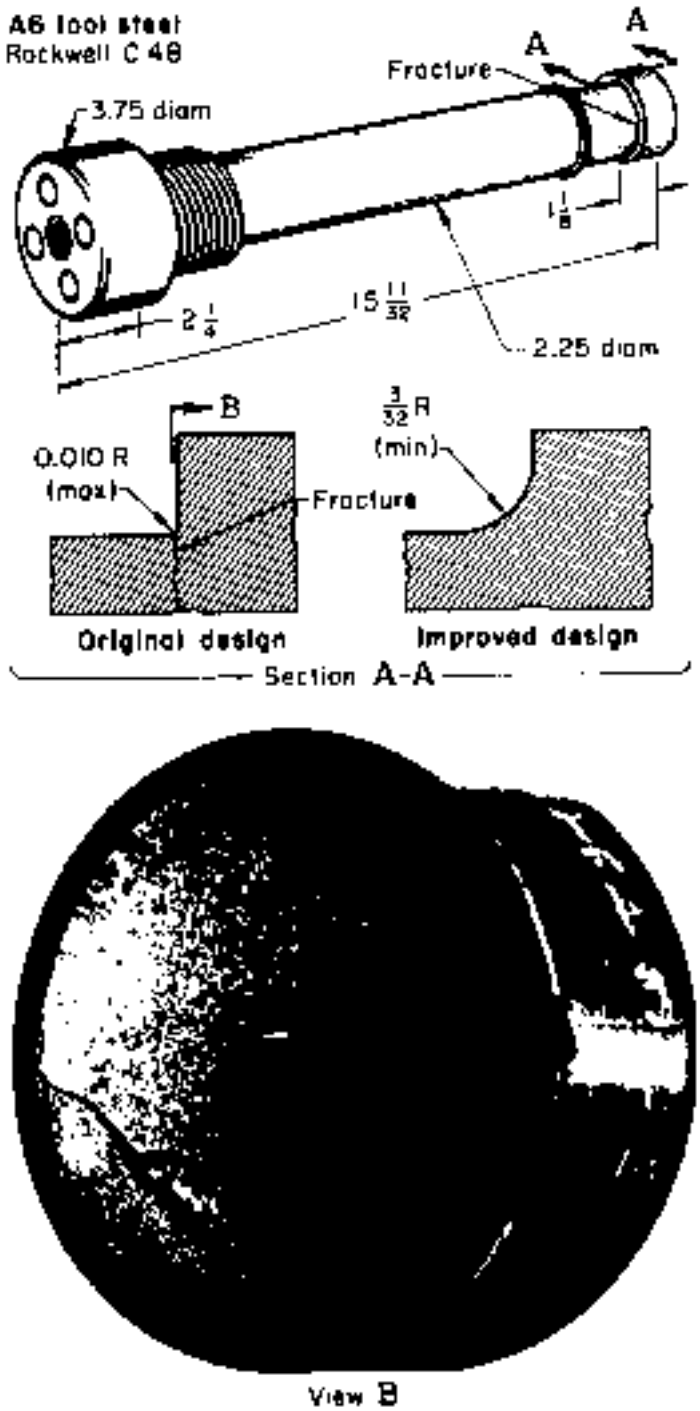


Fig. 3 A6 tool steel tube-bending-machine shaft that failed by fatigue fracture. Section A-A: Original and improved designs for fillet in failure region. Dimensions are in inches. View B: Fracture surface showing regions of fatigue-crack propagation and final fracture

The shafts were subjected to a tensile stress imposed by the clamping force and a bending stress resulting from the nature of the operation. Unidirectional-bending stresses were imposed on one shaft when a right-hand bend was made in the tubing and on the other shaft when a left-hand bend was made. Approximately 45 right-hand and 45 left-hand bends were made per hour on the machine; the total number of bends made before the shafts failed was not known. The tensile stress on the shafts was also cyclic, because the clamping force was removed after each bend was made.

Investigation. Analysis of the steel, using wet chemical and spectroscopic techniques, showed that the composition was within specifications. The average hardness of the steel was 48 HRC. A 1.3-cm (0.505-in.) diam tensile specimen removed from the center of one of the shafts failed in a brittle manner at a tensile stress of 1572 MPa (228 ksi).

The microstructure of the steel was fine, dispersed, tempered martensite with elongated stringers of manganese sulfide. Also present were spheroidized white particles that were identified as high-alloy complex carbides (M_6C) corresponding

to the double carbides Fe_4Mo_2C and Fe_4Cr_2C. Microscopic examination of the edge of the fracture surface at 100 and 1000× revealed some nonmetallic oxide-sulfide segregation.

Visual examination of the fracture surface revealed both a smooth area and a coarse, granular area (View B, Fig. 3). The dull, smooth area is typical of some fatigue fractures and resulted as the crack was opened and closed by the bending stress. Beach marks on the smooth area of the fracture surface also indicate fatigue fracture. The coarse, bright, crystalline-appearing area is the final-fracture zone. The smooth-textured fatigue zone is relatively large compared with the crystalline-textured final-fracture zone, which indicates that the shaft was subjected to a low overstress. The final-fracture surface at bottom shows that a one-way bending load was involved.

The fatigue crack was initiated in a 0.25-mm (0.010-in.) radius fillet at a change in section ("Original design," Section A-A, Fig. 3). Cracking was nucleated by a nonmetallic inclusion that intersected the surface at a critical location in the fillet.

Conclusions. The shafts fractured in fatigue as the result of a low-overstress, high-cycle unidirectional-bending load. The small radius of the fillet at the change in section resulted in a stress concentration that, in conjunction with the oxide-sulfide inclusion that intersected the surface of the fillet, initiated a crack.

Corrective Measures. New shafts were made with a 2.4-mm (0.09-in.) radius fillet at the critical change in section ("Improved design," Section A-A, Fig. 3). The larger-radius fillet minimized stress concentration in this region and prevented recurrence of failure.

Reversed-Bending Fatigue. When the applied bending moment is reversing (alternating), all points in the shaft are subjected alternately to tension stress and compression stress; while the points on one side of the plane of bending are in tension, the points on the opposite side are in compression. If the bending moment is of the same magnitude in either direction, two cracks of approximately equal length usually develop from origins diametrically opposite each other and often in the same transverse plane. If the bending moment is greater in one direction than in the other, the two cracks will differ in length.

Figure 4 shows typical fatigue marks on the fracture surface of a stationary (nonrotating) shaft subjected to a reversing bending moment evenly distributed along its length. The crack origins (arrows) are shown diametrically opposite each other, but sometimes they are slightly displaced by minor stress raisers. The pattern shown in Fig. 4(a) is typical of that for a single-diameter shaft with no stress concentration. The bending moment is equal in both directions.

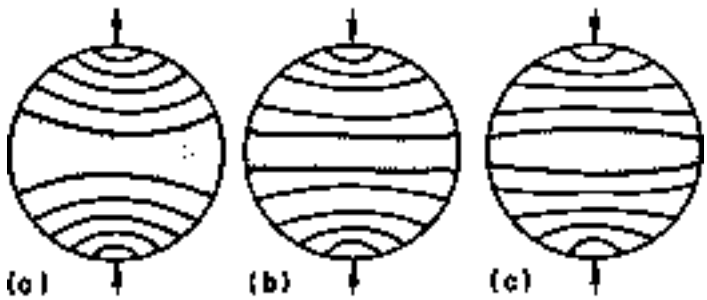


Fig. 4 Typical fatigue marks on the fracture surface of a uniformly loaded nonrotating shaft subjected to reversed-bending stresses. (a) No stress concentration. (b) Moderate stress concentration. (c) Severe stress concentration. Arrows indicate crack origins; shaded areas are final-fracture zones.

A large-radius fillet at a change in shaft diameter imposes a moderate stress concentration. Figure 4(b) shows the pattern on the surface of a fracture through such a fillet. A small-radius fillet at a change in diameter results in a severe stress concentration. Figure 4(c) shows the typical pattern on the surface of a fracture through a small-radius fillet. The reason for the fatigue-pattern changes is that the fatigue crack propagates faster in more severe stress concentrations at each end than in the interior.

Under the above loading conditions, each crack is subjected alternately to tensile and compressive stresses, with the result that the surfaces of the crack are forced into contact with one another during the compression cycle, and rubbing occurs. Rubbing may sometimes be sufficient to obliterate many of the characteristic marks, and the crack surfaces may become dull or polished.

Rotating-Bending Fatigue. The essential difference between a stationary shaft and a rotating shaft subjected to the same bending moment is that in a stationary shaft the tensile stress is confined to a portion of the periphery only. In a rotating shaft, every point on the periphery sustains a tensile stress, then a compressive stress, once every revolution. The relative magnitude of the stresses at different locations is determined by conditions of balance or imbalance imposed on the shaft.

Another important difference introduced by rotation is asymmetrical development of the crack front from a single origin. There is a marked tendency of the crack front to extend preferentially in a direction opposite to that of rotation. The crack

front usually swings around about 15° or more (Fig. 5a and c). A third difference arising from rotation is in the distribution of the initiation sites of a multiple-origin crack.

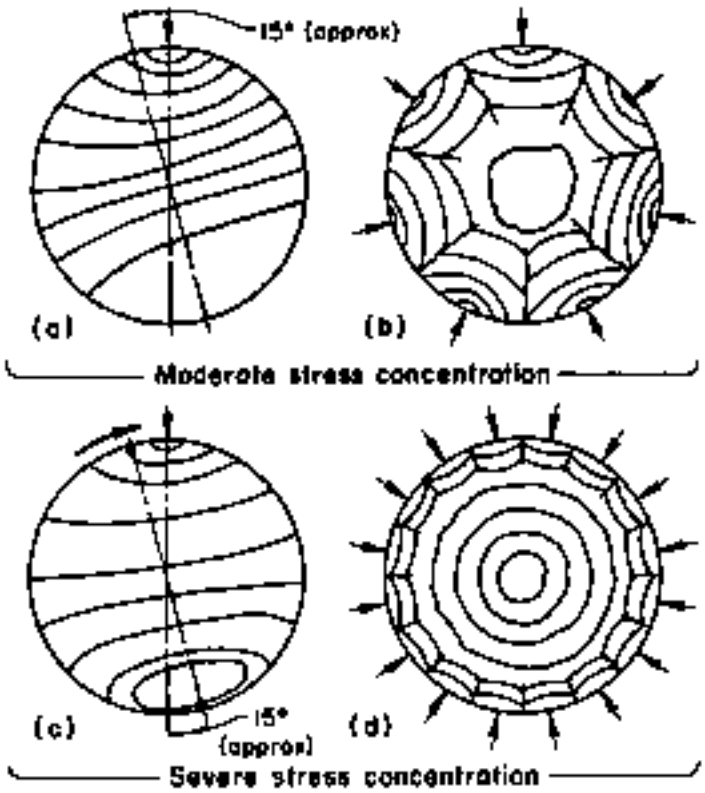


Fig. 5 Typical fatigue marks on the fracture surface of a uniformly loaded rotating shaft. Marks are produced from single and multiple origins (arrows) having moderate and severe stress concentration; shaded areas are final-fracture zones. Shaft rotation is clockwise.

In a nonrotating shaft subjected to unidirectional bending, the origins are located in the region of the maximum-tension zone (Fig. 2). In a nonrotating shaft subjected to reversed bending, the origins are diametrically opposite each other (Fig. 4). In rotary bending, however, every point on the shaft periphery is subjected to a tensile stress at each revolution; therefore, a crack may be initiated at any point on the periphery (Fig. 5b and d).

The crack surfaces are pressed together during the compressive component of the stress cycle, and mutual rubbing occurs. A common result of final fracture is that slight movement of one side of the crack relative to the other side frequently causes severe damage to the fracture surfaces and tends to obliterate many marks. However, although the high spots on one surface may rub the high spots on the other, the marks in the depressions are retained. Because the depressions are negative images of the damaged high spots on the opposing surface, they provide useful evidence; therefore, it is desirable to examine both parts of a cracked or fractured shaft.

The similarity in macroscopic appearance of fractures in shafts resulting from rotating-bending fatigue and from single overload torsional shear of a relatively ductile metal frequently results in misinterpretation. The fracture surface shown in Fig. 6(a) was the result of fatigue, as evidenced by the ratchet marks around the periphery and the pronounced beach marks. Under the low magnification of the fractograph shown in Fig. 6(b), beach marks are not visible, because they were obliterated by rubbing. The presence of ratchet marks around the periphery is also an indication of rotating-bending fatigue. However, the metal smearing apparent on the fracture surface and the twisting deformation of the shaft shown in Fig. 6(c) indicate torsional shear and would preclude mistaking this fracture for a fatigue fracture.

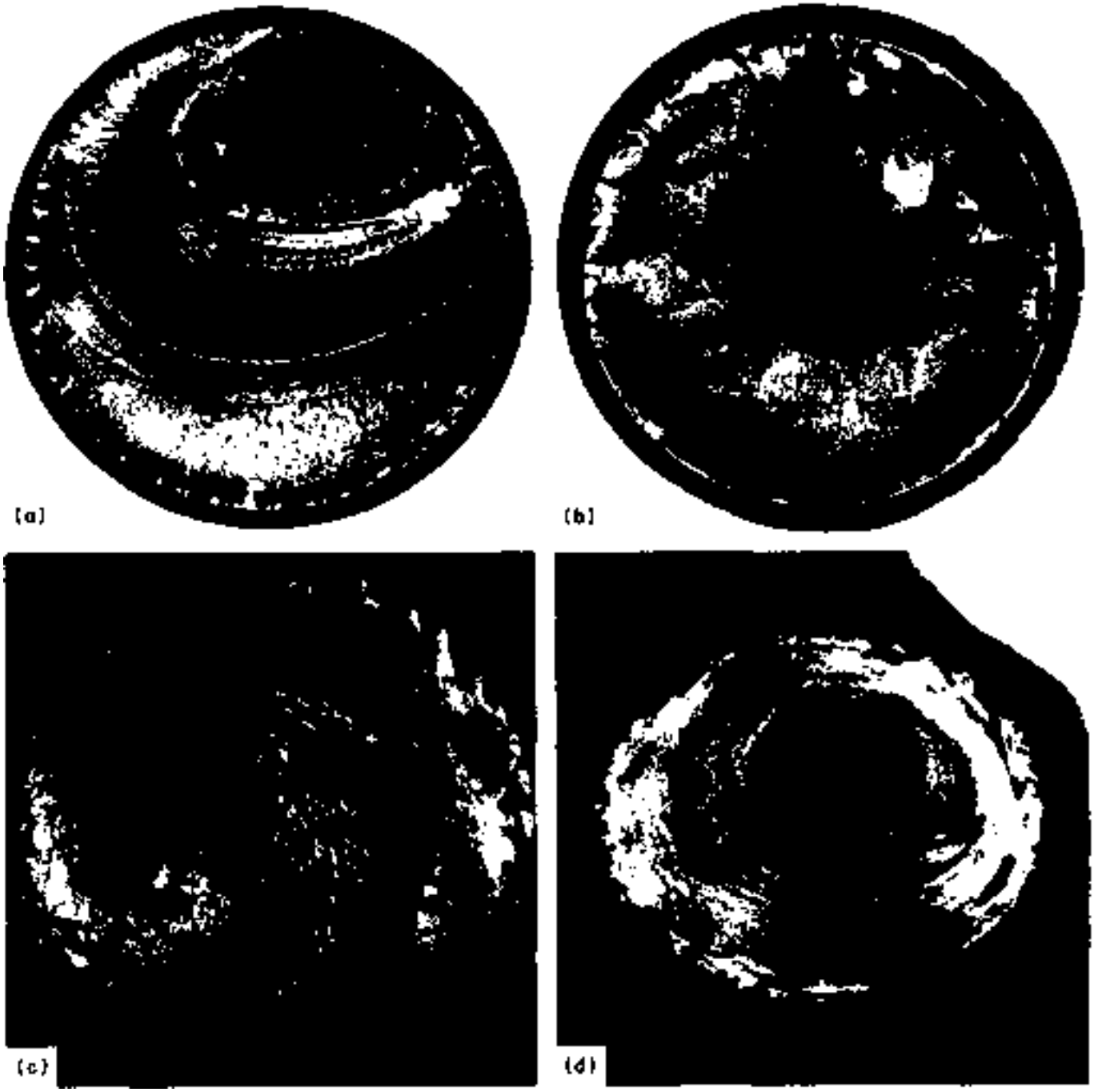


Fig. 6 Fracture surfaces of failed shafts. (a) and (b) Failure by fatigue. (c) and (d) Failure by torsional shear. See text for discussion.

The fracture shown in Fig. 6(d) also exhibits a superficial similarity to a fatigue fracture. However, it is evident that this fracture was the result of torsional shear, because the entire fracture surface has a smooth texture and no well-defined final-fracture area.

In splined shafts, fracture resulting from torsional shear is frequently accompanied by deformation of the splines not engaged by the mating part. However, the portion of the shaft not engaged by the mating part is sometimes unavailable for examination. When macroscopic examination affords only inconclusive evidence, use of an electron microscope may reveal fatigue striations or elongated shear dimples. Also, metallographic examination of a section through the fracture surface may reveal microdeformation from torsional shear in the rotary direction that would not be present in a fatigue fracture.

Torsional Fatigue. Fatigue cracks arising from torsional stresses also show beach marks and ridges. Longitudinal stress raisers are comparatively harmless under bending stresses, but are as important as circumferential stress raisers under torsional loading. This sensitivity of shafts loaded in torsion to longitudinal stress raisers is of considerable practical importance because inclusions in the shaft material are almost always parallel to the axis of rotation. It is not unusual for a

torsional-fatigue crack to originate at a longitudinal inclusion, a surface mark, or a spline or keyway corner and then to branch at about 45°.

When a stress raiser such as a circumferential groove is present, different states of stresses exist around the stress raiser, and the tensile stress is increased to as much as four times the shear stress. Therefore, the tensile stress on the 45° plane will exceed the tensile strength of the steel before the shear stress reaches the shear strength of the steel. Fracture occurs normal to the 45° tensile plane, producing a conical or star-shaped fracture surface.

Example 2: Torsional-Fatigue Fracture of a Large 4340 Steel Shaft That Was Subject to Cyclic Loading and Frequent Overloads. The 4340 steel shaft shown in Fig. 7(a), which was the driving member of a large rotor subject to cyclic loading and frequent overloads, broke after three weeks of operation. The shaft was also part of a gear train that reduced the rotational speed of the driven member. The driving shaft contained a shear groove at which the shaft should break if a sudden high overload occurred, thus preventing damage to an expensive gear mechanism. The rotor was subjected to severe chatter, which was an abnormal condition resulting from a series of continuous small overloads at a frequency of about three per second.

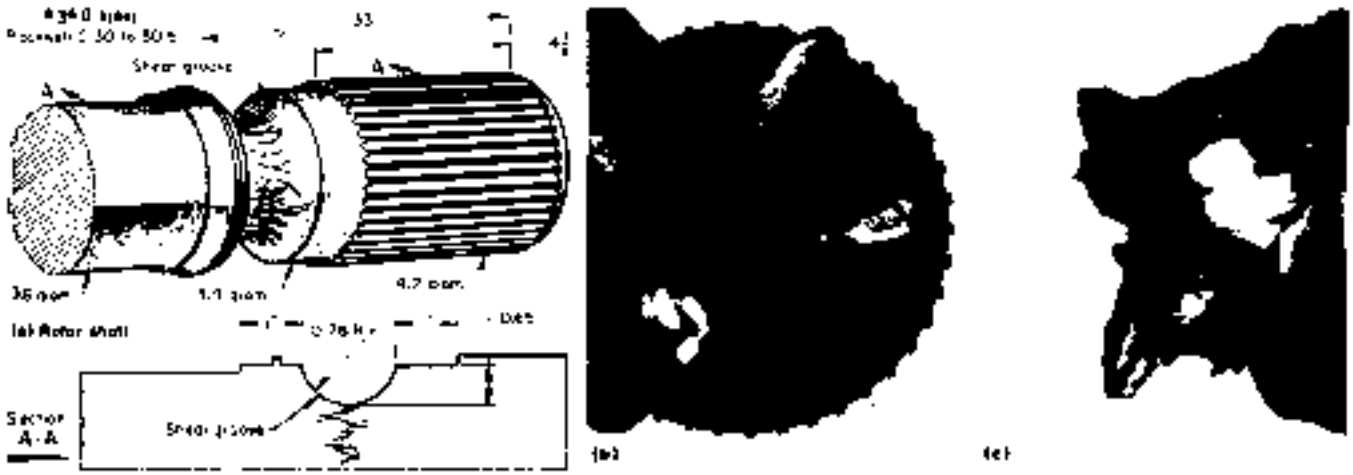


Fig. 7 4340 steel rotor shaft that failed by torsional fatigue. (a) Shear groove designed to protect gear mechanism from sudden overload. Dimensions are in inches. (b) Star-shaped pattern on a fracture surface of the shaft. (c) Longitudinal and transverse shear cracks on the surface of the shear groove, which resulted from high peak loads caused by chatter

Investigation. Examination disclosed that the shaft had broken at the shear groove and that the fracture surface contained a star-shaped pattern (Fig. 7b). Figure 7(c) shows the fracture surface with the pieces fitted back in place. The pieces were all nearly the same size and shape, and there were indications of fatigue, cleavage, and shear failure in approximately the same location on each piece. The cracks were oriented at approximately 45° to the axis of the shaft, which indicated that final fracture was caused by a tensile stress normal to the 45° plane and not by the longitudinal or transverse shear stress that had been expected to cause an overload failure.

Examination of the surfaces of one of the pieces of the broken shaft revealed small longitudinal and transverse shear cracks at the smallest diameter of the shear groove. Also, slight plastic flow had occurred in the metal adjacent to these cracks. Cracking occurred at many points in the groove in the shaft before several of the cracks grew to a critical size.

No surface irregularities were present in the shear groove at any of the shear cracks. The structure of the metal was normal, with a uniform hardness of 30 to 30.5 HRC across the section, indicating a strength in the expected range for quenched-and-tempered 4340 steel shafts. A hot-acid etch showed the steel to be free from pipe, segregation, or other irregularities.

Conclusion. The basic failure mechanism was fracture by torsional fatigue, which started at numerous surface shear cracks, both longitudinal and transverse, that developed in the periphery of the root of the shear groove. These shear cracks resulted from high peak loads caused by chatter. Stress concentrations developed in the regions of maximum shear, and fatigue cracks propagated in a direction perpendicular to the maximum tensile stress, thus forming the star pattern at 45° to the longitudinal axis of the shaft. The shear groove in the shaft, designed to prevent damage to the gear train, had performed its function, but at a lower overload level than intended.

Corrective Measures. The fatigue strength of the shaft was increased by shot peening the shear groove, and chatter in the machine was minimized.

The relative extent of development of two torsional-fatigue cracks mutually at right angles can indicate the magnitude of the torque reversals that have been applied. If the cracks are of approximately the same length, the indications are that the torque reversals have been of equal magnitude, but only if the cracks are in a comparatively early stage of development.

Beyond this stage, one crack usually takes the lead, and such inferences are no longer justified. If the shaft transmits a unidirectional torque but two cracks develop mutually at right angles, it can be presumed that the torque was of a reversing character. If a bending stress is applied to a shaft that is transmitting torque, the angle at which any fatigue crack develops will be modified. Therefore, if the angle differs significantly from 45° to the shaft axis, the presence of a bending stress is indicated.

Contact Fatigue

Contact fatigue occurs when components roll, or roll and slide, against each other under high contact pressure and cyclic loading. Pitting occurs after many repetitions of loading and is the result of metal fatigue from the imposed cyclic contact stresses. Factors that govern contact fatigue are contact stress, relative rolling/sliding, material properties, and metallurgical, physical, and chemical characteristics of the contacting surfaces, including the oil film that lubricates the surfaces.

The significant stress in rolling-contact fatigue is the maximum alternating shear stress that undergoes a reversal in direction during rolling. In pure rolling, as in antifriction bearings, this stress occurs slightly below the surface and can lead to the initiation of subsurface fatigue cracks. As these cracks propagate under the repeated loads, they reach the surface and produce cavities, or pits.

When sliding is superimposed on rolling, as in gear teeth, the tangential forces and thermal gradient caused by friction alter the magnitude and distribution of stresses in and below the contact area. The alternating shear stress is increased in magnitude and is moved nearer to the surface by friction resulting from the sliding action. Additional information is provided in the articles “Fatigue Failures” and “Failures of Rolling-Element Bearings” in this Volume.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Wear

Wear of metal parts is commonly classified into either of two categories: abrasive wear or adhesive wear.

Abrasive wear, the undesired removal of material by a cutting mechanism, can reduce the size and destroy the proper shape of a shaft. The shaft may then fail by another means, such as by fracture, or may cease to perform its designed function. Foreign particles, such as sand, dirt and other debris, in the lubricant can cause wear of a shaft.

Example 3: Wear Failure of a Fuel-Pump Drive Shaft Caused by the Presence of Sand, Metallic Particles, and Vibration. The fuel pump in a turbine-powered aircraft failed, resulting in damage to the aircraft. The pump is shown in Fig. 8(a) and (b).

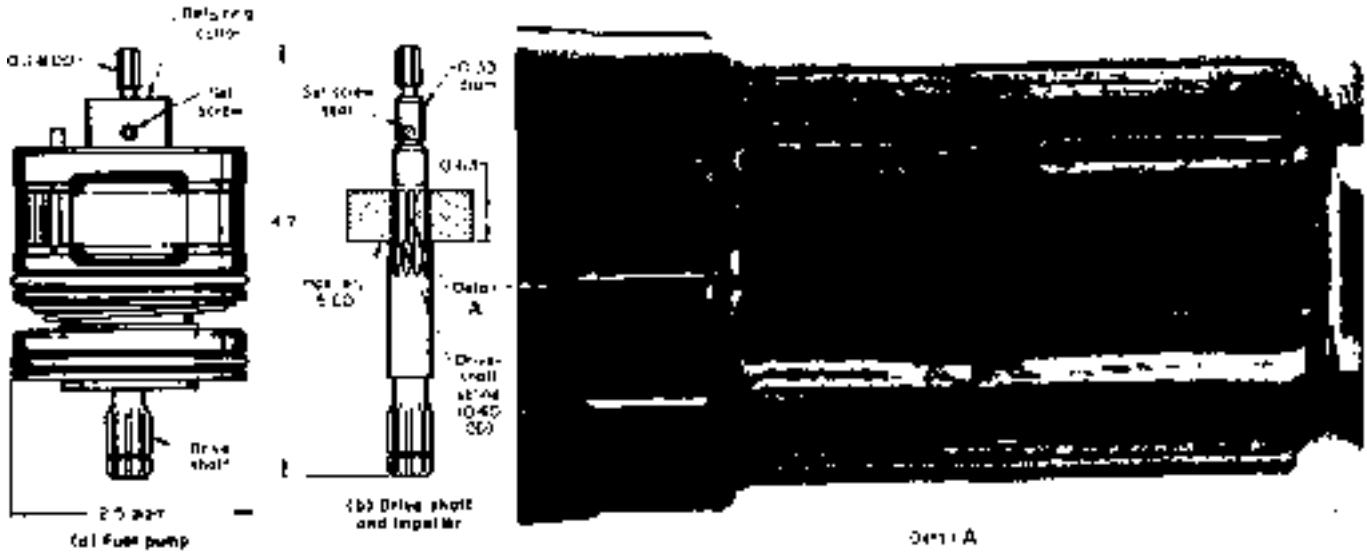


Fig. 8 Fuel pump that failed by vibration and abrasion. (a) Configuration and dimensions (given in inches). (b) Splines on the drive shaft and in the impeller were worn away by vibration in the presence of sand and metallic particles. Detail A: Enlarged view of failure area showing worn splines

This particular model of fuel pump had a history of wear failures and had been redesigned to incorporate a shaft of case-hardened steel (composition not reported). Vibration was common during operation, but generally was not excessive for the aircraft.

Investigation. The pump and the filter chamber were found to be dry and free of any debris or contamination, except for some accumulated deposit on the filter cartridge, when the pump was disassembled in the laboratory. The drive-shaft splines that engaged the impeller were almost completely destroyed down to the roots (Detail A, Fig. 8). Extensive damage to the splines was apparent because on subsequent reassembly the shaft could be rotated without rotating the impeller.

The pressure side of each spline tooth in the impeller also exhibited some damage. Relatively smooth cavities and undercutting of the flank on the pressure side of the spline teeth indicated that the damage had not been caused by wear from metallic contact between the splines but by an erosion or abrasion mechanism.

Hardness readings taken at several axial locations on the drive shaft showed a reasonably uniform hardness of approximately 570 HV. The impeller and the retaining ring each had a hardness of approximately 780 HV, and the parts surrounding the impeller (including the vanes) exhibited a hardness of approximately 630 HV. The microstructure of the metal in the impeller exhibited scattered porosity and carbide particles and appeared to be a sintered powder metallurgy compact.

Metallographic examination of a section through the damaged splines and of a section through the adjacent undamaged part of the same splines disclosed no material defects. The microstructure indicated that the shaft had been satisfactorily heat treated by quenching and tempering, but there was no evidence of case hardening on the spline surfaces, a treatment commonly given to shafts of this type.

The worn surfaces of the splines showed evidence of cold work at the edges. Also, there was a relatively smooth worn area at the center of each tooth that appeared to be free of cold work and that appeared to have been caused by an abrasive action. The damaged side of each spline appeared as an undulating outline with some undercutting rather than a jagged or deformed shape.

The residue on the filter cartridge was brown, and when viewed under a low-power microscope, particles of sand, paint, or plastic, fibers from the cartridge, brass, and steel could be identified. Application of a magnet to the sample showed that it contained a large amount of iron.

Chemical analysis indicated that the deposit contained about 20% sand, 30% iron, and 30% organic material (paint, plastic, and filter fibers). The reddish-brown color of the deposit suggested that some of the iron present was an oxide (rust), but this was not confirmed.

Discussion. Vibration in the fuel pump could be expected to initiate damage, particularly when combined with an abrasive action. Under these conditions, fretting or abrasive wear can be expected on sliding-contact surfaces that are not sufficiently abrasion resistant. The residue from the filter contained significant quantities of sand and iron; the iron probably originated from the damaged shaft.

The examination indicated that the splines on the drive shaft had been damaged by abrasion, which could have been caused by the combined effect of vibration and abrasives, such as sand and the metal particles removed from the splines.

The same action would also damage the internal splines of the impeller, but to a lesser extent because of its greater hardness and thus greater abrasion resistance. However, the internal splines exhibited significant damage, which appeared to have been produced by erosion. Thus, the internal splines suffered slight abrasive damage and somewhat more damage by erosion.

Conclusions. Failure of the shaft was the result of excessive wear on the splines caused by vibration and the abrasive action of sand and metal particles.

Recommendations. To increase resistance to wear and abrasion, the surfaces of the spline teeth should be case hardened. Although the drive shaft exhibited reasonably high strength, its resistance to wear and abrasion was inadequate for the conditions to which it was exposed.

Wire Wooling. Abrasive-wear failure of shafts by wire wooling has been observed under certain circumstances where contact occurs between the shaft and a stationary part, resulting in removal, by machining, of fine wire shavings that resemble steel wool. This type of failure has been found on turbine and turbine-generator shafts made of 3Cr-0.5Mo steels, on 12% Cr stainless steels, on 18Cr-8Ni stainless steels, and on nonchromium steels in the presence of certain chloride-containing oils. Wire wooling has also been observed on thrust bearings and on centrifugal compressor shafts.

Figure 9 shows the end of a 13.3-cm (5.25-in.) diam shaft that was worn by wire wooling. The worn surface was in contact with a labyrinth seal and was not a bearing surface. The shaft had operated for about $3\frac{1}{2}$ years before the damage was discovered. Although records were not available, the machine was known to have been opened at least once during that time, but damage under the seal had not been noticed. Figure 9(b) shows a micrograph of a section through the material found in the circumferential grooves in the shaft. The material was loose in the grooves; some was almost identical in appearance with steel wool, whereas other pieces were coarser and more like slivers.

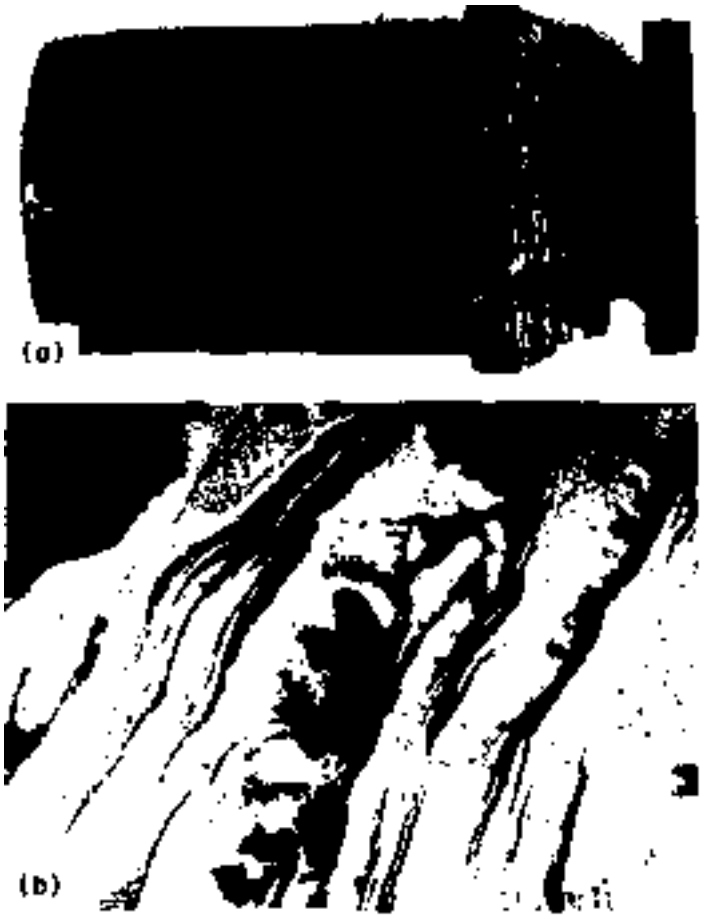


Fig. 9 Shaft that was severely damaged by wire wooling. (a) End of shaft. (b) Micrograph of an unetched section through loose material found in grooves worn in shaft by wire wooling. 50×

Although the mechanisms of wire wooling are not clearly understood, it is known that the process requires contact between a shaft (or shaft sleeve) and a labyrinth (or bearing), either directly or through buildup of deposits. If the deposit or the stationary part contains hard particles, fine slivers can be cut or spun off the shaft surface. As fine slivers or pieces come off, additional hard particles may be formed by reaction between the steel and the oil or gas present if the resultant

friction and heat are sufficient. Scabs or solid chunks of laminated or compacted slivers and other deposits are sometimes formed (Fig. 9b). Bearings or labyrinths of babbitt or copper alloys have been associated with this type of failure. Contact-area atmospheres, in addition to bearing oils, have contained air and methyl chloride.

Methods for prevention of wire wooling include changing the shaft material, using a softer bearing or labyrinth material, changing to a different oil, eliminating the deposits, and providing greater clearance.

Adhesive Wear of Shafts. Adhesive wear, sometimes called scoring, scuffing, galling, or seizing, is the result of microscopic welding at the interface between two mutually soluble metals, such as steel on steel. It frequently occurs on shafts where there is movement between the shaft and a mating part, such as a gear, wheel, or pulley.

Fretting on a shaft can be a source of serious damage. Production of a reddish-brown powder is characteristic of fretting on steel. Typical locations for fretting are at splined or keyed hubs, components that are press fitted or shrink fitted to a shaft, and clamped joints. Shot peening, glass-bead peening, and surface rolling are methods of reducing the possibility of fatigue fracture of shafts because of fretting of joints. Fretting is discussed in the article “Fretting Wear Failures” in this Volume.

Adhesive wear has a characteristic torn appearance because the surfaces actually weld together, then are torn apart by continued motion, creating a series of fractures on both surfaces. This indicates that metal-to-metal contact took place between clean, uncontaminated mating surfaces.

Because excessive frictional heat is generated, adhesive wear often can be identified by a change in the microstructure of the metal. For example, steel may be tempered or rehardened locally by the frictional heat generated. Additional information is provided in the article “Fundamentals of Wear Failures” in this Volume.

Influence of Bearings and Seals. If acceptable wear resistance is not obtained by optimizing selection of shaft material and heat treatment, consideration should be given to the sliding (sleeve) bearing and seal material and its compatibility with the shaft. Very often, the high-wear area of a shaft may be chromium plated or may be covered with a sleeve that can be discarded and replaced when worn (rather than throwing away an expensive shaft or replacing the bearings). These methods involve risks, and steps should be taken to avoid or offset their effects.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Brittle Fracture of Shafts

Brittle fractures are associated with the inability of certain materials to deform plastically in the presence of stress at the root of a sharp notch, particularly at low temperatures. Brittle fractures are characterized by sudden fracturing at extremely high rates of crack propagation, perhaps 1830 m/s (6000 ft/s) or more, with little evidence of distortion in the region of fracture initiation. This type of fracture is frequently characterized by marks known as herringbone or chevron patterns on the fracture surface. The chevrons point toward the origin of the fracture. Additional information on brittle fracture can be found in the article “Mechanisms and Appearances of Ductile and Brittle Fractures” in this Volume.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Ductile Fracture of Shafts

Ductile fractures, which result from microvoid coalescence, exhibit evidence of distortion (plastic flow) at the fracture surface similar to that observed in ordinary tensile-test or torsion-test specimens. When a shaft is fractured by a single application of a load greater than the strength of the shaft, there is usually considerable plastic deformation before fracture. This deformation is often readily apparent upon visual inspection of a shaft that fractured in tension, but is often not obvious when the shaft fractured in torsion. This ability of a material to deform plastically (permanently) is a property known as ductility. The appearance of the fracture surface of a

shaft that failed in a ductile manner is also a function of shaft shape, the type of stress to which the shaft was subjected, rate of loading, and, for many alloys, temperature. In general, ductility is decreased by increasing the strength of the metal by cold work or heat treatment, by the presence of notches, fillets, holes, scratches, inclusions, and porosity in a notch-sensitive material, by increasing the rate of loading, and for many alloys, by decreasing the temperature.

Ductile fracture of shafts occurs infrequently in normal service. However, ductile fractures may occur if service requirements are underestimated, if the materials used are not as strong as had been assumed, or if the shaft is subjected to a massive single overload, such as in an accident. Fabricating errors, such as using the wrong material or using material in the wrong heat treated condition (for example, annealed instead of quenched and tempered), can result in ductile fractures.

Distortion of Shafts

Distortion of a shaft can render the shaft incapable of serving its intended function.

Permanent distortion simply means that the applied stress has exceeded the yield strength (but not the tensile strength) of the material. If it is not feasible to modify the design of the shaft, the yield strength of the shaft material must be increased to withstand the applied stress. Yield strength may be increased either by using a stronger material or by heat treating the original material to a higher strength.

Creep, by definition, is time-dependent strain (distortion) occurring under stress imposed at elevated temperature, provided the operational load does not exceed the yield strength of the metal. If creep continues until fracture occurs, the part is said to have failed by stress rupture. Creep can result from any type of loading (tensile, torsion, compression, bending, and so on).

Some high-temperature applications, such as gas turbines and jet aircraft engines, require materials to operate under extreme conditions of temperature and stress with only a limited amount of deformation by creep. In other high-temperature applications, the permissible deformation is high and may not even be limited as long as rupture does not occur during the intended life of the part. For this type of service, stress-rupture data, rather than long-term creep data, are used for design.

Buckling, a third type of distortion failure, results from compressive instability. It can occur if a long slender rod or shaft collapses from compressive axial forces. The load required to cause buckling can be changed only by design changes, not by metallurgical changes, such as heat treatment, in a given type of metal.

Corrosion of Shafts

Most shafts are not subjected to severe reduction in life from general corrosion or chemical attack. Corrosion may occur as general surface pitting, may uniformly remove metal from the surface, or may uniformly cover the surface with scale or other corrosion products. Corrosion pits have a relatively minor effect on the load-carrying capacity of a shaft, but they do act as points of stress concentration at which fatigue cracks can originate.

A corrosive environment will greatly accelerate metal fatigue; even exposure of a metal to air results in a shorter fatigue life than that obtained under vacuum. Steel shafts exposed to salt water may fail prematurely by fatigue despite periodic, thorough cleaning. Aerated salt solutions usually attack metal surfaces at the weakest

points, such as scratches, cut edges, and points of high strain. To minimize corrosion fatigue, it is necessary to select a material that is resistant to corrosion in the service environment or to provide the shaft with a protective coating.

Most large shafts and piston rods are not subject to corrosion attack. However, because ship-propeller shafts are exposed to salt water, they are pressure rolled, which produces residual surface-compressive stresses and inhibits origination of fatigue cracks at corrosion pits. Also, rotating parts, such as centrifugal compressor impellers and gas-turbine disks and blades, often corrode. Centrifugal compressors frequently handle gases that contain moisture and small amounts of a corrosive gas or liquid. If corrosion attack occurs, a scale is often formed that may be left intact and increased by more corrosion, eroded off by entrained liquids (or solids), or thrown off from the rotating shaft.

Stress-corrosion cracking occurs as a resuit of corrosion and stress at the tip of a growing crack. Stress-corrosion cracking is often accompanied or preceded by surface pitting; however, general corrosion is often absent, and rapid, overall corrosion does not accompany stress-corrosion cracking.

The tensile-stress level necessary for stress-corrosion cracking is below the stress level required for fracture without corrosion. The critical stress may be well below the yield strength of the material, depending on the material and the corrosive conditions. Evidence of corrosion, although not always easy to find, should be present on the surface of a stress-corrosion-cracking fracture up to the start of final rupture.

All of the common materials used in shafts may undergo stress-corrosion cracking under certain specific conditions. Factors that influence stress-corrosion cracking, either directly or indirectly, include microstructure, yield strength, hardness, corrodent(s), concentration of corrodent(s), amounts and nature of water, pH, and applied and residual stresses, degree of cold working, and chemical composition of the base metal. Additional information is available in the article “Stress-Corrosion Cracking” in this Volume.

Corrosion fatigue results when corrosion and an alternating stress—neither of which is severe enough to cause failure by itself—occur simultaneously; this can cause failure. Once such a condition exists, shaft life will probably be greatly reduced. Corrosion-fatigue cracking is usually transgranular; branching of the main cracks occurs, although usually not as much as in stress-corrosion cracking. Corrosion products are generally present in the cracks, both at the tips and in regions nearer the origins. The article “Corrosive Wear Failures” in this Volume contains more detailed information on the effect of combined corrosion and fluctuating stress.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Common Stress Raisers in Shafts

Most service failures in shafts are largely attributable to some condition that intensifies stress. In local regions, the stress value is raised above a value at which the material is capable of withstanding the number of loading cycles that corresponds to a satisfactory service life. Only one small area needs to be repeatedly stressed above the fatigue strength of the material for a crack to be initiated. An apparently insignificant imperfection, such as a small surface irregularity, may severely reduce the fatigue strength of a shaft if the stress level at the imperfection is high. The most vulnerable zone in torsional and bending fatigue is the shaft surface; an abrupt change in surface configuration may have a damaging effect, depending on the orientation of the discontinuity to the direction of stress.

All but the simplest shafts contain oil holes, keyways, or changes in shaft diameter (threads, fillets, annular grooves, and so on). The transition from one diameter to another, the location and finish of an oil hole, and the type and shape of a keyway exert a marked influence on the magnitude of the resulting stress-concentration and fatigue-notch factors, which often range in numerical value from 1 to 5 and sometimes attain values of 10 or higher.

Types of Stress Raisers. The majority of stress raisers can be placed into one of the following general groups:

- *Group 1:* Nonuniformities in the shape of the shaft, such as steps at changes in diameter, broad integral collars, holes, abrupt corners, keyways, grooves, threads, splines, and press-fitted or shrink-fitted attachments
- *Group 2:* Surface discontinuities arising from fabrication practices or service damage, such as seams, nicks, notches, machining marks, identification marks, forging laps and seams, pitting, and corrosion
- *Group 3:* Internal discontinuities, such as porosity, shrinkage, gross nonmetallic inclusions, cracks, and voids

Most shaft failures are initiated at primary (group 1) stress raisers, but secondary (group 2 or 3) stress raisers may contribute to a failure. For example, a change in shaft diameter can result in stress intensification at the transition zone; if there is a surface irregularity or other discontinuity in this zone, the stress is sharply increased around the discontinuity.

Influence of Changes in Shaft Diameter

A change in shaft diameter concentrates the stresses at the change in diameter and in the smaller-diameter portion. The effects of an abrupt change and three gradual changes in section on stress concentration are shown schematically in Fig. 10. The sharp corner at the intersection of the shoulder and shaft in Fig. 10(a) concentrates the stresses at the corner as they pass from the large to the small diameter. The large-radius fillet shown in Fig. 10(d) permits the stresses to flow with a minimum of restriction. However, the fillet must be tangent with the smaller-diameter section, or a sharp intersection will result, overcoming the beneficial effect of the large-radius fillet.

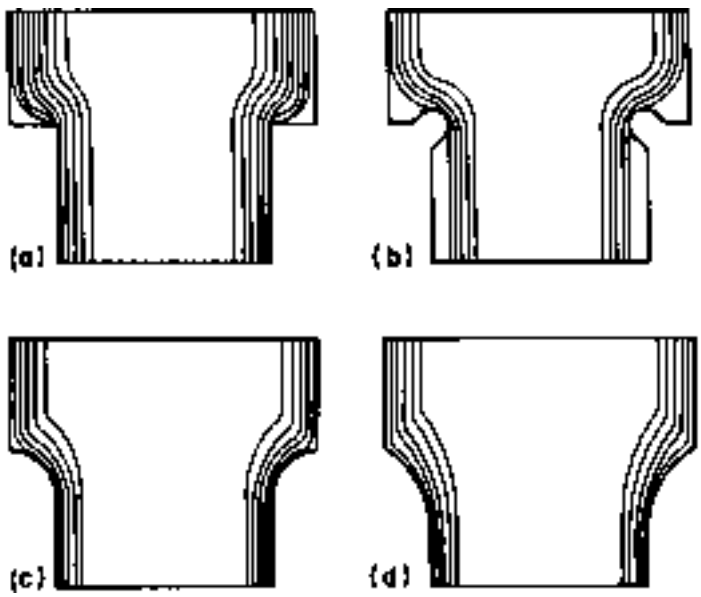


Fig. 10 Effect of size of fillet radius on stress concentration at a change in shaft diameter. See text for discussion.

Example 4: Fatigue Fracture of a 6150 Steel Main Shaft in a Coal Pulverizer. Unusual noises were noted by the operator of a ball-and-race coal pulverizer (shown schematically in Fig. 11a), and the unit was taken out of service to investigate the cause. The pulverizer had been in service for ten years.

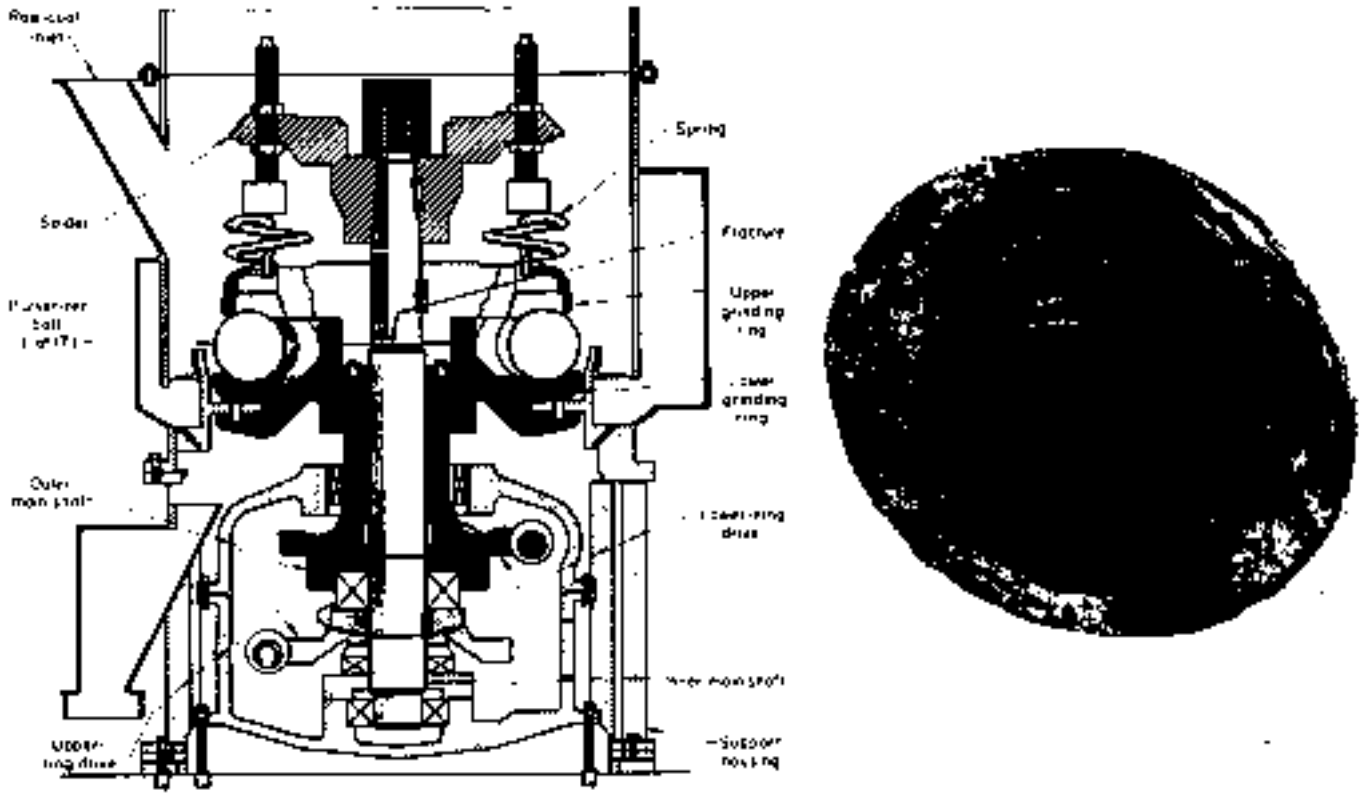


Fig. 11 6150 steel coal pulverizer shaft that failed by fatigue. (Left) Section through pulverizer showing the inner main shaft that fractured, repaired itself by friction welding, and fractured a second time. (Right) Photograph of the friction welded surface

The lower grinding ring of the pulverizer was attached to the outer main shaft (Fig. 11, left side). The upper grinding ring was suspended by springs from a spider that was attached to the main shaft. The weight of the upper grinding ring and the load imposed by the springs resulted in a total force of 8165 kg (18,000 lb) on the pulverizer balls. The shaft was made of 6150 steel normalized to a hardness of about 285 HB.

Investigation. Visual examination of the shaft revealed a circumferential crack in the main shaft just below the upper radial bearing at an abrupt change in shaft diameter. The shaft was set up in a lathe to machine out the crack for repair welding, and the smaller end (17.8 cm, or 7 in., in diameter) was found to be slightly eccentric with the remainder of the shaft. The crack did not disappear after a 6.3-mm (0.25-in.) deep cut, and the crack was opened by striking the small end of the shaft. The shaft broke about 1.3 cm (0.5 in.) from the crack. Examination of the fracture surface (Fig. 11, right side) revealed a previous fracture, almost perpendicular to the axis of the shaft, that resulted from torsional loading acting along a plane of maximum shear.

Although the shaft operated at the relatively low speed of 82 rpm, the weight of the upper portion of the shaft, the spider, and the upper ring caused the shaft to repair itself by friction welding in a band about 2.5 cm (1 in.) wide beginning at the periphery of the shaft. Thus, friction welding was confined to a small part of the 17.8-cm (7-in.) diam cross section. The center region of the fracture surface of the shaft (Fig. 11, right side) contained swirls of metal softened by frictional heat as a result of rotation of the lower part of the shaft. Welding of the plasticized outer ring of metal on the shaft and solidification of the swirls probably occurred while the pulverizer was shut down.

Examination of the machined surface revealed faint lines parallel to the visible crack. These faint lines generally followed bands containing large grains. The lines were thought to be fatigue cracks. The hardness of the shaft in the friction welded area was 38 to 49 HRC, and inward from the weld area, the hardness was 31 to 33 HRC.

Conclusion. The second fracture of the shaft was by fatigue, which resulted from an eccentric condition after the shaft was friction welded and from the inherent vibrations within the machine. Shafts in other coal pulverizers at the same plant fractured in the same region after similar service lives.

Corrective Measures. The shaft was repaired by welding a new 20.3-cm (8-in.) diam section to the older, longer section and machining to the required diameters. The fatigue cracks in the longer section were machined away before repair welding. Liquid-penetrant inspection was performed to ensure soundness of the metal in the older section. The repaired shaft was machined to provide a taper between the two diameters, rather than reproduce the sharp change in diameter in the original design. The taper section was carefully blended with the smaller diameter so that there were no sharp corners. The repaired coal-pulverizer shaft operated satisfactorily for more than five years. As shafts in other units failed, they were repaired by the same technique.

Example 5: Bending-Fatigue Fracture of an Inertia-Welded Alloy Steel Pushrod Originating at a Sharp Internal Corner and a Decarburized Surface. The pushrod shown in Fig. 12 fractured two weeks after it was installed in a mud pump operating at 160 rpm with a discharge pressure of 14.4 MPa (2.1 ksi). The pushrod was made by joining two pieces of bar stock by inertia welding. Each piece was rough bored or drilled to produce a wall thickness of about 1.3 cm (0.5 in.) at the surfaces to be inertia welded. The flange portion was made of 94B17 steel, and the shaft was made of 8620 steel. After welding, the rod was machined, carburized, and hardened to 60 HRC; the shaft was then chromium plated to within 1.3 cm (0.5 in.) of the flange.

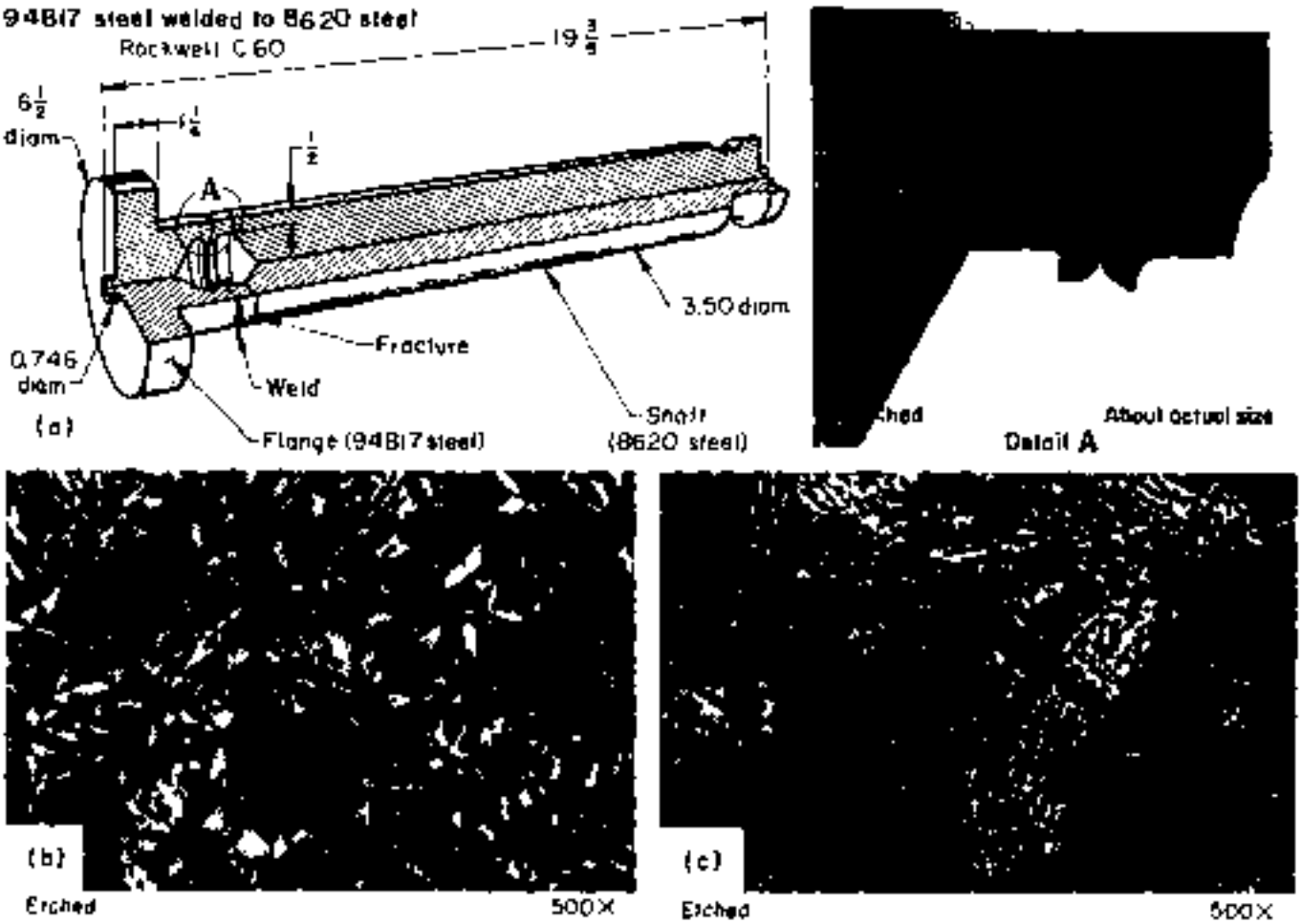


Fig. 12 Pushrod that fractured in bending fatigue after being fabricated by inertia welding. (a) Configuration and dimensions (given in inches). (b) and (c) Micrographs showing structure of decarburized inner surface and sound metal below the decarburized layer

Investigation. Visual examination of the pushrod disclosed that fracture had occurred in the shaft portion at the intersection of the 1.3-cm (0.5-in.) thick wall and the tapered surface at the bottom of the hole. This intersection was about 1.3 cm (0.5 in.) from the weld.

The fracture surface contained indications that a fatigue crack had initiated at the inner surface and progressed around the entire inner circumference, forming a narrow plateau, before propagating outward to final fracture. Operation of the pushrod after it had fractured obliterated some detail of the fatigue pattern on the fracture surface, but it still was evident that one-way bending stresses had influenced cracking. The amount of coarse texture on the fracture surface indicated that fatigue cracking had occurred for a short time.

Detail A in Fig. 12 shows a macrograph of a section through the weld and fracture region. A heavily decarburized layer was found on the inner surface of the flange portion, but was not detectable in the shaft portion. The inner surface had a poor finish, and there was a sharp corner at the intersection of the sidewall and bottom of the hole. The weld flash protruded from the inner wall, producing a small-radius fillet at the wall surface. Except for the metal extrusion, the weld zone was almost undetectable.

A metallographic specimen taken from near the fracture surface showed that the microstructure of the metal at the inner surface contained acicular ferrite and bainite, indicative of decarburization (Fig. 12b) The microstructure of the metal approximately 1.6 mm (0.0625 in.) from the inner surface contained low-carbon martensite and bainite (Fig. 12c).

Hardness of the flange was equivalent to 27 HRC at the bottom of the hole and 37 HRC on the inner-wall surface. The shaft had a hardness equivalent to 22 HRC at the bottom of the hole and 27 HRC at the inner-wall surface.

Sample parts having shapes and dimensions similar to those of the fractured pushrod were joined by inertia welding, then sectioned so that the weld area and inner surfaces could be examined. The sample weldments were not heat treated or machined. Sharp corners and rough surfaces were found that were similar to those on the fractured part (Detail A, Fig. 12). The cut surface was ground smooth and etched, which revealed the heat-affected zone of the weld but no indications of abnormal oxidation or decarburization of the inner surfaces.

Conclusion. The pushrod failed in fatigue primarily from one-way bending stresses. Cracking was initiated in a sharp, rough corner at the intersection of the sidewall and the bottom of a drilled hole that acted as a stress raiser. The stress raiser created by the abrupt section change was accentuated by the roughness of the drilled surface and by the decarburized layer. Decarburization occurred during heat treatment and was caused by the atmosphere trapped in the cavity formed by friction welding.

Corrective Measures. The design of the pushrod was changed to a one-piece forging. To produce a high-quality welded pushrod, it would be necessary to provide radii at the corners of the drilled holes, finish machine the inner surface, and drill a hole in the flange end of the rod to permit circulation of atmosphere during heat treatment.

Press-Fitted Members. Gears, pulleys, wheels, impellers, and similar components are often assembled on shafts by means of press fitting or shrink fitting, which can result in stress raisers under bending stress. Figure 13(a) shows typical stress flow lines in a plain shaft at a press-fitted member. Enlarging the shaft at the press-fitted component and using a large-radius fillet would produce a stress distribution such as that shown schematically in Fig. 13(b). A small-radius fillet at the shoulder would result in a stress pattern similar to that shown in Fig. 10(a).

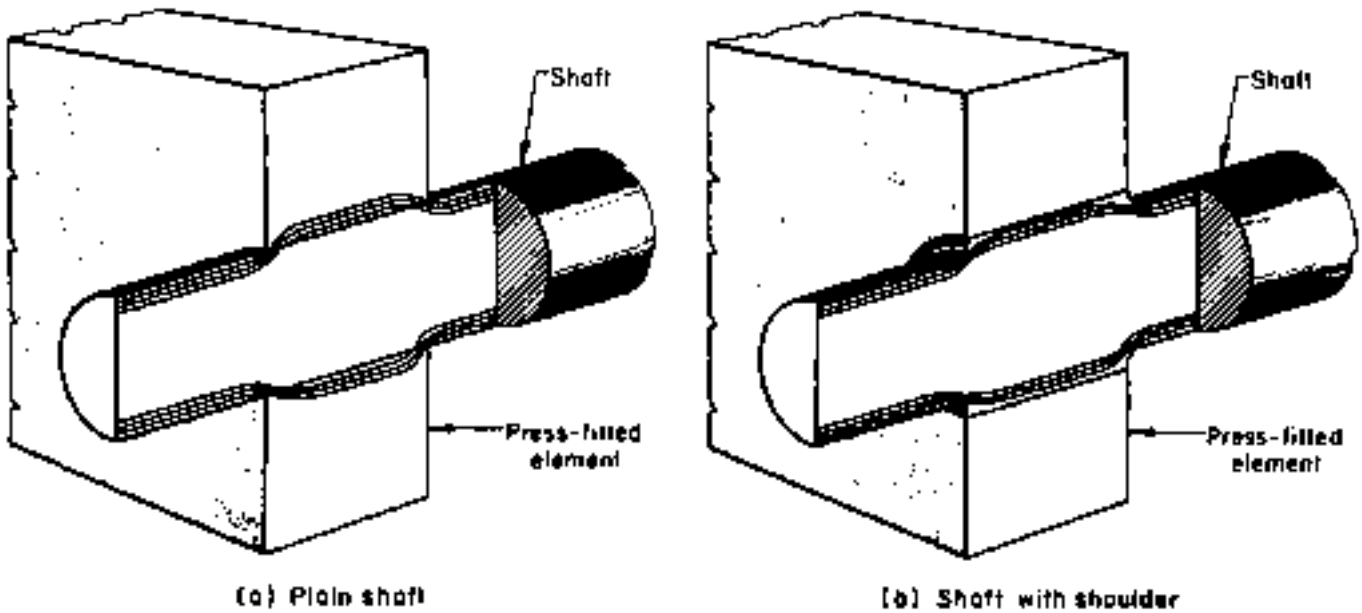


Fig. 13 Schematic illustration of stress distribution in two types of rotating shafts with press-fitted elements under a bending load.

An investigation was conducted to study the influence of absolute specimen dimensions on the fatigue strength of specimens with press-fitted bushings. The tests were carried out on specimens of St 35 carbon steel [0.18% C (max), 0.40% Mn (min), 0.35% Si (max), 0.05% P (max). 0.05% S (max)] with ultimate tensile strength of 559 MPa (81 ksi) and high-quality chromium-nickel-molybdenum alloy steel with ultimate tensile strength of 726 MPa (105 ksi) (similar to AISI 9840). Specimen blanks were cut from the surfaces of large forgings that were to be used for propeller shafts. Two groups of specimens were tested: (1) smooth cylindrical specimens 5, 12, 27, and 50 mm (0.2, 0.5, 1.1, and 2 in.) in diameter, with gage portions equal in length to four specimen diameters and shoulder radii equal to one diameter and (2) smooth cylindrical specimens 5, 12, 27, and 50 mm (0.2, 0.5, 1.1, and 2 in.) in diameter with press-fitted bushings. The outside diameter and length of each bushing were equal to two specimen diameters: thus, the bushings were similar in size to actual machine parts of this kind. The bushings were made of normalized (195 HB) St 34 steel [0.15% C (max), 0.08% P (max), and 0.06% S (max)] and a brass alloy of unreported composition. Press fitting was done to class 2 tolerances.

The specimens were tested in rotating bending at a constant frequency of 3000 rpm. The test base (in air) was 10^7 cycles for specimens without bushings and 5×10^7 cycles for specimens with bushings. The results (Table 1) showed that the fatigue limit of steel specimens was sharply reduced by the presence of press-fitted bushings and that the magnitude of this reduction increased with steel strength.

Table 1 Fatigue strength of steel specimens without bushings, with steel bushings, and with brass bushings

Specimen diameter		St 35 carbon steel						Chromium-nickel-molybdenum alloy steel					
		Without bushings		Steel bushings		Brass bushings		Without bushings		Steel bushings		Brass bushings	
mm	in.	MPa	ksi	MPa	ksi	MPa	ksi	MPa	ksi	MPa	ksi	MPa	ksi
5	0.2	265	38	157	23	147	21	319	46	196	28	147	21
12	0.5	255	37	157	23	142	20.6	314	45.5	167	24	152	22
27	1.1	221	32	123	18[(a)]	123	18	294	43	123	18	147	21
50	2.0	201	29	103	15	108	16	294	43	103	15	123	18

(a) 20-mm (0.8-in) diam specimen

The endurance of specimens with press-fitted bushings depends on specimen diameter to a greater extent than does the endurance of specimens without bushings. The effect of specimen diameter on the fatigue strength of both plain carbon steel and alloy steel specimens with press-fitted bushings was more pronounced when the bushings were made of steel than when they were made of brass.

The main cause of the reduction in the fatigue strength of the specimens with bushings was friction between the bushing and the shaft, which resulted from the cyclic loads in rotating bending, and the resulting wear of the shaft. The wear, which depends on the frictional force, that is, on the distribution of specific pressure, should vary (as did the pressure) along the shaft surface. When a shaft with a press-fitted bushing is bent. the maximum pressure should be near the bushing edge, at a point whose location will vary depending on the fit.

Friction produces wear of the shaft, leads to wear-induced surface roughness, and causes a local temperature increase. all of which promote nucleation and growth of cracks. Friction also leads to the destruction of oxide films, which passivate and strengthen metal parts, and activates metal surfaces as a result of plastic deformation. These factors facilitate chemical reaction between the metal surfaces and the working environment; the resulting damage is known as fretting.

The fatigue strength of specimens up to 20 mm (0.8 in.) in diameter was reduced more by the brass bushings than by the steel bushings. Conversely, the fatigue strength of specimens greater than 20 mm (0.8 in.) in diameter was reduced less by the brass bushings than by the steel bushings. The steel bushings, which are harder and more rigid than brass bushings, exhibited a higher friction coefficient when in contact with steel and produced more intense wear than the brass bushings. Other factors being equal, the environment had a more damaging effect on small-diameter specimens; the fatigue strength of large-diameter specimens was influenced predominantly by stress raisers produced as a result of friction-induced wear.

Longitudinal grooves in shafts, such as keyways and splines, are the origins of many service failures of shafts subjected to torsional stress. Most of these failures result from fatigue fracture where a small crack has initiated at a sharp corner because of stress concentrations. The crack gradually enlarges as cycles of service stress are repeated until the remaining section breaks. A sharp corner in a keyway can cause the local stress to be as much as ten times the average nominal stress.

Failures of this kind can be avoided by using a half-round keyway, which permits the use of a round key, or by using a generous fillet radius in the keyway. Good results are obtained by the use of fillets having radii equal to approximately one-half the depth of the keyway. A half-round keyway produces a local stress of only twice the average stress, thus providing greater load-carrying ability than that permitted by a square keyway. Many shafts with square keyways do not fracture in service because stresses are low or because fillets with generous radii are used.

Shafts with tapered, or conical, ends may also have keyways for alignment and/or redundancy. Integrity of the tapered joint relies on the friction fit between the wedgelike taper and the tapered hole in the hub of the part fitting to the shaft. It is essential to keep the fit very tight by using a threaded end and a locknut or other mechanism to avoid relying on the key and keyway to transmit the torque.

In an assembly in which a keyed member was loosely fitted to a shaft, nearly all the alternating torque was transmitted through the key, resulting in initiation of cracks at the bottom edge of the keyway of the shaft and producing a peeling type of fracture (Fig. 14a). Occasionally, peeling progresses entirely around the shaft and results in a sharp-edged shell like that shown in Fig. 14(b).

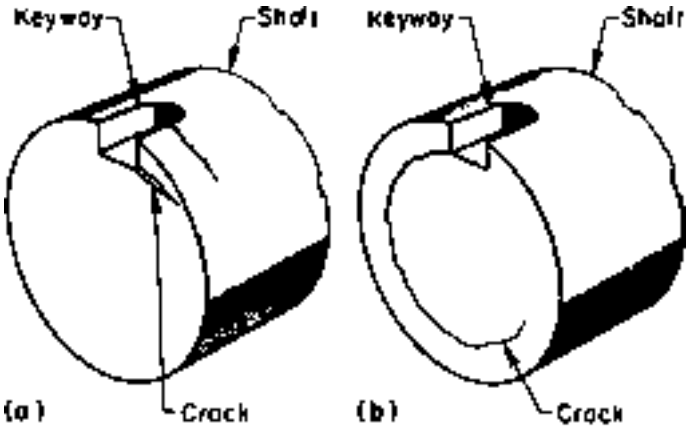


Fig. 14 Peeling-type cracks in shafts. (a) and (b) Cracks originated at keyways.

Figure 15 shows stress fields and corresponding torsional-fatigue cracks in shafts with key-ways or splines. In Fig. 15(a), the fillet in one corner of the keyway was radiused and the fillet in the other corner was sharp, which resulted in a single crack; this crack progressed approximately normal to the original stress field. In Fig. 15(b), both fillets in the keyway corners were sharp, which resulted in two cracks that did not follow the original stress field, a condition that results from the cross effect of cracks on the stress field.

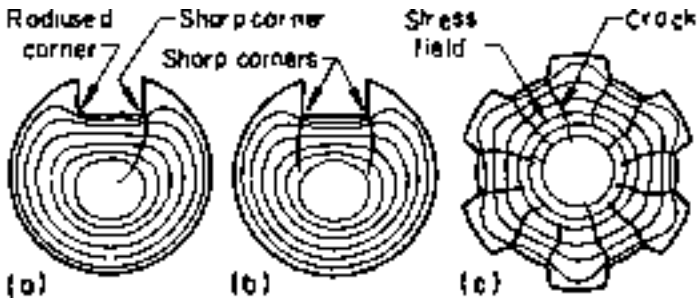


Fig. 15 Stress fields and corresponding torsional-fatigue cracks. (a) and (b) Shaft with keyway. (c) Shaft with splines

A splined shaft subjected to alternating torsion can crack along the bottom edges of the splines (Fig. 15c). This is another instance of a highly localized stress field strongly influencing crack development.

Example 6: Fatigue Cracking of 4340 Steel Compressor Shafts Because of Cyclic Stresses Induced by Gear-Type Couplings. A plant utilized high-horsepower electric motors to drive large compressors required for a manufacturing process. Eight compressors had shafts made of 4340 steel quenched and tempered to a hardness of 35 to 39 HRC and gear-type couplings. The compressors operated for short periods at power levels up to 3730 MW (5000 hp) and for longer periods at slightly above 2980 MW (4000 hp). Figure 16(a) shows the design of the shaft and the keyways. Two keys, spaced 180° apart, were expected to transmit the load so that an interference fit between shaft and coupling would not be required, thus simplifying maintenance. After only a few months of operation at above 2980 MW (4000 hp), six of the eight compressor shafts were found to be cracked in a keyway, and one of the six had fractured.

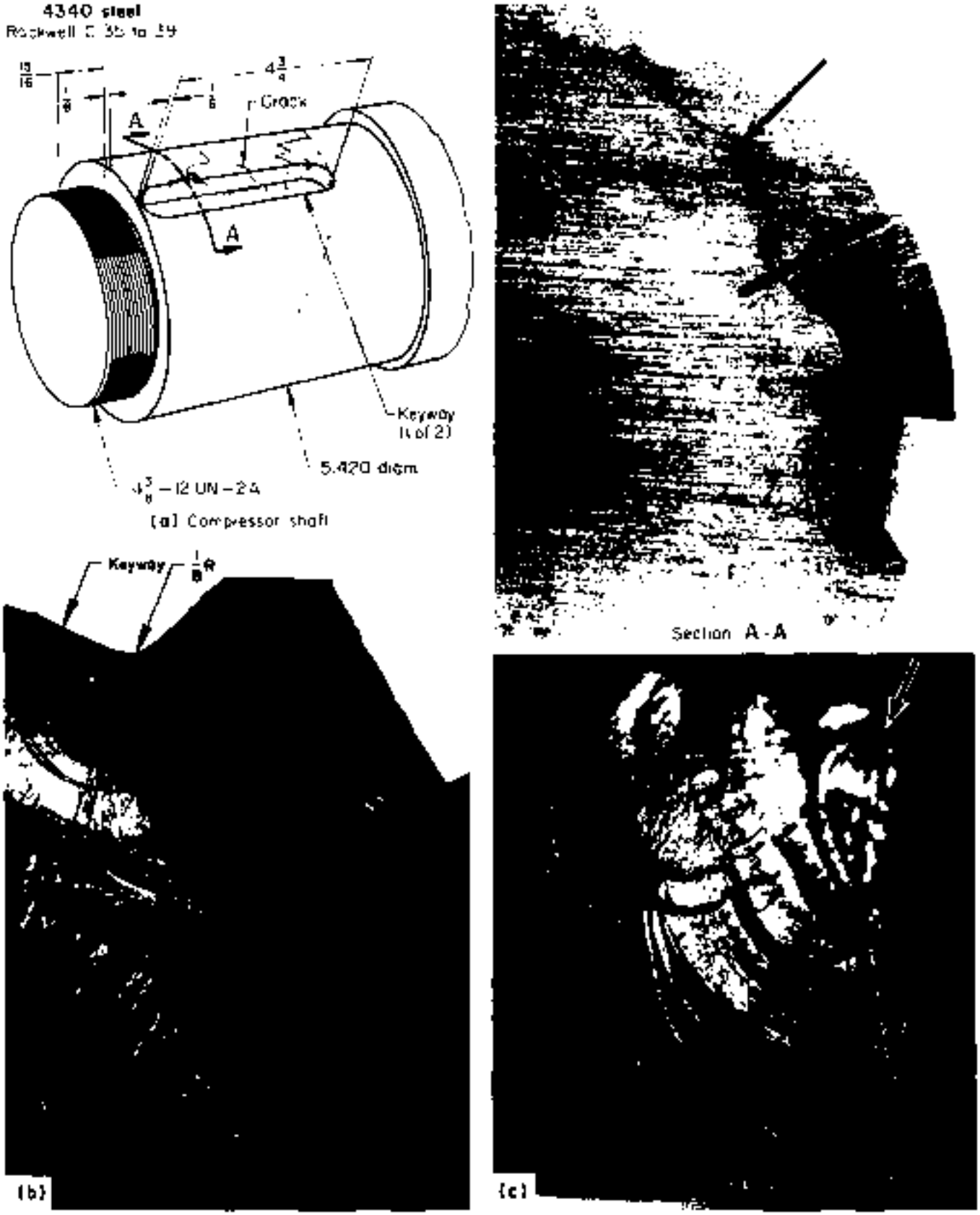


Fig. 16 4340 steel compressor shaft that fractured because of peeling-type fatigue cracking. (a) Configuration and dimensions (given in inches). (b) and (c) Photographs showing crack path, failure origin, and beach marks

Revised by Donald J. Wulpi, Metallurgical Consultant

Influence of Fabricating Practices

Surface discontinuities produced during manufacture, repair, or assembly (into a machine) of a shaft can become points of stress concentration and thus contribute to shaft failure. Operations or conditions that produce this type of stress raiser include:

- Manufacturing operations that introduce stress raisers, such as tool marks and scratches
- Manufacturing operations that introduce high tensile stresses in the surface, such as improper grinding, repair welding, electromachining, and arc burns
- Processes that introduce metal weakening, such as forging flow lines that are not parallel with the surface, hydrogen embrittlement from plating, or decarburization from heat treatment

Fatigue strength may be increased by imparting high compressive residual stresses to the surface of the shaft. This can be accomplished with such processes as surface rolling or burnishing, shot peening, tumbling, coining, or induction hardening.

Improper Machining. There are many ways in which improper machining can lead to shaft failures, and unless they are recognized, correction of service-failure problems can be difficult. The metal on the surface of a machined part can be cold worked and highly stressed to an appreciable depth (approximately 0.5 to 0.8 mm, or 0.020 to 0.030 in.). Occasionally, the heat generated in machining, particularly in grinding, is sufficient to heat a thin layer of the steel above the transformation temperature and thus cause martensitic hardening at the surface upon cooling. Stresses resulting from thermal expansion and contraction of the locally heated metal may even be great enough to cause cracking of the hardened surface layer (grinding cracks). Rough-machining operations can produce surface cracks and sharp corners, which concentrate stresses.

Example 7: Fatigue Fracture of a 4140 Steel Forged Crankshaft Resulting From Stress Raisers Created During Hot Trimming. Textile-machine crankshafts like that shown in Fig. 17 were usually forged from 1035 steel, but because of service conditions, the material was changed to 4140 steel. The forgings were made from 5.5-cm (2.15625-in.) diam bar stock by cutting to length, hot bending, upsetting, hot-trimming flash, hot pressing, and visually inspecting before shipping.

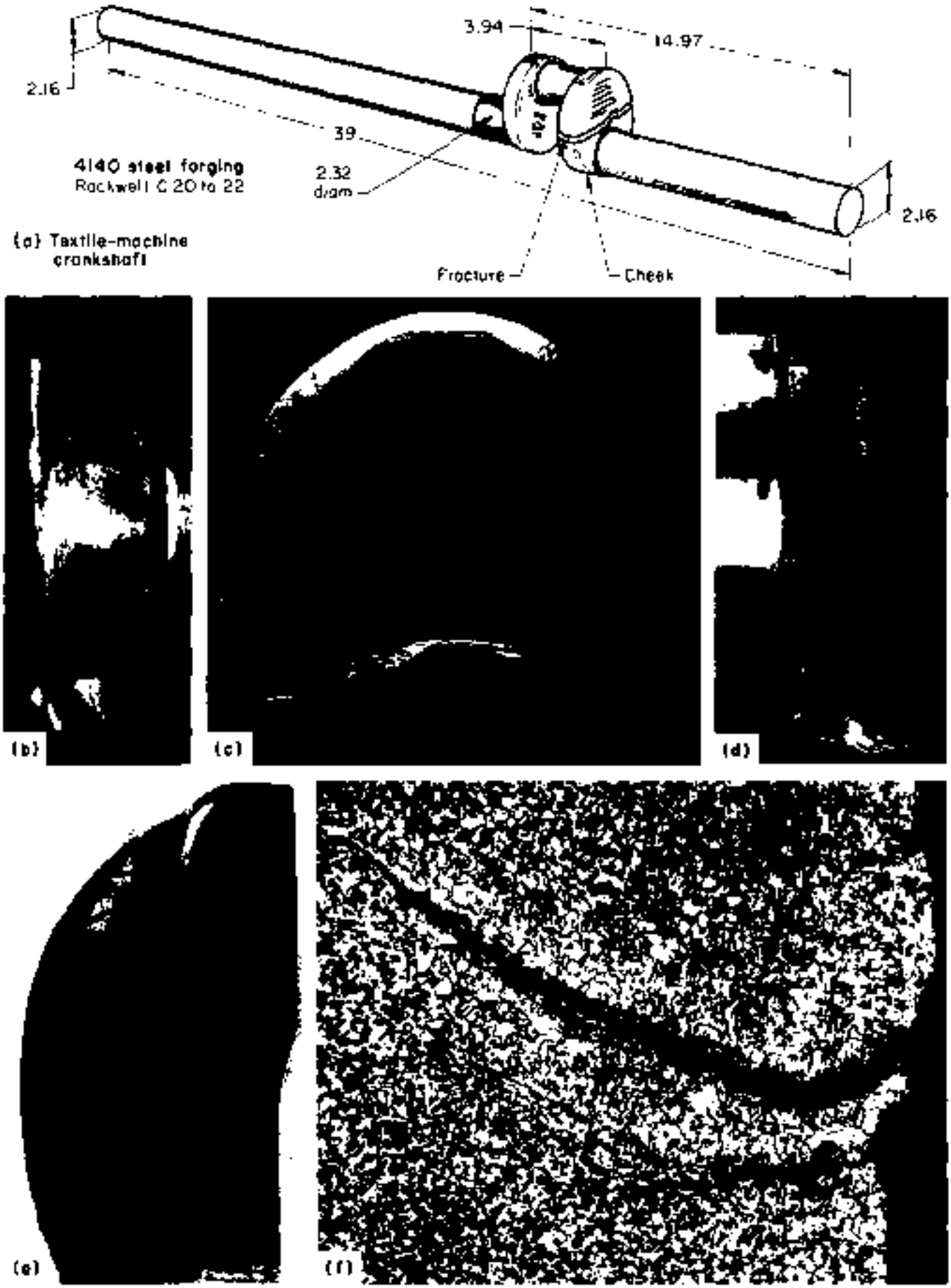


Fig. 17 Forged 4140 steel textile-machine crankshaft that fractured in fatigue originating at machining marks and forging defects. (a) Configuration and dimensions (given in

inches). (b) Fracture surface. (c) Hot trim marks. (d) Snag grinding marks. (e) Hot folds. (f) Section through a hot fold. Etched with 2% nital. 100×

The crankshafts were failing by transverse fracture of one cheek after one to three years of service. The expected life was 20 years of continuous service. One complete forging that had fractured (No. 1 in this Example) and a section containing the fractured cheek on the shorter shaft of another forging (No. 2) were sent to the laboratory so that the cause of failure could be determined.

Investigation. Visual examination of the fracture surfaces of both crankshafts revealed indications of fatigue failure; however, the origins were not readily visible (Fig. 17b). The surfaces had a clean fine-grain structure, but the edges were peened—evidently the result of damage after fracture.

The surfaces of the cheeks at the parting line contained rough grooves from hot trimming of the flash and from snag grinding (Fig. 17c and d). Forging 2 contained the most severe of such markings.

Longitudinal and transverse sections were prepared and etched with hydrochloric acid at a temperature of 71 to 77 °C (160 to 170 °F). The steel was of good quality and contained the normal amount of nonmetallic inclusions but no segregation or pipe.

Examination of an as-polished specimen revealed no intergranular oxidation. Etching the specimen with a 10% sulfuric acid and 10% nitric acid solution revealed no evidence of burning or overheating of the steel.

An area containing shallow surface folds was found on the outer face of one cheek of the throw on forging 1 (Fig. 17e). As shown in Fig. 17(f), the metal around one of the folds contained some ferrite, and the forged surface was slightly decarburized. Also, a fatigue crack had initiated in the fold and was propagating across the cheek. Examination of a section through the fracture surface disclosed cold working of the surface, which could have been the result of a rather extended period of fatigue cracking.

Chemical analysis of the two forgings found the metal to be 4140 steel, as specified. The hardness of the two forgings at the subsurface, midradius, and core ranged from 19 to 22 HRC. The specified hardness of the machined forging was unknown.

Tensile strength of the forgings ranged from 790 to 817 MPa (114.5 to 118.5 ksi), and yield strength at 0.2% offset was 580 to 620 MPa (84 to 90 ksi). Elongation in 3.8 cm (1.5 in.) was 20.6 to 22%, and reduction of area was 56.2 to 59.4%. These properties were representative of 4140 steel quenched and tempered to a hardness of 20 to 22 HRC. The general microstructure of the forgings was tempered bainite; the grain size was ASTM 6 to 8.

Conclusions. Fatigue cracking resulted in the transverse fracture of one cheek in each of the two crankshafts submitted for examination. In crankshaft 1, fatigue cracks were initiated at a shallow hot-work defect. A rough surface resulting from hot trimming or snag grinding of the forging flash was the point of initiation of fatigue cracks in crankshaft 2.

Corrective Measures. Before being machined, the forgings were normalized, hardened and tempered to 28 to 32 HRC to increase fatigue strength. The quenching procedure was changed to produce a more complete martensite transformation and to increase the ratio of yield strength to tensile strength. The surfaces were inspected by the magnetic-particle method, and shallow folds, notches, or extremely rough surfaces were removed by careful grinding.

Identification Marks. Excessive stresses may be introduced in shafts by stamped identification marks that indicate manufacturing date or lot number, steel heat number, size, or part number. The location of such a mark and the method by which it is made can be important. Identification marks should not be placed in areas of high bending or torsional stresses. For shafts, this often requires that they be located either on the end face or on an adjoining collar, but surface-finish requirements on the end of the shaft cannot be ignored if thrust loads are taken at that location. Stamping of marks with straight-line portions is the most likely to cause cracks, although characters with rounded contours also can cause cracking (Fig. 18). Stamping of metal shaft surfaces should be avoided because it is impossible to predict on which surface stamp marks will cause cracking in service (more information is provided in the article “Fatigue Failures” in this Volume.)

Fig. 18 Cracks in a shaft radiating from deep identification marks made with steel stamps. Cracking occurred during heat treatment of the shaft.

Example 8: Fatigue Fracture of a 4337 Steel Articulated Rod Originating at an Electroetched Numeral. The articulated rod in an aircraft engine fractured after being in operation for 138 h since the engine had been overhauled; total operating time was unknown. The rod was made from a 4337 steel (AMS 6412) forging and was quenched and tempered to 36 to 40 HRC.

Investigation. Visual examination disclosed that the rod had broken into two pieces, approximately 6.4 cm (2.5 in.) from the center of the piston-pin-bushing bore. The fracture had nucleated at an electroetched numeral 5 on one of the flange surfaces (arrow B, Fig. 19a). Around the nucleus was a series of concentric beach marks (arrows A and C, Fig. 19a) that extended almost the full width of the flange and about one-half the width of the web.

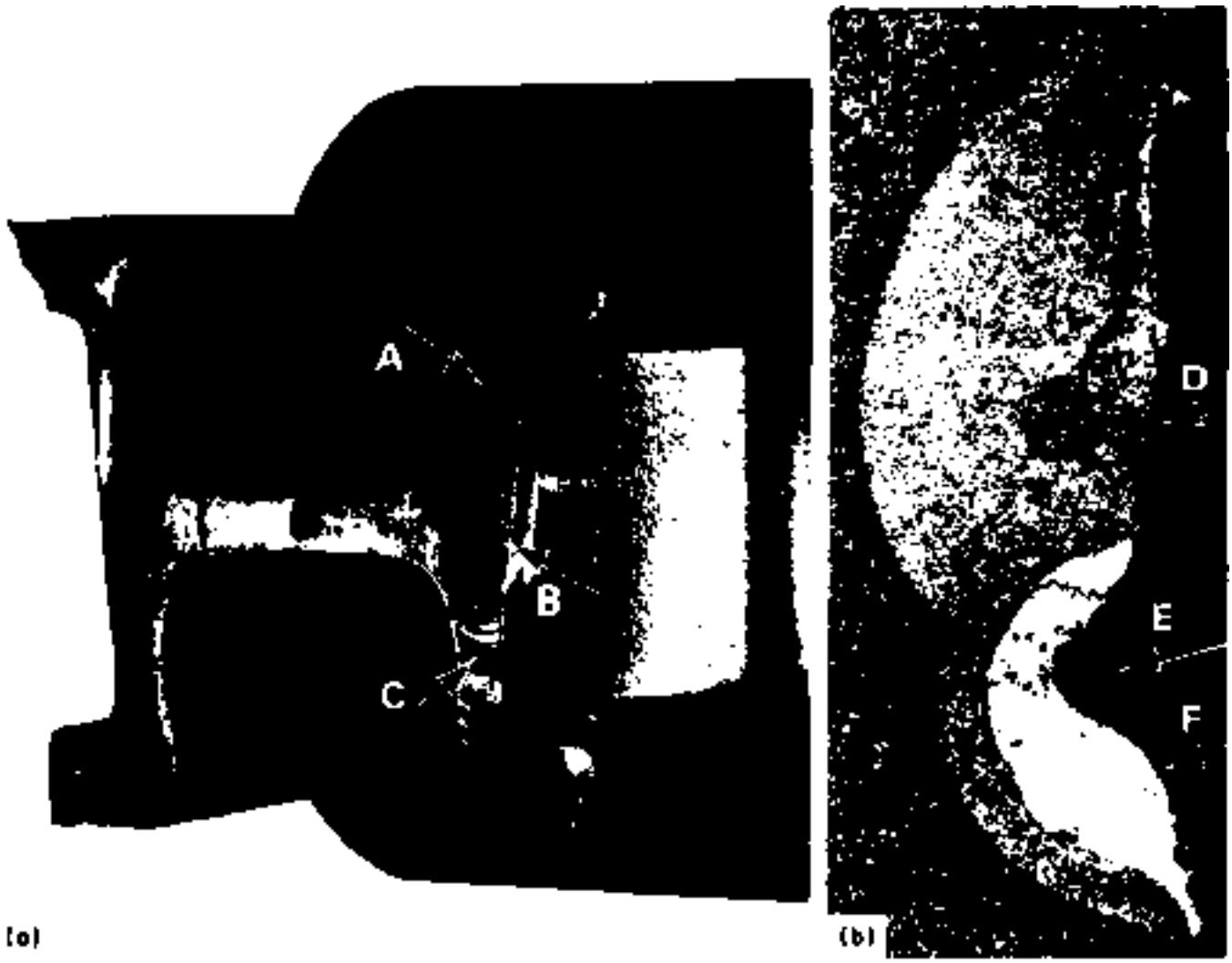


Fig. 19 Fracture surface of an articulated rod made from a 4337 steel forging. (a) Fracture origin (arrow B) at an electroetched numeral, and beach marks (arrows A and C). (b) Micrograph through fracture origin showing remelted zone (arrow D), notch produced by electroetching (arrow E), and untempered martensite (arrow F). Etched with 2% nital. 100×

Metallographic examination of a polished-and-etched section through the fracture origin revealed a notch (arrow E, Fig. 19b) that was caused by arc erosion during electroetching. The microstructure of the metal at the origin consisted of a remelted zone (arrow D, Fig. 19b) and a layer of untempered martensite 0.4 mm (0.015 in.) deep (arrow F, Fig. 19b). Small cracks were observed in the remelt area and were the result of thermal stresses caused by the high temperatures developed during electro-etching.

The hardness of the untempered martensite was 56 HRC, and the hardness of the tempered-martensite core was 40 HRC. The hardness of the rod (40 HRC) and the microstructure of the tempered martensite remote from the electro-etched area indicated that the material had been properly heat treated.

Conclusions. Fatigue fracture of the rod was caused by a metallurgical notch that resulted from electroetching of an identification number on the flange.

Corrective Measures. Marking of the articulated rods by electroetching was discontinued because of the detrimental effect of electroetching such highly stressed parts.

Residual Surface Stresses. Most machining operations produce notches and may also cause residual tensile stresses on the surface of the workpiece. Grinding can cause local high-temperature heating followed by very rapid cooling as the surface of the workpiece leaves the grinding-wheel-contact area. Grinding cracks can result, and these cracks can provide points of fatigue-crack initiation. However, grinding under properly controlled conditions provides a surface with a minimal amount of residual stress.

Grinding cracks are sometimes visible in oblique light, but are often so tight that they are impossible to see. They can be detected readily by magnetic-particle inspection, by fluorescent liquid-penetrant inspection, or by etching in cold dilute nitric acid. If grinding cracks are not detected before the shaft is put in service, the cracks may enlarge such that the shaft can fail by fatigue or brittle fracture. Grinding cracks often show a characteristic pattern; light grinding cracks occur as parallel cracks at 90° to the direction of grinding, whereas heavy grinding cracks have a rectangular network pattern.

Residual surface stresses also are produced by electrical discharge machining (EDM), which is metallurgically analogous to torch (flame) cutting but on a much smaller scale. The cutting effect is produced by a succession of sparks between the electrode and the workpiece. Each spark heats a small volume of metal to a temperature well above its melting point. Most of the molten metal is removed from the part by the action of the spark and the surrounding liquid dielectric, but a thin white surface layer remains on the workpiece and resolidifies as it is quenched by the surrounding mass of metal. The thickness of the white (untempered martensite) layer varies from 0.005 mm (0.0002 in.) on finishing cuts to 0.08 mm (0.003 in.) on roughing cuts. This untempered martensite will be very hard and brittle, possibly cracking and leading to fracture.

Beneath the quenched surface layer is another layer of steel that is heated during EDM; thus, if the workpiece was originally in the hardened condition, the heated layer is tempered. The lowest tempered hardness is adjacent to the quenched layer, and the hardness gradually rises in subsequent layers until the original base hardness is reached. The tempered layer is similar metallurgically to the heat-affected zone in welding and is impossible to avoid. Electrical discharge machining often produces an as-quenched martensitic layer of base metal that was not molten, but only heated above the critical temperature to a depth about the same as the white layer. Frequently this layer cannot be distinguished as separate from the remelted white layer. The tempered zone, when hardened steel is cut by EDM, varies in depth from about 0.05 mm (0.002 in.) for finishing cuts up to 0.8 mm (0.030 in.) for roughing cuts.

Example 9: Fatigue Fracture of a Rolling-Tool Mandrel Initiated at Cracks Formed by Machining of a Hole. The mandrel shown in Fig. 20 was part of a rolling tool used for mechanically joining two tubes before they were installed in a nuclear reactor. The operation consisted of expanding the end of a zirconium tube into a stainless steel cylinder having an inside diameter slightly larger than the outside diameter of the zirconium tube.

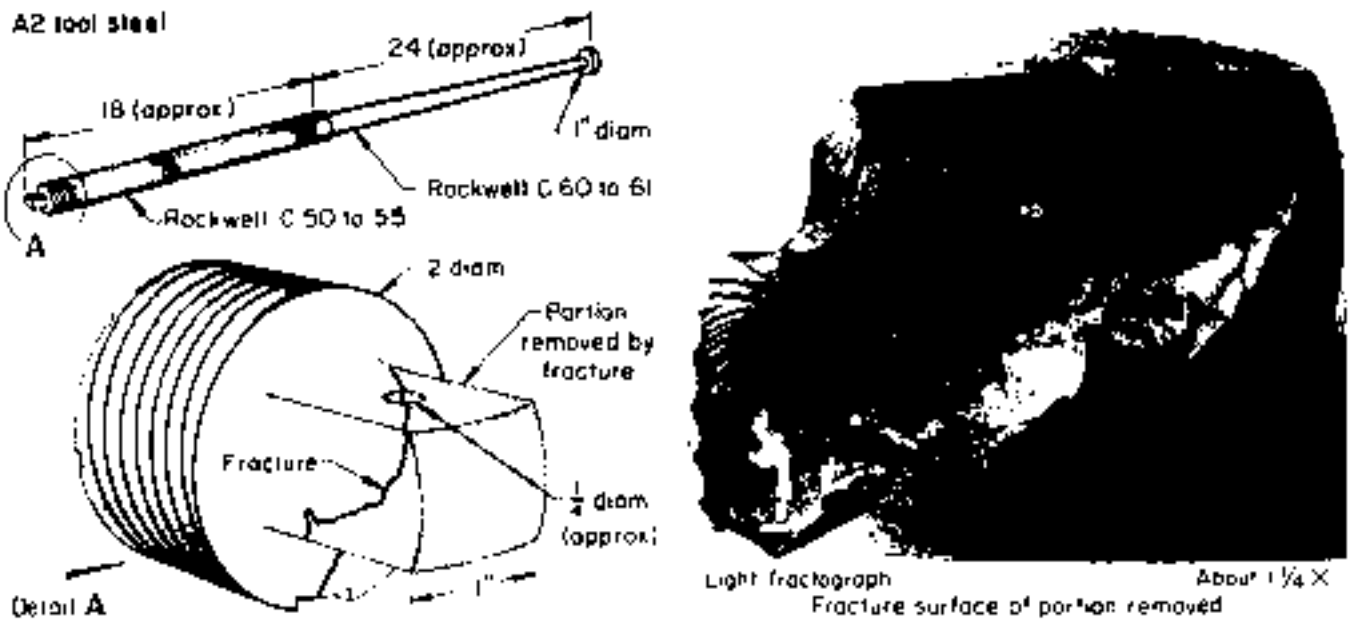


Fig. 20 A2 tool steel mandrel for a tube-expanding tool. Fracture originated at a 6.3-mm (0.25-n.) diam hole in the square end that was drilled by EDM. The fractograph shows a crack pattern on the fracture surface that originated at the hole.

Difficulty was experienced in withdrawing the tool, and a 6.4-mm (0.25-n.) diam hole was drilled by EDM through the square end of the hardened mandrel. The mandrel, which fractured after making five rolled joints, was made of A2 tool steel. The tapered end was hardened to 60 to 61 HRC, and the remainder of the mandrel to 50 to 55 HRC.

Investigation. Fracture had occurred at approximately 45° through the 6.4-mm (0.25-in.) diam hole in the square end and progressed into the threaded section to form a pyramid-shape fragment (Detail A, Fig. 20). The fracture surfaces exhibited brittle fracture characteristics but with clearly defined beach marks. The fracture pattern was characteristic of a torsional fatigue fracture.

The fracture originated on both sides and near the top of the hole in the square end, as shown in the fractograph in Fig. 20 (the poor surface finish of the hole is visible). Examination at approximately 10× revealed that the rough surface was the result of the metal having been melted.

Metallographic examination (at 250×) of specimens taken through the fracture origin revealed that melting had occurred around the hole, resulting in an irregular zone of untempered martensite with cracks radiating from the surface of the hole. The core material away from the hole exhibited a microstructure of fine tempered martensite containing some carbide particles. The martensitic zones around the hole had a hardness of 68 to 70 HRC. The core structure had a hardness of 60 to 61 HRC.

Conclusion. Failure of the mandrel was the result of torsional fatigue initiated by cracks formed by the EDM process used to drill the hole in the square end. Propagation of the fatigue crack was accelerated by the hardness of the material, which was considered exceptionally high for this application.

Corrective Measures. The hole through the square end of the mandrel was incorporated into the design of the tool and was drilled and reamed before heat treatment. The specified hardness of the threaded portion and square end of the mandrel was changed to 45 to 50 HRC, from the original hardness of 50 to 55 HRC. The specified hardness of the tapered end remained at 60 to 61 HRC.

Damage During Assembly. Operations performed during assembly can sometimes introduce misalignment of components, which can be detrimental to the performance of shafts. Proper alignment of bearings, seals, couplings, pedestals, and foundations has an influence on vibration and bending loads and thus on the fatigue performance of shafts.

Misalignment or even a change in alignment within tolerance limits in conjunction with a stress raiser, such as a deep scratch or a large nonmetallic inclusion, may initiate a fatigue crack. Figure 21(a) shows the fracture surface of a crankshaft for a diesel engine that had been operating for several years. The engine was overhauled, and the bearing journals were chromium plated. The crankshaft failed after the engine had operated only a few hundred hours.

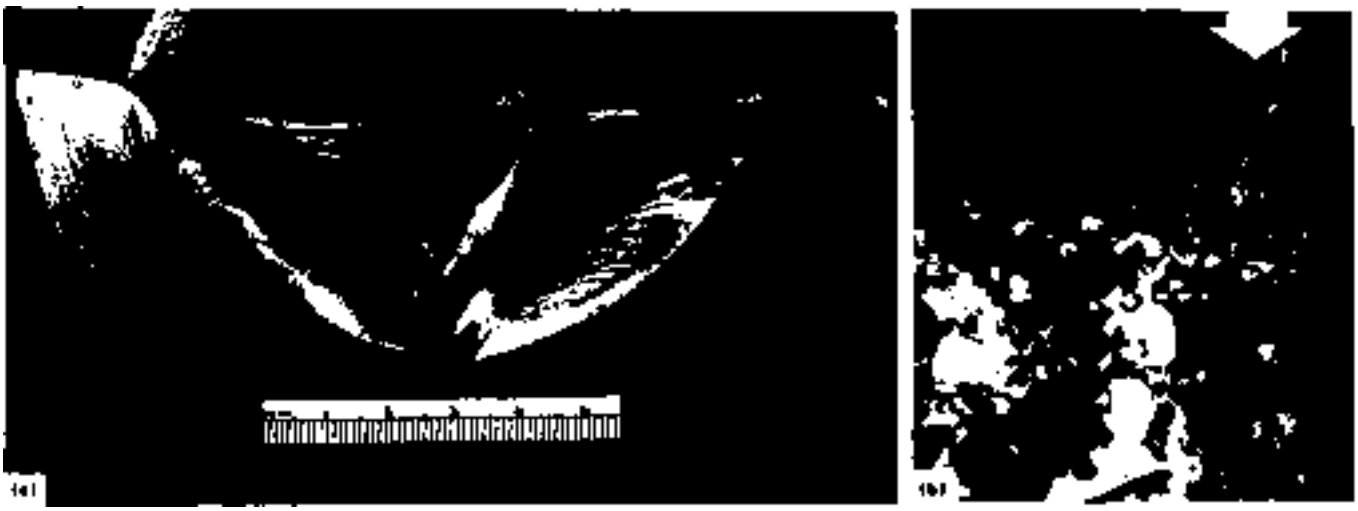


Fig. 21 Diesel-engine crankshaft that broke because of misalignment. (a) Fatigue marks on the fracture surface. (b) Micrograph of a section through the fracture origin showing a small crack (arrow) and some inclusions. Etched with nital. 500×

No scratches or tool marks were found at the fracture origin. Metallographic examination of a section through the fracture surface revealed a very fine crack parallel to the fracture surface and a line of inclusions (Fig. 21b).

Misalignment of the extension shaft with the drive shaft of the assembly shown in Fig. 22 resulted in bending-fatigue fracture of the shaft at a change in shaft diameter. The drive shaft protruded from a gearbox that was mounted rigidly on the base of the machine, but the extension shaft was supported in a pillow-block bearing. Tightening the pillow block deflected the extension shaft downward, imparting significant bending stress to the drive shaft.

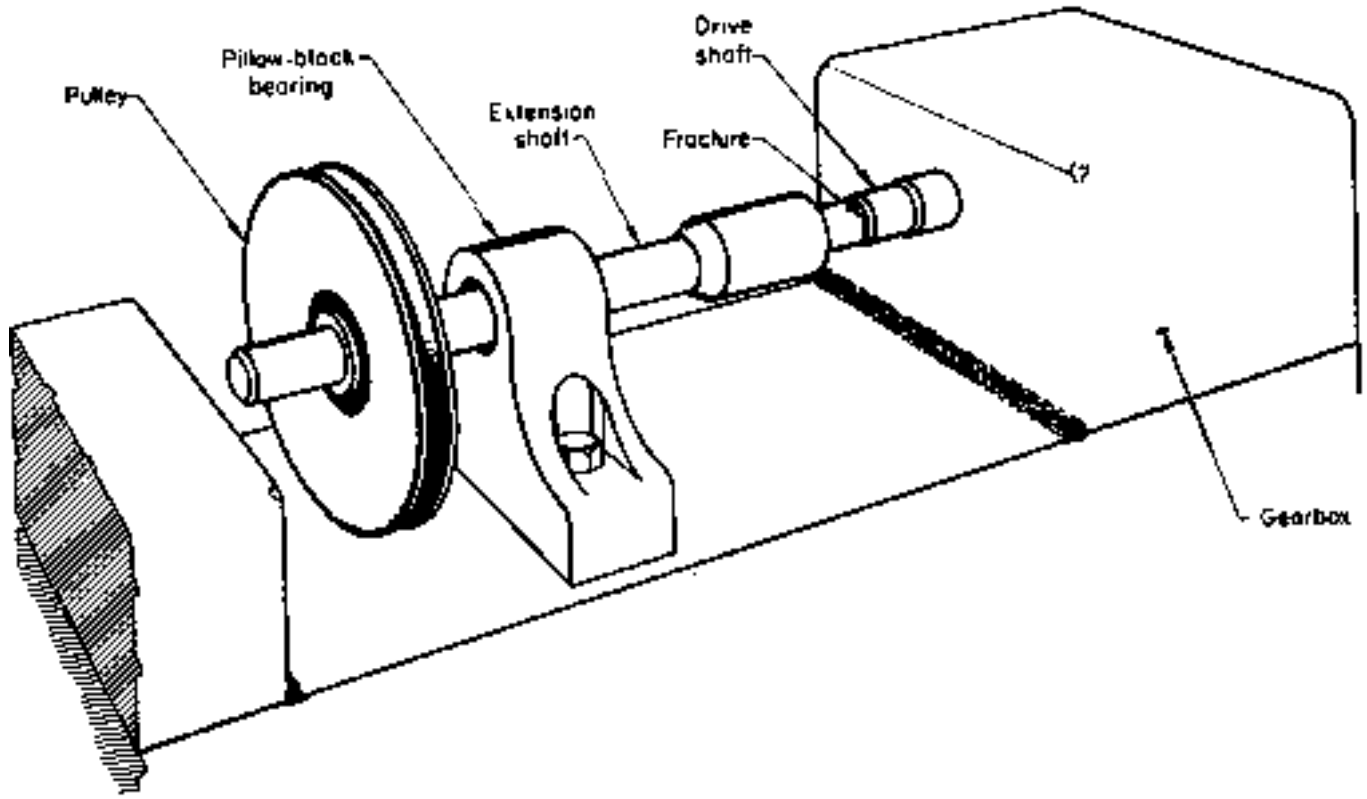


Fig. 22 Shaft assembly in which the height of the pillow-block bearing caused misalignment of the extension shaft with the drive shaft, resulting in bending-fatigue fracture.

Heat Treatment. The mechanical properties of the material in most shafts are developed through heat treatment or cold drawing. For quenched-and-tempered shafts less than 7.6 cm (3 in.) in diameter, a predominantly martensitic microstructure tempered to a surface hardness of 235 HB or higher is typical.

Where the properties of the shaft material throughout its entire cross section are important, use of the lowest austenitizing temperature consistent with the strength required is usually good practice. The higher the fracture toughness of a shaft (high K_{Ic}), the larger the crack needed to cause fracture.

For most shafts, where surface stresses are high and crack initiation is of concern, surface decarburization can be a major problem. For very large shafts, rotational speeds are often high enough that internal centrifugal stresses are of major concern. For such shafts, heat treatments are designed to give optimum strength and toughness uniformly across the full section.

Example 10: Fatigue Fracture of an 8640 Steel Shaft From a Fuel-Injection-Pump Governor Because of Insufficient Fatigue Strength. The shaft shown in Fig. 23 was from a fuel-injection-pump governor that controlled the speed of a diesel engine used in trucks and tractors. Shafts in newly installed governors began breaking after only a few days of operation. Specifications required the shaft to be made of 8640 steel heat treated to a hardness of 32 to 36 HRC.

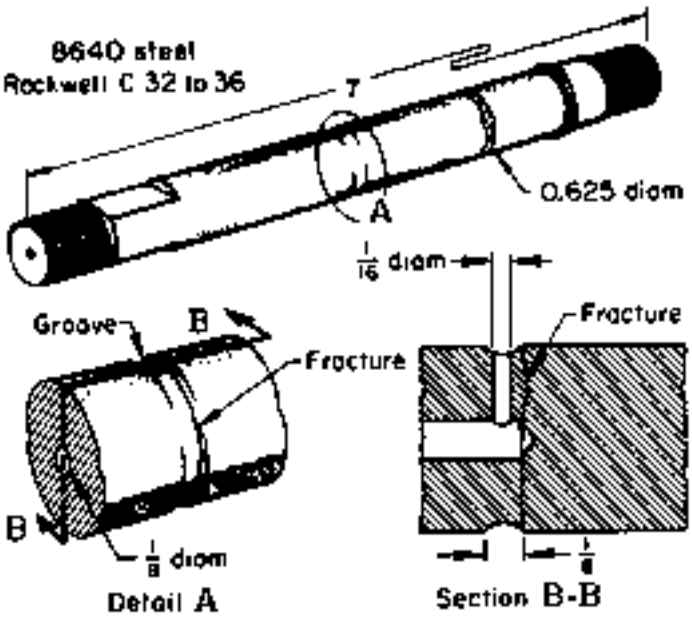


Fig. 23 8640 steel shaft for a fuel-injection-pump governor that fractured by fatigue through a lubrication hole. Fatigue life of the shaft was increased by nitriding the critical surface. Dimensions given in inches

The shaft had a cross hole and a groove that were part of a force-feed lubricating system for a sleeve that moved longitudinally on the shaft to control the amount of fuel delivered to the engine. The shaft rotated at relatively high speed and was subjected to shock loading; therefore, the mechanism that drove the shaft included a slip clutch designed to eliminate transmission of shock to the governor shaft.

Investigation. Visual examination of the broken shaft showed that fracture had initiated in the sharp corner at the bottom of the 3.2-mm (0.125-in.) diam longitudinal hole (Section B-B, Fig. 23). Beach marks were observed on the fracture surfaces. The shafts had been made of the prescribed metal and heat treated as specified.

Further investigation disclosed that, in an effort to reduce costs, the slip clutch had been eliminated from the drive mechanism, thus removing the cushioning effect that had been provided for the governor shaft. Restoration of the original design in a short time was not feasible.

Shafts were taken from stock and fatigue tested in a rotating-beam machine in a manner that would concentrate the stress at the circumferential groove. By this procedure, the fatigue limit for the shaft (survival for 10^7 cycles) was found to be about 480 MPa (70 ksi). This fatigue limit was insufficient for the application because shafts were fracturing after operating a few hours to a few days.

To improve the fatigue life of the shafts, finished shafts were taken from stock and nitrided for 10 h at 515 °C (960 °F). The nitrided shafts were then fatigue tested. Results of the tests indicated that the nitriding treatment had increased the fatigue limit from a minimum of 480 MPa (70 ksi) to a minimum of 760 MPa (110 ksi). Tests showed this strength to be satisfactory for the application.

Conclusions. Fatigue fracture of the shaft was the result of increased vibration and shock loading after the slip clutch had been eliminated from the drive mechanism. The increased vibration produced stresses that exceeded the fatigue strength of the metal in the shaft.

Corrective Measures. The shafts were nitrided, and the portion of the surface covered by the sleeve was lightly buffed after nitriding. No further changes were made in the manufacture of these parts. No failures occurred during several months of operation, and the use of nitrided shafts became standard practice. Nitriding the shafts was more economical than restoring the slip clutch.

Example 11: Fracture of a 1040 Steel Fan Shaft Resulting From Use of an Improper Material. The fan drive support shaft shown in Fig. 24 fractured after 3600 km (2240 miles) of service (minimum expected life, 6440 km, or 4000 miles). Specifications required that the shaft be made of cold-drawn 1040 to 1045 steel with a minimum yield strength of 586 MPa (85 ksi). The fractured shaft was sent to the laboratory to determine if failure had resulted from inherent stress raisers in the design or from nonconformance to specifications.

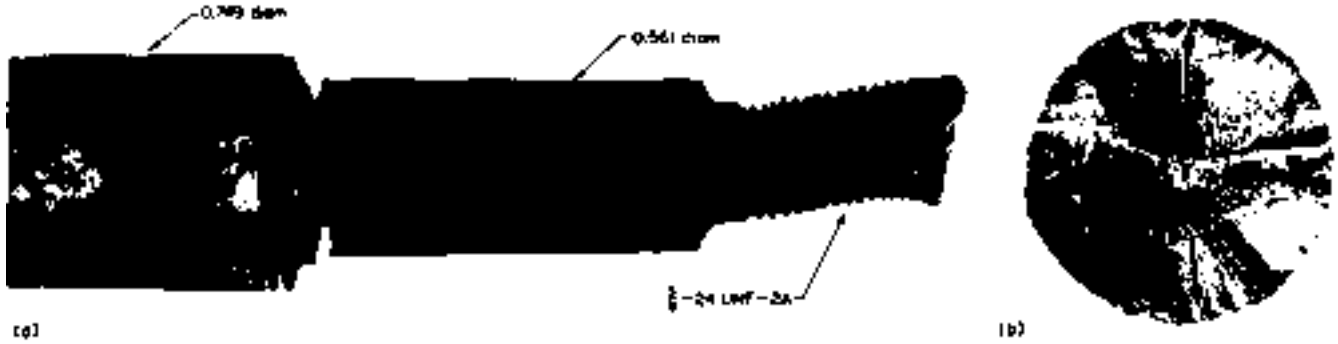


Fig. 24 1040 steel fan shaft that fractured in reversed-bending fatigue. (a) Overall view of shaft. Dimensions given in inches. (b) Fracture surface showing diametrically opposed origins (arrows)

Investigation. Visual examination of the shaft revealed that the fracture had been initiated near the fillet at an abrupt change in shaft diameter. Examination of the fracture surface disclosed that cracks had originated at two locations approximately 180 ° apart on the outer surface of the shaft (arrows, Fig. 24b) and propagated toward the center. Final fracture occurred near the center of the shaft. The fracture surface exhibited features typical of reversed-bending fatigue (Fig. 24b). Analysis of the material in the shaft showed that it conformed with the composition of 1040 steel.

The mechanical properties of a tensile specimen machined from the center of the shaft were: tensile strength, 631 MPa (91.5 ksi); yield strength, 369 MPa (53.5 ksi); and elongation, 27%. The yield strength was much lower than the specified minimum of 586 MPa (85 ksi). Hardness of the shaft was 179 HB.

Metallographic examination disclosed that the microstructure was predominantly equiaxed ferrite and pearlite, indicating that the material was in either the hot-worked or normalized condition. The grain size was ASTM 6 to 7, which is a fine-grain structure. No nonmetallic stringers or segregation were visible. Severe corrosion, which probably occurred after fracture, was found on the fracture surfaces of the shaft.

Conclusions. The shaft failed in reversed-bending fatigue. Fracture was initiated near a fillet at an abrupt change in shaft diameter, which is considered a region of maximum stress. The metal did not conform to specifications; it was either hot rolled or normalized and had a yield strength of only 369 MPa (53.5 ksi).

Recommendations. The development of a quenched-and-tempered microstructure would result in an improvement of the fatigue strength of the shaft. A 40% increase in the fatigue limit could be effected by quenching and tempering the steel to a hardness of 30 to 37 HRC (286 to 344 HB) after machining. Shot peening the fillet after machining would provide residual compressive stresses at the surface, which would inhibit the formation of fatigue cracks and increase the fatigue limit.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Influence of Metallurgical Factors

The fatigue properties of a material depend primarily on microstructure, inclusion content, hardness, tensile strength, distribution of residual stresses, and severity of the stress concentrators that are present.

Internal discontinuities, such as porosity, large inclusions, laminations, forging bursts, flakes, and centerline pipe, will act as stress concentrators under certain conditions and may originate fatigue fracture. To understand the effect of discontinuities, it is necessary to realize that fracture can originate at any location—surface or interior—where the stress first exceeds material strength. The stress gradient must be considered in torsion and bending because the stress is maximum at the surface, but is zero at the center or neutral axis. In tension, however, the stress is essentially uniform across the section.

If discontinuities such as those noted above occur in a region highly stressed in tension by bending or torsional loading, fatigue cracking may be initiated. However, if the discontinuities are in a low-stress region, such as near a neutral axis, they will be less harmful. Similarly, a shaft stressed by repeated high tensile loading must be free from serious

imperfections, for there is no neutral axis; any imperfection can be a stress concentrator and can be the origin of fatigue cracking if the stress is high with respect to the strength.

Example 12: Fatigue Cracking of a Forged 4337 Steel Master Connecting Rod Because of Nonmetallic Inclusions. Routine inspection of a reciprocating aircraft engine revealed cracks in the master connecting rod. Cracks were observed in the channel-shaped section consisting of the knuckle-pin flanges and the bearing-bore wall. The rods were forged from 4337 (AMS 6412) steel and heat treated to a specified hardness of 36 to 40 HRC.

Investigation. Visual examination revealed H-shaped cracks in the wall between the knuckle-pin flanges (Fig. 25a). The cracks originated as circumferential cracks, then propagated transversely into the bearing-bore wall. Magnetic-particle and x-ray inspection before sectioning did not detect any inclusions in the master rod.

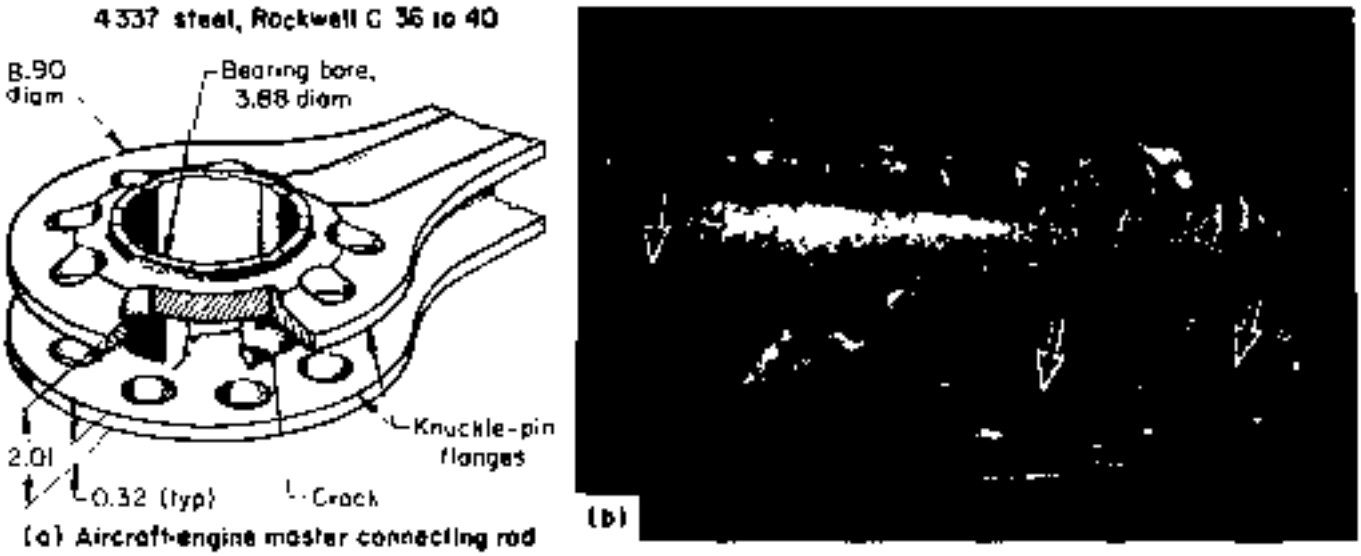


Fig. 25 Forged 4337 steel master connecting rod for a reciprocating aircraft engine that failed by fatigue cracking in the bore section between the flanges. (a) Configuration and dimensions (given in inches). (b) Fractograph showing inclusions (arrows) and fatigue beach marks

Macroscopic examination of one of the fracture surfaces revealed three large inclusions lying approximately parallel to the grain direction and fatigue beach marks around two of the inclusions. The inclusions and beach marks are shown in Fig. 25(b).

Microscopic examination of a section through the fracture origin showed large nonmetallic inclusions that consisted of heavy concentrations of aluminum oxide (Al_2O_3). These inclusions were of the type generally associated with ingot segregation patterns.

The hardness of the rods, 36 to 40 HRC, and the microstructure of the heat-treated alloy steel were satisfactory for the application.

A preliminary stress analysis indicated that the stresses in the area of cracking, under normal operating conditions, were relatively low compared with other areas of the rod, such as in the shank and the knuckle-pin straps.

Conclusions. The rod failed in fatigue in the bore wall between the knuckle-pin flanges. Fatigue was initiated by the stress-raising effect of large nonmetallic inclusions. The nonmetallic inclusions were not detected by routine magnetic-particle or x-ray inspection because of their orientation.

Recommendations. The forging vendors were notified that nonmetallic inclusions of a size in excess of that expected in aircraft-quality steel were found in the master connecting rods. Forging techniques that provided increased working of the material between the knuckle-pin flanges to break up the large nonmetallic inclusions were not successful. A nondestructive-testing procedure for detection of large nonmetallic inclusions was established.

Example 13: Fatigue Cracking of a 1040 Steel Crankshaft Because of Excessive Segregation of Nonmetallic Inclusions. The crankshaft in a reciprocating engine had been in operation for less than one year when the engine was shut down for repairs. Examination of the engine components disclosed that the crankshaft had suffered fairly severe cracking. The crankshaft was sent to the laboratory for a complete examination.

Investigation. The journals of the main and crankpin bearings were inspected by the magnetic-particle method. At least four of the main-bearing journals had three to six indications of discontinuities 1.5 to 9.5 mm (0.0625 to 0.375 in.) long. Another main-bearing journal had approximately 20 similar but generally shorter indications. One main-bearing journal had a crack along the fillet and almost entirely through the web (Fig. 26). Another main-bearing journal had a 1.3-cm

(0.5-in.) long crack in the fillet. Several of the crankpin journals had cracks 10 to 12.7 cm (4 to 5 in.) long, primarily in the fillets.

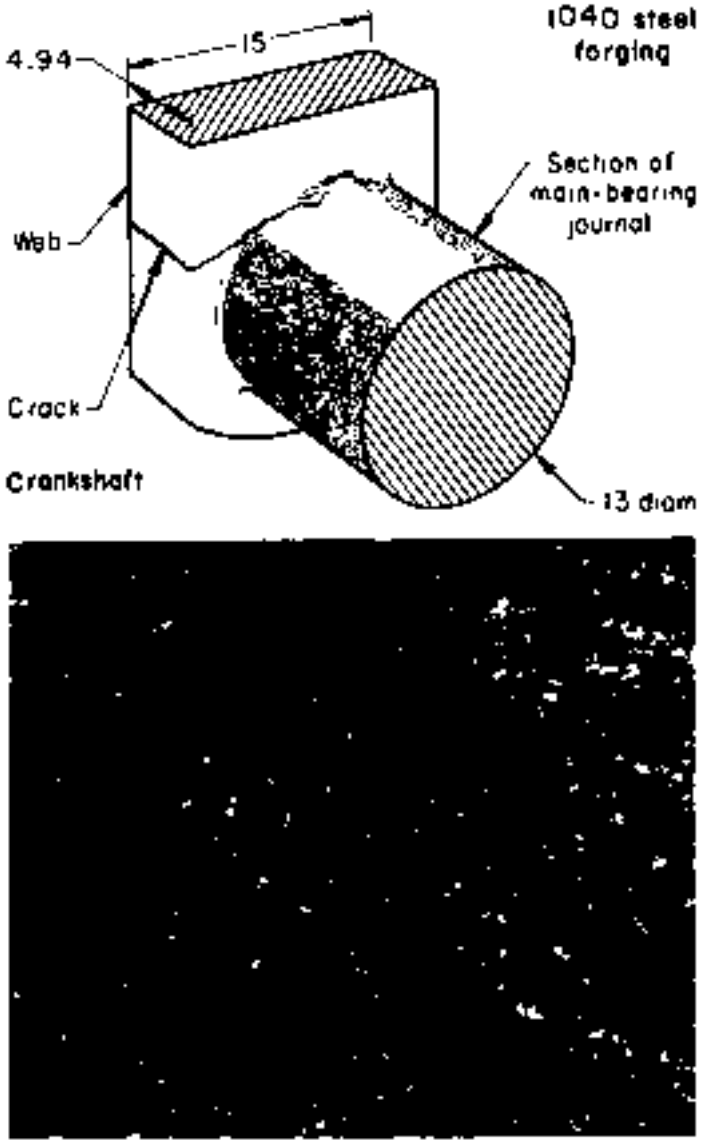


Fig. 26 Forged 1040 steel main-bearing journal that failed in fatigue. Top: Section showing cracks originating at coarse sulfide inclusions. Dimensions given in inches. Bottom: Macrograph of a 5%-nital-etched section showing the segregated inclusions (dark areas). 4×

A metallographic section was taken through the No. 4 main-bearing journal at the primary crack. The surface was macroetched, which disclosed numerous large, coarse segregates identified as sulfide inclusions (Fig. 26).

Macroscopic examination of the crack surface disclosed indications of fatigue cracking with low-stress high-cycle characteristics. Sulfide inclusions were present in the region where cracking originated.

The crankshaft was made from a 1040 steel forging. The mechanical properties of the metal met the specifications.

Conclusions. The crankshaft failed in fatigue at the main-bearing and crankpin-bearing journals because of excessive segregation of sulfide inclusions, which acted as stress raisers at which fatigue cracks initiated.

Corrective Measures. Ultrasonic inspection was used in addition to magnetic-particle inspection to detect discontinuities.

Surface Discontinuities. During primary and secondary mill operations, a variety of surface imperfections often result from hot plastic working of material when lapping, folding, or turbulent metal flow is experienced. The resultant surface discontinuities are called laps, seams, and cold shuts. Similar discontinuities also are produced in cold-working operations, such as fillet and thread rolling. Other surface imperfections develop from embedding of foreign material under high pressures during the working process; for example, oxides, slivers, or chips of the base material are occasionally rolled or forged into the surface.

Most of these discontinuities are present in the metal before final processing and are open to the surface. Standard nondestructive testing procedures, such as liquid-penetrant and magnetic-particle inspection, will readily reveal the presence of most surface discontinuities. If not detected, discontinuities may serve as sites for corrosion or crack initiation during fabrication, in addition to their deleterious effect on fatigue strength.

Because fatigue-crack initiation is the controlling factor in the life of most small shafts, freedom from surface imperfections becomes progressively more important in more severe applications. Similarly, internal imperfections, especially those that are near the surface, will cause fatigue cracks to grow under cyclic loading and result in failure when critical size is attained. Service life can be significantly shortened when such imperfections cause premature crack initiation in shafts designed on the basis of conventional fatigue-life considerations. Surface or subsurface imperfections can cause brittle fracture of a shaft after a very short service life when the shaft is operating below the ductile-to-brittle transition temperature. When the operating temperature is above the transition temperature or when the imperfection is small relative to the critical flaw size, especially when the cyclic-loading stress range is not large, service life may not be affected.

Example 14: Brittle Fracture of Splines on Induction-Hardened 1151 Steel Rotor Shafts Caused by a Seam in the Material. Splined rotor shafts like that shown in Fig. 27 were used on small electric motors. It was found that one spline was missing from each of several shafts before the motors were put into service. The shafts were made of 1151 steel, and the surfaces of the splines were induction hardened to 58 to 62 HRC. Several shafts were sent to the laboratory for examination to determine the cause of failure.

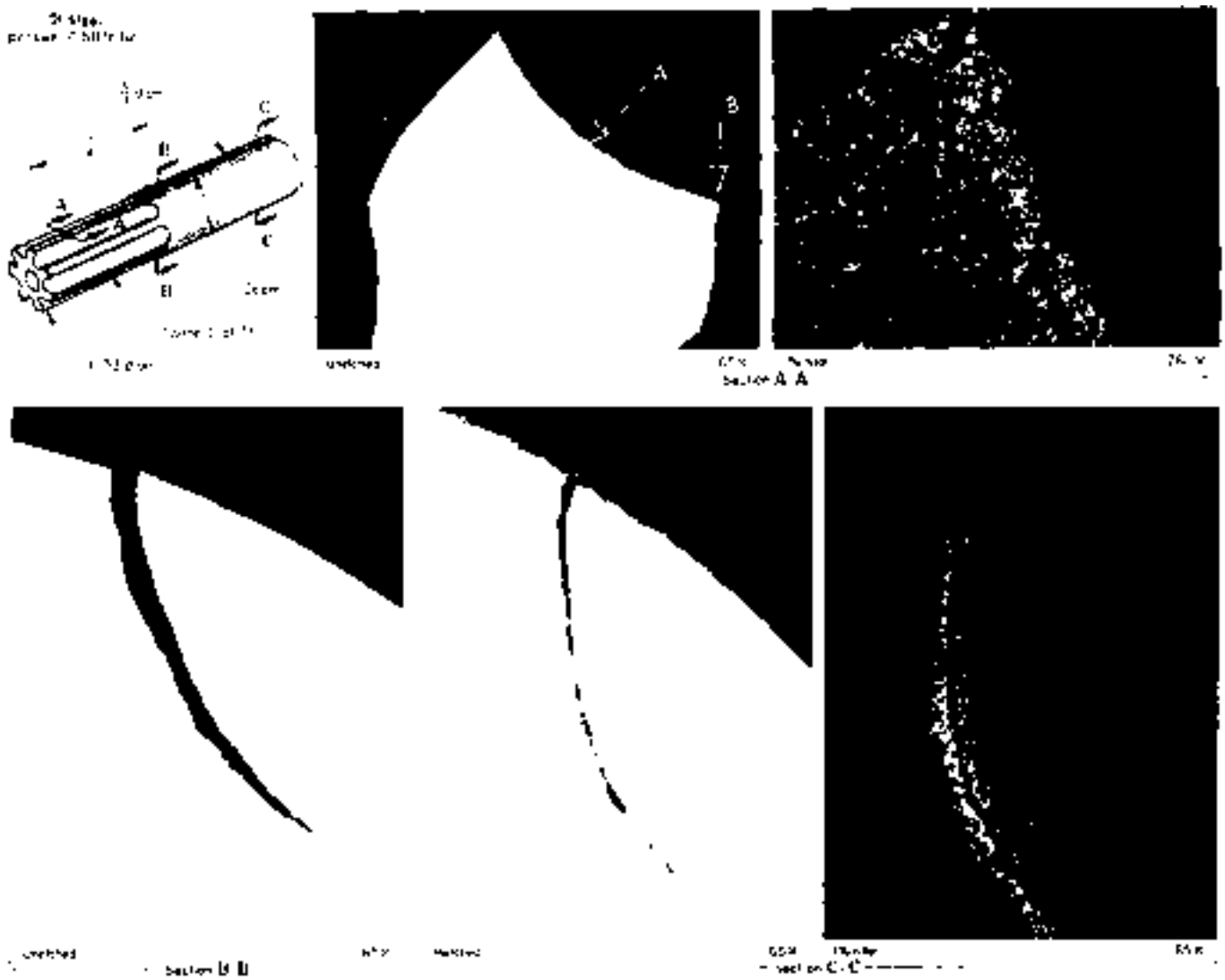


Fig. 27 Induction-hardened 1151 steel rotor shaft in which a spline fractured because of a seam. Top left: Configuration and dimensions (given in inches). Section A-A: Micrographs of section through broken spline, showing shape of fracture (arrow A), root of seam (arrow B), and decarburized surface. Sections B-B and C-C: Micrographs of

section through seam showing regions where metal was affected and not affected by induction heating

Visual examination. The shafts were examined both visually and with the aid of a stereoscopic microscope. This examination revealed apparent peeling of splines on the induction-hardened end of each rotor shaft. Only one tooth on each shaft had broken off. Examination of the fracture surfaces revealed that they were dark in color, which indicated that cracking had occurred before or during oil quenching after induction heating.

The shafts were inspected by the fluorescent liquid-penetrant method for surface cracks or imperfections. This inspection revealed a longitudinal discontinuity extending from the fracture surface of the tooth through the heat-affected zone at the end of the spline and into the unheated portion of the shaft (Fig. 27). This defect exhibited the characteristic directionality of a seam.

Metallographic Examination. Specimens were removed at three locations along each of the fractured shafts and prepared for metallographic examination.

A micrograph (65 ×) of an unetched section through a fractured tooth (Section A-A, left, Fig. 27) showed that the fracture surface (arrow A, Fig. 27) was concave and had an appearance characteristic of a seam. The tooth separated from the shaft when a crack originated at the root of the seam (arrow B, Fig. 27) and propagated at 90° to the seam toward the root fillet. Etching the specimen in 1% nital and examining it at a magnification of 260 × disclosed partial decarburization of the surface (Section A-A, right, Fig. 27). The microstructure of the metal in this region of the shaft was characteristic of a free-machining medium-carbon steel.

Examination (at 65 ×) of an unetched section through the region at the end of the splines that was affected by induction heating (Section B-B, Fig. 27) showed that the seam had been opened as the result of thermal stresses produced during induction hardening. The curvature of the surface at the seam was similar to that on the fractured tooth.

Examination (at 65 ×) of an unetched section through the shaft in an area unaffected by induction heating (Section C-C, left, Fig. 27) disclosed the presence of a crack extending to a depth of approximately 0.6 mm (0.025 in.). The root of the crack contained oxides typical of those found in seams. Etching the specimen in 1% nital and examining it at 65 × revealed partial decarburization along the crack surfaces similar to that found on the fracture surface of the tooth (Section C-C, right, Fig. 27). The curvature of the crack surface was similar to that observed in the heat-affected zone and on the fracture surface of the spline tooth.

Decarburization of the crack surfaces and oxides in the root of the crack indicated that the seam had been present before the shaft was heat treated and are characteristic of seams produced during the manufacture of steel billets, bars, rods, and wires.

The average hardness across the tooth surface was 60 HRC, which was within the specified range.

Conclusions. The shafts failed by brittle fracture because of deep seams produced during processing of the steel. The seams acted as stress raisers during induction hardening.

Recommendations. Specifications should require that the shaft material be free of seams and other surface imperfections.

Grain size, composition, and microstructure of a shaft material largely define the strength and toughness of that material and thus the performance of a shaft of that material. Generally, a fine-grain low-temperature transformation product (martensite or bainite) is desirable. However, where elevated temperatures are involved, a coarse grain size may be preferred. Grain size, composition, and microstructure affect material toughness at any given strength level; only when these factors combine to produce low toughness do they contribute to brittle fracture of a shaft.

Toughness and transition temperature are generally more important in large-diameter shafts than in small-diameter shafts. The life of a small-diameter shaft is usually determined both by the time required for a fatigue crack to initiate and by the rate of crack propagation.Large-diameter shafts are more likely to contain flaws of critical size. Thus, the life of a large-diameter shaft is often determined solely by the rate of propagation of cracks originating at pre-existing flaws. For large-diameter shafts with high internal stresses, the minimum level of toughness that can be expected in the shaft at minimum operating temperature must be known in order to determine the maximum size of flaw that can be tolerated.

Temper embrittlement of certain shaft steels, especially the nickel-chromium, nickel-chromium-molybdenum, and nickel-chromium-molybdenum-vanadium alloy steels used in larger rotor shafts, results from prolonged exposure at temperatures of 350 to 575 °C (660 to 1070 °F) or from slow cooling after exposure above this temperature range. In most instances, a short exposure time within, or rapid cooling through, this temperature range will minimize temper embrittlement. However, when heat treating thick sections, this procedure may not be possible, and temper embrittlement may occur. Fortunately, temper embrittlement can be reversed by retempering above the temperature range of 350 to 575 °C (660 to 1070 °F) followed by rapid cooling through this range to restore toughness (details of various types of embrittlement are provided in the article “Mechanisms and Appearances of Ductile and Brittle Fractures” in this Volume).

Temper embrittlement decreases impact toughness by increasing the ductile-to-brittle transition temperature and decreases the amount of ductile fracture; thus, it is of considerable concern for larger shafts. For steels that exhibit a sensitivity to temper embrittlement, the degree of embrittlement that develops is largely dependent on composition, heat treatment, and service conditions.

With respect to composition, molybdenum in amounts from 0.20 to 0.30% significantly retards temper embrittlement. Greater amounts of molybdenum yield no additional improvement. The effect of molybdenum on suppressing temper embrittlement decreases as the purity of the steel increases. The total content of impurities (sulfur plus phosphorus plus nonferrous elements plus gases), expressed in atomic parts per million, ranges in value from about 1500 ppm in steels of conventional purity (corresponding to the usual amounts found in air-arc-melted steel) to 1000 ppm in very clean steel (corresponding to vacuum-melted steel) and to about 500 ppm in the extraclean steels (corresponding to vacuum-melted steel using a very pure charge). Hence, when relatively impure steel is used (with over 1500 ppm impurities and over 0.01% P), the unique role of molybdenum in minimizing temper embrittlement is very important. Molybdenum does not appear to be a necessary alloying element in the production of high-purity steels (under 500 ppm impurities and under 0.001% P), which are not susceptible to temper embrittlement.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Effect of Surface Coatings

Metallic coatings of several types and compositions are often used to protect shaft surfaces from wear and corrosion and to repair worn shafts. The types of coatings include, but are not limited to, electroplating, metal spraying (with or without subsequent fusing), catalytic deposition, and metal overlaying. Typical coating materials are chromium, nickel, iron, alloy steels, aluminum alloys, copper alloys, and chromium-nickel-iron alloys. Care must be taken in applying these coatings, because harmful residual stresses and/or stress concentrators may be inadvertently left in or produced in the shaft.

Many electroplating processes leave harmful residual stresses in the plating and therefore in the adjacent base metal. Chromium, iron, and nickel platings generally contain high residual tensile stresses, which reduce the fatigue strength of the base metal of a shaft. Residual tensile stresses reduce the corrosion resistance of both the plating and the base metal. Piston rods and other parts made of high-strength steel and plated with metallic coatings can crack and fracture at unexpectedly low stress levels in simple fatigue.

During acid cleaning and electroplating, hydrogen is usually produced and may be absorbed into the base metal. If the hydrogen is not removed by a subsequent heat treatment, severe embrittlement of the base metal may occur, especially in base metals with hardnesses above approximately 35 HRC. High-strength highly stressed parts can crack and fracture as a result of hydrogen embrittlement. Failure by hydrogen embrittlement is even more likely to occur if the shaft contains high residual tensile stresses before being plated. More information on failures caused by hydrogen embrittlement can be found in the article “Hydrogen Damage and Embrittlement” in this Volume.

Metal spraying to obtain specific surface properties is sometimes used on shafts and, if proper bonding is obtained, is usually satisfactory. However, spraying heavy thicknesses of metals that have fairly high coefficients of thermal expansion on large-diameter shafts results in high residual tensile stresses; in the as-sprayed condition, the metal sometimes cracks. If subsequent heating is required to bond the coating to the base metal, high residual stresses and/or thermal cracking of the base metal or coating may occur.

Sprayed metal is mechanically weak and, if overloaded in compression, tends to spread laterally, thus disturbing its adhesion to the underlying base metal. Sprayed metal adheres well to crankshaft main journals but not to crackpin journals, which are more heavily loaded than crankshaft main journals and are subject to shock loading. The sprayed metal is entirely dependent on a mechanical keying action for its adherence to the base metal. A shallow V-shaped groove, such as a shallow small-pitch thread, is inadequate and has an undesirable stress-raising effect. The most satisfactory method of surface preparation is shot or grit blasting, which must be done in a manner so that the surface is roughened and not simply peened. The stress-raising effects of blasting are negligible because of the small size of the indentations. However, adhesion of sprayed metal to a surface prepared by grit blasting may not always be satisfactory under severe loading conditions, such as those to which a crankpin is subjected in service.

Example 15: Fatigue Fracture of a Crankshaft That Had Been Reconditioned by Metal Spraying.* During a general overhaul of a four-cylinder engine, the crankpin journals and 7.6-cm (3-in.) diam main-beating journals were built up by metal spraying. Four weeks later, the crankshaft broke through the crankpin farthest from the flywheel (driving) end.

Investigation. Examination of the fracture surface revealed that cracking had originated in the fillet at the “inside” surface of the crankpin (arrow A, Fig. 28a). The crack had propagated in fatigue perpendicular to the crankpin axis and across approximately 80% of the crankpin before final rupture occurred.

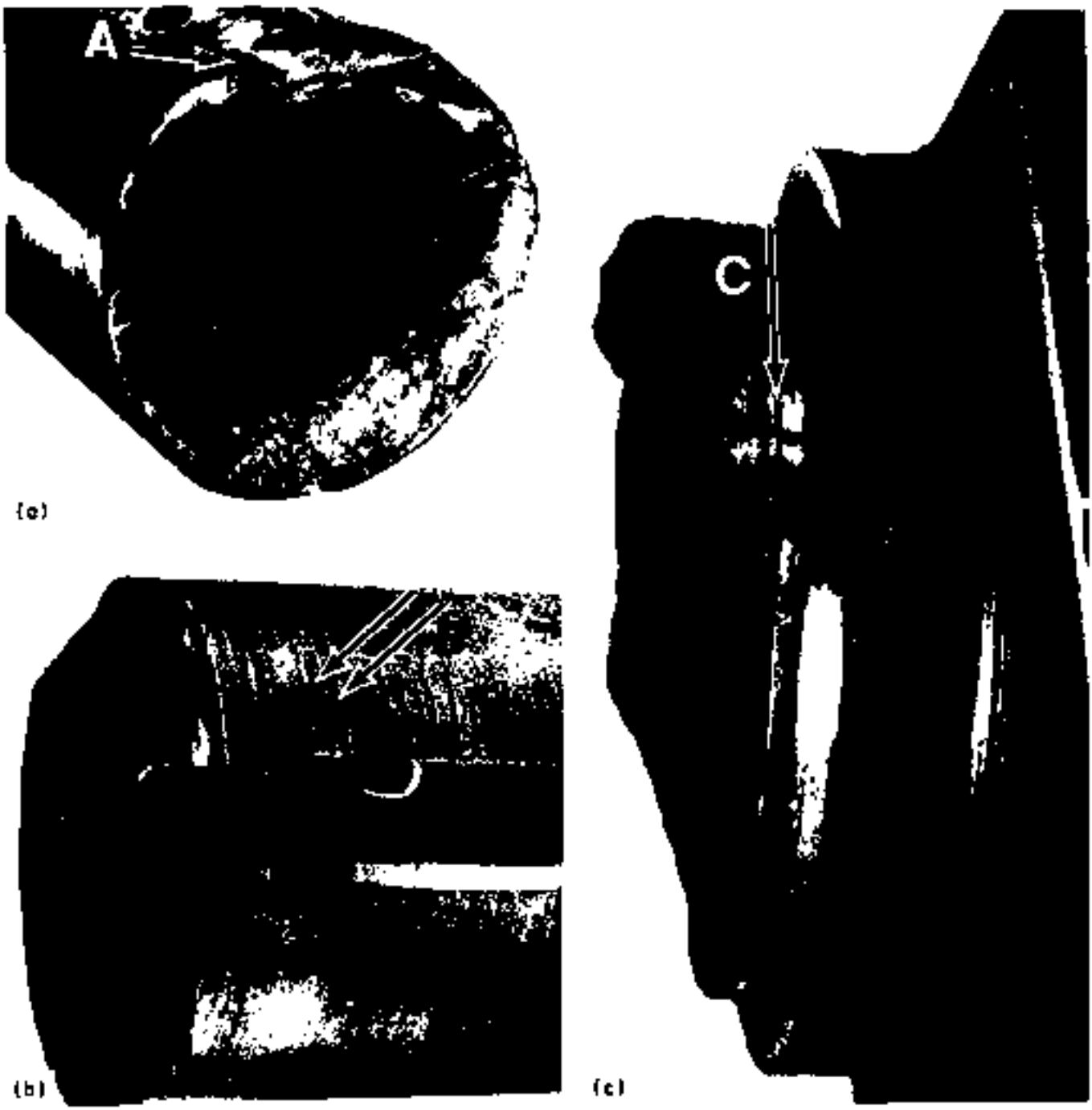


Fig. 28 Crankpin of a crankshaft that fractured in fatigue after the crankpin journals and main-bearing journals were rebuilt by metal spraying. (a) Fracture surface showing fracture origin (arrow A). (b) Machined grooves and punch marks (arrow B) on surface of crankpin. (c) Groove in crankpin, where cracking was initiated

It was found that adhesion of the sprayed-metal coating to the fractured crankpin was very poor and that the coating could be detached easily by tapping a chisel inserted between the coating and the pin. Portions of the sprayed coating were removed from all the pins and journals so that the surface preparation carried out before metal spraying could be examined. In each of two pins, a helical groove approximately 0.8 mm (0.03 in.) wide and deep, with a pitch of 10 per inch, had been cut in the surface (arrow B, Fig. 28b). The grooving extended over one-half of the pin periphery only; the remaining portion of the surface had been indiscriminately indented with a tooth punch (Fig. 28b).

From one pin, a large portion of the sprayed metal, extending the whole width of the crank-pin and covering one-third of its circumference, came away in one piece. Examination indicated that the tongues of metal that filled the grooves had broken away from the main layer at the level of the original pin surface. The average thickness of the sprayed layer was 0.8 mm (0.03 in.). The hardness of the crankpin was only 163 HB, indicating that it was readily machinable.

However, the surfaces of the main journals had not been grooved, but appeared to have been roughened by shot or grit blasting before spraying. The sprayed coating was more firmly adherent to these surfaces than to those of the pins.

With regard to the sequence of failure, there was a possibility that the crack might have been developing at the time restoration of the shaft was carried out. Careful examination showed that this was unlikely, because the origin of the crack was situated at the bottom of one of the turned grooves and the crack followed the groove about halfway around the pin periphery. Figure 28(c) is a view of a portion of the crankpin, showing that the crack occurred in the groove. The continuation of the groove beyond the point where the crack changed direction is shown at arrow C in Fig. 28(c). The crack was not located near the region where the fillet was tangent with the shaft surface, as are the majority of such cracks in crankshafts, but developed about 3.2 mm (0.125 in.) away in the cylindrical region of the pin, which also suggests that its location was determined by the pre-existing machined groove.

Conclusions. The shaft failed in fatigue. Cracking originated in a groove that was machined in the crankpin to help retain the sprayed metal.

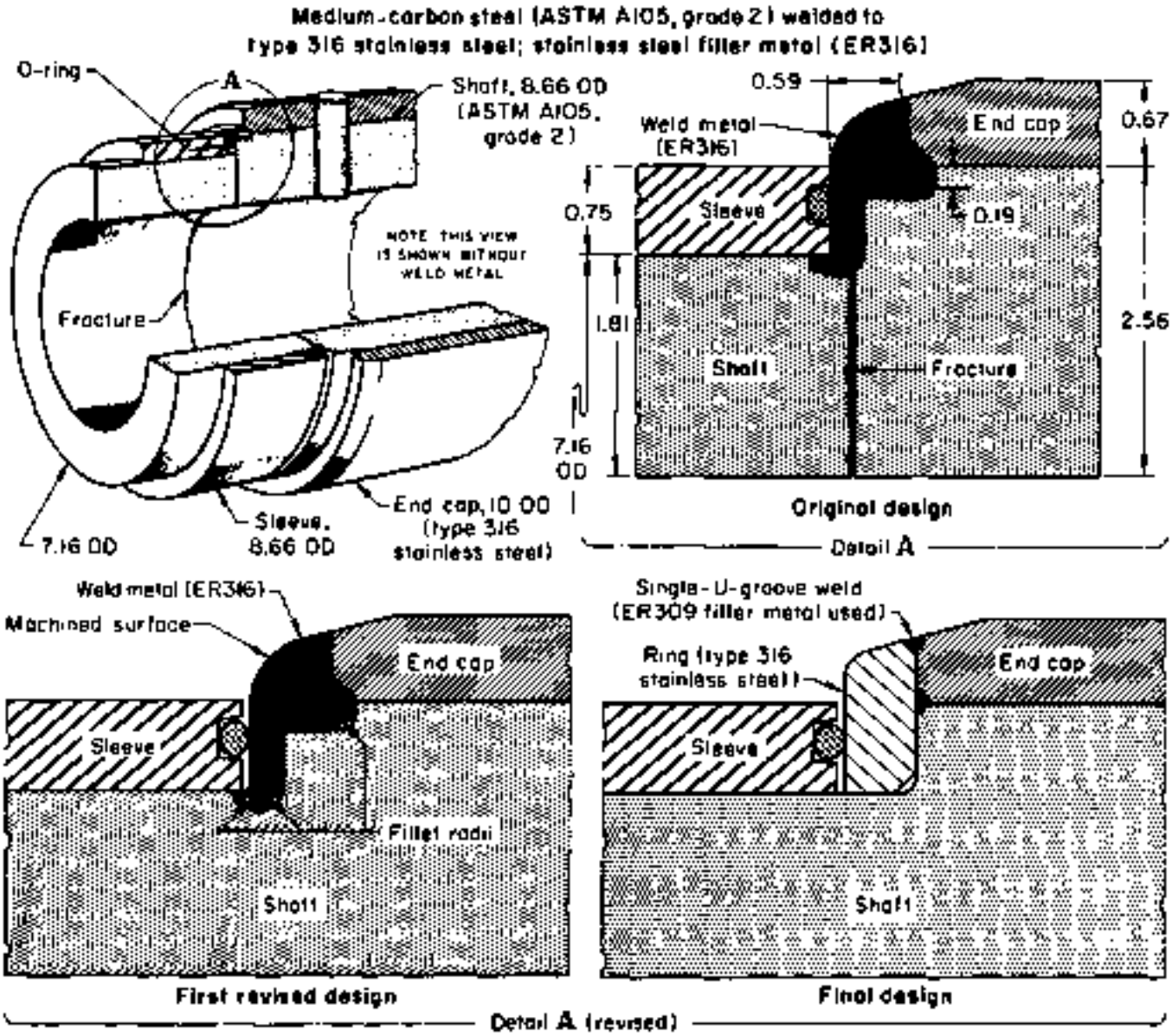
Recommendations. Surfaces should be prepared for metal spraying by shot or grit blasting in such a manner that they are roughened, not simply peened. Metal spraying is generally satisfactory for rebuilding of the surfaces of main-bearing journals, but not for the surfaces of crankpins that are subject to shock loading.

Shaft Repair by Welding. Shafts made of annealed materials can usually be repaired without difficulty using well-established welding practices. Generally, these practices include matching of base-metal and filler-metal compositions, proper preparation of the region to be welded, and thermal procedures that include preheating, postheating, and stress relieving. The repaired shaft is then heat treated in the conventional manner.

Welding of hardened materials involves compromises and should be avoided whenever possible. For instance, the heat-affected zone usually contains areas with hardness 10 to 20 points (HRC) lower than that of the base metal, thereby causing stresses in the shaft due to a so-called metallurgical notch. To minimize these stresses, the material should be slowly cooled after welding, then tempered immediately using a temperature slightly lower than the original tempering temperature.

Weld overlays resulting in a diameter larger than the adjacent shaft diameter should have a tapered transition. This would minimize the stress concentration at an otherwise abrupt change of section that may coincide with the heat-affected zone of a weld. Any subsequent grinding of the weld area to produce the desired dimensions should be done carefully, using a soft grinding wheel and numerous light passes to avoid overheating that would cause grinding cracks.

Example 16: Ductile Fracture of a Forged Steel Shaft at a Change in Section and at a Stainless Steel Weld. The stub-shaft assembly shown in Fig. 29, which was part of the agitator shaft in a PVC (polyvinyl chloride) reactor, fractured in service after a nut that retained a loose sleeve around the smaller-diameter section of the shaft had been tightened several times to reduce leakage. Removal of the stainless steel sleeve revealed that tightening of the retaining nut had forced the end of the sleeve against the machined stainless steel shoulder on the shaft and had permanently flattened the O-ring seal.



Element	Chemical analysis of steels in shaft assembly					Nominal composition of type 316 stainless steel[(a)]
	ASTM A105, grade 2, steel Shaft		Type 316 stainless steel End cap			
	Typical	At fusion line	At fusion line	At center of weld	In end cap	
Carbon	0.456	…	0.55	0.054	0.037	0.08[(b)]
Manganese	0.25	0.18	0.20	0.73	0.53	2.0
Phosphorus	…	…	…	0.010	0.016	0.045
Sulfur	0.017	…	…	0.016	0.022	0.03
Silicon	0.35	0.35	0.25	0.29	0.607	1.00
Nickel	0.10	0.75	10.00	10.16	11.06	10–14
Chromium	0.068	2.00	16.00	19.43	19.46	16–18[(c)]
Molybdenum	0.023	0.65	2.00	2.86	2.88	2–3
Iron	rem	rem	rem	rem	rem	rem

(a) Maximum content unless range is given.
(b) Carbon content is 0.03% max for type 316L stainless steel.
(c) Chromium content is 18 to 20% for type 316 stainless steel filler metal.

Fig. 29 Stub-shaft assembly, for agitator in a polyvinyl chloride reactor, that failed by ductile fracture. Top left: Configuration and dimensions (given in inches). Detail A: Sections through failure area showing original design, first revised design, and final design

The shaft was made of ASTM A105, grade 2 steel, and the larger-diameter section was covered with a type 316 stainless steel end cap. The cap was secured at each end by welding, using type ER316 stainless steel filler metal (Detail A, "Original design," Fig. 29). The radial surface at the change in diameter was a weld overlay, a continuation of the weld metal securing the cap to the shaft. Transverse fracture of the shaft occurred near this abrupt change in shaft diameter.

The stub shaft was removed from the agitator so that a replacement stub shaft could be installed. Both sections of the fractured shaft were sent to the laboratory for analysis.

A new stub shaft was made by welding a type 316 stainless steel cap to a forged steel shaft that had a chemical content of 0.25% C, 0.25% Mo, and 1% Cr and a tensile strength of 600 MPa (87 ksi). The grooves machined in the replacement shaft for the weld overlay had radii instead of sharp corners (Detail A, "Original design" and "First revised design," Fig. 29).

Investigation. Examination revealed that the forged steel shaft had fractured at approximately 90° to the shaft axis (Detail A, "Original design," Fig. 29). The face of the outboard section of the shaft had a groove approximately 6.4 mm (0.25 in.) deep by 7.9 mm (0.3 in.) wide around its circumference. This step followed the shape of the weld metal on the mating piece. The fracture surface along the weld metal contained no forged steel, indicating that the failure was in the weld metal and not in the heat-affected zone of the forged steel shaft. Worn and smeared metal on both mating surfaces indicated that the stub shaft had continued to rotate after it had fractured.

Inspection of the PVC reactor revealed that the drive coupling between the shaft and motor was purposely misaligned 1.3 mm (0.05 in.) in the vertical direction with the shaft. This misalignment was done to counter the misalignment caused by bow of the concrete slab base, which occurred when the reactor was loaded.

Cracks in the heat-affected zone or cracks in the martensite transition zone at the fusion line were possible, but were not evident from microscopic inspection of the weld area. Presence of the martensite transition zone offered a path of easy crack propagation.

Chemical analyses of the steel shaft and the stainless steel end cap at various locations are given in the table that accompanies Fig. 29. Analyses of metal at the fusion line were made from scrapings and drill shavings that included metal on either side of the fusion line; therefore, the values are average values for the general area rather than actual values for the metal at the fusion line.

The chromium, nickel, and molybdenum contents of the weld metal were lower than those of the filler metal, whereas the contents of the same elements in the heat-affected zone of the shaft were higher than those of the base metal, indicating that the weld metal had been diluted by the carbon steel shaft.

The stainless steel used for the end cap was similar to type 316 stainless steel except for a higher chromium content. The carbon content of the forged steel used for the shaft was greater than the maximum specified for ASTM A105, grade 2.

Microstructure. The base metal that was not affected by the heat of welding had a typical pearlite structure, with ferrite surrounding each grain. The microstructure changed as the fusion line was approached: the amount of ferrite decreased, and the microstructure became martensitic.

The hardness of the metal in the steel shaft near the fusion line was 358 HV. Because a maximum hardness of approximately 520 HV can be obtained with a carbon steel of 0.45 to 0.50% C content, the martensite may have been tempered by the welding heat during subsequent weld passes.

Discussion. The sharp comers at the root of the weld, which were shown on the original shaft drawing, were not evident on the shaft, and there was no evidence of a sharp comer on either fracture surface. Therefore, it can be assumed that the fracture was initiated at the surface (the welding practices employed and penetration of the weld metal may have removed evidence of the sharp corners).

Some features on the fracture surface were removed by rotation after fracture, which was unfortunate because the surface may have contained indications of whether fracture originated on the surface in the heat-affected zone or at the fusion line of the weld. An initial crack on the surface of the shaft could have been caused by inherent defects in the weld or by tensile stresses resulting from rotation of the misaligned shaft.

Conclusions. The forged steel shaft failed by ductile fracture. The fracture surface was approximately 90° to the axis of the shaft. The fracture was the result of:

- Reduction of the effective cross-sectional area of the shaft by a crack that followed the fusion line of a circumferential weld that served as an overlay and as a means of securing one end of a stainless steel cap
- A stress concentration from the notch effect of the crack created at the root of the weld
- Tensile stresses induced by misalignment of the shaft

- In addition to the normal torsional and tensile stresses of the shaft, a tensile stress that was applied to this area by the sleeve being forced against the machined surface of the weld when the retaining nut was tightened excessively.

Corrective Measures. The replacement shaft for the immediate failure followed the first revised design shown in Detail A in Fig. 29. Shafts for new agitators were made using the final design in detail A in Fig. 29. Both the end cap and the ring were made of type 316 stainless steel. Welding was done with type ER309 stainless steel filler metal, and the end cap and end ring were preheated to 150 °C (300 °F). The cap was joined to the shaft using a single U-groove weld, which was machined flush before the stainless steel ring was installed. Another single U-groove weld was used to join the ring to the end cap. This design eliminated welding at an inside corner, reducing the stress-concentration factor at that area.
Manganese and zinc phosphate are conversion coatings capable of producing stress concentrations on hardened ferrous material. This condition results from failure to follow proper processing procedures, which include control of bath composition, bath temperature, and processing time. Camshaft lobes and bearings are sometimes coated with manganese phosphate to assist when lubrication is marginal and during break-in. If one or more control parameters are not observed, the surface becomes grossly etched and pitted. This intergranular surface attack leads to flaking in service and ultimately to complete deterioration of the highly stressed lobe surface.

Footnote

* Adapted from "Fatigue Failure of a Metal-Sprayed Crankshaft," *Failure Analysis: The British Engine Technical Reports,* American Society for Metals, 1981, p 131–133

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Introduction

MECHANICAL DEVICES usually contain some moving parts. Bearings are usually provided where a specific spatial relationship (alignment) must be maintained between the parts or where a force is to be transmitted from one part to the other. Mechanical bearings are usually made of selected material and in special geometries in order to effect easy and reliable motion for a specific task. The major types are rolling-element (ball or roller) bearings and sliding bearings. This article will discuss sliding bearings.

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Classification of Sliding Bearings

Sliding bearings are usually conformal; that is, the components of a sliding pair fit together fairly closely. The surfaces may be flat, spherical, or cylindrical. Flat surfaces include the slideways of machine tools and the thrust bearings that limit the axial movement of rotating shafts. Spherical surfaces allow rotary motion around more than one axis, such as that provided by the hip joint of the human skeleton. The cylindrical surface pair (the shaft and journal pair) is the most common type in machines. These are also called plain bearings or simple bearings.
Sliding surfaces may be in motion for long periods of time, as in clocks. Some start and stop frequently, but always slide along one axis, as in engines; others oscillate, as in door hinges. They may operate dry (photo film-

making machinery) or lubricated with available moisture (some water pumps) or a specifically applied lubricant. It is usually the case that the more severe the duty of bearings the more precisely they will be made and the more likely there will be a support (lubrication) system to ensure long life and to prevent catastrophic failure.

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Bearing Materials

The lowest cost bearings are bushings of molded polymers (plastics) in which steel shafts rotate. The lowest cost polymers are the nylons and acetals. At low sliding speeds, they usually wear by abrasion due to dirt or the roughness of the surface of the shaft.

Polymers have two disadvantages—their low thermal conductivity and their tendency to cold flow and become distorted if loaded very heavily. If polymers are used at high speeds, the surfaces may melt, possibly even with some of the molten polymer being squeezed out. There are three solutions to the thermal problems. One is to operate the bearing with a low contact pressure, p, and at a low speed, V. A recommended pV limit is 0.2 MPa · m/s (Ref 1). Another is to lay a thin film of polymer on a metal backing to increase the heat-flow rate. A third is to fill the polymer with a material that has high thermal conductivity. The third solution to the thermal problem also offers resistance to cold flow. For these purposes, the polymer can be filled and reinforced with particles or fibers of glass, metal, and minerals. However, these fillers usually cause more wear of the shaft than the polymer does.

For low friction, the more expensive polytetrafluoroethylene (PTFE) is available. Polytetrafluoroethylene is sometimes added to acetals to reduce friction. Graphite and molybdenum disulfide (MoS_2) are also added; although they decrease friction, these solid lubricants usually do not prevent wear.

Another low-cost bushing material is porous bronze, which can be impregnated with lubricant. For a little higher cost, cast iron bearings can be used, particularly for large and flat surfaces.

Polymers, gray cast iron (with large amounts of graphite), and bronzes (see the section “Bronze Bearing Materials” in this article) are selected as bearing materials because of their ability to survive even though they may be made to poor tolerances or may be operated with occasional lubricant starvation or without maintenance. Poor treatment of these bearings will often appear as adhesive grooving and galling. Where these materials are run against a harder counter-surface, there may be metal transfer to that surface.

Almost all materials, such as carbon, ceramics, white cast iron, and metal alloys of all kinds, can be used as bearings. These materials may be chosen primarily for corrosion resistance, for prevention of sparking, for resistance to high volumes of abrasives, or simply because they are available. High-performance bearings are usually made with considerable care and are comparatively more expensive.

High-Performance Bearings. The soft metals, in various combinations, are used in high-performance bearings, and they are virtually always lubricated. The soft metals are sometimes used in thicknesses up to 6 mm (0.24 in.), but to increase fatigue resistance, the soft metals are coated or plated on a steel (or other stronger metal) backing in layers as thin as 25 μm.

The major groups of soft metal bearing materials are:

- *Tin-base or lead-base babbitts or white metals:* These materials are good for embedding hard contaminant particles and for resistance to galling, but fatigue quickly. Lead alloys also corrode readily
- *Copper-lead alloys:* These are superior to the white metals for corrosion resistance in many applications
- *Aluminum alloys that contain tin and lead:* The alloying elements do not mix, thus forming small solid droplets in the aluminum that smear over the aluminum surface at times of inadequate lubrication. The aluminum alloys are superior in corrosion resistance

Babbitts are of two types: tin-base with 4 to 8% Cu and 4 to 8% Sb (SAE 11 and 12; ASTM B23, grades 1, 2, 3, and 11) and lead-base with a maximum of 10.7% Sn and 9.5 to 17.5% Sb (SAE 13–16; ASTM B23, grades 7, 8, 13, and 15). These babbitts generally cause the least damage to steel shafts when operated with inadequate lubrication or with contaminants. Tin-base babbitt is more corrosion resistant than lead-base babbitt and is easier to bond to a steel backing. Babbitts are inherently weak and strongly affected by high temperatures. For this reason, thin layers of babbitt are often applied to hacking materials, such as bronze, steel, or cast iron.

Copper-lead bearing alloys usually contain between 20 and 30% Pb, up to 3% Sn, and may contain up to 5.5% Ag and less than 0.1% Zn. Typical alloys are SAE 48, 49, 480, and 481 (CDA alloys C98200, C98400, C98600, and C98800, respectively). These alloys have all been used as main bearings and connecting-rod bearings of engines. However, in recent years, corrosion has been found to be a problem with these alloys. They are, therefore, usually covered with a thin layer of lead-tin, lead-tin-copper, or lead-indium alloy to reduce the corrosion of the substrate. In some instances, a nickel barrier about 1 μm thick may be placed between the substrate and the overlay to prevent diffusion of tin from the overlay to the copper in the substrate.

Bronze bearing materials are either leaded bronze, tin bronze, or aluminum bronze. Strength and high-temperature resistance increase in the order lead to tin to aluminum, but embeddability and scuff resistance decrease in that order.

The leaded bronzes contain from 5 to 25% Pb and up to 10% Sn. The high-lead alloy is soft and has a temperature limit of about 230 to 260 °C (445 to 500 °F). The addition of tin increases the fatigue resistance and hardness. Zinc in place of tin improves ductility but decreases strength and hardness. Some wrought leaded bearing bronzes are SAE alloys 791 to 794, 797, and 799. The alloy SAE 795 has especially good fatigue resistance. Some castable leaded bearing bronzes are SAE 40 (CDA C83600), SAE 660 (CDA C93200), SAE 66 (CDA C93500), SAE 64 (CDA C93700), CDA C94100, and CDA C94300.

Tin bronzes contain more than 20% Sn, with less than 10% Pb added to aid machinability. Small amounts of zinc, nickel, or both are sometimes added to increase strength.

Aluminum bronzes have iron, nickel, silicon, and manganese as the principal alloying elements. The hardness of these alloys may range to 240 HB. They have excellent shock resistance and will operate at comparatively high temperatures. They tend to scuff more readily than the leaded bronzes. Typical alloys are CDA C86200, C86300, C95200, C95300, C95400, and C95500.

Aluminum alloys other than bronzes contain about 7% Sn and 1% Cu, with 1 or 2% Si or Mg, plus some lead and cadmium. The copper is retained in solid solution in the aluminum, but the other alloys form soft-solid particles in the matrix. These alloys are usually relatively weak and require strong backing. They do not require an overlay for protection from corrosion; however, an overlay is often provided for start-up protection. Typical alloys are 850.0-T5, A850.0-T5, and B850.0-T5 (formerly 750-T5, A750-T5, and B750-T5).

Reference cited in this section

1. G.C. Pratt, Materials for Plain Bearings, *Int. Met. Rev.,* Vol 18 (No. 17), 1973

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Fluid-Film Lubrication

Fluid films can be provided in a bearing by surface-tension retention of a fluid in a gap, by pumping fluid into a contact region (hydrostatic lubrication), or by hydrodynamic action.

Surface Tension. If a drop of liquid is placed on a flat surface and then another flat surface is laid upon the wetted surface, some, but not all, of the liquid will be squeezed out. Surface tension, the same force that makes liquid rise in a very small diameter glass tube, makes complete exclusion of the liquid impossible. The amount that will be retained in the gap between two surfaces is related to the wettability of the liquid (lubricant) on the surface of interest. If a drop of the lubricant spreads out completely and spontaneously on the surface, that lubricant will also run out of the bearing

completely. If the drop of lubricant stands up as water does on a waxed surface, retention will be low. If the drop has a base diameter of about twice the height of the drop, retention will be satisfactory. The drop-spreading test is convenient for selecting materials and lubricants for applications in which small quantities of lubricant are applied throughout the life of the product. This method is frequently used for flat surfaces, rolling-element bearings, low-speed bushings, and the myriad of surfaces that are lubricated by the infrequent drop of oil.

Hydrostatic Lubrication. Two sliding surfaces can be separated by pumping a fluid into the contact region at a sufficient pressure. The contact region must be properly designed so that there is a symmetric or balanced contact-pressure distribution. A large volume of fluid will separate the sliding surfaces a great distance, thereby producing a low resistance to movement. However, the energy required to pump the fluid must be considered in the overall economy of the bearing system. Another factor to consider is that hydrostatic lubrication requires a pump, which may of course fail and ruin a bearing. The principles of hydrostatic and hydrodynamic bearings are detailed in Ref 2, 3, and 4.

Hydrodynamic Lubrication. If one surface slides along another at a moderately high speed and if the shape of the leading edge of the moving surface is such that fluid can be gathered under the sliding surface, the two surfaces will be separated and will slide easily. This is hydrodynamic lubrication.

Figure 1(a) shows the start-up of a shaft in a bearing. A shaft at rest, with a downward load on it, will contact the bearing at the bottom. As the shaft begins rotating, it climbs one side of the bearing, but it eventually begins to slide. If the shaft and bearing are immersed in oil, the sliding shaft will drag oil underneath itself to begin forming the hydrodynamic wedge. It is not a visible wedge because the entire system is immersed. Rather, it is a pressurized region in which the pressure lifts the shaft. When the wedge of fluid is fully developed, the shaft takes the position shown in Fig. 1(b), with a minimum separation *h.* The fluid-film pressure builds behind the location of minimum separation between the shaft and bearing, taking the shaft surface as the reference.

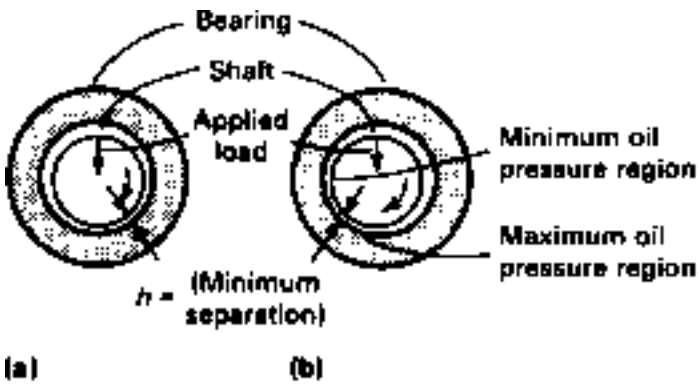


Fig. 1 Schematic of the transition from inadequate (boundary) lubrication at start-up to hydrodynamic lubrication at full speed. (a) Steady load at start-up. (b) Steady load at full speed

The mathematics of the fluid wedge are embodied in equations of hydrodynamics. Some of the variables of interest for a shaft of diameter *D* rotating in a journal bearing of inner diameter *D* + 2*c* are:

$$\frac{hN}{p}\left(\frac{D}{c}\right)^2\left(\frac{L}{D}\right)^2 = \frac{(1-\epsilon^2)^2}{\epsilon p\sqrt{(1+0.62\epsilon^2)}} \qquad \text{(Eq 1)}$$

where η is the lubricant viscosity (which will change with temperature and pressure), *N* is the shaft speed (in rpm), $p = W/DL$ or the average pressure between the shaft and bearing due to the applied load *W*, *L* is the length of contact between the shaft and bearing, *c* is the clearance, and $\epsilon = 1 - h/c$ (*h* is the minimum separation). The term on the left is one form of the Sommerfeld number and is sometimes referred to as the bearing characteristic. It is clear from Eq 1, at least, that bearings with the same Sommerfeld number will operate with the same ϵ .

Many calculations have followed upon the work of Sommerfeld in attempts to optimize bearing design. The result is a number of guidelines, the most important of which relate to efficiency and stability. For efficiency, Ref 2 recommends operating with the ratio $h/c \approx 0.3$. The consequence of this recommendation would be a particular set of values for the adjustable variables $\eta N/p$ for a given bearing. These values can be readily calculated, and this is done in design. The procedure is rather detailed and requires some expertise in dimensions and units (see Ref 2, 3, 4). It is sufficient for this discussion to suggest how to recognize the likely regimes of operation of bearings from the damage done to them.

Miscellaneous Methods of Lubrication. Low-speed bearings do not need large quantities of lubricant flowing through to remove heat. Therefore, some are lubricated by a drip-feed device, some by a wick that carries lubricant from a reservoir to the shaft surface, and some use grease. One design for a horizontal shaft uses a large ring that rests on top of the shaft, dips into a supply of lubricant, and carries lubricant to the shaft as the shaft rotates.

References cited in this section

2. F.T. Barwell, *Bearing Systems*, Oxford Press, 1979

3. A. Cameron, *The Principles of Lubrication*, John Wiley & Sons, 1966

4. D.F. Wilcock and E.R. Booser, *Bearing Design and Application*, McGraw-Hill, 1957

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Shaft Whirl

Equation 1 was described in the context of a shaft that was carrying a constant downward load. It could also apply to a weightless shaft that is out of balance. In this case, the contact is restricted to one location on the shaft, but rotates on the inner surface of the bearing. The fluid wedge develops as in the former case except that it does so ahead of the location of minimum separation between the shaft and bearing.

An interesting situation develops when a shaft has a small load on it and when an intermittent load is also imposed such that the position of the shaft in the bearing changes. The shaft will move around in the bearing as if it were carrying an unbalanced load, which makes the location of minimum separation move in the direction of shaft motion. This amounts to a receding of the wedge-forming surface, which decreases h momentarily until the wedge ahead of minimum separation develops. However, because this film cannot be sustained unless there is a constant out-of-balance load, the film soon diminishes as well. The result is a whirl of the shaft in the direction of shaft rotation but at half the shaft speed. The value of h is much smaller than for either a constant load or a constant out-of-balance load. Whirl can be prevented by reducing, $\eta N/p$ or by reducing the load-carrying capacity of the bearing.

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Load-Carrying Capacity, Surface Roughness, Contaminants, and Grooves

The maximum load-carrying capacity of a bearing, according to Eq 1, would appear to be when $h = 0$, but this is not practical. First of all, the above equations are based on perfectly smooth sliding surfaces. Second, in practice some account should be taken of the inevitable dirt, wear debris, and contaminants in lubricated systems.

One practical limit on the minimum value of h is the point at which the asperities or small bumps on all sliding surfaces begin to interact. This is usually expressed in terms of the lambda ratio, $\Lambda = h/\sqrt{s_1^2 + s_2^2}$. In this equation, the root mean square (rms) surface roughness of the shaft is σ_1, and that of the bearing surface is σ_2. Where $\Lambda > 1$, there should be no contact and thus no wear, but such bearings are not economical for uses other than in large capital machinery, such as electric generators. Most other bearings usually operate with Λ near 1 and even less. The consequence of operating with $\Lambda < 1$, as in the case of overload or where the viscosity of the lubricant is low because of overheating, is that the surface topographical features (bumps) begin to interact through a very thin local fluid film. This interaction is popularly referred to as metal-metal contact, which is not accurate. In any case, this interaction causes deformation and fracture of the beating surface on both micro- and macroscopic scales.

If a bearing is operating in the thick fluid-film regime and if $\eta N/p$ decreases, h will increase. Eventually, as Λ decreases, the friction increases. A widely used method of representing this transition is the plot of the coefficient of friction, f,

versus η*N*/*p* (Fig. 2). This type of curve is often referred to (although not strictly accurately) as the Stribeck or McKee-Petroff curve.

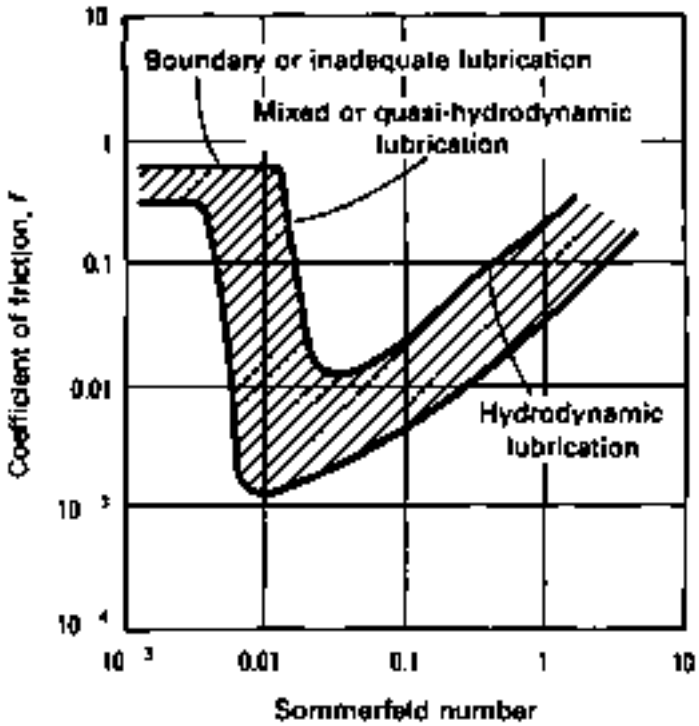


Fig. 2 Approximate relationship of coefficient of friction to Sommerfeld number, η*N*/*p*.

The curve implies that the minimum *f* and the transition to inadequate hydrodynamic lubrication occur at one critical Sommerfeld number, which is not the case. Any single curve is for a single value of Λ as well as for particular values of *D, c,* and *L.* However, the curve is illustrative and is widely used. The type of damage seen in bearings is related to these factors. A whirling shaft may be operating to the left of the transition in Fig. 2, although the design parameters would indicate adequate fluid-film lubrication.

Grooves and depressions are often provided in bearing surfaces to supply or feed lubricant to the load-carrying regions. It is often stated that such grooves are reservoirs for lubricant, but they can also be pathways for lubricant to escape. They function best when lubricant is pumped into them at the pressure prevailing at their location in the bearing. In a hydrodynamic bearing carrying a steady load, the grooves should be located in the half of the beating behind the location of the minimum fluid-film thickness. Grooves may be circumferential, axial, or spiral in orientation. If they are supplied with lubricant from a pump, they should not be open ended because pressure cannot be established in such grooves. The circumferential groove in the center of the bearing is best for the highest performance beating supplied with large quantities of lubricant. Spiral grooves are often preferred where the supply volume is low, but axial grooves are easier to make.

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

---

Squeeze-Film Lubrication

The above discussion relates to a near steady state of bearing operation. In some instances, a shock load may be imposed on a bearing in which an equilibrium fluid film was established. The value of *h* cannot decrease instantaneously: Time is required to decrease the fluid-film thickness. Thus, a bearing will operate for a few microseconds in a state that is sometimes referred to as the squeeze-film regime. The fluid is squeezed out at a rate that is proportional to $1/h^2$. There are few guidelines for calculating the effects of such shock loads.

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Elasto-hydrodynamics

There is another class of surfaces for which adequacy of a fluid film is calculated by equations of elasto-hydrodynamics. These are cases in which the contacting surfaces are of counterformal shape, as, for example, with a ball on a bearing race, and the contact surfaces are deformed elastically an amount that is a significant fraction of the fluid-film thickness. The principles of lubrication are the same as for the hydrodynamic case, but the governing equations are much more complicated

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Lubricants

In hydrodynamic theory, only the viscosity of the lubricant is considered. Actually, most lubricants are chemically complex. They often react with the sliding surfaces to some extent, and they change in use. For example, the pH of a lubricant usually changes during use, which changes both the chemical reactivity and the viscosity. Some lubricants have fatty acids added in order to achieve desirable friction properties. Others have additives that are chemically reactive under extreme pressure to prevent scuffing. These chemical reactions are usually carefully controlled to minimize detrimental effects.

Each type of additive in oil produces selected effects that are too numerous to tabulate here. One effect is to alter friction. Figure 3 shows schematically the effect of three general classes of oils: without additive, with fatty acid added, and with additives of the general types known as antiwear and extreme-pressure (EP) additives. Each is responsive to temperature in a manner different from the others. The temperature increase may be effected by general heating or by localized friction effects.

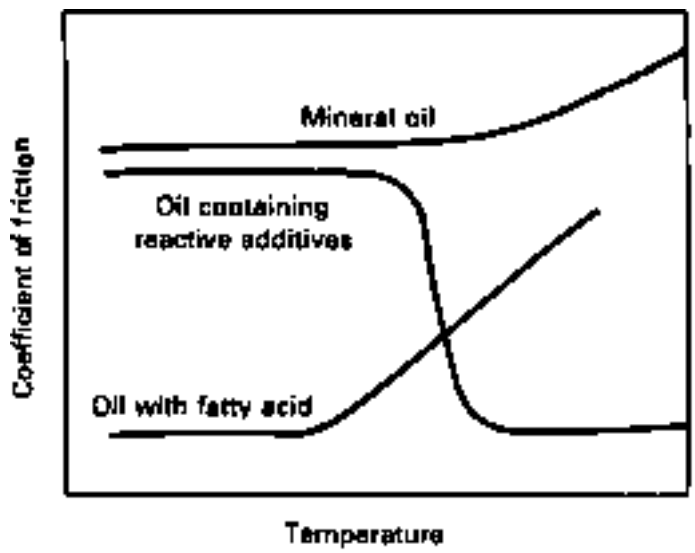


Fig. 3 The friction of sliding bearing pairs for three different lubricants.

There are some graphic representations of the effects of additives to be found in the literature, but there is no method of determining how any one additive will work except to test it. It is thought that the fatty acids attach to a sliding surface without reacting chemically with the surface. The antiwear and EP additives are thought to form a thin film of reaction products that has low friction.

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Debris and Contaminant Particles

Virtually all bearings contain particles that were not deliberately placed there. Some of the particles are left from the manufacturing process; some are dirt; some are the products of chemical reaction, such as oxidation; and some are wear-debris particles. Hydrostatic bearings are flushed continuously by the flow required to carry a load; hydrodynamic bearings are often flushed by a flow of lubricant to remove heat, which develops mostly in the fluid film itself due to viscous shear. It is convenient, then, to supply filtered lubricant to keep the contact areas reasonably clean.

The effect of contaminants in bearings depends on the hardness of the contaminant, the number of such particles, and their size. In general, most dirt, usually SiO_2 and related minerals, is harder than most shafts and bearings and will wear them away. One solution is to make the bearing very soft so that particularly offensive contaminant particles will become embedded in the metal and not cause damage. The alternative is to make the surfaces so hard that they resist abrasion. In general, a particle will abrade another surface if it is 1.3 times (or one Mohs number) harder than the other material.

Contaminant particles have been found to be of a size range from less than one micron to tens of microns. For perspective, hydrodynamic bearings are usually designed with a clearance that is between 0.1 and 0.2% of the shaft diameter. Thus, for a 50-mm (2-in.) diam shaft, the clearance will be between 50 and 100 μm. The design value of minimum film, h, will be between about 15 and 30 μm at normal load and down to 5 μm when under very heavy load. By using automated particle count devices, it has been found that used engine oil may contain, for example, 31 particles in the size range of 30 to 40 μm, 850 particles in the size range of 15 to 20 μm, and 58,469 particles in the size range of 5 to 10 μm. Under normal conditions, the small particles flow through the bearing with no effect. The larger ones can be filtered out without difficulty, but the midrange particles may not be adequately removed. As a rule, those particles of about the dimension of h will cause the most wear.

A surface can be examined to determine whether contaminants may be a major source of wear. This is done by comparing the surface roughness of the worn surface with that of a new surface. For example, a crankpin surface may begin with a roughness of 0.25 μm rms. If the crankpin is worn but has the same roughness as before, it has probably been worn by large contaminant particles of the order of 30 μm. If the crankpin is worn but has a roughness of 0.05 μm rms, it has been worn with contaminants of about 6 μm in diameter.

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Failure Analysis Procedure

A general procedure for failure analysis of sliding bearings involves the following steps:

- Do not clean the machine or bearing before the first examination
- Obtain a history of operation of the bearing, lubricants used, other failures in the system, and histories of bearings in similar machines. It is often helpful to get comments on machine behavior from technicians and mechanics instead of from engineers and managers
- Examine the bearing in its normal operating position, and look for evidence of abnormal environment and surrounding hearing and damage. Photograph the machine, both close up and overall

- Examine all parts, wear debris, decomposed lubricant, and so on, during disassembly. Retain samples of debris and lubricant, and retain both halves of a sliding pair
- Use a low-power (binocular, if possible) optical microscope for observing the uncleaned parts, and try to establish the progression of damage
- Use a higher-power optical microscope if something interesting appears and it is not yet appropriate to clean the parts thoroughly
- If the parts can be cleaned, it is usually most convenient to use the scanning electron microscope, even for magnifications down to 20 ×, because the greater depth of field of the scanning electron microscope is far less likely to obscure details on a rough surface than an optical microscope at the same magnification
- X-ray microanalysis is a convenient tool for detecting material transfer, the composition of embedded particles, and loss of layers of multilayer bearings
- Look for evidence of fretting, poor bonding between layers in the bearing, oxidation, plastic flow, fatigue, high temperature, scuffing, and embedded particles (often found in regions of low pressure in the bearing)
- Analyze the lubricant for debris particles and by spectrographic methods to measure the concentration of dissolved elements, and compare the results with a normal lubricant. The lubricant supplier may have access to the characteristics of such normal lubricant. Determine the possible sources of high amounts of dissolved elements

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Bearing Failures

Wear Mechanisms in General. There is a limited number of modes or mechanisms by which small amounts of material can be removed from a larger body. These are ductile fracture, brittle fracture, high-cycle (ordinary) fatigue, low-cycle (plastic) fatigue, melting, and chemical dissolution. Hydrogen embrittlement is an example of a hybrid mechanism in which ordinarily ductile material may fracture in a brittle manner. It is often helpful to try to classify the type of observed wear in terms of the modes of material loss, because it is in these terms that materials are selected to resist wear. However, the emphasis in this article is on the analysis of bearings that have failed or are failing so that solutions may be found for the problem.

Fatigue. Segments of a long-lived bearing may fail by fatigue. A bearing may have many cycles of varying load on it, as in engines, or may have an unbalanced load, as in other machinery. Fatigue failure will be hastened by high stresses and by a corrosive environment.

Fatigue cracking of bimetal bearings occurs in stages, beginning with a hen-track pattern on the bearing surface. As fatigue progresses, the soft metal separates from the hard backing in a pattern shown in Fig. 4. This may occur even though lubrication is adequate, but operation with very thin fluid films could cause heating and weakening of the bearing material.

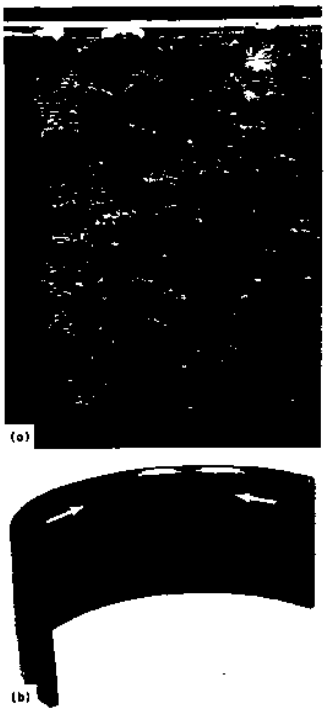


Fig. 4 Sliding bearing surfaces that failed by fatigue. (a) Wormhole appearance in a lead-base babbitt bearing. (b) Flaking (between arrows) of overlay in a trimetal bearing

The fatigue limit of engine bearings is a major topic. The performance of such bearings in service is very difficult to quantify, but there are practical guidelines for selecting material types based on applied stress. Table 1 lists such information.

Table 1 Peak stresses for normal fatigue life of various bearing materials used in automotive applications[(a)]

Bearing metal	Peak stress (max)	
	MPa	psi
Solid-metal bearings		
Aluminum alloy SAE 770-T101 4.1 mm (0.162 in.) thick	34.5	5000
Bimetal bearings		
Tin babbitt 0.6 mm (0.022 in.) thick on steel	8.3	1200
Lead babbitt 0.6 mm (0.022 in.) thick on steel	9.7	1400
Tin babbitt 0.1 mm (0.004 in.) thick on steel	15.2	2200
Lead babbitt 0.1 mm (0.004 in.) thick on steel	16.5	2400
Copper-lead 0.6 mm (0.022 in.) thick on steel	34.5	5000
Aluminum alloy SAE 780 0.8 mm (0.03 in.) thick on steel	41.4	6000
Trimetal bearings		
Lead-tin 0.025 mm (0.001 in.) thick on copper-lead 0.6 mm (0.022 in.) thick on steel	51.7	7500
Lead-tin 0.025 mm (0.001 in.) thick on silver 0.3 mm (0.013 in.) thick on steel	51.7	7500
Lead-tin 0,025 mm (0.001 in.) thick on aluminum alloy SAE 780 0.8 mm (0.03 in.) thick on steel	55.2	8000
Lead-tin-copper 0.4 mm (0.015 in.) thick on aluminum alloy SAE 781 0.4 mm (0.015 in.) thick on steel	69	10,000

(a) Fatigue strength of bearing materials was determined under laboratory conditions of excellent alignment, excellent oil flow, a high degree of cleanness, ideal clearance, and normal operating temperature of 80 to 105 °C (180 to 220 °F). Normal fatigue life of bearings is considered to be 161,000 km (100,000 miles) in gasoline automobile engines and 402,000 km (250,000 miles) in diesel truck engines.

Electrical Erosion. Some bearings, particularly those in steam turbines, may have an electrical potential applied across them, usually inadvertently. When the gap spacing in the bearing is small and the applied potential is high enough, sparks will pass through the oil film and cause erosion. The bearing surfaces will not be reflective, but will appear frosted (Fig. 5).

Fig. 5 Electrical wear on the surface of a thrust bearing from a steam turbine. The microscopic pitting is caused by sparking from electric current. Actual size

Fretting. When oscillatory sliding occurs with an amplitude of the order of 0.05 μm, fretting occurs. This may happen on the outside of the bearing backing, against the housing of the machine. The bearings may not have fit tightly, or the housing of the machine may have been too flexible, or the bearing may have been assembled with some dirt in it. A fretted surface is shown in Fig. 6.

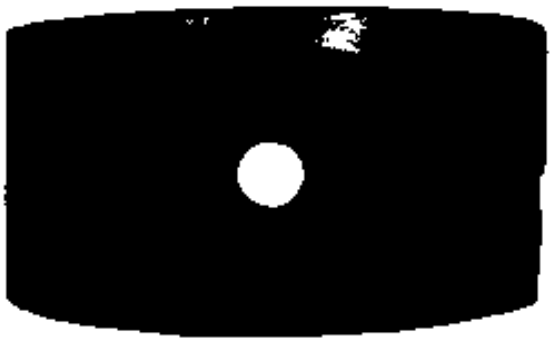


Fig. 6 Fretting on the outside surface of steel backing of a sliding bearing.

Corrosion. Several products of degradation of oil (alcohols, aldehydes, ketones, and acids), contaminants (such as water), and some antiwear additives in oil cause corrosion. A corrosive environment may selectively attack an alloy, removing only one element. In some environments, lead and zinc may be more readily dissolved than copper or aluminum, for example, leaving a surface such as that shown in Fig. 7. The remaining copper matrix is unsupported and can readily fatigue and fracture away.

Fig. 7 Micrograph of a section through a copper-lead alloy bearing that failed by deleading. Light area at the upper surface is the copper matrix that remained after the alloy was depleted of lead. As-polished. 100×

Example 1: Corrosion of Copper-Lead Alloy Sliding Bearings by Sulfur Compounds. A connecting-rod shell bearing from a six-cylinder gasoline engine was returned to the factory for examination. The bearing was made of copper-lead alloy SAE 485 bonded to a low-carbon steel backing. The recommended lubrication oil was used in the engine. The length and type of service was not reported.

Investigation. Material loss was measurable over most of the bearing halves, but the steel backing of the upper bearing half in particular was exposed in a wide region at the centerline. No dirt or other foreign particles were found. Shallow pits and a darkened region were visible, and both were partly covered with a brittle waxlike substance that was identified to be a mixture of copper and lead sulfides.

Conclusion. The sulfides were probably formed by the action of sulfuric acids in the oil on the copper and lead in the bearing. The sulfides are soft and easily removed from a surface, and this removal constitutes a wearing away of the bearing.

Recommendations. The best solution to the problem is to change oil with greater frequency to prevent the buildup of sulfur compounds. If the problem persists, bearing halves that have corrosion-resistant overlay materials should be used.

Corrosion Fatigue. Under some conditions of corrosion, a material that would not fail without the corrosion can fail by fatigue.

Example 2: Corrosion-Fatigue Failure of a Bearing in a Marine Environment. A support bearing of a hydrofoil vessel failed after only 220 h of operation. The bearing (Fig. 8) consisted of an outer ring made of chromium-plated AISI type 416 stainless steel and an inner ring with a spherical outer surface made of AISI type 440C stainless steel, with a plastic material between the two. The plastic material was bonded to the outer ring.

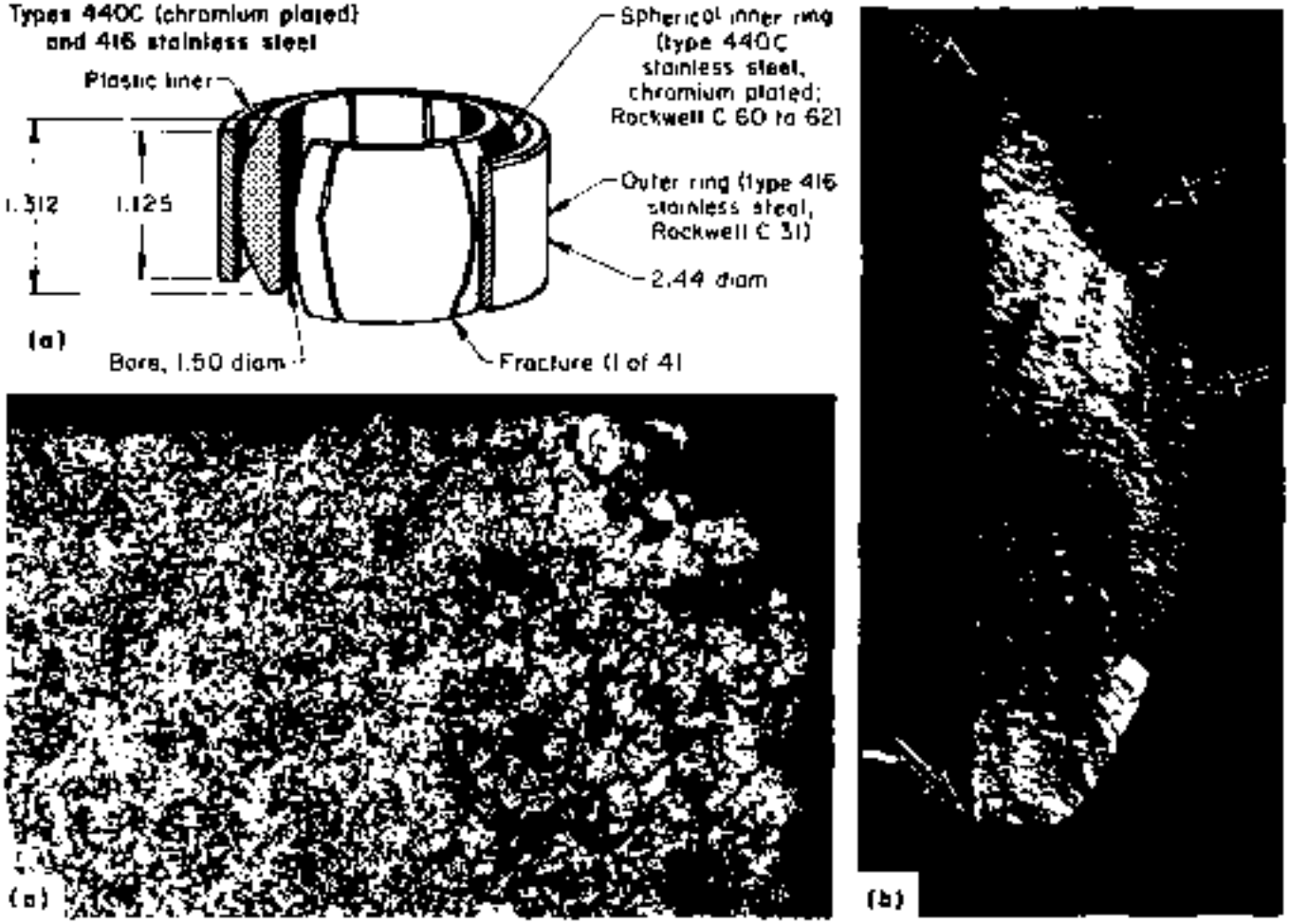


Fig. 8 Plastic-lined stainless steel spherical bearing for a hydrofoil that failed by corrosion fatigue. (a) Construction of bearing and location of fractures. Dimensions given in inches. (b) Fracture surface showing multiple fatigue origins (arrows) at edge of bore and on the spherical surface. 2.5×. (c) Picral-etched section through one of the fatigue origins at the edge of the bore showing intergranular corrosion (arrow A) and intergranular attack (arrow B) on the fracture surface. 250×

Investigation. The inner ring had failed in four places, as shown, and several less serious cracks had begun as well. The plastic material had disappeared, allowing the two metallic rings to come in contact with each other. Examination of the fracture surfaces of the inner ring showed that it had failed by fatigue and that the fatigue had initiated in corrosion pits. The fracture was mostly transgranular, and it was corroded by seawater.

Conclusions. Corrosion fatigue was the cause of failure. Corrosion pitting of the spherical inner ring produced local stress concentrations and formed cracks under repeated stressing. The cyclic stressing and corrosion in the cracks led to the failure on the inner ring.

Recommendation. The inner and outer rings should both be made from 17-4 PH (AISI type 630) stainless steel, which is more resistant to corrosive marine environments than either type 440C or type 416 stainless steel.

Cavitation. In certain areas of fluid flow along a solid boundary, high pressures and low pressures can develop. If the negative pressure is high, cavities can form in the fluid. If the pressure then increases as the cavity travels to an area of higher pressure, the cavities collapse, and the surrounding fluid impinges on the solid surface. This may produce large stresses in the solid, and if the stress is repeated, the solid may fail by fatigue. Cavitation often occurs in the unloaded region of a bearing or on the surface of high-speed gear teeth. Its appearance is shown in Fig. 9.

Fig. 9 Typical cavitation damage on the surface of a sliding bearing.

Cavitation may be reduced by using fatigue-resistant metals, by increasing the lubricant viscosity, and by increasing the pressure in the system. Entrained air or water in oil can increase the likelihood of cavitation. More information of this phenomenon is available in the article “Liquid-Impact Erosion ” in this Volume.

Foreign Particles. As discussed earlier, dirt, chips, and abrasive grit cause wear by abrasion. This is a problem in all applications of sliding bearings. The foreign matter in office machines is often eraser residue, which is composed of 50% glass beads by weight and clay from high-quality paper. Some of this material can be kept out with felt wipers and covers, but it is sometimes better to make the bearings of hardened, sintered iron rather than bronze. One insidious effect of foreign particles is that they may scratch or score the shaft surface rather severely, which adds more foreign particles to the system (Fig. 10).

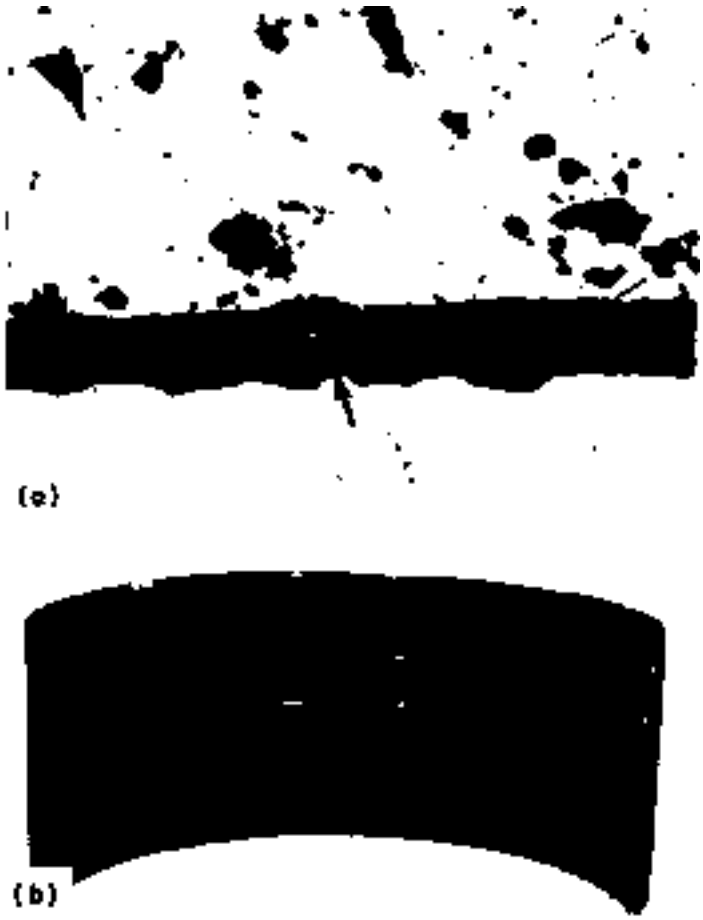


Fig. 10 Sliding bearings damaged by abrasion by foreign particles. (a) Unetched specimen showing a large foreign particle (arrow) at the interface of a bearing and shaft. 110×. (b) Bearing surface scored by metal chips

Misalignment. Parts are sometimes improperly manufactured, as in the case of a burr left from drilling (Fig. 11). Another example of improper manufacture is when the shaft is not cylindrical (Fig. 12). Some bearings are improperly assembled and fracture very early in use. Some bearing-holder systems deflect excessively, as shown in Fig. 13. The result of misalignment is a distorted wear pattern, as shown in Fig. 14.

Fig. 11 Burrs around a hole drilled through the wall of a spherical bearing at a lubrication groove. The burrs contacted the mating surface, resulting in rough operation of the bearing.

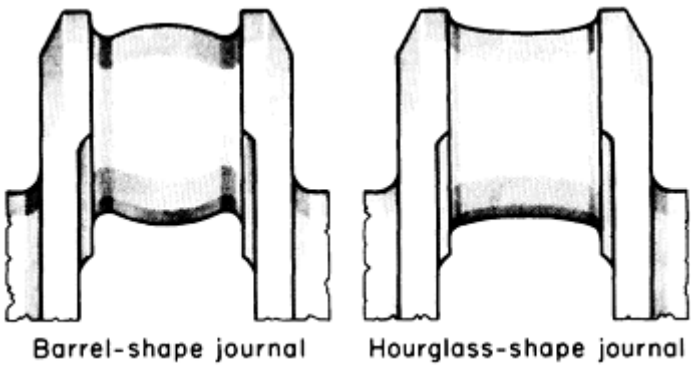


Fig. 12 Barrel-shape and hourglass-shape journals. These types of journals can penetrate the oil film and result in bearing failure because of increased loading in localized areas.

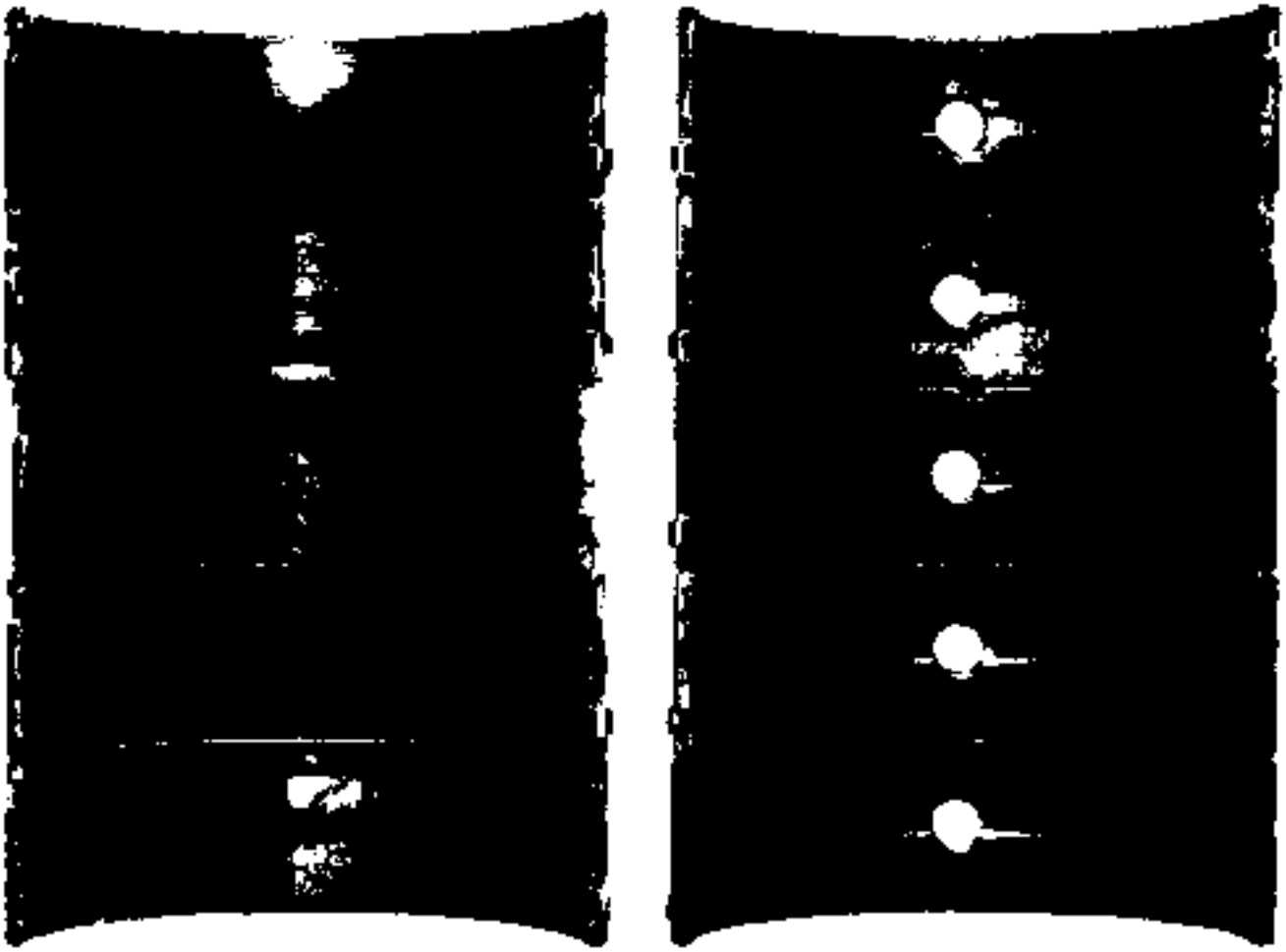


Fig. 13 Lower (left) and upper (right) halves of trimetal bearings for a diesel engine that failed because of distortion of the crankcase. The center three bearing halves sustained greater damage than the end halves.

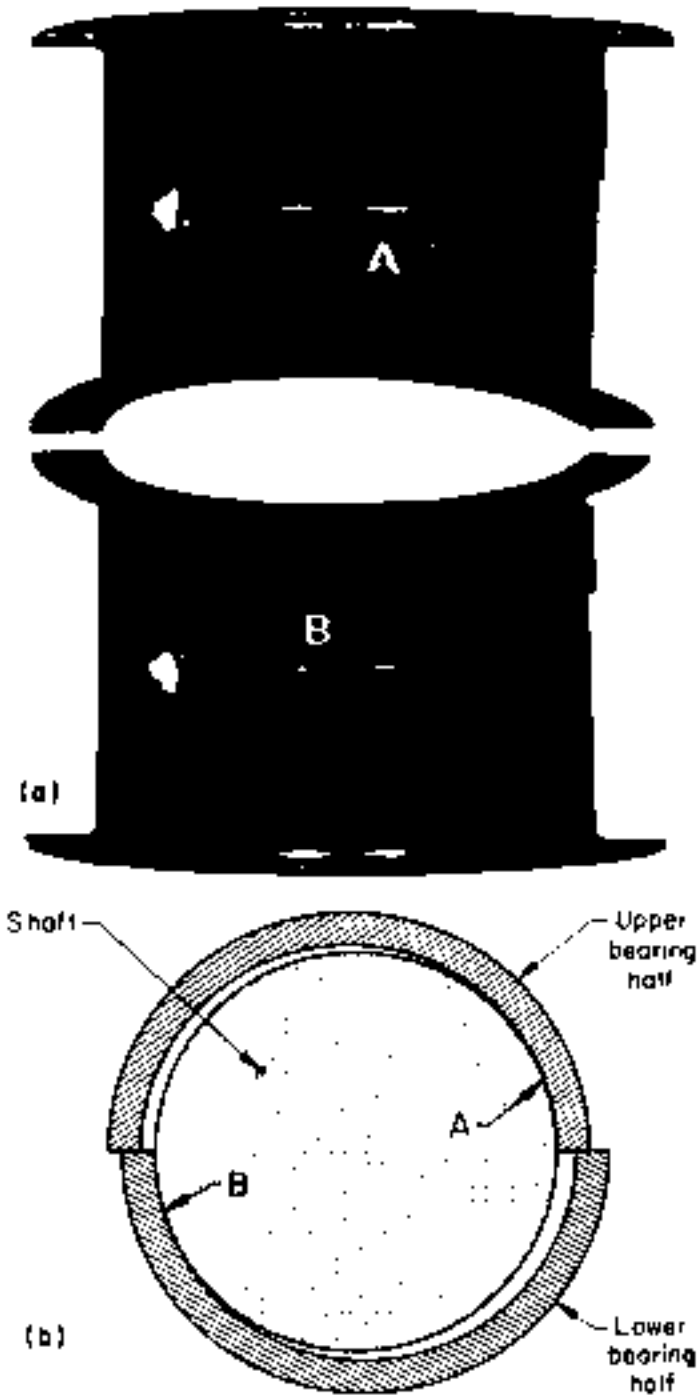


Fig. 14 Bearing halves (a) that failed by fatigue resulting from localized overloading after bearing cap shifted position. (b) Note the damaged areas (arrows A and B) on diametrically opposite sides of the bearing halves where they contacted the shaft after the cap shifted.

Long bearings are more likely to be misaligned than short bearings. Sharp edge bearings are likely to fail on the edge, particularly when there is a heavy overhanging load on the shaft.

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

References

1. G.C. Pratt, Materials for Plain Bearings, *Int. Met. Rev.*, Vol 18 (No. 17), 1973
2. F.T. Barwell, *Bearing Systems*, Oxford Press, 1979
3. A. Cameron, *The Principles of Lubrication*, John Wiley & Sons, 1966
4. D.F. Wilcock and E.R. Booser, *Bearing Design and Application*, McGraw-Hill, 1957

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Introduction

ROLLING-ELEMENT BEARINGS use rolling elements (either balls or rollers) interposed between two raceways, and relative motion is permitted by the rotation of these elements. Bearing raceways that conform closely to the shape of the rolling elements are normally used to house the rolling elements.The rolling elements are usually positioned within the bearing by a retainer, cage, or separator; in ball bearings of the filling-slot type and in needle bearings, they occupy the available space, locating themselves by contact with each other.

Ball bearings can be divided into three categories: radial contact, angular contact, and thrust. Radial-contact ball bearings are designed for applications in which loading is primarily radial with only low axial (thrust) loads. Angular-contact bearings are used in applications that involve combinations of radial loads and high axial loads and require precise axial positioning of shafts. Thrust bearings are used primarily in applications involving axial loads.

Roller bearings have higher load capacities than ball bearings for a given envelope size and are usually used in moderate-speed heavy-duty applications. However, in recent years, improved materials and special designs have allowed use of cylindrical and tapered-roller bearings in high-speed applications. The principal types of roller bearings are cylindrical, needle, tapered, and spherical.

Cylindrical-roller bearings have rollers with approximate length-to-diameter ratios of 1:1 to 3:1. Needle-roller bearings have cylindrical rollers (needles) with greater length-to-diameter ratios (approximately 4:1 to 8:1). The rolling elements of tapered-roller bearings are truncated cones. Spherical-roller bearings are available with both barrel- and hourglass-shape rollers. Figure 1 shows the principal components of ball and roller bearings.

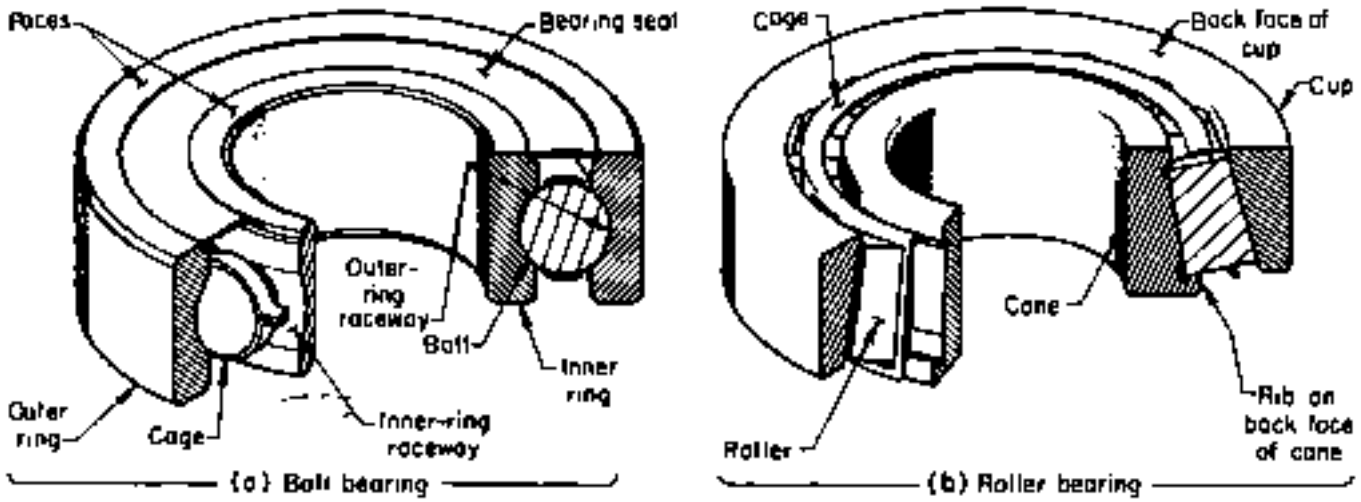


Fig. 1 Principal components of rolling-element bearings.

Bearing Materials

A significant improvement in air-melted bearing steels has occurred in recent years. Bearing steels are currently made by electric-furnace melting, ladle refining, and casting by either bottom or top pouring of ingots or continuous casting. The air-melted steel quality is approaching that of electroslag remelting (ESR) or vacuum arc remelting (VAR). However, steel from these secondary refining processes is still required for bearings in high-reliability applications, such as aircraft engines.

The bearing industry has used through-hardened material, such as 52100, for ball bearings since early introduction in Europe and carburized materials, such as 8620, for roller bearings since the turn of the century in the United States. A commonly accepted minimum surface hardness for most bearing components is 58 HRC. The carburizing grades have a core hardness range of 25 to 45 HRC.

At surface-hardness values below the minimum 58 HRC, resistance to pitting fatigue is reduced, and the possibility of brinelling (denting) of bearing raceways is increased. Because hardness decreases with increasing operating temperature, the conventional materials for ball and roller bearings can be used only to temperatures of approximately 150 °C (300 °F). Although ball bearings made of high-temperature materials, such as M50 (Fe-0.80C-4Cr-1V-4.25Mo), or roller bearings made of CBS1000M (Fe-0.13C-0.5Mn-0.5Si-1.05Cr-3Ni-4.5Mo-0.4V) are usable to approximately 315 °C (600 °F), the practical limit is actually determined by the breakdown temperature of the lubricant, which is 205 to 230 °C (400 to 450 °F) for the synthetic lubricants that are widely used at elevated temperatures.

Molybdenum high-speed tool steels, such as M1, M2, and M10, are suitable for use to about 425 °C (800 °F) in oxidizing environments. Grades M1 and M2 maintain satisfactory hardnesses to about 480 °C (900 °F), but the oxidation resistance of these steels becomes marginal after a long exposure at this temperature. An important weakness of these highly alloyed materials is their fracture toughness. Also, regardless of operating temperature, bearings require adequate lubrication for satisfactory operation.

For bearings that operate in moderately corrosive environments, AISI type 440C stainless steel should be considered. Its maximum obtainable hardness is about 62 HRC, and it is recommended for use at temperatures below 175 °C (350 °F). However, the dynamic-load capacity of bearings made from type 440C stainless steel is not expected to be comparable to that of bearings made from 52100 steel. The carbide structure of 440C is coarser, the hardness generally lower, and the fracture toughness about one-half that of 52100 steel.

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Bearing-Load Ratings

Bearing-load ratings are based on results of laboratory rolling-contact fatigue tests that have been conducted under conditions as near ideal as possible. Any departure from these reasonably ideal conditions, such as misalignment, vibration, shock loading, insufficient or inefficient lubrication, extremes of temperature, or contamination, will reduce life expectancy and may cause a type of failure other than by rolling-contact fatigue.

The life of a bearing is expressed as the number of revolutions, or the number of hours at a given speed, a bearing will complete before developing fatigue spalling. Life may vary from bearing to bearing, but conforms to a statistically predictable pattern for large numbers of bearings of the same size and type operating under the same conditions. The L_{10} rating life of a group of such bearings is defined as the number of revolutions (or hours at a given constant speed) at which at least 90% of the tested bearings will survive. Similarly, the life reached or exceeded by 50% of the tested bearings is called the L_{50} life, or median life. The L_{50} life is about 3.5 to 5 times the L_{10} life.

The basic dynamic-load rating for a radial rolling-element bearing is the calculated constant radial load that 90% of a group of apparently identical bearings with stationary outer rings can theoretically endure for one million revolutions of the inner ring. The basic load rating is a reference value only because, as a load, it usually exceeds the yield limit of the bearing material. To avoid harmful plastic deformation, static-load capacities are published along with dynamic-load ratings in most catalogs.

Formulas used by most manufacturers to calculate basic load ratings of rolling-element bearings are given in Anti-Friction Bearing Manufacturers Association (AFBMA) standards. The basic load-rating life may be modified by life-adjustment factors, such as reliability, material, lubrication, alignment, temperature, and other environmental factors. Additional information regarding use of these formulas is provided in the AFBMA-ANSI Standards 9 and 11 and in literature from bearing manufacturers.

Experimental data have provided a simple relationship between load and bearing life. The bearing life, L_{10}, in millions of revolutions equals the ratio of bearing rating, C, to applications load, P, raised to an exponent, y, expressed as $L_{10} = (C/P)^y$, where $y = 3$ for ball bearings and $y = \frac{10}{3}$ for roller bearings. The exponential character of this relationship between basic load and bearing life indicates that, for any given speed, a change in load may have a substantial effect on life in hours.

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Examination of Failed Bearings

If a bearing fails to meet its predicted life requirement, the analyst must discover the cause of damage that led to failure and recommend measures that will eliminate or control this damage. The influence of uncontrolled or unknown factors that can overshadow the effect of the controlled variable in tests must be determined.

The analyst should first obtain complete information about the application and operating environment and, if possible, get samples showing the earliest detectable stages of damage. The complete bearing including the separator should be obtained rather than the failed part only. Second, the bearings and, if possible, the machine should be visually examined to determine the general condition of all related parts and the specific condition of the bearing components so that the type of damage can be identified. The bearing and the installation should be photographed to document the appearance of the damaged area. Such pictures are valuable for preserving the bearing condition prior to cleaning for future reference. Parts of this step may be carried out by an experienced field engineer. Third, samples of lubricant and debris should be collected for analysis. Fourth, the bearings should be thoroughly examined at low magnification (3 to 50×) to classify the type of damage. Significant areas may be photographed. Fifth, representative areas associated with the damage should be selected for metallographic study at high magnification (100× or higher) to verify the results of visual examination. A scanning electron microscope, if available, is an important tool that can provide important analysis and record photographs over a wide range of magnification. An entire small bearing can be placed on some microscope stages. Sixth, available information on load, speed, lubrication, operating temperature, and other environmental factors should be considered in comparing actual and expected life.

The analyst needs information about the application for comparison with previous experience to expedite analysis. The following information on the application and environment should be available:

- Description of the application, including the method of mounting the bearing
- Speed of rotation and whether rotation is constant or intermittent or of variable speed
- Lubrication system, lubricant, and filtering, if used

- Temperature of environment, lubricant, and bearing components
- Potential sources of debris
- Potential sources of electrical current passing through the bearing
- Potential sources of water and other corrosive fluids
- Load spectrum and sources of potential-unknown loads
- Misalignment of deflections that influence load distribution within the bearing system

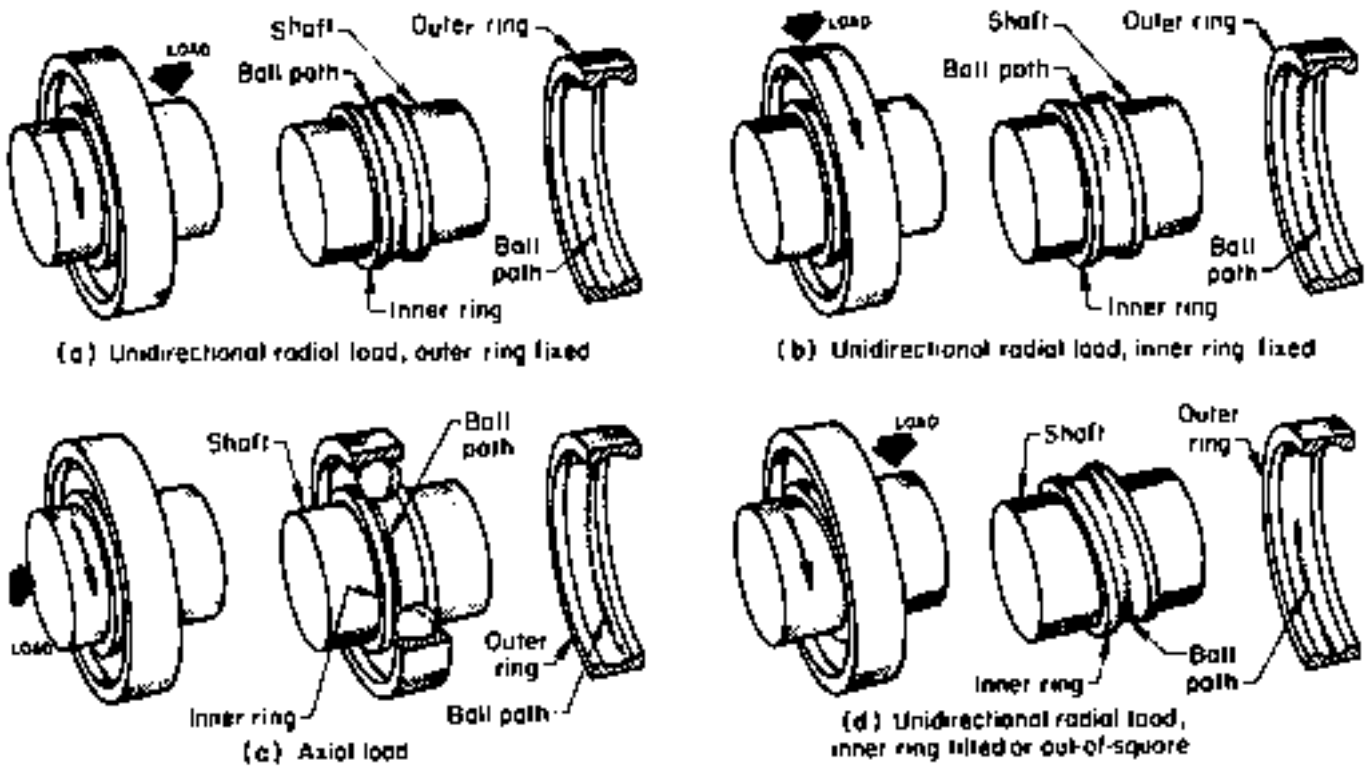
Not all of this information may be needed for every analysis, but it is desirable because the necessary data cannot be established until considerable analytical work has been done. Later acquisition of pertinent data is frequently difficult or impossible. Generally, the data come from the field representative, the bearing user, or both. The analyst should note comments by the field representative or bearing user regarding overheating, excessive noise, frequency of replacement, vibration, looseness, and resistance to shaft rotation.

Before a failed bearing is removed from its housing, the orientation of the outer race with respect to the housing and that of the inner race with respect to the shaft should be marked so that any distress pattern noted in subsequent inspection can be related to the conditions that might have produced it. Frequently, such orientation markings can be compared with similar markings applied at the time of original assembly.

Ball or Roller-Path Patterns. Damage to bearings usually results from subjection to loads or conditions other than those for which the bearings were designed. For example, misalignment or improper fit can result in loading that differs considerably, both in magnitude and direction, from that anticipated by the designer. Determination of such abnormal conditions by inspection of the location and distribution of damage on bearing components is often helpful. Most types of damage to ball bearings will be located on the path of ball travel. Because each type of loading produces its own characteristic ball path, the conditions of loading under which the damage occurs can be determined from an inspection of this ball path. Similar observations can be made on roller bearings.

Satisfactory operation of a bearing that has functioned under radial loads can be easily recognized. In radial ball bearings that have been properly mounted, operated under good load conditions, and kept clean and properly lubricated, the paths of the balls on the highly polished raceways appear as dulled surfaces (similar to lapped surfaces) on which microscopic grinding scratches have been smoothed out. No appreciable amount of material has been removed from the surface of the raceways or balls, as indicated by the fact that there is no measurable decrease in the diameter of the balls, although the entire surface has been dulled. Other indications of satisfactory operation are uniformity; exact parallelism with the side of the raceway, which indicates correct alignment; and centering of the ball path in the raceway, which indicates that loading of the bearing has been purely radial.

With unidirectional radial loading and a fixed outer ring (Fig. 2a), the load is normally carried on slightly less than half of the outer-ring raceway. The inner ring rotates through the loaded region; thus, the ball path on the innerring raceway extends around its entire circumference. Ball paths produced by unidirectional radial loading of a fixed inner-ring bearing are shown in Fig. 2(b). The load is carried by slightly less than half of the inner-ring raceway; the outer ring rotates through the load region and develops a ball path that extends around the entire outer-ring raceway.



(a) Unidirectional radial load, outer ring fixed

(b) Unidirectional radial load, inner ring fixed

(c) Axial load

(d) Unidirectional radial load, inner ring tilted or out-of-square

Fig. 2 Ball paths produced on raceways of ball bearings by axial and unidirectional radial loads. Small arrows indicate rotating elements. See text for discussion. Source: Ref 1

The ball paths produced by unidirectional radial loading of a bearing whose fixed outer ring has crept rotationally in its housing are centered in the raceways, but the path on the outer raceway is longer than that shown in Fig. 2(a) by an amount equal to the length of the arc through which rotational creep has occurred. Any rotational creep that occurs is usually sufficient to make the ball path around the outer raceway of the bearing continuous. Evidence of such creep rotation may be visible on the surfaces at which creep occurs.

A radial-type load can be imposed on a bearing by the manner in which the bearing is mounted. Too tight a press fit of the outer ring in its housing or of the inner ring on its shaft will result in a more uniform ball path around the circumference of each raceway than would be seen with adequate internal clearance.

The ball paths produced by pure axial (thrust) loading are shown in Fig. 2(c). In an angular-contact bearing, the distance that the paths are displaced from the center of the raceways is greater than for a radial-contact bearing carrying an axial load, because of the greater angle of contact in the radial type. If the axial load is applied continuously, the balls do not have an opportunity to change their axis of rotation, and circumferential banding of the balls results (Fig. 2c).

If both radial and axial loads act together on a bearing, the path produced will depend on the relative magnitude of each load. If the axial load is large compared with the radial load, the paths will be similar to those produced by pure axial loading. However, if the radial load is greater than the axial load, the path on the inner raceway will be wider than the path on the outer raceway, and the outer-raceway path will be tilted by an amount depending on the relative magnitudes of the radial and axial loads.

Another type of ball path is caused by misalignments, such as bearings being out of line, outer rings being tilted or out of square, or inner rings being tilted or out of square. Tilting of the outer ring in its housing causes the ring to become slightly oval, which results in a squeeze across one diameter and a relief across the diameter at right angles to the first. This causes narrowing of the ball path at the opposite sides of the diameter across which the relief occurs. If the tilt is not great enough to take up all the clearance, there will be no ball path on the upper (nonloaded) half of the outer raceway.

Figure 2(d) shows a bearing whose inner ring was tilted or out of square with respect to the axis of rotation. Again, the amount of tilt and the amount of bearing clearance affect ball-path characteristics. The paths shown are produced when the load is taken by the bottom portion of the bearing, that is, when some clearance still exists at the top of the bearing after tilting. The inner-ring ball path is at an angle in the raceway as shown and becomes narrower at two opposite points on the raceway. This is due to the ovality caused by pressing the ring on the shaft in a tilted manner. If the tilting were great enough to take up all the clearance at the top of the bearing, the path on the outer raceway would extend around the entire circumference.

Reference cited in this section

1. H.N. Kaufman, Classification of Bearing Damage, in *Interpreting Service Damage in Rolling Type Bearings—A Manual on Ball and Roller Bearing Damage,* American Society for Lubrication Engineers, Chicago, April 1958

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure Types

Failure of rolling-element bearings can occur for a variety of reasons. Accurate determination of the cause of a bearing failure depends to a large extent on the ability of the analyst to recognize and distinguish among the various types of failures. In most instances, this recognition will enable the analyst to determine the primary cause of failure and to make suitable recommendations for eliminating the cause.

The major factors that, singly or in combination, may lead to premature failure in service include incorrect fitting, excessive preloading during installation, insufficient or unsuitable lubrication, overloading, impact loading, vibration, excessive operating or environmental temperature, contamination by abrasive matter, entry of harmful liquids, and stray electric currents. The deleterious effects resulting from the above factors are as follows, though not necessarily in the

same related order: flaking or pitting (fatigue), cracks or fractures, rotational creep, smearing, wear, softening, indentation, fluting, and corrosion (Ref 2).

Two or more failure mechanisms can be active simultaneously and can thus be in competition with one another to terminate the life of the bearing. Also, a mechanism that is active for one period in the life of a bearing can lead to or can even be supplemented by another mechanism, which then produces failure. Thus, in some instances, a single mechanism will be obvious; in others, indications of several mechanisms will be evident, making exact determination of the primary cause difficult. When more than one mechanism has been active, proper determination depends not only on careful examination of the failed components but also on analysis of the material and on the manufacturing, installation, and operating history of the bearing.

Before discussing the characteristics of the various types of failure in detail, it is pertinent to mention that a rolling-element bearing does not have an unlimited life—a fundamental fact that does not appear to be generally appreciated. Even if a bearing is run under the recommended conditions of load, speed, and lubrication and is protected against adverse external influences that would otherwise tend to reduce its life, failure will ultimately result by some process, such as fatigue, wear, or corrosion.

When a bearing is loaded, elastic deformation occurs at the zones of contact between the rolling elements (balls or rollers) and the raceways. Although the stresses involved appear to be compressive only, just below the surface there are high shear stresses and low tensile stresses. Subsurface stresses are described in more detail later in this article. Repeated stressing in shear resulting from rotation of the bearing ultimately leads to the initiation of fatigue cracks, and fragments of metal become detached to produce the effect described as pitting or flaking. This type of failure is similar to that found on the flanks of gear teeth. In addition to load, the pitting or flaking of material from the surface of a bearing race depends on the lubricant viscosity and/or film thickness. Various theories have been proposed that attribute the spalling to hydrostatic pressure of the lubricant being forced into surface cracks to propagate them. Another theory states that the propagation is caused by the tractive forces in the contact causing the subsurface shear stresses to be at the surface.

Experimental data indicate that a rolling-element bearing, considered as a composite whole, does not appear to possess a fatigue limit, such as is found with the materials from which it is constructed; that is, a rolling-element bearing has no lowest value of loading below which failure will not occur no matter how many stress cycles are applied. Typical characteristics and causes of several types of failures are given in Table 1.

Table 1 Characteristics and causes of some common failures of rolling-element bearings

Characteristic and/or location	Cause and remarks
Wear	
Dull raceway surfaces	Coarse abrasive matter in bearing
Shiny raceway surfaces	Fine abrasive matter in bearing
Dull indentations in raceway at same spacing as rolling elements	See *Fretting* and *Indentation*
Wear, together with discoloration	Running without lubricant or with lubricant that has hardened
Wear at points of contact with cage	Inertia forces acting on cage and insufficient lubricant
Wear of ball cage	Inner or outer rings out of square; abrasive matter in bearing
Internal looseness from wear of raceway surfaces and rolling elements	Abrasive matter in bearing because of insufficient or inefficient lubrication or ineffective filtration of lubricant
Uneven wear in ball path	Vibration in presence of abrasive matter
Shaft wear or wear in bearing or housing bore	Rotational creep because of loose fit
Electrical Pitting	
Small craters uniformly distributed over ball paths	Continuous passage of electric current
Small craters randomly oriented or in bandlike order along ball paths	Intermittent passage of electric current
Fluting	
Ridges with burnt craters	Vibration together with electrical current passing through rotating bearings
Ridges without burnt craters	Vibration together with wear or excessive overload on rotating bearings
Smearing	
Cage material smeared onto the rolling elements	Jamming of rolling elements in cage pockets; too high speed; inefficient lubrication; inertia forces
Smeared peripheral streaks in bore or on periphery of outer ring	Rotational creep from inner ring being loose on shaft or outer ring being loose in housing bore
Smeared peripheral streaks on end face	Rotational creep as above, or sliding under pressure against shoulder or other

of ring	surface
Axial smear marks on raceway and/or rolling elements of cylindrical and spherical roller bearings	Mounting under load; inner and outer rings forcibly assembled out of square with each other; axial displacement between inner and outer rings while under load
Axial smear marks on spherical raceway of self-aligning bearings	Angular movements of shaft while bearings are stationary under load
Axial smear marks on spherical raceway of self-aligning thrust bearings	Radial movement in unloaded bearings; improper mounting; inefficient lubrication; faulty manufacture of parts adjacent to bearing
Peripheral streaks on raceways, cage pockets, or rolling elements	Inefficient lubrication or light loads and high speeds
Spiral streaks on ends of rollers or locating flange	Inefficient lubrication under axial load; running under no load because of large radial clearance
Fretting	
Red or black oxide spots, usually with shiny borders, in bore or on periphery of outer ring	Incomplete contact and slight movement or vibration between ring and its seat
Red oxide on surface where ring contacts shoulder	Slight movement due to shaft flexure
Dull indentations on raceway at same spacing as rolling elements	Vibration in stationary bearing, particularly in presence of abrasive particles; referred to as false brinelling
Corrosion	
Local spots or pits on raceways, same spacing as rolling elements	Moisture or acid in bearing while stationary for extended periods
Spots or pits on surfaces	Corrosive lubricant or free water in lubricant; moisture on unprotected surfaces; corrosive atmosphere
Indentation	
Shiny indentations on raceway at same spacing as rolling elements	Improper mounting procedure, hammer blows during mounting, or excessive load on stationary bearings; referred to as true brinelling
Dull indentations on raceway at same spacing as rolling elements	Vibration in stationary bearing, particularly in presence of abrasive particles; referred to as false brinelling
Indentations near filling slot or other isolated area on raceway	Excessive external force when assembling rolling elements
Irregular indentations all over rolling elements	Abrasive products or other foreign matter
Flaking	
At isolated areas on raceway	Early stage of fatigue originating at indentations from foreign matter, bruise on raceway, or corrosion
All around one raceway	Progressive fatigue from overload or insufficient clearance because of inner-ring expansion or outer-ring contraction
At diametrically opposite points on radial bearings	Ring oval through distortion by mounting on out-of-round shaft or in out-of-round housing bore
On only one side of raceway surface	Improper mounting or excessive axial load
At one end of roller raceway	Ring misaligned with cylindrical roller bearings
Equally spaced apart as the spacing of rolling elements	Either true brinelling or false brinelling, depending on appearance of surface; see *Indentation*
Oblique flaking on raceway of rotating shaft	Misalignment or deflection of shaft; ring out of square
Oblique flaking on raceway of stationary shaft	Shaft deflection or ring out of square
On rolling element	Forced assembly; overloaded or insufficient lubrication
Eccentric pitting on raceway of thrust bearing	Eccentric mounting or loading
Cracking and Fracturing	
Cracks through ring	Fit too tight; nonuniform seating surface; deformation or ovality of housing; rotational creep or fretting
Axial cracks in inner-ring bore or on periphery of outer ring	Rotational creep or fretting

Circumferential cracks in rings	Deformation of housing; nonuniform seating surface; excessive overload
Radial cracks on end face of rings	Smearing resulting from rotational creep
Radial cracks on end face of rotating ring	Fouling or rubbing on housing or shoulder during operation
Fractured flange on ring of roller bearing	Mounting pressure unevenly distributed around flange; hammer blows during mounting
Crack or fracture at root of cage tongues or across cage pocket	In bearings generally: insufficient lubricant; too high speed or inertia forces; smearing; fracture of rolling elements; misalignment. In thrust bearings: ring mounted eccentrically or out of square; one row of balls not under load
Miscellaneous	
Enlargement of inner and outer rings	Operation for prolonged periods at high or low temperature; aging of ring material
Superficial discoloration of material without reduction in hardness	Lubricant affected by temperature rise; film deposited by extreme-pressure (EP) lubricants
Abnormal temperature increase	Low lubricant viscosity; too much lubricant, resulting in churning; insufficient internal clearance; speed too high; excessive load; incipient failure
Pronounced high noise level	Raceways indented by blows during mounting or by vibration when bearings are stationary
Piping or metallic sound	Insufficient internal clearance; inadequate lubrication
Irregular noise	Foreign matter in bearing; incipient flaking or other surface discontinuities on the raceways; acoustical properties of adjacent parts

Source: Ref 13

References cited in this section

2. F.R. Hutchings, “A Survey of the Causes of Failure of Rolling Bearings,” Technical Report Vol VI, British Engine Boiler & Electrical Insurance Co., Ltd., Manchester, 1965, p 54–75

13. T.E. Tallian, Rolling Contact Failure Control Through Lubrication, *Proc. Inst. Mech. Eng. (London),* Vol 182 (Pt 3A), 1967–68, p 205

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure by Wear

Rolling-element bearings are designed on the principle of rolling contact rather than sliding contact; frictional effects, although low, are not negligible, and lubrication is essential. If a bearing is adequately and correctly lubricated, wear should not occur. In practice, however, the entry of dirt, hard particles, or corrosive fluids will initiate wear, causing increases in the running clearances of the various parts of the bearing, which may lead to noisy operation and early failure.

Significant amounts of material can be removed from contact surfaces under relatively mild operating conditions if (1) abrasive contaminants are present in the contact area, (2) there is highly inadequate lubrication, or (3) the contact surfaces are subjected to a high rate of sliding. This failure mode is characterized by the presence of wear debris in the lubricant and by evidence of abrasion on the contact surfaces of the raceways and rolling elements. Generally, the surface character is impaired only gradually, and some rolling-element bearings can sustain considerable wear before they become unsuitable for operation. Wear failure can be evaluated by dimensional or weight measurement of the components, by observation of loose wear particles near the contact area, or by spectrographic analysis of the lubricant for wear particles.

Dirt. Ball bearings are particularly sensitive to dirt or foreign matter, which is always more or less abrasive, because of the very high unit pressure between the balls and raceways and because of the rolling action of the balls, which entraps foreign particles, especially if they are small. Foreign matter may get into the bearing during initial assembly of the machine, during repairs, or by seepage from the atmosphere into the bearing housing during operation of the machine. Dirt can also enter bearings as adulterants in the lubricant.

The character of the damage caused in bearings by different types of foreign matter varies considerably. Foreign matter that is fine, or soft enough to be ground fine by the rolling action of the balls or rollers, will have the same effect as that resulting from the presence of a fine abrasive or lapping material. Both raceways and rolling elements become worn, and the bearings become loose and noisy. The lapping action increases rapidly as the fine steel debris from the bearing surfaces adds more lapping material.

A full-complement (uncaged) drawn-cup needle-roller bearing that failed because of wear by fine abrasive material in the bearing is illustrated in Fig. 3(a). Wear decreased each roller diameter by more than 0.05 mm (0.002 in.), and a block of rollers eventually turned in the cup. The flow of lubricant (and abrasive) through the bearing was sufficient to forestall development of heat, although polished flats are much in evidence on the rollers.

Fig. 3 Bearings that failed because of wear by abrasive material in the bearing. (a) Needle-roller bearing. Note that flats have been worn on the rollers. (b) Abrasive wear caused by natural diamond dust (≤5 μm) that was deliberately introduced into the lubricant in the laboratory. Deep grooving of the raceways and excessive wear of the rib and roller ends is evident. Similar wear can occur from hard third particles, such as bearing or gear materials, sand, or machining chips. Source: Ref 4

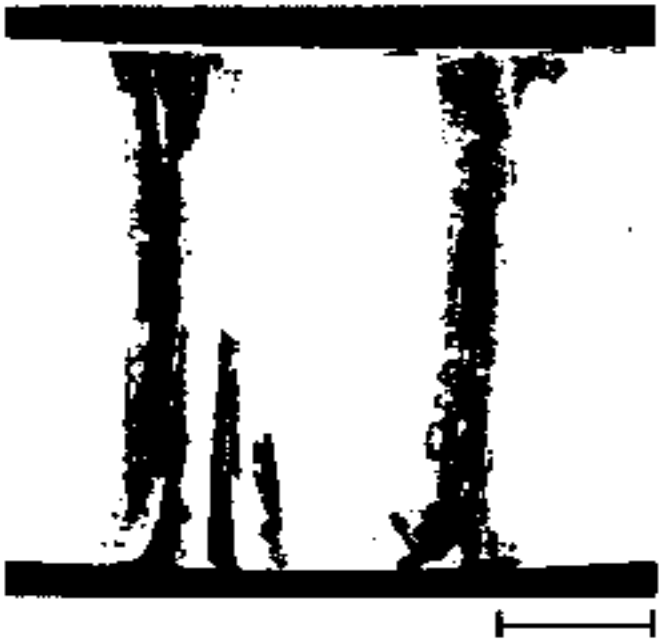
Hard particles, usually nonmetallic, of any size that will scratch, cut, or lap the surfaces of a bearing must be regarded as abrasive. Figure 3(b) shows the effects of natural diamond particles (≤5 μm) that were added to the lubricant in a laboratory test. The wear, known as third-particle abrasive wear, was very extensive on all load-carrying surfaces. Some substances that might not normally be considered as hard or even abrasive, such as iron oxide powder, are excellent lapping compounds.

Hard, coarse foreign particles or iron, steel, or other metals introduced during assembly of the machine produce small depressions considerably different from those produced by overload failure, acid etching, or corrosion. Jamming of such hard particles between rolling elements and raceways may cause the inner ring to turn on the shaft or the outer ring to turn in the housing.

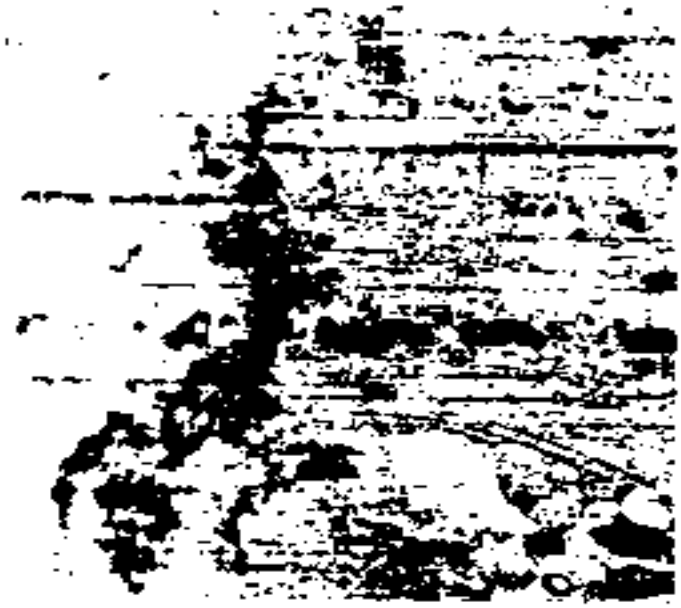
Corrosive Fluids. Water, acid, or other corrosive fluids, including corrosives formed by deterioration of the lubricant, produce a type of failure that is characterized by a reddish-brown coating and very small etch pits over the entire exposed surfaces of the raceways. Frequently, such etching is not evident on the ball path because the rolling action of the balls pushes the lubricant, loaded with corrosive, away from the ball path. The corrosive oxides formed act as lapping agents that cause wear and produce a dull gray color on the balls and the ball paths, in contrast to the reddish-brown color of the remainder of the surface.

Figure 4(a) shows a failure traceable to vibration in the presence of dirt and moisture. General rusting and deep pits can be seen in this drawn cup, as well as polished indentations resulting from abrasive action under vibratory conditions. The roller spacing is indicated by the spacing of the polished indentations.

(a)



(b)



(c)

Fig. 4 Bearings that were damaged by corrosive fluids. (a) Drawn cup for a needle-roller bearing damaged by vibration in the presence of dirt and moisture. Roller spacing is indicated by polished indentations. (b) and (c) Corrosion or water etch on bearing components resulting from a defective seal in a grease-lubricated bearing application. (b) 1.5×. (c) 75×. Source: Ref 4

Some lubricants decompose after extended service without replacement or replenishment, forming acidic by-products. In the presence of moisture, bearing surfaces may be badly attacked by what is referred to as black-acid etching. Such attack produces numerous stress raisers that, under subsequent heavy service loads, can lead to spalling. Figure 4(b) shows an example of water-etched raceways of tapered-roller bearings.

Electrical pitting is produced by passage of electrical current between two surfaces. In such applications as electrical equipment and railway cars and locomotives, there is a possibility of electric current passing through the bearings. When the current is broken at the contact surfaces between raceways and rolling elements, arcing or sparking occurs, producing high temperatures and localized damage. The overall damage is proportional to the number and size of localized points.

A peculiar characteristic of the passage of electric current for prolonged periods of time is the fluting that sometimes occurs. Fluting is pitting in which cavities occur in a regular pattern, forming grooves (flutes). It can apparently occur in any type of bearing. Flutes sometimes develop considerable depth, producing noise, vibration, and eventual fatigue from local overstressing. The cause of fluting is not definitely known. However, it is believed that arcing takes place on the trailing edges of the rolling elements, resulting in local vaporization of metal and the production of craters and pits. This spark-erosion effect frequently gives rise to rapid vibrations that assist in producing the characteristic fluting effect. In some instances, under high magnification, it is possible to observe the burned or fused metal associated with individual craters. Figure 5 shows an example of fluting, including a micrograph of the hundreds of individual arc spots that help create the flutes. The effects of the passage of electric current are usually shown on raceways, but they may occasionally be found on rolling elements, particularly if sparking occurred while the bearing was stationary.

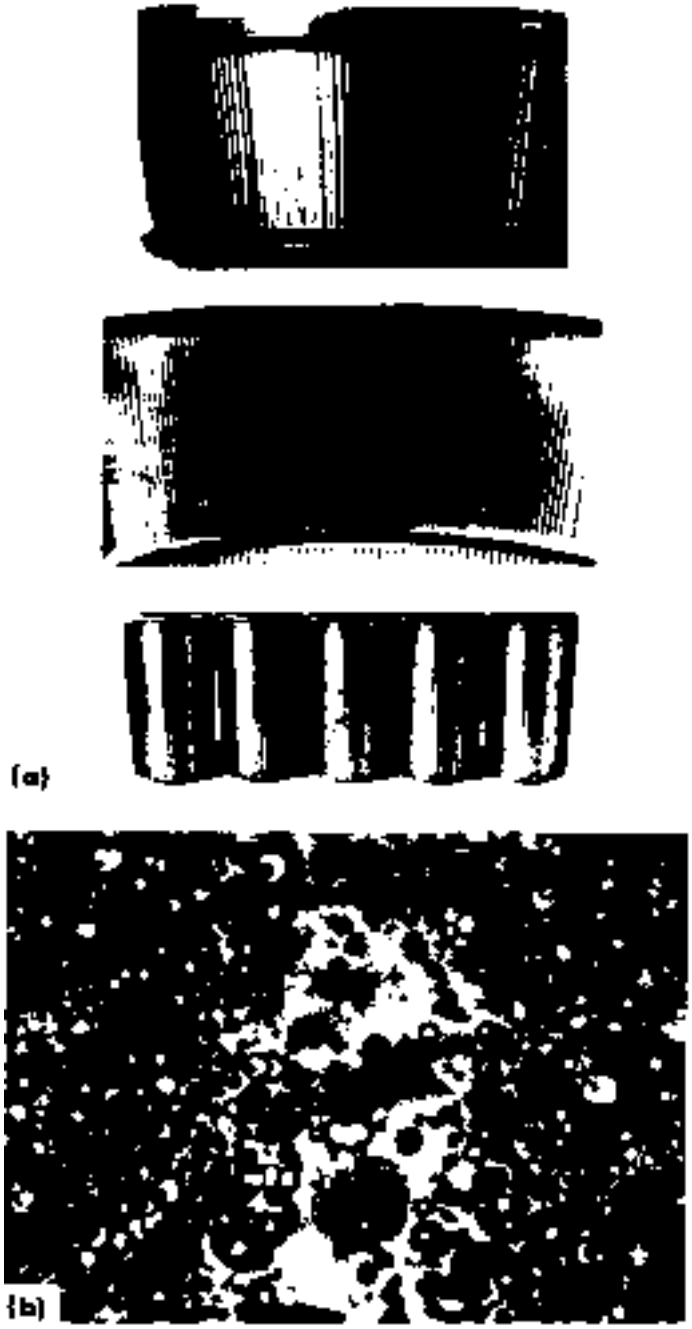


Fig. 5 Tapered-roller bearing damaged by electrical pitting. (a) Fluting damage caused by continuous passage of electrical current. (b) A roller from (a) polished on the outside diameter and etched with nital to show the many individual arc marks that led to the destruction of the raceway surfaces. Source: Ref 4

If fluting is found in bearings in rotating electrical machinery, it is reasonable to suspect that it resulted from passage of electric current associated with defective insulation or induction effects. Remedial measures to eliminate such stray currents include incorporation of insulation within bearing housings or at bearing pedestals and use of suitable short-circuiting connections between the rings or cones.

Pitting of bearing elements by electrical action causes noisy bearing operation. In the following example, use of an electrically non-conductive grease prevented proper grounding.

Example 1: Pitting Failure of Ball Bearings in an Electric Motor by Static Electrical Discharges. The electric motor in an office machine was producing intermittent noise. Ball bearings were suspected as the source of the noise and were removed from the motor and sent to the laboratory for examination.

The bearings were made of type 440C stainless steel hardened to 60 HRC and measured 41 mm OD × 19 mm ID × 12.7 mm wide (1.625 × 0.750 × 0.500 in.). Specifications required the bearings to be lubricated with a grease that could conduct electricity. Bearing life was approximately 200 h.

Investigation. The ball-bearing rings were sectioned for examination in an optical microscope. The surfaces of the balls were observed in a scanning electron microscope. A number of spots were found on the inner-ring raceway (Fig. 6a). An enlarged view of one spot (Fig. 6b) shows more clearly an area from which metal was removed and evidence of a heat-affected zone (HAZ) around the area.

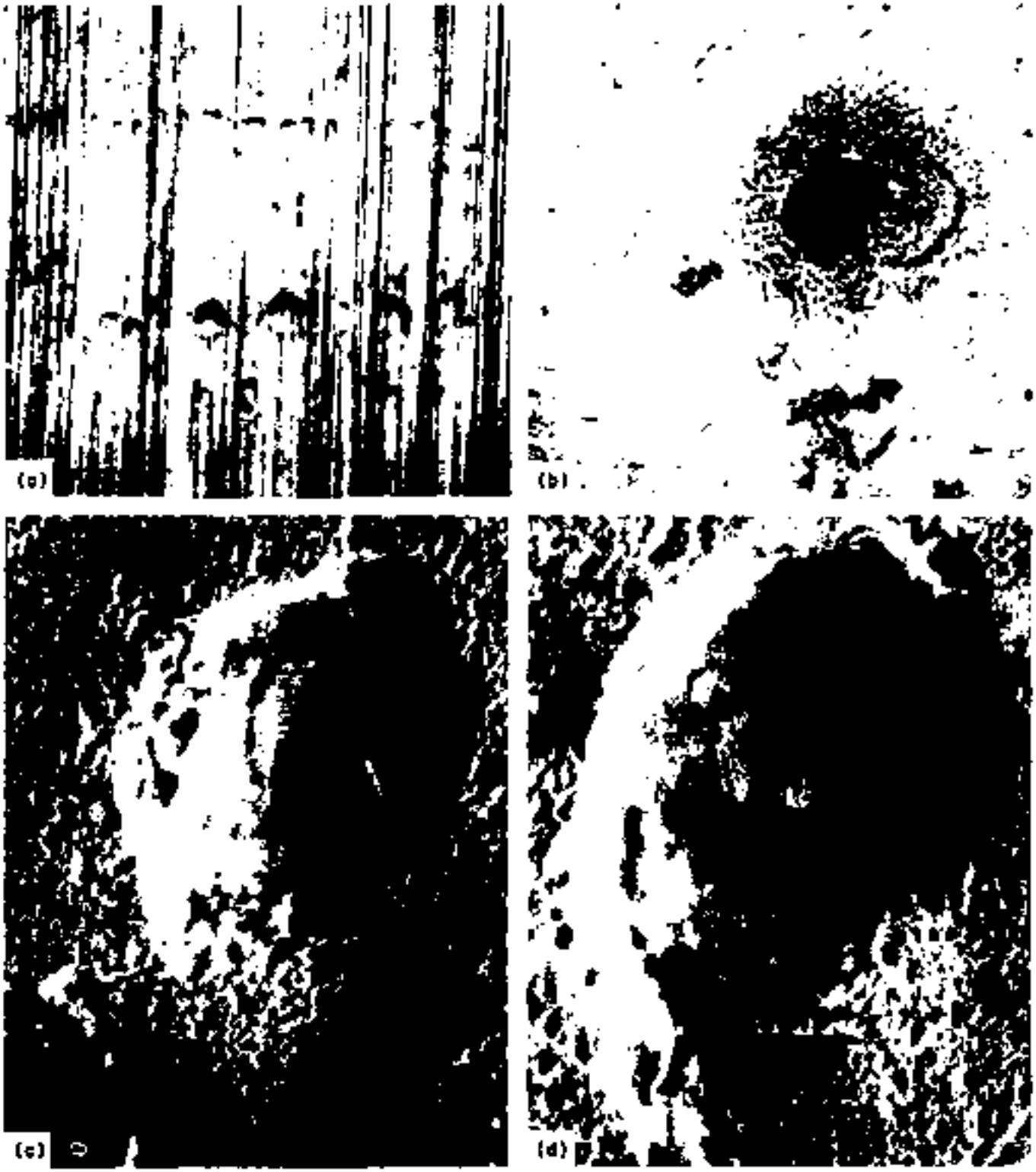


Fig. 6 Weld spots on contact surfaces of a type 440C stainless steel ball bearing. The spots are caused by static electrical discharges resulting from use of an electrically nonconductive grease. (a) and (b) Photographs of inner-raceway surface. 10 and 100×, respectively. (c) and (d) SEM micrographs of ball surfaces. Both 1000×

Figure 6(c) shows a spot on a ball where metal was removed. The metal in the area around the spot showed evidence of having been melted. Another spot on a ball is shown in Fig. 6(d). This ball was welded to the raceway in the region shown at the top on Fig. 6(d) and was separated from the raceway as it rotated. Melting of the metal and welding were indicated by small round beads found in deep indentations in both the ball and the inner-ring raceway. The metal was melted and joined together for only a brief period of time and was then pulled apart by the force of rotation of the ball and the inner-ring raceway.

Because welded areas were found on the raceways of both the inner and outer rings and were randomly spaced and not continuous, it was thought that welding was the result of short electrical discharges between the bearing raceways and balls. The electrical discharges were suspected to have resulted from use of an electrically nonconductive lubricant in the bearings, which permitted a static charge to be accumulated and discharged. To confirm this, a motor made by the same manufacturer was obtained from stock for testing. The electrical resistance between the rotor and the motor frame was measured and was found to be 2 to 3 MΩ while the motor was running. A set of bearings lubricated with a grease known to conduct electricity was installed in the motor. The resistance between rotor and motor frame with this arrangement was only 0.01 to 0.02 MΩ. When not running, the motor had a resistance of less than 0.5Ω (5×10^{-7} MΩ).

As a further check on the electrical conductivity of the lubricant in the failed motor, the resistance across two copper plates was measured using each type of grease as a separator. The result indicated that the grease in the failed bearings was a poor conductor of electricity.

Conclusions. Intermittent electrical discharges across the nonconductive lubricant in the ball bearings resulted in melting and welding of metal at several locations. Because welding was momentary, pits were produced in the ball and ring surfaces.

Corrective Measures. The lubricant was replaced by an electrically conductive grease. This permitted proper grounding of the rotor and eliminated welding between balls and race-ways.

Example 2: Failure of a Low-Alloy Steel Bearing in an Electric Motor Because of Stray Electric Currents. The roller bearing shown in Fig. 7 was one of two that were mounted in an electric motor/gearbox assembly. Rough operation of the bearing was observed, and the unit was shut down. The bearing was removed and submitted for laboratory analysis.

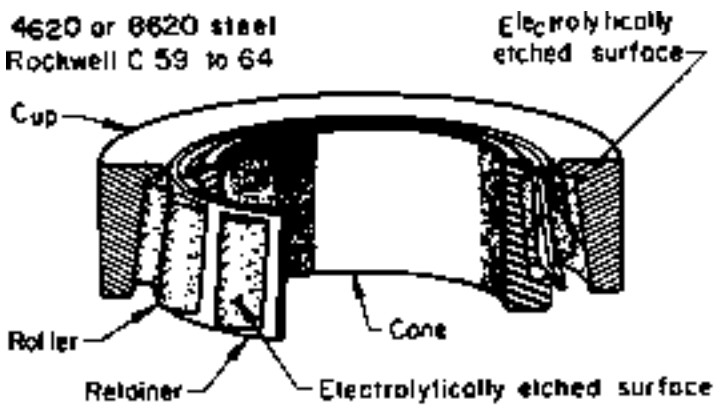


Fig. 7 Low-alloy steel roller bearing from an improperly grounded electric motor that was pitted and etched by electrolytic action of stray electric currents in the presence of moisture.

The bearing components were made of a low-alloy steel, such as 4620 or 8620. The cup, cone, and rollers were carburized and hardened and tempered to 59 to 64 HRC.

Investigation. Visual examination of the cup, cone, and rollers revealed that their contact surfaces were uniformly electrolytically etched. This could have occurred only as a result of an electric current and the presence of an electrolyte, such as moisture, between the cup and roller surfaces of the bearing. The action was similar to anodic etching, as in electroetching of metallographic specimens, only more severe.

Conclusion. Failure of the bearing was the result of electrolytic etching of the contact surfaces by stray currents in the electric motor, which was not properly grounded. The etching was caused by uncontrolled input of current for a long period of time and the presence of ample electrolyte (moisture).

Corrective Measures. The motor was electrically grounded, and the bearing was insulated for protection from stray currents. Both bearings in the assembly were then sealed to keep out moisture.

Cumulative material transfer, a type of failure sometimes called smearing or galling, results from considerable sliding between bearing contact surfaces. It is characterized by the formation of welded junctions between the contact surfaces. This welding usually occurs on a localized microscopic scale and results in removal of material from one or both bearing surfaces by a tearing action. Wear particles may form, which are sometimes work hardened and cause abrasive wear. However, metal buildup usually occurs on one or more of the bearing surfaces.

Smearing failures appear to be confined to contact surfaces that have undergone considerable sliding. Smearing is generally a self-aggravating condition and, if allowed to progress, often is followed by a temperature-imbalance failure resulting from the high friction between the smeared surfaces. If the operating conditions that lead to smearing are relaxed, the failure may be arrested, and there will be some plastic smoothing and possibly wearing away of the transferred material, which may permit continued operation. The transferred metal appears as a bright fracturelike patch.

One cause of sliding is hardening of grease lubricant due to deterioration or contamination by dirt or other foreign matter. As a result of hardening of grease, rotation of the balls or rollers within the cage may be effectively prevented, and in some instances, the bearing can freeze, for example, to such an extent that it will prevent rotation of an electric motor.

Sliding of rollers and subsequent smearing may also occur in a bearing if wear has taken place to a degree sufficient to permit skewing of the rollers so that their axes of rotation are no longer parallel to that of the shaft. Smearing, in a characteristic cycloidal pattern, may be frequently observed on the end faces of rollers as a result of sliding contact between the end faces and their retaining flanges. The possibility of smearing, not the fatigue properties of the material, limits the amount of axial loading that a roller bearing fitted with flanges on both cones can withstand. Inadequate lubrication or a lubrication failure can be the cause of smearing of cage material on rolling elements or adhesive wear between the roller end and cone rib (Fig. 8).

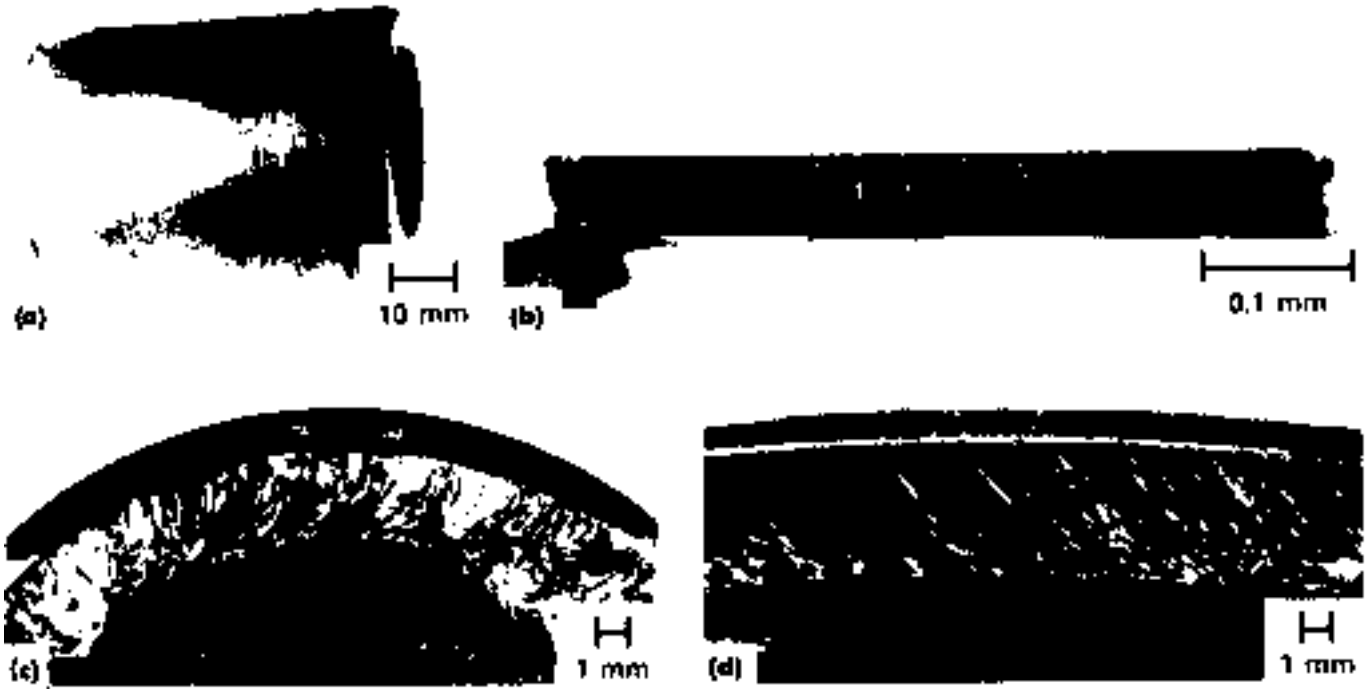


Fig. 8 Examples of adhesive-type wear caused by inadequate lubrication. (a) Metal pickup on a roller outside diameter from sliding contact on the cage shown in (b). Approximately actual size. (c) The end of the same roller showing the scoring damage from the rolling-sliding contact of the thrust rib shown in (d). 3×. Source: Ref 4

Rotational creep, when the term is applied to bearing-failure analysis, refers to an effect in which relative motion takes place between a press-fitted bearing ring and its shaft or housing. Rotational creep occurs principally on the inner ring and may result from the ring being initially loose on the shaft as a result of fitting errors or from plastic or elastic deformation due to either abnormal loads or severe imbalance in service. Under rotational-creep conditions, the assembly behaves like a friction gear, with rotation resulting from the very slight difference in the circumferences of the shaft and the inner ring.

Rotational creep may also occur in thrust bearings if the shaft shoulder against which the ring abuts is not perpendicular to the shaft axis. The relative motion that occurs produces heavy score marks and polished regions, both in the bore of the inner ring and on the shaft. The end face of the ring, which is in frictional contact with the shoulder on the shaft, shows a characteristic scuffing effect typified by heavy circumferential score marks and, in severe cases, by short radial cracks. This effect is similar to local seizure, and it is believed that the cracking arises from thermal stresses induced by intermittent local heating. In some instances, cracks can extend completely through the ring.

Scuffing effects, in addition to those resulting from rotational creep, may occur if a bearing locks solid. Scuff marks may be present on an outer ring, particularly if the bearing rotates as a whole, and a high polish is generally produced on the outer cylindrical surface where it rubs against the housing.

Reference cited in this section

4. R.L. Widner and W.E. Littmann, *Bearing Damage Analysis,* NBS 423, National Bureau of Standards, Washington, April 1976

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure by Fretting

Fretting is variously referred to as fretting corrosion, vibrational false brinelling, friction oxidation, and even chafing or wear corrosion. Such damage is common where there is vibration or low-radial angle oscillation between the bearing elements such that relative slip of the involved bearing-element surface takes place. Corrosion, which is part of two of the terms for fretting given above, describes the resultant wear product, red iron oxide (Fe_2O_3)—the usual evidence of fretting of opposing steel contact surfaces. However, ordinary red rust ($Fe_2O_3 \cdot H_2O$) looks the same. The by-product of fretting may also be black iron oxide (FeO or Fe_3O_4).

Types of Fretting. Fretting may be of two types. Type 1 is the contact corrosion (fit rust) that takes place between the bore of the bearing and the shaft or between the outside surface of the bearing and the bore of the housing. Type 2 is fretting damage within the bearing-contact area. This type of fretting is frequently referred to as false brinelling.

False brinelling is caused by vibrations or oscillations over a few degrees of arc between rolling elements and raceways in a nonrotating bearing. At the contact areas between the rolling elements and raceways, lubricant is squeezed out, resulting in metal-to-metal contact and localized wear. False brinelling does not occur during normal running, but is found in the bearings of machines subjected to vibration while at rest.

There are several features that permit false brinelling to be distinguished from true brinelling (see the section “Brinelling” in this article). False brinelling is found on those regions of a bearing that are subjected to the heaviest loads, for example, at the 6 o'clock position in a bearing supporting a horizontal shaft; the effect diminishes in a fairly regular manner on either side of this position.

False brinelling, although most often found in machines on standby duty subjected to vibration from nearby running machinery, has also been found to occur during transit. Transport by rail or truck is conducive to this form of damage. Even transport by sea does not guarantee immunity, particularly where pronounced vibrations from diesel engines may be present. False brinelling has also been found in the roller-bearing axle boxes of railway cars that have been standing for some time in sidings and that are subjected to vibrations from passing trains.

In many instances, false brinelling remains undetected until the machine is started, at which time noisy operation is immediately evident. The phenomenon is thought to be a manifestation of fretting. The minute movements initiated by vibration cause wear of the contact surfaces, and the fine particles produced rapidly oxidize and result ultimately in production of characteristic grooves, with the oxide acting as an abrasive. To prevent false brinelling in bearings of standby equipment, it may be necessary to arrange for continuous slow rotation of shafts while nearby machines are running.

For a given angular rotation of a shaft resulting from vibration, slightly greater movements of rolling elements relative to raceways will occur in a bearing with a greater number of elements. These movements may be greater than those at which false brinelling develops. Therefore, where false brinelling is not severe, changing to a bearing with a larger number of rolling elements or to one of the needle-roller type may be sufficient.

When equipment is to be transported, one of the following preventive measures should be adopted: (1) dismantle the machine, (2) remove the rolling-element bearings and fit temporary wooden packings in their place, or (3) use clamps to lock rotors of electric motors rigidly to the frames and, at the same time, relieve the bearing of the dead-weight load. In other instances, the machinery could be packed so that the axis of the rotating portion is in the vertical plane, thereby preventing damage by stationary indentation.

Recommendations for reducing fretting in rolling-element bearings include:

- Keep radial play in the bearings at the lowest practical value
- Increase the angle of oscillation (if possible) to secure roller or ball overlap in order to drag fresh lubricant into the area. If the surfaces can be separated by lubricant, fretting of the metal cannot occur
- Relubricate frequently to purge the red iron oxide debris and reinstate the lubricating film

- Use a larger bearing of higher capacity to reduce contact loads
- Increase the hardness of the elements as much as possible. The commercial antifriction bearing already has fully hardened components. For best results, the shaft, if used as an inner raceway, should be hardened to 58 HRC minimum
- Use a grease that has been specially formulated to provide maximum feeding of lubricant to areas susceptible to fretting damage. If the bearing is oil lubricated, flood it if possible

The corrugated surfaces produced by fretting or false brinelling form stress raisers that, under subsequent conditions of rotation, may produce excessive noise and may cause premature spalling by rolling-contact fatigue.
Figure 9 shows a portion of a shaft that served as the inner raceway for a drawn-cup needle-roller bearing. The rollers left deep and clearly defined impressions on the shaft. The damage was identified as fretting, or false brinelling, because of the dull surface with little or no trace of the original surface finish remaining at the bottoms of the indentations, although the shaft was also brinelled by a very heavy overload.

Fig. 9 Severe damage from fretting (false brinelling) on the surface of a shaft that served as the inner raceway for a needle-roller bearing.

Example 3: Fretting Failure of Raceways on 52100 Steel Rings of an Automotive Front-Wheel Bearing. The front-wheel outer angular-contact ball bearing shown in Fig. 10 generated considerable noise shortly after delivery of the vehicle. The entire bearing assembly was removed and submitted to the laboratory for failure analysis.

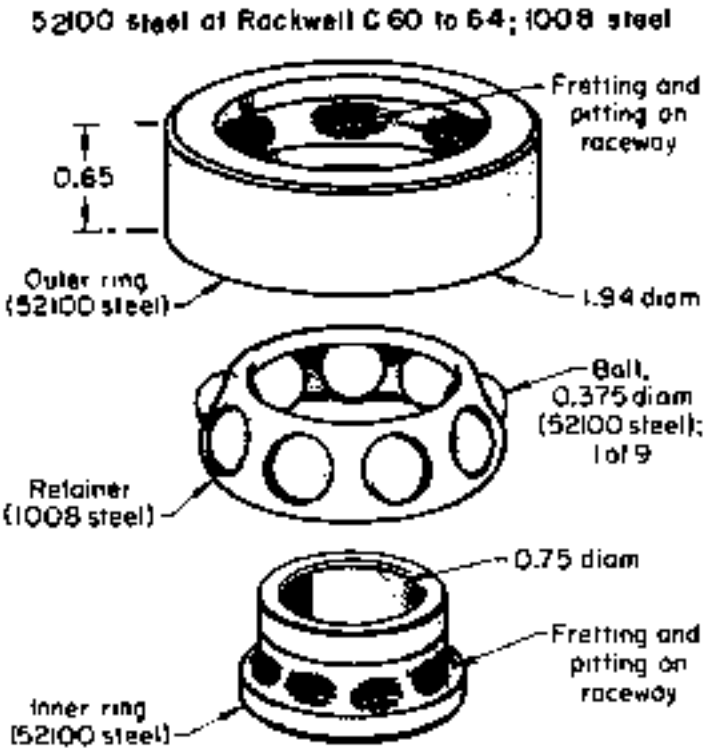


Fig. 10 Automotive front-wheel bearing that failed by fretting of raceways on inner and outer 52100 steel rings. Dimensions given in inches

The inner and outer rings were made of seamless cold-drawn 52100 steel tubing, the balls were forged from 52100 steel, and the retainer was stamped from 1008 steel strip. The inner ring, outer ring, and balls were austenitized at 845 °C (about 1550 °F), oil quenched, and tempered to a hardness of 60 to 64 HRC.

Investigation. Visual examination of the outer raceway revealed severe fretting and pitting in the ball-contact areas at spacings equivalent to those of the balls in the retainer. The areas were elongated, indicating that only a slight oscillation of the ring or wheel had occurred. Similarly spaced but less severely damaged areas were observed on the inner raceway.
Conclusions. Failure was caused by fretting due to vibration of the stationary vehicle position without bearing rotation. These vibrations were incurred during transportation of the vehicle. This bearing had experienced little if any service, but showed conclusive evidence of fretting between the balls and raceway.
Recommendations. Methods of securing the vehicle during transportation should be improved to eliminate vibrations.

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure by Corrosion

Corrosion of any or all components of a bearing can result from the entry of water or other liquids or from deterioration of the lubricant. The formation of deleterious decomposition products in lubricants is discussed in the section “Corrosive Fluids” in this article.
General corrosion of the hardened bearing surfaces often takes the form of minute pitting, which leads to noisy operation and provides surface discontinuities at which cracks can originate. Corrosion of bearings can take place during nonoperating periods or even when the bearings are being stored waiting installation, although corrosion in storage should not occur if the bearings are properly packaged.
Corrosion damage may not be externally visible and may show itself only by subsequent noisy operation. In this respect, it is generally recommended that the protective lubricant in which the bearings have been packed by the manufacturer not be removed unless, in the case of very small bearings, it is desirable to replace the

lubricant by one having lower viscosity. The presence of a small amount of moisture in a lubricant, although perhaps not sufficient to cause general corrosion, may result in a lowering of the fatigue life of a bearing ring.

If a cold bearing is degreased or washed in a cold, volatile solvent, moisture will condense on the cold bearing surfaces. The moisture may not be noticeable, but if the bearing is then dipped in cold oil or is covered with a cold slushing compound, the moisture cannot evaporate and will expend itself in corroding the steel. If the bearing is immersed in hot oil or slushing compound of suitable temperature and assumes the temperature of the bath, the moisture will be driven off, and no corrosion should occur. The presence of water or acids in the lubricant of a running bearing will sometimes cause lines of corrosion to develop in the direction the lubricant is squeezed out from between the rolling elements and raceways.

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure by Plastic Flow

Bearing components, such as balls, rollers, and rings, are generally made of high-strength steels with a hardness of 58 to 64 HRC. However, even at this hardness, overloading of metal components in rolling contact will cause plastic flow. The geometric relationship of the contact surfaces can be altered by cold flow or by unstable overheating.

Cold Plastic Flow. The primary effect of plastic flow on the operation of a bearing is geometric distortion. The contact surfaces can be damaged by formation of indentations if the bearing is stationary or by distortion if the bearing is rotating. Plastic flow on a small scale occurs under normal contact loads and operating temperatures. This has been recognized in defining a static capacity for rolling-element bearings. The absence of all plastic flow is not required, but flow must be limited to a tolerable level. Plastic flow in a rolling-element bearing may be evaluated by measuring the distortion of the contact area.

Rolling-element bearings can be damaged by indentations of several forms, each from a different cause: (1) shallow, random marks from rolled-in particles (Fig. 11), (2) shallow indentations in the raceways corresponding to regions in contact with rolling elements when the bearing is stationary, and (3) well-defined, uniformly spaced indentations.

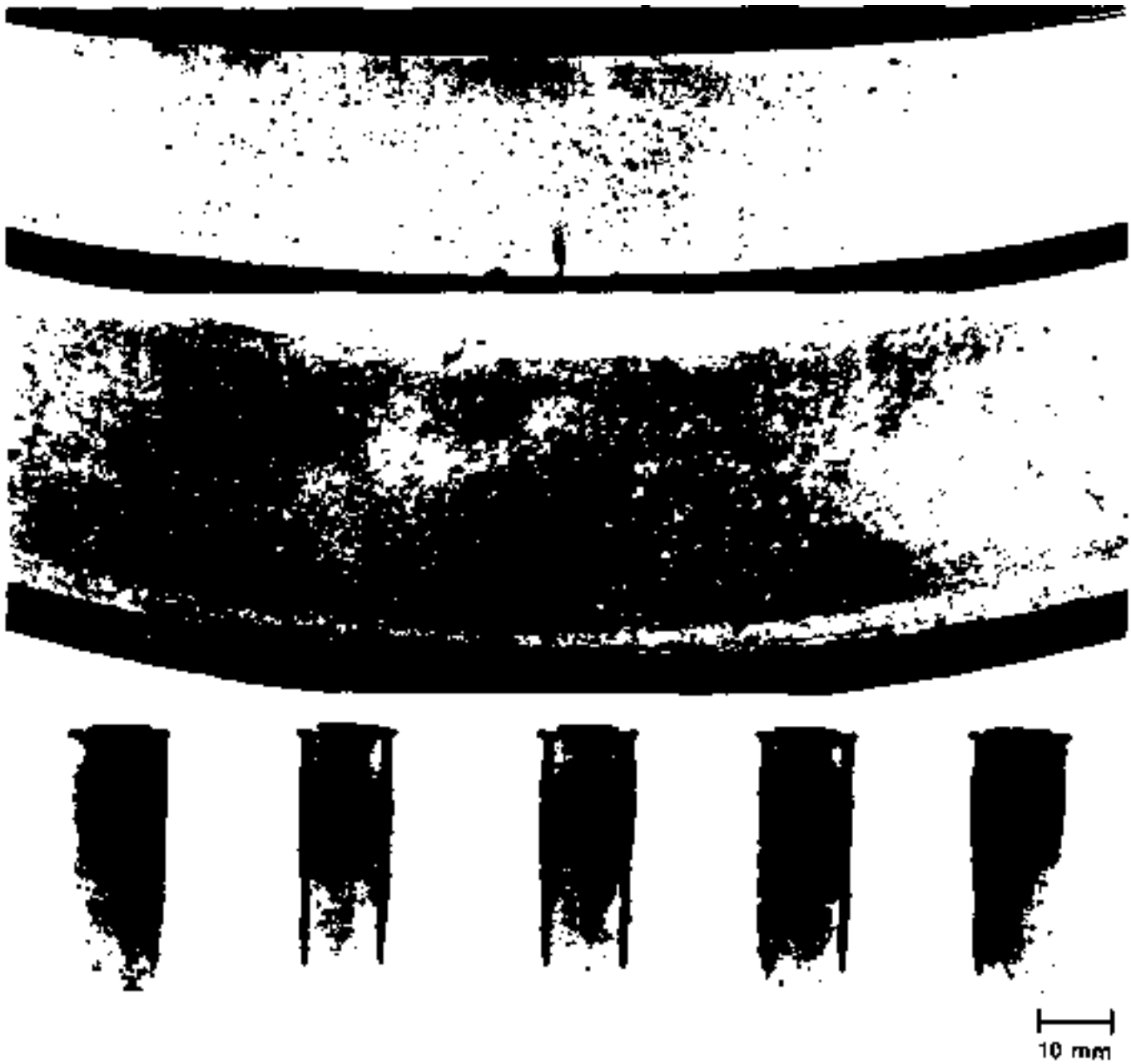


Fig. 11 Denting damage to cone, cup, and rollers from debris entering the bearing from spalling in another component, through defective seals, or from improperly cleaned housings.

The first form of indentation results in rolling-contact fatigue, and the second causes bulk damage (damage outside the contact zone). The third form of indentation is a type of cold plastic flow. The well-defined indentations spaced at a uniform pitch generally result from balls or rollers having been inadvertently forced into the raceways and often arise from carelessness or the implementation of incorrect procedures in withdrawing or installing rings. In the most common situation, the force required to assemble the bearing is applied indirectly through the balls or rollers instead of directly to the ring itself. Careless handling of the machine in transit can also damage bearings, particularly where impact loading has been sustained. Similar, regularly spaced marks may be produced on the raceways of a bearing that has been forcibly assembled with the axis of one ring at a slight angle to the other. These effects (to which the term brinelling is sometimes applied) usually manifest themselves as noisy operation.

Brinelling, sometimes referred to as true brinelling or denting, is permanent deformation produced by excessive pressure or impact loading of a stationary bearing. Hollows or dents are produced by plastic flow of metal. Such indentations in raceways will be spaced in correspondence to ball or roller spacing, thus differentiating true brinelling from dents made by debris, such as dirt or chips, which usually are spaced randomly. The final grinding marks made during the manufacture of raceways can be seen, virtually undisturbed, within the true brinelled areas of Fig. 12. Improper mounting procedures, such as forcing of a tight-fitting or cocked outer ring into a housing, are common causes of true brinelling.

Improper handling, such as dropping or pounding of bearings, will cause true brinelling. True brinelling can also be caused by vibrations in ultrasonic cleaning.

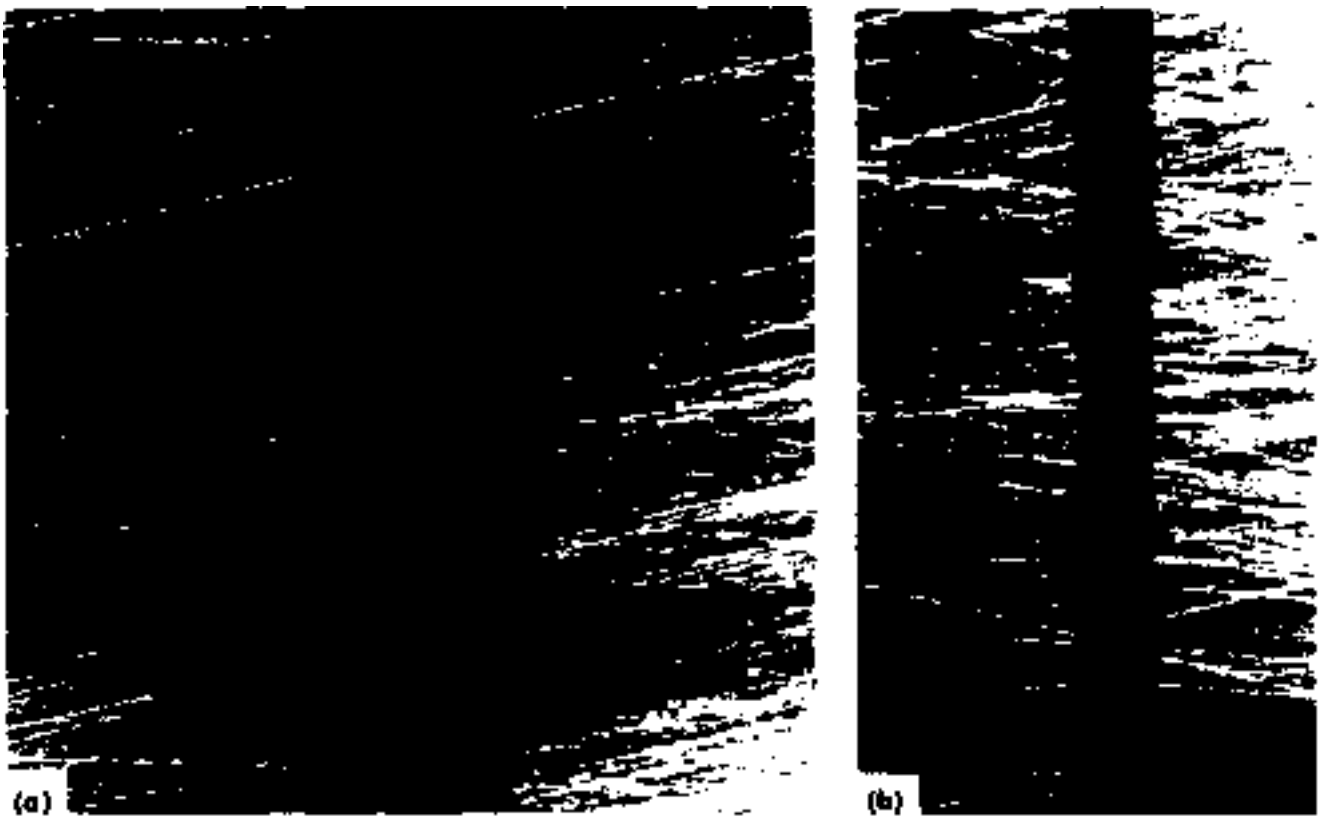


Fig. 12 Comparison of true brinelling with false brinelling. (a) In true brinelling, the surface texture is essentially undisturbed. Note grinding marks in the impression. (b) In false brinelling (fretting), the surface is worn away or material is transferred. No grinding marks are visible in the impression. Both 15×. Source: Ref 4

Example 4: True Brinelling of Ball-Bearing Raceways During Ultrasonic Cleaning. During the early stages of production, randomly selected dictating-machine drive mechanisms, which contained small ball bearings, were found to exhibit unacceptable fluctuations in drive output. This seemed to indicate that the bearing raceways were being true brinelled before or during installation of the bearings.

Investigation. The preinstallation practices and the procedures for installing the bearings were carefully studied to determine the cause of failure. New control practices were instigated with no net gain in bearing performance.

One preinstallation procedure involved removing the bearing lubricant applied by the bearing manufacturer and relubricating the bearings with another type of lubricant. The bearings were ultrasonically cleaned in trichlorethylene to ensure extreme cleanness. Careful examination of the bearing raceways at a moderate magnification revealed equally spaced indentations resembling true brinelling. Further examination showed that the ultrasonic cleaning technique was improper in that the ultrasonic energy transmitted to the balls brinelled the raceways enough to cause fluctuations in machine output.

Conclusion. The bearing raceways were true brinelled during ultrasonic cleaning.

Corrective Measures. The lubricant was removed from the bearings by solvent-vapor cleaning instead of by ultrasonic cleaning.

Spalling that originated at true-brinelling indentations in the surface of a shaft is shown in Fig. 13. The shaft surface in the failed area served as the inner raceway of a bearing.

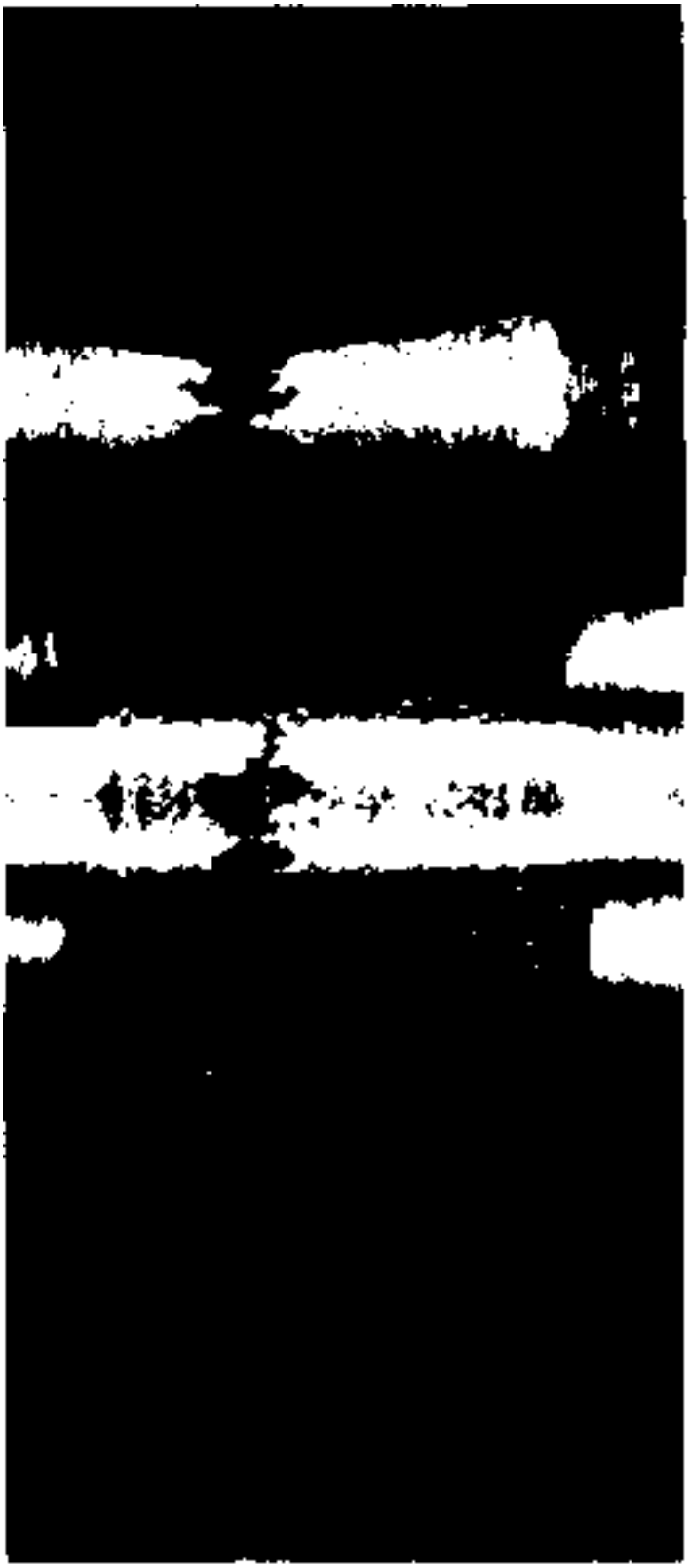


Fig. 13 Surface of a shaft, which served as an inner bearing raceway, that failed by spalling initiated at true-brinelling indentations.

Softening is a plastic-flow phenomenon observed in rolling-element bearings with unstable thermal balance resulting from the generation of more heat in the bearing than is being removed. The maximum permitted temperature in bearings depends on the material, but is generally quoted as approximately 120 °C (250 °F). An increase in operating temperature, if undetected, may result in lubricant failure and seizure of the bearing or in softening of the bearing steel with a reduction in L_{10}. Gross overheating of bearings above the temperature at which the rolling elements and rings were tempered during manufacture will result in rapid softening of these parts with subsequent plastic deformation. Occasionally, if the applied torque is sufficient to overcome the inherent resistance of a seized bearing or to cause the shaft to rotate within the inner

ring, the heat generated may result in temperatures approaching a dull red heat, leading to loss of strength in the shaft followed by rapid failure from shear.

Gross destruction of heat-softened components can sometimes be recognized by discoloration and a predominantly plastically deformed appearance (Fig. 14). Therefore, it may be different in appearance from fracture that results from cracking. However, temperature-imbalance failures are often so devastating that little is left of the bearing from which to identify a failure source.

Fig. 14 An example of burnup with plastic flow in a tapered-roller bearing. This type of failure may result from loss of lubrication or gross overload. The damage begins as heat generation followed by scoring, and if the lubricant is not replenished or the load reduced, the excessive heat buildup results in hot plastic flow. Source: Ref 4

Reference cited in this section

4. R.L. Widner and W.E. Littmann, *Bearing Damage Analysis*, NBS 423, National Bureau of Standards, Washington, April 1976

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure by Rolling-Contact Fatigue

Failures caused by rolling-contact fatigue are the result of cyclic shear stresses developed at or near bearing-contact surfaces during operation. The effects on the shear-stress field below the surface of a bearing raceway as a cylindrical roller travels across the surface are shown schematically, from right to left, in Fig. 15. The roller is shown at position 1 in Fig. 15(a). In the material directly below the roller, the shear stresses along the *x, y,* and *z* axes are equal to zero and are not shown in the stress diagram below the roller, but the maximum shear stress, τ_{max}, is indicated on planes at 45° to the coordinate axes. The material just to the left of the roller (position 2, Fig. 15a) is subjected to shear stresses, τ_{zy} and τ_{yz}, in the directions of the coordinate axes. In position 3, Fig. 15(a), the shear stresses are zero. In Fig. 15(b), the roller is shown in position 2. The 45° shear stress in the material below the roller is at a maximum, and τ_{zy} and τ_{yz} are zero. The material just to the right and left of position 2 is subjected to stresses τ_{zy} and τ_{yz}, which have finite values.

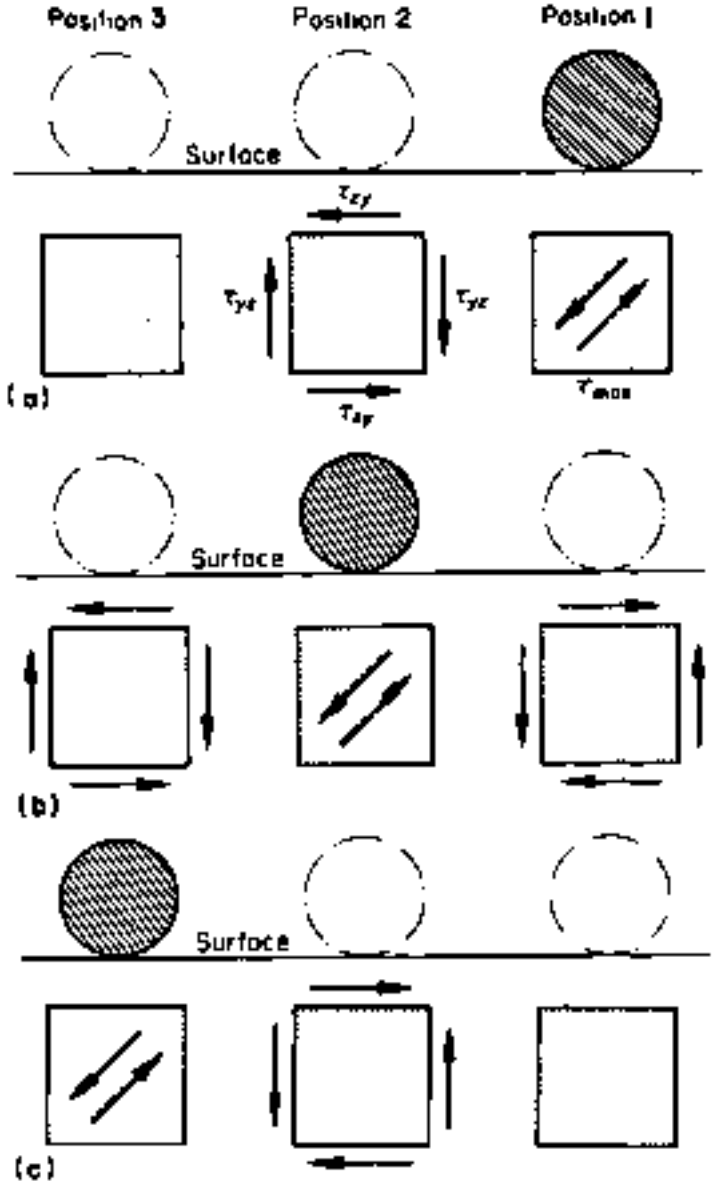


Fig. 15 Shear stresses produced by a cylindrical roller below the surface of o bearing raceway. Source: Ref 5

Figure 15(c) depicts the roller in position 3, and τ_{max} is shown as being in the material below the roller. The shear stresses along the major axes at position 2 once again have finite values, while the stress at position 1 has dropped to zero.

It will be noted after studying Fig. 15(b) that the shear-stress field reverses itself at position 2. This reversal is believed to be the primary contributor to bearing failure and is the phenomenon on which the Lundberg-Palmgren theory (Ref 5) is based. A maximum value for τ_{zy} is reached at approximately the same distance below the surface as is τ_{max}. The absolute value of this maximum, because it extends from τ_{zy} to $-\tau_{zy}$, becomes $|2\tau_{zy}|$ and thus is greater than the maximum shear stress, τ_{max}.

The probability of survival of a bearing is assumed to be a function of the stressed volume, alternating shear stress, number of stress repetitions, and depth of alternating shear stress. A discussion of the theoretical relationships involved is provided in Ref 6. The products of the above assumptions are mathematical relationships that allow life predictions based on geometric properties of the bearing.

Alternating shear stresses reach maximum values at a distance below the surface determined by the load on the bearing as well as by other design factors. Subsurface cracks are usually initiated in regions of surface contact where these alternating shear stresses acting in the material are at a maximum. Pure rolling loads generate maximum shear stresses slightly below the surfaces of the contacting components. With the addition of frictional or sliding forces, maximum shear stresses move closer to the surface. Sliding forces of sufficient magnitude can cause failure to originate at the surface and can significantly reduce bearing life.

Subsurface-initiated fatigue cracks usually propagate parallel to the surface and then abruptly to the surface. They are easily distinguishable provided subsequent operation does not destroy the identifying features. As metal spalls out, steep-

sided flat-bottomed pits remain. These result in progressive deterioration of the material by one or more cumulative damage mechanisms that eventually cause initiation and propagation of fatigue cracks. In some failures, the initial stages are characterized by polished contact surfaces in which small pits are often observed; this damage may be serious enough to preclude further satisfactory bearing operation. Such surface distress, if allowed to continue, can also lead to spalling, in which metal fragments break free from the components, leaving cavities in the contact surfaces. In other instances, subsurface cracks are initiated with little or no observable surface deterioration until the final stages of the process, when cracking and spalling become evident. The mechanisms of rolling-contact fatigue are discussed in more detail in the articles “Fatigue Failures” and “Fretting Wear Failures” in this Volume.

Rolling-contact fatigue failures can be separated into two general categories: surface initiated and subsurface initiated. These two categories may be further subdivided according to the appearance and location of the fatigue spalling, as well as the factors that led to crack initiation. These factors may be related to material, manufacture, mounting, operation, lubrication, and care of the bearing. The recognition of the interplay of such factors is achieved through fractography.

Surface-Initiated Fatigue. Fatigue cracks may be initiated at or very near the contact surfaces of bearing components as the result of several causes. Geometric stress concentrations, for example, can produce abnormally high stresses in localized regions of bearing raceways because of improper contact configuration, misalignment, or distortion. Eventually, such situations can lead to surface deterioration, crack initiation, and spalling. Figure 16 shows the damage on the outer raceway of a roller bearing due to surface deterioration and spalling that resulted from overloading of the rollers at one end. Inclusions and other defects are seldom responsible for surface damage, although they sometimes act as sites for crack initiation. In some failures, cracks are formed at the surface; in others, slightly below it.

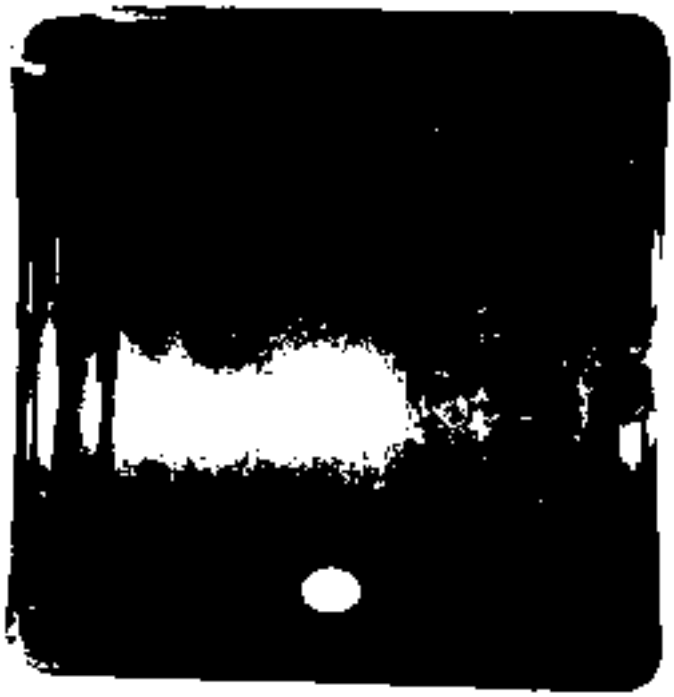


Fig. 16 Damage from surface deterioration and spalling in the drawn-cup outer raceway of a needle-roller bearing because the rollers were overloaded at one end.

Corrosion pits, handling scratches, surface inclusions, and surface dents can also cause stress concentrations and lead to surface-initiated rolling-contact fatigue cracks. Surfaces of fractures (spalls) resulting from such origins usually have an arrowhead type of appearance and point in the direction of load approach. The cracks are open to the surface at the earliest stages of development and thus can be reservoirs for the lubricant. Crack propagation may then be accelerated by hydraulic pressure developed in the cracks during each stress cycle. Corrosive contaminants such as water and organic acids resulting from lubricant deterioration, may also expedite crack propagation if they infiltrate the cracks.

Minor surface discontinuities, such as peeling, are usually associated with ineffective lubrication. Shallow flakes of material are removed from the contact zone, whereas relatively deep cavities are produced by spoiling. Micropits may form on the surface and can initiate fatigue cracks. Crack propagation is usually at an acute angle to the contact surface, but results in a spall cavity that is parallel to the contact surface with only shallow penetration (Fig. 17).

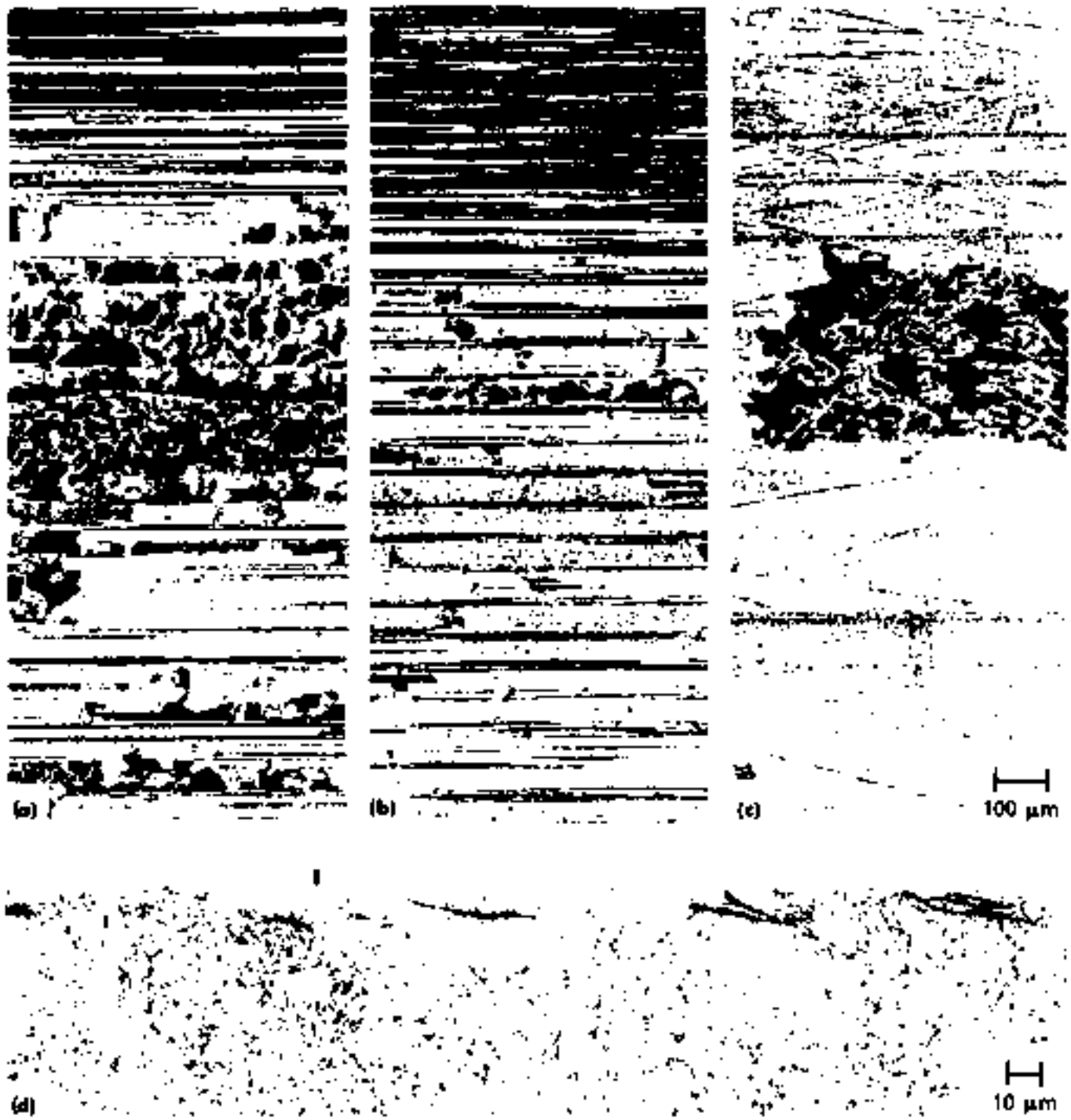


Fig. 17 Microspalling (peeling) on a tapered-roller bearing caused by a thin lubricant film compared to the composite surface roughness. (a) Cup showing fatigue on the peaks of surface texture. (b) Cone showing fatigue on the peaks of surface texture. (c) Roller with a general spalled area. All 65×. (d) Transverse section through a slightly spalled area on a roller showing typical shallow cracking. Nital etch. 350×. Source: Ref 4

Damage involving large regions of a bearing surface indicates marginal oil-film thickness. Damage limited to areas surrounding surface irregularities, such as scratches and other depressions, results from localized metal-to-metal contact; if surface deterioration is allowed to continue, deep spalling will eventually occur.

In an application in which a portion of the surface of a shaft served as a roller-bearing inner raceway, flaking on the raceway (Fig. 18) contributed to failure of the shaft. The lubricating oil was very light, and the film was of insufficient thickness to separate the raceway and the rollers. Fine flaking propagated along the ridges of the surface finish; as the ridges flaked away, the valleys began to flake, and surface breakup became general.

Fig. 18 Fine flaking damage on the surface of a shaft that served as a roller-bearing inner raceway. The flaking originated along the ridges of the surface finish of the shaft.

Example 5: Contact-Fatigue Failure of a Raceway for a Thrust Bearing. The service life of a production gearbox had decreased drastically from that normally encountered. The axial load on a bevel gear was taken by a thrust-type roller bearing in which a ground surface on the back of the bevel gear served as a raceway. The gear was made of 8620 steel and was carburized and case hardened to 60 HRC; case depth was 0.9 mm (0.036 in.). The outside diameter of the bevel gear was 9.2 cm ($3\frac{5}{8}$ in.). The thrust bearing had a 3.8-cm ($1\frac{1}{2}$-in.) inside diameter and a 5.6-cm ($2\frac{3}{16}$-in.) outside diameter.

Investigation. The gearbox was dismantled, and inspection of the bevel gear disclosed spalling damage on the ground bearing raceway at five equally spaced zones.

The gear was mounted on an arbor, and the bearing raceway was checked for runout. The raceway was found to undulate to the extent of 0.008 mm (0.0003 in.) total indicator reading (TIR). A spalled area was observed at each of the high points.

The outside surface of the gear hub was a 3.8-cm ($1\frac{1}{2}$-in.) diam raceway for a radial-contact roller bearing. Examination of this surface disclosed a five-lobe contour having a 0.007-mm (0.00028-in.) (TIR) radial deviation. A chart trace on a rotary profile recorder was made of the two surfaces. The high surfaces on the radial-bearing raceway coincided exactly with the beginning of the spalled area on the thrust-bearing raceway and therefore the high points on that raceway.

Magnetic-particle inspection of the thrust-bearing raceway revealed the presence of numerous cracks that resembled grinding cracks. Microscopic examination of a section through the thrust-bearing raceway also disclosed numerous fissures that resembled grinding cracks.

The thrust bearing operated under a heavy load, which was intensified by the local support at the five high zones. This high nonuniform loading, in conjunction with grinding cracks, was sufficient to produce spalling.

Corrective Measures. The spindle of the grinding machine was reconditioned to eliminate the five equally spaced undulations. Heat treatment of this gear was more carefully controlled to minimize retained austenite. Fatigue life of the bearing returned to normal with these changes.

Flaking is preceded by formation of pits, which appear initially on one of the raceways rather than on the balls or rollers. In the early stage of flaking, isolated pits are formed, but these join rapidly, forming lines or bands of confluent pits and progressively destroying the raceways. The process of destruction is self-accelerating because the pits themselves, as well as the particles of metal that become embedded in the raceways, cause vibration and noisy running. Impact loading at the edges of the pits produces further deterioration. In addition, metallic particles in the lubricant increase the overall wear rate.

As a general rule, fatigue flaking first becomes evident in the stationary ring of a bearing. For a horizontal shaft supported by bearings at each end, with the inner rings rotating and the outer rings stationary (Fig. 2a), loading is radial and the highest stress occurs opposite the point of loading. On each outer (stationary) raceway, there is a specific zone that is

subjected to a cycle of maximum stress every time a rolling element passes over it, but because of the rotation of the inner ring, the maximum stress is exerted at a different region of the inner raceway with every cycle. In circumstances in which the outer ring revolves and the inner ring is stationary (Fig. 2b), the converse is true. For the same speed of rotation, a bearing having a fixed inner ring has a lower life expectancy than a bearing in which the inner ring rotates.

The cause of flaking can frequently be determined by its location on a raceway. Ideally, the load in a roller bearing is uniformly distributed across the rollers and raceways. Flaking in a narrow ring in a plane corresponding to one end of the rollers would indicate grossly uneven loading and would suggest that the axes of the two raceways are not parallel. Flaking restricted to one portion of a stationary raceway is usually caused by overloading from dead weight, excessive belt tension, gear-tooth reaction, or misalignment of driving couplings. Service life can be extended by partially rotating the stationary ring at appropriate intervals, shorter than those at which flaking was found to occur.

Shaft-bearing assemblies required to withstand a significant amount of axial loading usually incorporate angular-contact bearings arranged in opposition and initially adjusted to a certain degree of preload. In this type of application, flaking can be caused by the stresses induced by an increase in preload resulting from thermal expansion as the operating temperature is increased.

Flaking damage may also be distributed around a raceway at intervals corresponding approximately to the distance between the balls or rollers. In these instances, flaking may be initiated at indentations (true-brinelling marks) produced by impact when the bearing is mounted. Another cause for flaking located at roller spacing is corrosion pitting. Both of these modes of damage were described earlier.

Example 6: Rolling-Contact Fatigue Failure of Type 440C Stainless Steel Radial-Contact Ball Bearings Because of Excessive Axial Load. The radial-contact ball bearings supporting a computer microdrum became noisy and were removed for examination. A sample of the lubricant used with the bearings was sent to the laboratory with four sets of bearings.

Two sizes of bearings were used for the microdrum: 20 mm ID × 42 mm OD × 12 mm wide (0.8 × 1.7 × 0.5 in.), and 35 mm ID × 62 mm OD × 14 mm wide (1.4 × 2.4 × 0.6 in.). The bearings were made of type 440C stainless steel and were hardened to 60 to 62 HRC. For accurate positioning of the microdrum, which rotated at 3600 rpm, a spring washer that applied a 22.6-kg (50-lb) axial load on the smaller bearing was installed in contact with the inner ring.

Investigation. The seals of one set of bearings (one small and one large) were removed so that the inner surfaces of the bearings and seals could be examined. Also, samples of the grease in the bearings were taken for analysis and comparison with the sample submitted with the bearing. Comparison of infrared spectrophotometer patterns indicated that the two greases were similar, but not of the same composition.

The bearings were soaked and washed in a petroleum solvent until they were absolutely clean of grease. The solution was evaporated by heating. The residue contained particles that were attracted to a magnet. These magnetic particles were recleaned and mounted for viewing in a scanning electron microscope. Observation at a magnification of 100× indicated that the particles had flaked off the outer-raceway surface and were not foreign contaminants (Fig. 19a).

Fig. 19 SEM views of the surface of the outer-ring raceway of a type 440C stainless steel radial-contact ball bearing that failed by rolling-contact fatigue and of particles found in the lubricant. (a) Lubricant-residue particles that flaked off the raceway. 100×. (b) Flaked surface showing nearness of flaking to edge of raceway (arrow). 200×. (c) Flaked surface. 1000×

The surfaces of the inner and outer raceways were examined in a scanning electron microscope. Smearing, true-brinelling marks, and evidence of flaking were found off-center on one side of the outer-ring raceway. This pattern is commonly found in a radial ball bearing subjected to an axial load that moves the contact area to one side of the raceway. Views of

the flaked surface in the outer-ring raceway are shown in Fig. 19(b) and 19(c). Surface damage to the inner raceway was less severe.

The true-brinelling marks on the raceways were caused by excessive loading when the bearing was not rotating or during installation. These marks were not prominent and were smeared when the raceway surfaces were flaked.

A new bearing of the smaller size was installed in a test stand with a preload of 22.6 kg (50 lb). After 4000 h of operation, the temperature of the bearing was 44 °C (111 °F), and the bearing was noisy. Normal operating temperature for these bearings did not exceed 31 °C (88 °F).

Conclusions. The bearings failed in rolling-contact fatigue, as indicated by flaking in the raceway surface. The 22.6-kg (50-lb) preload moved the zone of contact between the balls and raceway to one side of the raceway, reducing the normal contact area. Metallic particles found in the lubricant in the bearing were from the outer-ring raceway. Some flaking occurred at the true-brinelling marks in the raceways.

Corrective Measures. The preload was reduced to 13 kg (28 lb) using a different spring washer. This eliminated the noise, reduced the operating temperature to normal, and extended bearing life to an acceptable level.

Spalling failure is the result of fatigue and is recognizable by craterlike cavities in the surfaces of raceways or rolling elements. The cavities have sharp edges, generally steep walls, and more or less flat bottoms. Spalling may be obliterated by further destruction of the part or by rolling of surrounding metal into the cavities. Spalling can originate either at or below the contact surface. Figure 20 shows severe spalling damage on the end of a shaft that served as a raceway for a roller bearing. The spalling did not extend all the way around the shaft, only about 225°. The general pattern of spalling damage indicated good alignment and load distribution. However, failure was premature. Linear marks in the damaged surface indicated that the shaft was of poor-quality material. The raceway surface was underlaid with inclusions, which served as stress concentrators for fracture initiation.

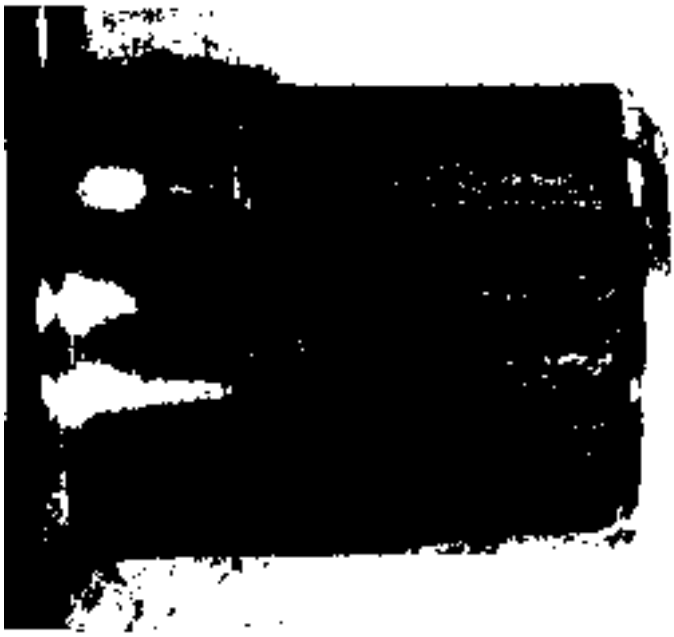


Fig. 20 Spalling damage on the end of a shaft that served as roller-bearing raceway. The spalling was initiated at subsurface inclusions.

Subsurface-Initiated Fatigue. Subsurface fatigue results from the shear stresses described above acting in the region below the contact surface of the rollers or balls on the bearing races. Nonmetallic inclusions act as stress concentrators and are the cause for the most common mode of subsurface-initiated fatigue.

Microstructural alterations that occur at high-stress levels or very high lives at more moderate stress levels have been judged to cause planes of weakness resulting in fatigue spalling. Subcase fatigue in case-carburized components is also related to highly stressed applications.

Effect of Inclusions. Nonmetallic inclusions in the load-carrying areas of bearing elements can reduce resistance to contact-fatigue fracture. An inclusion is foreign material with properties different from those of the matrix. Nonmetallic inclusions are compounds, sometimes complex compounds, that are present in all steels as a result of steelmaking practices, additions for machinability, or additions for strengthening or grain-size control. The primary inclusion types are oxides, nitrides, sulfides, and carbides.

The most detrimental inclusions are the hard (compared to the steel matrix) nondeformable types with an angular shape (oxides, nitrides, and some carbides) that are sometimes loosely bonded to the steel matrix. Oxide inclusions have the greatest influence on rolling-contact fatigue because of their size and frequency and their poor bond with the steel matrix.

Oxide inclusions result from the steelmaking process, which requires the use of aluminum or silicon for deoxidation of the molten steel during the refining of steel. Aluminum is also required to ensure that a steel is fine grained and remains that way during subsequent thermal processing. Modern bearing steelmaking processing is aimed at elimination of large oxide inclusions and minimizing the frequency and severity of the smaller ones.

Sulfur is added to steel to improve its machinability. The sulfur forms inclusions by combining with manganese, an alloying element added for hardenability. Fortunately, manganese sulfide inclusions are deformable and form an intimate bond with the steel matrix. Historically, they have not been found to influence rolling-contact fatigue life. In wrought products, the inclusion formations are usually discontinuous or semicontinuous stringers oriented parallel to the direction of working. Thus, the most adverse effects of nonmetallic inclusions occur when the stress direction is perpendicular to the lay of the stringer. Under cyclic or fluctuating load, inclusions locally intensify stresses to a degree depending on the shape, size, hardness, and distribution of the inclusions. Initiation and propagation of fatigue cracks from a nonmetallic inclusion origin does not automatically mean the bearing has been prematurely damaged. The true bearing load and the material specification for the particular application must be known.

Example 7: Fracture of Ball-Bearing Components by Rolling-Contact Fatigue Because of Subsurface Nonmetallic Inclusions. The pilot of an aircraft reported illumination of the transmission oil-pressure light and an accompanying drop in pressure on the oil-pressure gage. The aircraft was grounded so that the transmission could be checked.

Investigation. Teardown analysis of the transmission revealed no discrepancies in assembly of the bearings and related components. The oil strainer contained numerous fragments of bronze similar to that used in the bearing cage.

The center bearing of the transmission input-shaft ball-bearing stack had a broken cage and one ball that had been split into several pieces along paths resembling the seams of a baseball. Also observed were several scored balls and flaking damage in the raceways of the inner and outer rings. Total operating time for this bearing stack was 770 h.

The origin of flaking in the raceway of the rotating inner ring was identified by the following distinguishing features (Fig. 21a). The origin (area in rectangle), oriented axially in the raceway, was dark and discolored, and was flanked by areas of markedly different-textured flaking damage, that is, shallow fine-texture flaking damage extending from the origin in the direction of inner-ring rotation (arrow A) and deep coarse-texture flaking damage extending from the origin in the direction opposite to the rotation of the inner ring (arrow B). Both areas had a convex texture in relation to the axial origin.

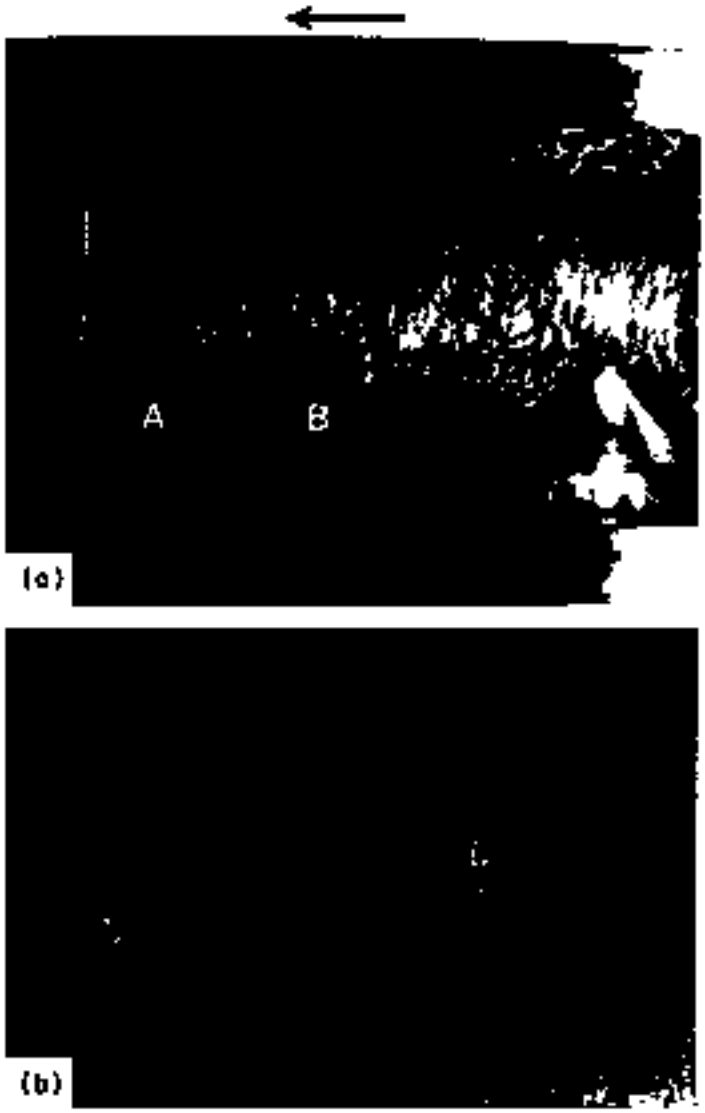


Fig. 21 Inner-ring raceway of an aircraft-transmission ball bearing that failed by rolling-contact fatigue because of subsurface nonmetallic inclusions. (a) Macrograph of inner-ring raceway showing fine-texture flaking damage (arrow A), coarse-texture flaking damage (arrow B), and origin of flaking (rectangle). 3×. (b) Micrograph of section through inner ring at origin of flaking showing inclusions (arrow C). 440×

A section parallel to the axially oriented origin was cut at the origin area. The specimen was ground and polished for metallographic examination. As shown in Fig. 21(b), stringers of nonmetallic inclusions (arrow C) were revealed at the origin. The chemical composition, microstructure, and hardness of the bearing elements were checked and found to be acceptable.

Conclusions. Failure of the bearing occurred through a contact-fatigue mechanism (flaking) activated by the presence of subsurface nonmetallic inclusions (stringers).

Microstructural alterations are another type of inhomogeneity sometimes found in bearing materials. Although inclusions and other kinds of discontinuities are initially present in bearing materials, microstructural alterations are formed during actual operation of the bearing. These alterations result from localized plastic deformation caused by the action of subsurface cyclic stresses. When observed in a microscope, microstructural alterations appear as a white-etching constituent, either in a characteristic patchy form known as butterflies (Fig. 22) or in the form of narrow parallel bands at an acute angle to a contact surface (Fig. 23a) beneath a spall. At present, it is uncertain to what extent such alterations affect the fatigue life of bearings. Although fatigue cracks have been observed in or very near alterations and sometimes propagate along them, it is not certain that alterations actually cause fatigue-crack initiation. Nevertheless, microstructural alterations are associated with rolling-contact fatigue and may yet prove to be responsible for the initiation of cracks in certain instances.

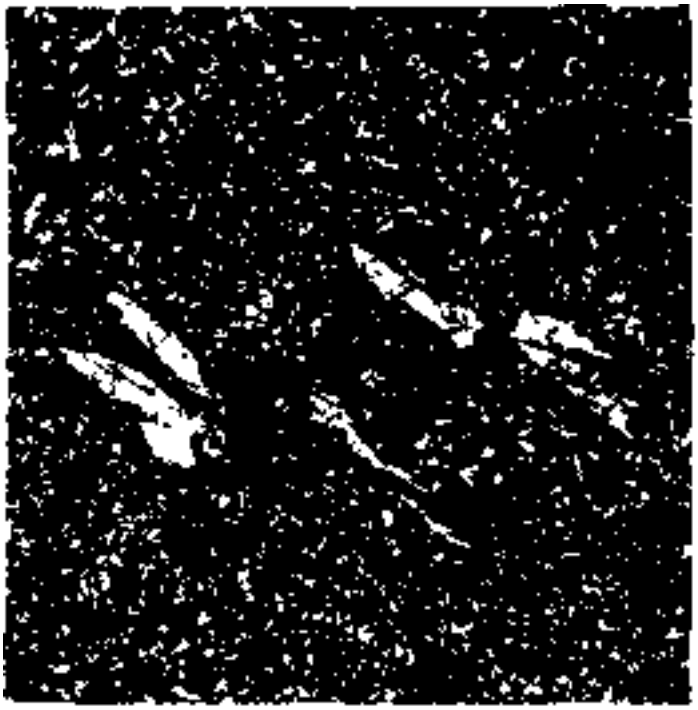


Fig. 22 Stress butterflies (microstructural alterations) in a steel bearing ring. 4% nital etch. Approximately 425×

Fig. 23 Microstructural alterations in a 52100 steel bearing ring. (a) Micrograph of picral-etched section through a spalled area showing elongated bandlike microstructural alterations. 200×. At the top is a profile of the bottom of the spalled area. (b) Macrograph of part of the spalled area. 7$\frac{1}{4}$×. Note parallel ridges that were formed because subsurface cracking partly followed microstructural alterations.

Microstructural alterations create planes of weakness along which cracks can propagate. Figure 23 illustrates a spalled area that was produced by rolling-contact fatigue in which cracks propagated alternately along and between bandlike microstructural alterations inclined at 23° to the surface.

Subcase fatigue is also the result of subsurface-initiated cracking in case-hardened components. Cracks are initiated in or below the lower-carbon portion of the case near the junction of case and core. Once formed, such cracks usually propagate parallel to the case-hardened region until branching cracks propagate to the surface and form spalling cavities. This type of failure is rarely encountered in normal service; rather, it appears under gross-overload conditions, such as may be encountered in accelerated tests. Increasing case depth usually minimizes subcase fatigue.

References cited in this section

4. R.L. Widner and W.E. Littmann, *Bearing Damage Analysis*, NBS 423, National Bureau of Standards, Washington, April 1976

5. A.G. Lundberg and A. Palmgren, Dynamic Capacity of Roller Bearings, *Acta Polytech.*, No. 210, 1952

6. A. Mendelsen, *Plasticity: Theory and Applications*, Macmillan, 1968

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Failure by Damage

Bulk-damage failures are the result of damage to bearing components outside the contact zone. Although evidence of material deterioration may be observed on the rolling-contact surfaces, the most severe damage is found at other locations on bearings. Deterioration of the bearing results from one or more causes, including overloading, overheating, bulk fatigue, fretting, and permanent dimensional changes.

Fractured rings, rollers, balls, and cages are examples of bulk-damage failures. Such failures may be the result of overloading the bearing beyond its design limits, but they can also occur because of misalignment, improper mounting, and improper fitting of a ring on a shaft or in a housing. Severe overloading is usually accompanied by permanent dimensional changes in the bearing as evidenced by distortion of the components. Figure 24 shows a drawn-cup needle-roller bearing that was mounted side by side with another in a sleeve, or tire. The bearing failed because of gross overload. The core material, being relatively soft, suffered bulk plastic flow. As the cup increased in width, circumferential stretch cracks developed on the outside surface, and the oil hole became elongated. The cup wall thinned out sufficiently that radial play became excessive and, coupled with noisy operation, resulted in termination of the service life of the bearing. A similar but perhaps more dramatic example of bulk damage due to high compact loads on a stationary tapered-roller bearing cone is shown in Fig. 25.

Fig. 24 Drawn-cup needle-roller bearing that failed by gross overload. As the cup increased in width under overload, the oil hole became elongated, and circumferential cracks developed in the outer surface.

Fig. 25 Bulk damage to a stationary tapered-roller bearing cone resulting from gross impact loading that yielded the cone material and cracked the case-carburized surface. Source: Ref 7

Cracks and Fractures. Significant cracking in rings and rolling elements takes several forms and results from several causes. Although fracture can result from fatigue, fatigue that leads to fracture differs from the very localized rolling-contact fatigue cracking that causes pitting, flaking, and spalling.

Circumferential cracks that develop in the outer ring of a bearing may indicate of a lack of uniform support from a housing not truly cylindrical but shaped like a barrel or an hourglass. Under these conditions, flexing can take place under load, leading to the development of fatigue cracks that, in a ball bearing, are usually found at the bottom of the ball raceway. This form of cracking ultimately leads to separation of the ring into two or more pieces. Crowned races can be used to compensate for bending in the shaft.

Cracking of inner rings frequently occurs in service. Such cracking generally takes place in an axial direction and may be caused by an abnormally high hoop stress resulting from an excessive interference fit on the shaft. On the other hand, a loose fit can lead to movement between the ring and the shaft, giving rise to radial cracking on the end faces—an effect discussed in the section “Rotational Creep” in this article.

Cracking of any component of a bearing can result from gross overloading, but may also be associated with other effects, such as overheating, wear, and flaking. Fatigue cracks are sometimes initiated at surface locations where pitting, flaking, and spalling have occurred. In other instances, cracks may start from regions where there is evidence of fretting, which has an adverse effect on fatigue endurance.

Cracks that lead to detachment of small, roughly semicircular pieces from the retaining flanges (back-face ribs) of roller bearings may be caused by excessive loading due to (1) incorrect assembly and dismantling procedures, (2) general abuse in service, or (3) abnormally severe axial impact, which brings the ends of the rollers into heavy contact with the cup or cone back-face rib.

Fracture and cracking of balls and rollers occur occasionally. Balls generally fail by splitting into portions of approximately equal size. Beach marks on the fracture surfaces indicate normal fatigue cracking, that is, not cracking due to rolling-contact fatigue; the cracks often originate in regions of incipient flaking on the outer surface. Splitting of balls is initiated primarily in the area of fiber flow essentially perpendicular to the rolling surface. Fiber orientation arises because balls are manufactured by upsetting slugs of round stock between hemispherical dies.

A similar form of failure is shown by rollers. Most rollers fracture axially into two equal portions, but others, although extensively cracked, remain intact. The region of final fracture is approximately circular and is situated on the central axis of the roller adjacent to the one end face. The basic cause of this particular form of cracking is not obvious, but it most likely arises from gross overloading in both the radial and axial directions.

Rollers subjected to considerable end thrust, with one face being forced into heavy contact with one of the flanges of the inner raceway, frequently fail by smearing and wear. Final fracture is usually located adjacent to the smeared end, but may be near the opposite end face.

Another form of failure involves fatigue cracking in a diametral plane and in a helical direction at 45° to the axis of the roller. The pattern of cracking is symmetrical, with a second, helical crack on the remote side of the roller. Usually, cracking in the diametral plane develops first, and the cracks turn when they reach the opposite edge of the roller and proceed in a helical direction. In general, such cracking develops as a result of excessive compressive loads, which produce tensile stresses at right angles to the direction of loading. In addition, torsional stresses are involved, as indicated by the helical cracks. In theory, a roller should not be subjected to torsional stresses, but such stresses can result from misalignment or inherent dimensional discrepancies, particularly if the load on the inner and outer rings is carried by opposite ends of the rollers.

Reference cited in this section

7. M.R. Hoeprich and R.L. Widner, "Environmental Factors and Bearing Damages," SAE Paper 800678, Society of Automotive Engineers, Warrendale, PA, April 1980

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Effects of Fabrication Practices

Correct assembly of a rolling-element bearing into a machine, as well as correct design and fabrication of the machine elements contiguous with the bearing, are essential for satisfactory operation. Satisfactory life and operation of a bearing depend on correct mounting design, accuracy of the machine elements that support the bearing, cleanness of the bearing assembly, and proper mounting on the shaft and in the housing.

Mounting Design. The rotating ring of a rolling-element bearing is generally press fitted onto a shaft or into a housing, while a sliding fit is maintained on the other ring. Tightness of fit between a bearing ring and a shaft or housing varies with the use to which the bearing will be subjected. The type of fit depends on the magnitude and direction of the load (radial or axial) and on whether the load will be applied to the inner or the outer ring or whether application of load to both members may occur. Shaft and housing tolerances have been standardized by ANSI and AFBMA.

The standard internal looseness in a bearing is sufficient to allow for recommended tightness of fit between the inner ring and the shaft or between the outer ring and the housing. If tight fits are required in mounting both rings or if the inner ring will be substantially warmer than the outer ring during operation, the internal clearance in the unmounted bearing must be greater than normal to compensate for press fitting or thermal expansion. A smaller internal clearance than normal is used when radial and axial displacements must be minimized.

Selection of a bearing with inadequate internal clearance for conditions where external heat is conducted through the shaft can result in overheating or noisy operation due to expansion of the inner ring. An oversize shaft to undersize bore results in an excessive press fit and reduces the internal bearing clearance sufficiently to make the bearing tight, causing the bearing to overheat or become noisy.

Rotational creep will occur when the inner ring is loose on the shaft or the outer ring is loose in the housing. This will result in overheating, excessive wear, and contact erosion (or fretting) between the ring and the shaft or housing.

Shaft Shoulders and Fillets. The height of the shaft shoulder should be about half the thickness of the inner ring at the face. If the shaft shoulder is too low, the corner radius of the inner ring will push up against the shaft shoulder (Fig. 26a). Inadequate support of the bearing at the shaft shoulder can result in bending of the shaft or cocking of the bearing, which can cause overheating of the bearing, vibration, hard turning of the shaft, and general unsatisfactory operation of the equipment.

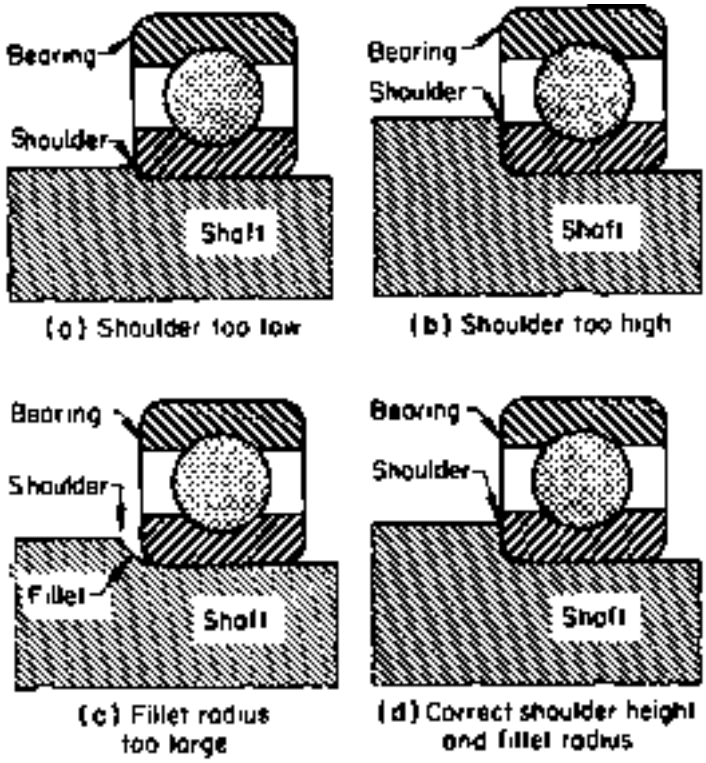


Fig. 26 Incorrect and correct shaft-shoulder heights and fillet radii for mounting rolling-element bearings on shafts. (a), (b), and (c) Incorrect designs. (d) Correct design. See text for discussion.

A shaft shoulder that is too high (Fig. 26b) results in rubbing or distortion of the bearing seals. Also, it is sometimes necessary to remove the bearing from the shaft. This operation requires that a puller be placed against the inner ring, and a surface must be left free on which the puller can operate. A slight corner break should be used on the shaft shoulder and on the bearing seat to avoid raising burrs that would interfere with assembly of the bearing.

Either an undercut or a fillet is used between the bearing-seat surface and the shaft shoulder. An undercut is usually the simplest design at the corner between the shaft seat and the shoulder provided the weakening effect of the undercut can be tolerated. Frequently, however, a fillet must be used; the shaft fillet radius should be less than the bearing bore corner radius indicated in the bearing catalog for the particular bearing being considered. With too large a fillet radius, the face of the bearing will not seat against the shaft shoulder (Fig. 26c), and misalignment of the bearing, bending of the shaft, inadequate support, and improper shaft and bearing location may result.

The proper relationship between the diameter of the shaft and the outside diameter of the inner ring and between the shaft fillet radius and the corner radius on the inner ring are shown in Fig. 26(d). Generally, these same conditions apply to the housing shoulder and fillet radius.

Shaft Alignment and Deflection. Alignment of shaft, bearings, bearing seats, and the machine itself should be within certain tolerances to provide efficient mechanical operation. Misalignment can originate from any of the four general situations shown in Fig. 27. Carefully calibrated micrometers and dial indicators, with the appropriate mounting and clamping accessories, are used to evaluate misalignment.

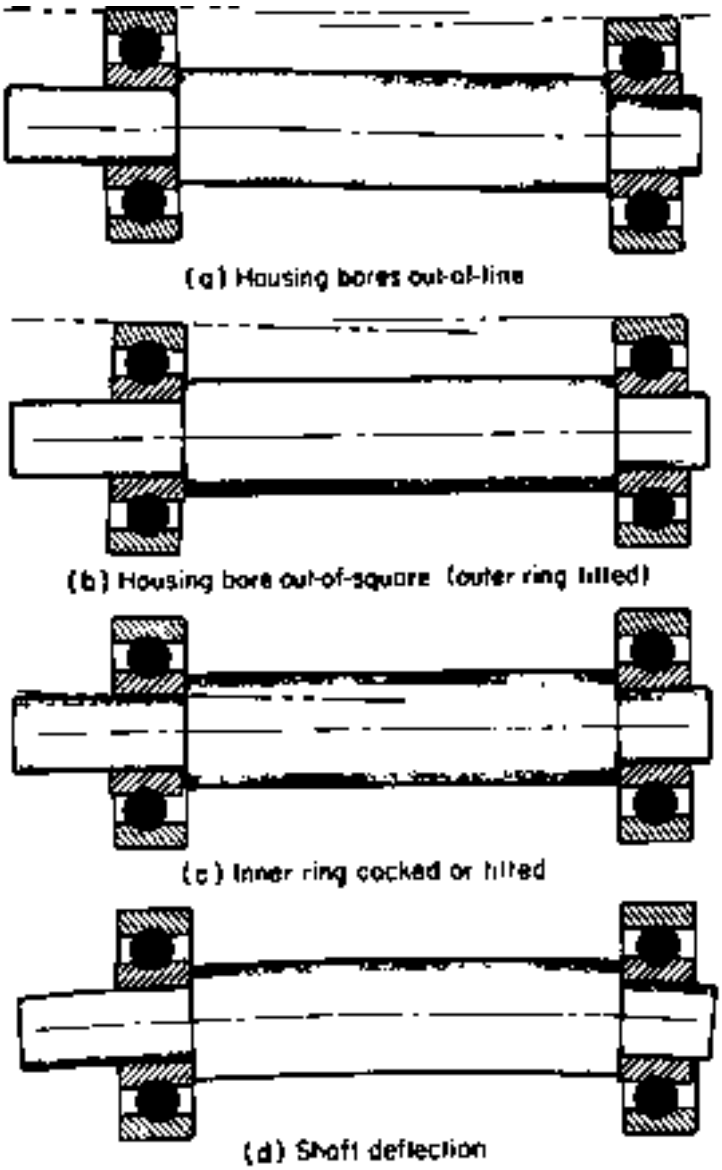


Fig. 27 Four types of misalignment of rolling-element bearings.

In actual engineering applications, misalignment does not always cause failure. A single-row deep-groove ball bearing can with-stand $\frac{1}{4}°$ angular misalignment without cramping or binding. This is less than 1.6 mm ($\frac{1}{16}$ in.) per foot, and a 30-cm (1-ft) long shaft out of line by as much as 1.6 mm ($\frac{1}{16}$ in.) is unusual. Figure 27(a) illustrates tilting of inner rings when housing bores are out of line. Much more prevalent is an out-of-square bearing-housing bore (Fig. 27b) or housing shoulder; these can cause tilting of the outer ring. A cocked or tilted inner bearing ring (Fig. 27c) is also common. The same effect is created by out-of-square lockwashers and locknuts holding the bearing inner ring. On a 2.5-cm (1-in.) diam shaft, a shoulder being out of square by 0.1 mm (0.005 in.) corresponds to $\frac{1}{4}°$ misalignment; thus, difficulties with tilted inner and outer bearing rings are not uncommon. Shaft deflection (Fig. 27d) is rarely critical, because many shafts would break before their angularity at the bearing reached $\frac{1}{4}°$. All these conditions should be checked if the ball path on a raceway indicates misalignment; if any of these conditions exist, changes in bearing clearances may be required, or inspection procedures with improved measurement techniques might be necessary.

Misalignment is indicated by a ball path not parallel to the edge of the raceway (Fig. 2d). When the misaligned path is on the outer raceway, such as in applications where the inner ring rotates, the bore of the housing is not parallel to the shaft axis (Fig. 27b). If the path on the inner raceway is not parallel with the edge of the raceway, it is generally because the inner ring is cocked on the shaft (Fig. 27c), the shaft shoulder is not square with the bearing seat, or the shaft is bent (Fig. 27d).

Misalignment of either ring of a deep-groove bearing will impose an additional load on the bearing. This additional load, together with the normal load, causes overstressing and overheating, resulting in early failure, as in the following example.

Example 8: Failure of a Bearing for a Jet Engine Because of Misalignment Between the Bearing and a Shaft. An engine on a jet aircraft produced excessive vibration in flight and was immediately shut down.

Dismantling the engine revealed complete failure (fracture) of the cage in one of the two main-shaft ball bearings. The ball bearings were both of the single-row deep-groove type with split inner rings, manufactured for use in jet-aircraft engines. The bearings were made of vacuum-melted 52100 steel and were designed to operate at a maximum temperature of 175 °C (350 °F). The bearings were side by side in the engine.

Investigation. Hardness surveys were conducted on the inner and outer rings of both bearings. Hardness of the inner and outer rings of the bearing having the unfailed cage averaged 58 and 59 HRC, respectively, which is consistent with the specified hardness of 58 to 64 HRC. The inner ring of the bearing with the failed cage had a hardness of 55 HRC, and the outer ring 57 HRC. At 175 °C (350 °F), these parts would have a hot hardness of 50 HRC, perhaps indicative of overtempering.

All components of both bearings were visually inspected for evidence of excessive heating, structural damage, and misalignment. Figure 28(a) shows a photograph of the bearing with the fractured cage. Damage caused by severe scoring, scuffing and plastic deformation of the material surrounding the ball pockets of the cage is shown in Fig. 28(b). The cage was also cracked between pockets at several locations around the periphery; one of these cracks (arrow) is visible between the two pockets at top in Fig. 28(b).

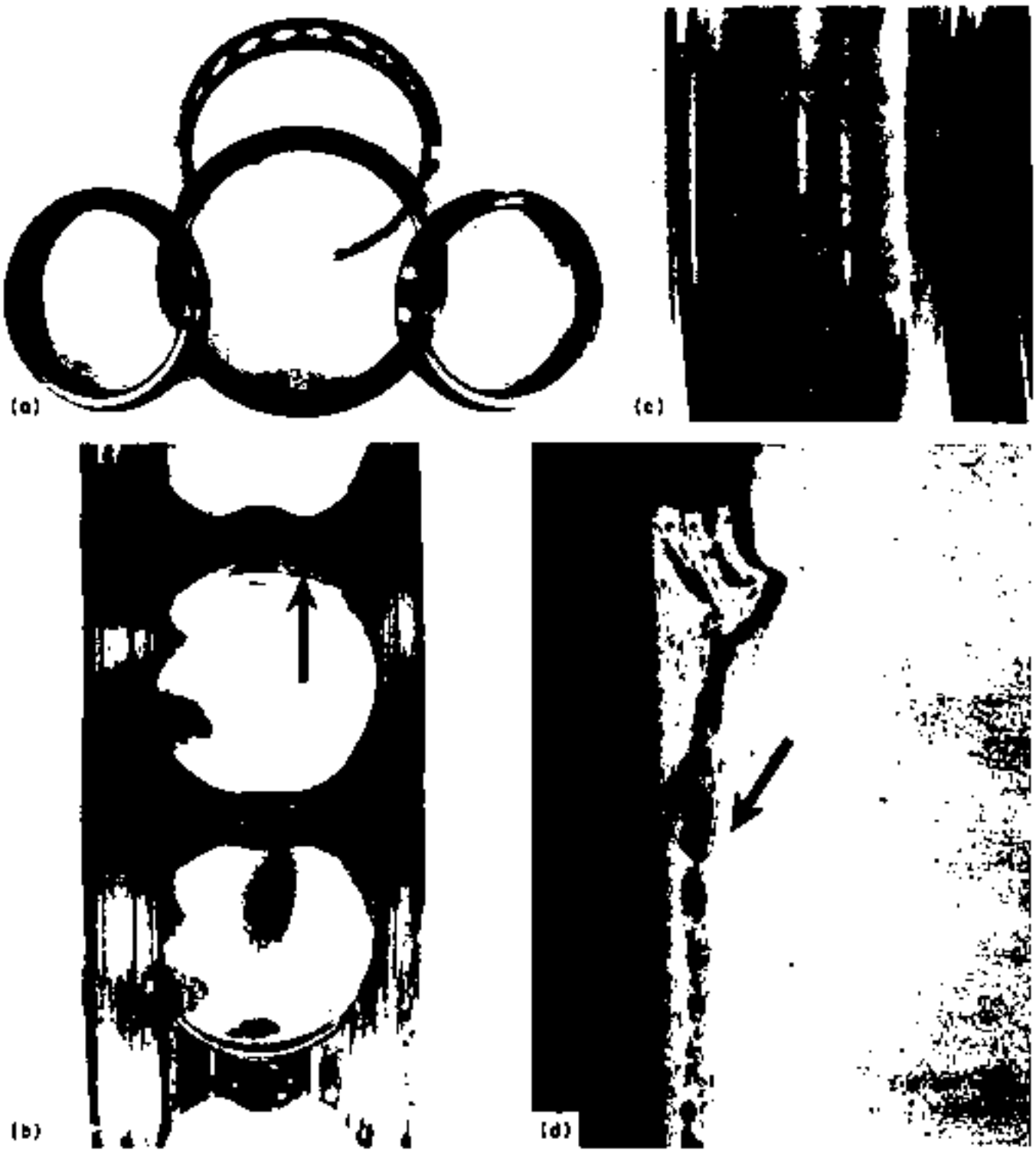


Fig. 28 52100 steel jet-engine ball bearing that failed because of overheating resulting from misalignment. (a) Photograph of bearing components showing fractured cage. (b) Enlarged view of cage showing damage caused by scoring, scuffing, and plastic deformation around ball pockets and a circumferential crack (arrow). (c) Segment of ball groove in outer raceway, showing off-center damage from misalignment. (d) Micrograph of section through spalling cavity in inner raceway of unfailed bearing showing inclusion (arrow) about 0.013 cm (0.005 in.) below surface

Examination of the outer raceway of the bearing with the failed cage revealed severe damage to approximately 20% of the load-bearing surface. As shown in Fig. 28(c), more damage occurred on one shoulder of the groove than on the other, indicating misalignment of the bearing during at least part of its operation. Further evidence of misalignment was found in the ball pockets of the cage, which were unevenly worn. This misalignment caused severe stressing of the cage material,

resulting in fracture of the cage. Discoloration of the bearing components indicated that the bearing had been heated to temperatures in excess of the design limit. The cause of misalignment is uncertain, but may have been improper mounting of the bearing on the main shaft.

The second bearing exhibited no damage other than several small spalling cavities in the inner-ring raceway. Metallographic examination of a circumferential section through the largest cavity revealed an elongated subsurface inclusion (arrow, Fig. 28d). This inclusion was in the region of maximum shear stress in the bearing (approximately 0.1 mm, or 0.005 in., below the surface), which indicated that it had initiated the spalling cavity.

Conclusions. Failure of the cage resulted from overheating and excessive stressing of the cage caused by misalignment of the bearing on the main shaft. The bearing that failed was slightly tempered by overheating during operation.

The small spalling cavities found on the inner ring of the backup bearing were caused by an elongated subsurface inclusion located several thousandths of an inch below the surface and in the region of maximum shear stress. These cavities were not large enough to contribute to vibration in the engine at the time it was shut down.

Abuse before or during mounting can cause bearing raceways to become ball dented. Bearings operated in this condition are very noisy. Ball denting, or true brinelling, is caused by excessive pressure exerted on the raceways through the balls when the bearing is not rotating (see the section “Brinelling” in this article). Denting commonly occurs in small bearings that are pressed into the housing while already mounted on the shaft. If the outer ring is tight or becomes cocked in the housing and pressure to force it into position is applied through the inner ring, indentations are produced.

Impact in an axial direction can force the balls against the edge of the groove, causing dents and nicks. Brinelling can be easily distinguished from dents produced by chips or dirt because ball dents are usually spaced the same distance apart as the balls. Ball dents can be produced in the bottom of a groove if sufficient impact loads are transmitted through the balls in a radial direction.

Mounting bearings on a shaft by applying blows or force to the outer raceway will also cause denting. If the force is sufficient, the raceway may crack. With less force, dents are produced near the center of the raceway, causing rough and noisy operation. As operation continues, pitting followed by flaking will develop at the individual dents. A nick on a ball or a raceway resulting from impact by some sharp object causes noise and vibration, and pitting is likely to follow.

Metal chips and dirt can adhere to and enter bearings that are placed on dirty benches or floors. Such contaminants can cause noise, denting, and locking of one or more of the rolling elements, which ultimately produces failure. Heating bearings with a torch to facilitate mounting or removal is hazardous. A severe temperature gradient is produced that may cause cracking or warping, and excessive heating will lower the hardness, change dimensions, and induce early failure.

Punctured and bent bearing shields interfere with bearing performance and eventually cause bearing damage. Shield damage usually results from use of an improper tool in mounting or removing a bearing from a shaft or spindle.

Broken ball retainers in ball bearings may be the result of misalignment during installation. Excessive shaft deflection can also be responsible for broken retainers. A ball-path pattern alternating from one side of a raceway to the other will be in evidence when a retainer has failed.

When bearings are pressed onto a shaft, the bore must be in perfect alignment with the shaft surface before force is applied. If a misaligned bearing is forced onto a shaft, the inner ring may be distorted or cracked. Forcing a misaligned hardened ring onto a shaft can burr or score the shaft seat. An arbor press is the most satisfactory means of applying the mounting force; use of impact tools should be avoided.

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Heat Treatment and Hardness of Bearing Components

The most common through-hardened ball bearing materials are 52100 low-alloy steel and M50 high-speed tool steel. Table 2 lists average hardness, amount of retained austenite, and typical austenite grain size for eight ball bearing materials (three heat-treatment lots of each material).

Table 2 Variations among heat-treatment lots in average Rockwell C hardness, amount of retained austenite, and austenite grain size, for eight bearing materials

Bearing material	Average hardness, Rockwell C	Retained austenite, vol %	Austenite grain size (ASTM E 112)
52100 steel			
Lot A	62.5	4.90	13
Lot B	62.0	4.10	13
Lot C	62.5	0.80	13
Halmo			
Lot A	60.8	0.60	8
Lot B	60.8	1.00	8
Lot C	61.1	1.70	8
T1 tool steel			
Lot A	61.4	7.30	11
Lot B	61.4	5.20	9
Lot C	61.0	9.50	10
M42 tool steel			
Lot A	61.8	1.00	9
Lot B	61.3	4.40	10
Lot C	61.3	4.90	8
M50 tool steel			
Lot A	62.6	1.90	10.3
Lot B	62.2	2.90	9
Lot C	62.3	1.50	10
M10 tool steel			
Lot A	62.2	1.10	9
Lot B	62.0	2.40	6
Lot C	61.8	1.60	6
M1 tool steel			
Lot A	63.3	2.90	10
Lot B	63.4	3.30	9
Lot C	63.5	1.00	8
M2 tool steel			
Lot A	63.4	1.70	6
Lot B	63.4	2.40	10
Lot C	63.4	2.30	9

Source: Ref 8

Hardness is an important variable in rolling-contact fatigue and can significantly affect bearing life (see the section “Bearing Materials” in this article). In general, bearing hardness should be at least 58 HRC for adequate bearing life. Investigations have shown that, within limits, rolling-contact fatigue life increases as hardness of the bearing components is increased (Ref 9).

Significant differences in hardness between rings and balls can affect bearing life. Experimental results (Ref 10) indicate that balls should be approximately 1 to 2 Rockwell C points harder than rings.

Because bearings operate at elevated temperatures and because bearing life depends on hardness, the hardness of the bearing material at operating temperature is significant. Data on short-term hot hardness (Ref 11) indicate that there is a significant difference between bearing components of low-alloy steels, such as 52100, and those of high-speed tool steels, such as M50, in ability to maintain hardness as temperature increases.

Most bearing alloys fall within the low-alloy or high-speed tool steel categories; however, a heat treatment that is optimum for a material in a tool is not necessarily optimum for the same material when used in a bearing component. Unfortunately, most heat treatments for these bearing materials are based on the heat treatments that have been developed for the materials as tools.

Example 9: Bearing Failure Caused by Improper Heat Treatment of Outer-Ring Raceway. A large bearing from a radar antenna was replaced because of deformation, surface cracking, and spalling on the raceway of the outer ring. Figure 29(a) shows a sectional view of the bearing.

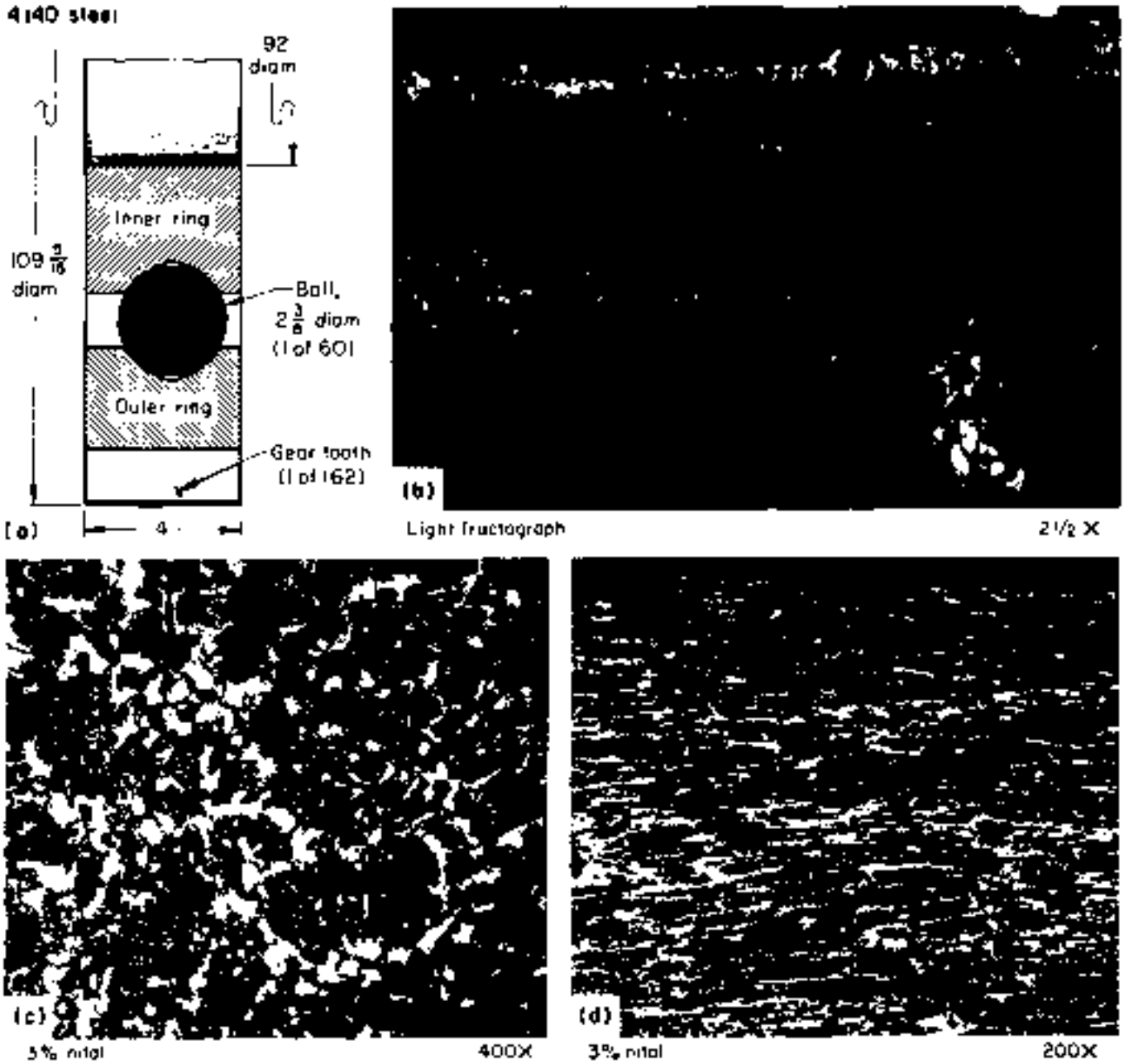


Fig. 29 Large-diameter 4140 steel radar-antenna bearing that failed because of improper heat treatment of outer-ring raceway. (a) Configuration and dimensions (given in inches). (b) Fractograph showing typical damage on outer-ring raceway. (c) Micrograph of section through metal in outer ring adjacent to raceway showing ferrite, scattered patches of pearlite, and tempered martensite. (d) Micrograph of section through outer ring at raceway showing grains elongated by metal movement (raceway surface is at top)

Specifications required that the rings be made of 4140 steel. The raceway surfaces were to be flame hardened to 55 HRC minimum and 50 HRC 3.2 mm ($\frac{1}{8}$ in.) below the surface. Other surfaces of the rings were to have a hardness of 24 to 28 HRC. The bearing capacity at 1.67 rpm was a radial load of 713,000 kg (1,572,000 lb) and an axial (thrust) load of 459,000 kg (1,012,800 lb). Samples of the inner and outer rings 15 to 20 cm (6 to 8 in.) long were sent to the laboratory for examination.

Investigation. Examination revealed that the raceway of the outer ring was damaged by deformation, surface cracking, and spalling (Fig. 29b). The raceway of the inner ring exhibited little or no damage.

Wet chemical analysis of the material in the inner and outer rings showed it to be 4140 steel. Molybdenum contents were 0.26% in the inner ring and 0.31% in the outer ring, both slightly above the specified range of 0.15 to 0.25%. However, this deviation is not significant and would not affect strength properties or hardenability.

A cross-sectional specimen about 9.5 mm ($\frac{3}{8}$ in.) thick was taken through the center of each sample; after cutting, both surfaces of each specimen were ground smooth. Hardness of the raceway of the outer ring was 29.8 to 11.7 HRC. A horizontal traverse through the center of the ring showed a hardness of 26.1 to 18.7 HRC. A vertical traverse 4.1 cm ($1\frac{5}{8}$ in.) from the outer surface showed a hardness of 25.2 to 18.9 HRC. The low hardness values for the outer-ring raceway indicated that it had not been properly flame hardened. The hardness of the top and outer surfaces of the rings was within specifications. The hardness traverse of the inner ring raceway showed a hardness of 46.8 to 54.8 HRC, which was below the specified hardness of 55 HRC minimum.

The specimens were etched in 3% nital to differentiate between hardened and unhardened areas. The inner ring showed a well-developed hardened case at the raceway; the outer ring did not.

Metallographic examination of an etched specimen showed that the structure of the inner ring adjacent to the raceway was a mixture of tempered martensite and some ferrite. This suggested that heat treatment was insufficient to make the structure completely austenitic before quenching; therefore, the full hardening capability of the steel was not attained. The microstructure of the material away from the hardened surface was a mixture of finely divided pearlite and ferrite, which resulted in the metal being relatively soft.

The microstructure of the material in the outer ring adjacent to the raceway (Fig. 29c) was a mixture of white ferrite, scattered patches of pearlite, and martensite, which showed that the steel had been improperly austenitized, producing very low hardness.

Displacement of metal on the outer raceway is shown in Fig. 29(d). The grain structure had been elongated, indicating metal movement. Rolled-out and embedded metal particles were also found in the surface.

Hardenability of the metal in the outer ring was checked by heating a specimen 25 × 25 × 6.4 mm ($1 \times 1 \times \frac{1}{4}$ in.) along one edge with an oxyacetylene torch to 870 to 900 °C (1600 to 1650 °F), then quenching in water. Hardness in the heated area was 57 to 60 HRC; in the unheated area, 23 HRC.

Conclusions. Failure of the raceway surface of the outer ring of the bearing was the result of incomplete austenitization. The steel in the inner-ring raceway had been hardened to slightly below the specified hardness of 55 HRC minimum, but the outer-ring raceway had a maximum hardness of only 29.8 HRC. The raceway surface in the outer ring was not properly heated by the flame-hardening process; therefore, subsequent quenching and tempering operations, if any, would have had virtually no effect on hardness.

References cited in this section

8. J.L. Chevalier and E.V. Zaretsky, *Effect of Carbide Size, Area, and Density on Rolling-Element Fatigue,* NASA TN D-6835, National Aeronautics and Space Administration, Washington, June 1972

9. T.L. Carter, E.V. Zaretsky, and W.J. Anderson, *Effect of Hardness and Other Mechanical Properties on Rolling-Contact Fatigue Life of Four High-Temperature Bearing Steels,* NASA TN D-270, National Aeronautics and Space Administration, Washington, March 1966

10. E.V. Zaretsky, R.J. Parker, and W.J. Anderson, Component Hardness Difference and Their Effect on Bearing Fatigue, *J. Lubr. Technol. (Trans. ASME),* Jan 1967, p 47

11. J. Chevalier, M. Dietrich, and E. Zaretsky, *Short-Term Hot Hardness Characteristic of Rolling-Element Steels,* NASA TN D-6632, National Aeronautics and Space Administration, Washington

Revised by Ronald L. Widner, The Timken Company

Lubrication of Rolling-Element Bearings*

The functions of lubricants for rolling-element bearings are to provide a thin film of oil between bearing contact surfaces to reduce metal contact, to prevent excessive heating by acting as a coolant, to remove wear debris and other foreign substances from the contact surfaces, and to help seal out dust, dirt, and other environmental contaminants. Experience has shown that reductions in bearing life will occur if a lubricant is ineffective in performing any of these functions. Lubricants must have the proper physical and chemical characteristics under the conditions of operation. A program of regular inspection and maintenance is necessary to ensure that both the lubricant and its distribution system, as well as other mechanical components, remain in satisfactory condition.

Thin-Film Lubrication. When the proper amount of lubricant is supplied to a bearing, a thin film is formed on the bearing surfaces. During operation, extremely small and randomly distributed surface irregularities penetrate the thin film, causing metal-to-metal contact. The operational life of a bearing is inversely related to the number of such contacts per revolution. Other factors being equal, the fewer the contacts made per cycle, the longer the life of the bearing.

Under load, the area of contact between a ball and its raceway is elliptical in shape (Fig. 30a). The zone of contact between a cylindrical roller and its raceway is rectangular (Fig. 30b), whereas a tapered roller forms an area that is wider at one end (Fig. 30c). As load is applied, the elements deform elastically, and contact is spread over an increasingly larger area. With a suitable lubricant between the surfaces, such an increase in contact area and the increase in fluid viscosity caused by the load result in the formation of a pressurized (elastohydrodynamic) film that separates the surfaces during operation (Ref 13).

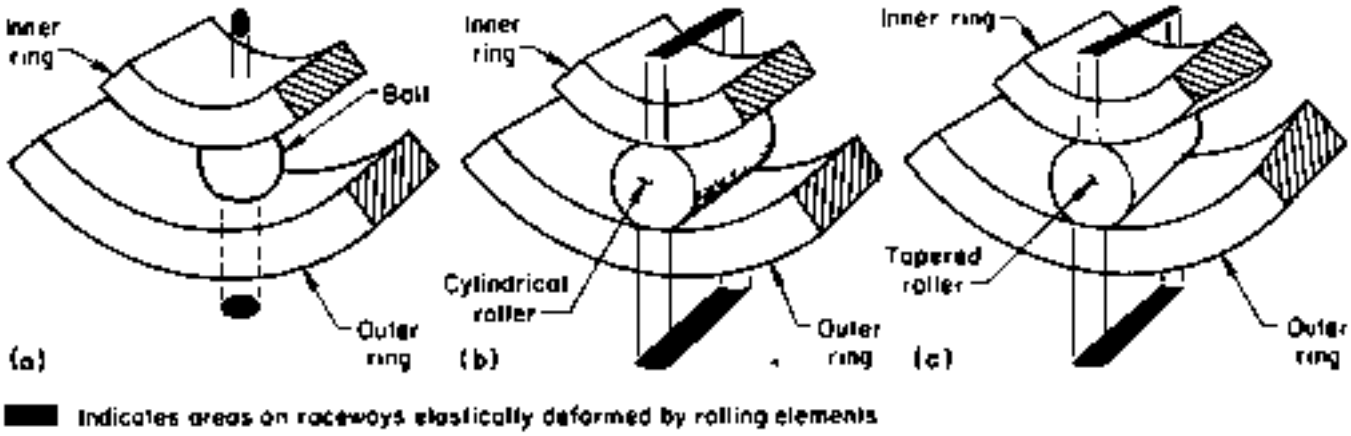


Fig. 30 Schematic representation of areas on raceways deformed by a ball (a), a cylindrical roller (b), and a tapered roller (c). Source: Ref 12

Elastohydrodynamic films are only a few micro-inches in thickness. Film thickness can be increased by increasing the rotational velocity of the bearing, increasing the viscosity of the oil, or improving the conformity of rolling elements to their raceways. Increasing either temperature or load decreases film thickness. Consequently, it is detrimental to bearing life to operate a bearing at too low a speed, too high a load or temperature, or with an oil of insufficient viscosity.

The surface finish of bearing elements is also an important consideration in elastohydrodynamic lubrication. In addition to the periodic dimensional deviations resulting from machining, bearing surfaces contain nonperiodic irregularities, including scratches, debris, and indentations with raised edges. Ordinarily, these features are quite small and have a statistical size distribution; they range in height from a few micro-inches to a few hundredths of a micro-inch or less. However, elastohydrodynamic films are also only a few micro-inches in thickness, and for a given film thickness, some of the surface features protrude through the film and contact the opposing surface. Under conditions of insufficient film thickness, the number of contacts per revolution is excessive, and bearing life suffers accordingly.

Temperature. Sources of heat in bearings include frictional contact between component surfaces, hysteresis losses because of cyclic stressing of the metal, and fluid friction caused by displacement and churning of the lubricant. Sources external to the bearing may also contribute to heating, as in hot-rolling mills, paper-mill dryers, and jet-aircraft engines.

Lubricating oils and greases are available for applications over a wide range of temperatures. As shown in Fig. 31, oils of higher viscosity are better suited for bearing applications in which rotational velocites are low and operating temperatures are high. Oils of lower viscosity are more suitable for bearings operated at higher speeds and lower operating temperatures. Petroleum-base oils and greases are available for both subzero and high-temperature operations. Many lubricants are formulated to function over a wide temperature range. For example, certain greases can be used at temperatures from -55 to 175 °C (-65 to 350 °F) (Ref 15). Under certain conditions, greases made with nonsoap thickeners can operate at temperatures up to 260 °C (500 °F). If oil-type lubricants are required for continuous high-temperature operation, as in jet-engine applications, specially formulated ester-base fluids may be required.

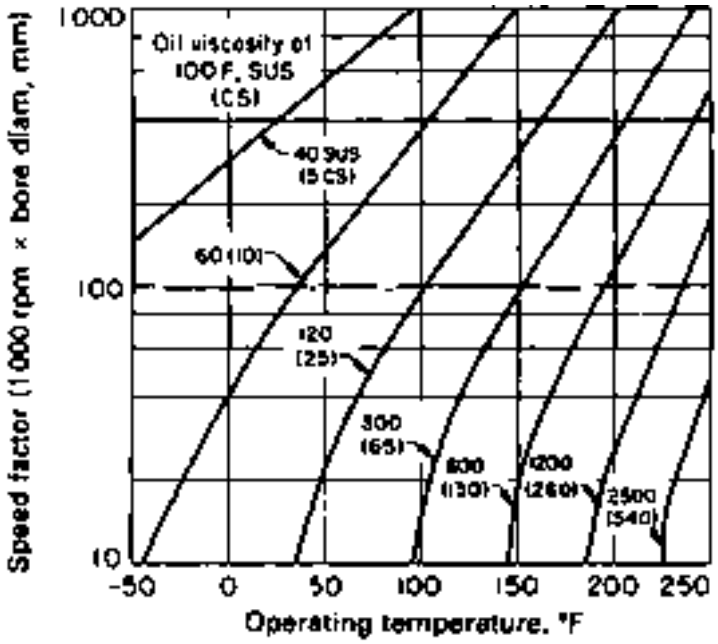


Fig. 31 Viscosities of several lubricants for ball and roller bearings, as related to operating speed and temperature. Source: Ref 14

Contamination. Contaminants that may enter a bearing include dust, dirt, fine metallic particles, water, and acid fumes. Lubricants help to keep such contaminants out of the bearing and restrict the action of those that reach the raceway and rolling-element surfaces. Hydrophobic oil films and antirust additives inhibit the effects of water contamination. Lubricating oils may also be used to flush foreign particles from the contact zone; these particles are removed from the oil by filtration.

Abrasive dust or dirt in a bearing causes wear of bearing components. The resulting increase of internal clearance causes radial and axial play, reduced accuracy and rigidity of shaft positioning, and increased vibration and noise. Furthermore, metallic particles and other hard debris can become embedded in contact surfaces. Repeated impact of rolling elements on these particles causes dents and scratches that deteriorate the surface finish. The resulting raised or sharp edges produce highly localized stress concentrations that can lead to early fatigue failure. Adequate lubrication minimizes the effects of such abrasive particles by establishing films of sufficient thickness to allow easy passage of the contaminants between bearing surfaces.

Corrosion can result if condensation occurs in an idle bearing, or if water enters the housing during operation. Rust particles can flake off the components, forming surface pits; in addition, an abrasive lapping compound may be formed by mixing of the particles with the lubricant. Water contamination can also cause corrosion fatigue, stress-corrosion cracking (SCC), galvanic corrosion, and hydrogen embrittlement of high-strength bearing steels (Ref 16). Corrosion from other sources, such as acid fumes, can produce effects similar to those caused by water. Although corrosion inhibitors are included in most lubricant formulations, moisture and other corrosive elements must be excluded from the bearing and the lubricating system.

Housings of open-type bearings are provided with shaft seals to control the flow of lubricant. Outward flow of a lubricant aids in preventing dirt, water, and other contaminants from entering the bearing. Shaft seals may be of the clearance or rubbing-contact type. Seals must be examined periodically to inspect for excessive wear and to evaluate effectiveness in preventing contamination. Slingers can sometimes be installed on shafts for centrifugal removal of water or other liquids.

Example 10: Failure of Bearings Because of Wear of Labyrinth Seals. A large number of drive-shaft hanger bearings failed after 300 to 400 h in service. The bearings normally lasted 600 h or more. The shaft was made of aluminum 2024-T3 tubing 2.5 cm (1 in.) in diameter and 1.2 mm (0.049 in.) in wall thickness. The shafts were supported by labyrinth-sealed single-row radial ball bearings of ABEC-1 tolerances. The bearings had spot-welded two-piece retainers and were lubricated with 1.5 ± 0.2 g of a paste-type mineral-oil lubricant or a grease conforming to MIL-G-81322. The lubricant contained molybdenum disulfide and polytetrafluoroethylene particles having a size of about 5 μm. The grease contained a thickening agent and synthetic hydrocarbons.

Investigation. Examination of the failed bearings revealed that the rings were blackened, deformed, and smeared. The balls were embedded in the inner-ring raceway, which had been softened by the elevated temperatures reached during the failure. The retainers were broken, and the seals were worn and bent out of shape.

Examination indicated that the following factors contributed to the failures:

- Insufficient lubrication
- Contamination of the bearings by gritty particles (dirt)
- Intrusion of a corrosive agent (water)
- Corrosion pitting of rings and balls
- Contact-fatigue mechanisms

Bearings that had been in service about 300 h were examined. These bearings exhibited light corrosion pits throughout the rings and balls. However, pitting occurred primarily in the areas of contact between balls and raceways. Areas surrounding these pits contained evidence of contact fatigue. The bearings were blackened by the high temperatures, but the amount of coke residue or burned lubricant was minimal. The seals exhibited wear, which had permitted penetration of gritty particles; water and other corrosive agents; and leakage of lubricant out of the bearing.

Maintenance reports contained notes regarding grease slinging by these bearings. This was substantiated by the dry conditions of the failed bearings.

New bearings were tested under simulated environmental conditions. Half of these bearings had labyrinth seals and were lubricated with the mineral-oil paste; the remainder had a positive rubbing seal and the MIL-G-81322 grease. The bearings with the positive rubbing seals and MIL-G-81322 grease lubricant had 30 to 100% longer life than those with the labyrinth seals and mineral-oil-paste lubricant.

Conclusions. Wear of the labyrinth seals permitted the lubricant to flow out of the bearing and dirt and corrosive agents to enter, resulting in overheating.

Corrective Measures. Bearings containing a positive rubbing seal and a MIL-G-81322 grease lubricant were installed. These bearings had satisfactory life.

In certain applications, contamination problems require a change in the method of lubrication. For example, where bath oiling is used and considerable trouble is encountered with dirt, it may be advisable to change to circulation oiling. The lubricant can then be cleaned by passing it through a full-flow filter. Where dust, moist air, or acid fumes are a problem, changing to an oil-mist system may be beneficial. This type of system can maintain clean, dry air under positive pressure in the housing and thus prevent contaminants from entering. In all instances, passages for supply and drainage of oil must be free of dirt and other clogging debris.

Effective Oil and Grease Lubrication. General guides for determining if oil or grease should be used as a lubricant are listed in Table 3. Lubricating oils must be of suitable viscosity for the particular application to minimize the frequency of contact between the bearing surfaces, to reduce frictional heating from all sources, and to protect against wear. Adequate film strength is necessary to resist the wiping action between components, particularly in tapered-roller bearings carrying heavy axial and radial loads. Lubricating oils should have the highest possible resistance to oxidation to prevent the formation and accumulation of sludge during long periods of operation. Corrosion inhibitors are needed to provide maximum protection when corrosive agents are present.

Table 3 Guides for choosing between grease and oil lubrication for bearings

Operating condition	Grease	Oil
Temperature	Below 93–121 °C (200–250 °F)	Above 93–121 °C (200–250 °F)
Speed factor (bore, mm, × rpm)	Below 200,000–300,000	Above 200,000–300,000
Load	Low to moderate	High, where oil is necessary for cooling
Long periods without attention	Yes	No, rate of leakage and other losses higher
Central oil supply for other machine elements	No	Yes
Dirty conditions	Yes, if proper design seals out contaminants	Yes, if circulated and filtered
Lowest torque	No, soap structure offers resistance	Yes, except excessive amount of oil and churning should be avoided. Use oil mist.

Source: Ref 14

Greases suitable for bearing applications should be of high quality to last for long periods of service. In addition to the requirements listed for oils, greases should have proper structural stability to resist both softening and stiffening in service and should have the appropriate consistency for the method of application. Greases must have the proper slump characteristics at operating temperature, or controlled oil-bleeding properties, for adequate penetration into the small clearances between separators and raceways. Greases used in permanently packed bearings must be exceptionally stable.

Although the housings of some bearings are packed by hand, most have grease fittings. Except when it is necessary to exclude water by complete filling with grease, no housing should be packed so full that the constant churning by the rolling elements causes unnecessary lubricant friction and high operating temperatures. Excessive lubrication, like lubricant starvation of bearings, is detrimental to optimum performance and long bearing life.

Footnote

* This section is a condensation of an unpublished article by C.J. Polk, Mobil Research and Development Corp.

References cited in this section

12. *Anti-Friction Bearings and Their Lubrication,* Technical Bulletin 3-92-002, Mobil Oil Corp., New York, 1963

13. T.E. Tallian, Rolling Contact Failure Control Through Lubrication, *Proc. Inst. Mech. Eng. (London),* Vol 182 (Pt 3A), 1967–68, p 205

14. D.F. Wilcock and E.R. Booser, *Bearing Design and Application,* 1st ed., McGraw-Hill, 1957

15. Rolling Element Bearings, *Mach. Des.,* Vol 44, June 22, 1972, p 59–75

16. I.M. Felsen, R.W. McQuaid, and J.A. Marzani, Effect of Seawater on the Fatigue Life and Failure Distribution of Flood-Lubricated Angular Contact Ball Bearings, *Am. Soc. Lubr. Eng. Trans.,* Vol 15, 1972, p 8–17

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

References

1. H.N. Kaufman, Classification of Bearing Damage, in *Interpreting Service Damage in Rolling Type Bearings—A Manual on Ball and Roller Bearing Damage,* American Society for Lubrication Engineers, Chicago, April 1958

2. F.R. Hutchings, “A Survey of the Causes of Failure of Rolling Bearings,” Technical Report Vol VI, British Engine Boiler & Electrical Insurance Co., Ltd., Manchester, 1965, p 54–75

3. R.K. Allan, *Rolling Bearings,* Sir Isaac Pitman & Sons, Ltd., 1945

4. R.L. Widner and W.E. Littmann, *Bearing Damage Analysis,* NBS 423, National Bureau of Standards, Washington, April 1976

5. A.G. Lundberg and A. Palmgren, Dynamic Capacity of Roller Bearings, *Acta Polytech.,* No. 210, 1952

6. A. Mendelsen, *Plasticity: Theory and Applications,* Macmillan, 1968

7. M.R. Hoeprich and R.L. Widner, “Environmental Factors and Bearing Damages,” SAE Paper 800678, Society of Automotive Engineers, Warrendale, PA, April 1980

8. J.L. Chevalier and E.V. Zaretsky, *Effect of Carbide Size, Area, and Density on Rolling-Element Fatigue,* NASA TN D-6835, National Aeronautics and Space Administration, Washington, June 1972

9. T.L. Carter, E.V. Zaretsky, and W.J. Anderson, *Effect of Hardness and Other Mechanical Properties on Rolling-Contact Fatigue Life of Four High-Temperature Bearing Steels,* NASA TN D-270, National Aeronautics and Space Administration, Washington, March 1966

10. E.V. Zaretsky, R.J. Parker, and W.J. Anderson, Component Hardness Difference and Their Effect on Bearing Fatigue, *J. Lubr. Technol. (Trans. ASME),* Jan 1967, p 47

11. J. Chevalier, M. Dietrich, and E. Zaretsky, *Short-Term Hot Hardness Characteristic of Rolling-Element Steels,* NASA TN D-6632, National Aeronautics and Space Administration, Washington

12. *Anti-Friction Bearings and Their Lubrication,* Technical Bulletin 3-92-002, Mobil Oil Corp., New York, 1963

13. T.E. Tallian, Rolling Contact Failure Control Through Lubrication, *Proc. Inst. Mech. Eng. (London),* Vol 182 (Pt 3A), 1967–68, p 205

14. D.F. Wilcock and E.R. Booser, *Bearing Design and Application,* 1st ed., McGraw-Hill, 1957

15. Rolling Element Bearings, *Mach. Des.,* Vol 44, June 22, 1972, p 59–75

16. I.M. Felsen, R.W. McQuaid, and J.A. Marzani, Effect of Seawater on the Fatigue Life and Failure Distribution of Flood-Lubricated Angular Contact Ball Bearings, *Am. Soc. Lubr. Eng. Trans.,* Vol 15, 1972, p 8–17

Failures of Rolling-Element Bearings

Revised by Ronald L. Widner, The Timken Company

Selected References

- E.N. Bamberger, Effect of Materials—Metallurgy Viewpoint, in *Interdisciplinary Approach to the Lubrication of Concentrated Contacts,* NASA SP-237, National Aeronautics and Space Administration, Washington, July 1969, p 409
- H.R. Bear and R.H. Butler, *Preliminary Metallographic Studies of Ball Fatigue Under Rolling-Contact Conditions,* NASA TN3925, National Aeronautics and Space Administration, Washington, 1957
- J.T. Burwell, Jr., Survey of Possible Wear Mechanisms, *Wear,* Vol 1, 1957–58, p 119–141
- R.E. Cantley, "The Effect of Water in Lubricating Oil on Bearing Fatigue Life," Preprints 76-AM-7B-1, American Society of Lubrication Engineers, Chicago, May 1976
- T.L. Carter, R.H. Butler, H.R. Bear, and W.J. Anderson, Investigation of Factors Governing Fatigue Life with the Rolling-Contact Fatigue Spin Rig, *Am. Soc. Lubr. Eng. Trans.,* Vol 1 (No. 1), 1958, p 23–32
- D.V. Culp and R.L. Widner, "The Effect of Fire Resistant Hydraulic Fluids on Tapered Roller Bearing Fatigue Life," SAE Paper 770748, Society of Automotive Engineers, Warrendale, PA, Sept 1977
- D.V. Culp and J.D. Stover, "Bearing Fatigue Life Tests in a Synthetic Traction Lubricant," Preprint 75-AM-1B-1, American Society of Lubrication Engineers, Chicago, May 1975
- C.H. Danner, Fatigue Life of Tapered Roller Bearings Under Minimal Lubricant Films, *Trans. ASLE,* Vol 13, 1970
- W.J. Davies and K.L. Day, Surface Fatigue in Ball Bearings, Roller Bearings and Gears in Aircraft Engines, in *Fatigue in Rolling Contact,* Institute of Mechanical Engineers, London, 1963, p 23–40
- T.A. Harris, Predicting Bearing Performance Under Nonstandard Operating Conditions, *Mach. Des.,* Aug 17, 1967, p 158–162
- J.J.C. Hoo, *Rolling Contact Fatigue Testing of Bearing Steels,* STP 771, ASTM, Philadelphia, 1982
- P.L. Hurricks, The Mechanism of Fretting—A Review, *Wear,* Vol 15, 1970, p 389–409
- R.F. Jacobs, Sleeve-Bearing Fatigue, *Mach. Des.,* Dec 19, 1963, p 134–139
- C.F. Jatczak, Specialty Carburizing Steels for Elevated Temperature Service, *Met. Prog.,* Vol 113 (No. 4), April 1978

- N.H. Kaufman and H.O. Walp, *Interpreting Service Damage in Rolling Type Bearings—A Manual on Ball and Roller Bearing Damage,* American Society of Lubrication Engineers, Chicago, 1953
- B.W. Kelly, Lubrication of Concentrated Contacts—The Practical Problem, Interdisciplinary Approach to the Lubrication of Concentrated Contacts, NASA symposium held at Rensselaer Polytechnic Institute, Troy, NY, July 15–17, 1969
- R.L. Leibensperger and T.M. Brittain, Shear Stresses Below Asperities in Hertzian Contact as Measured by Photoelasticity, *J. Lubr. Technol.,* July 1973, p 277–286
- W.E. Littmann, R.L. Widner, J. Stover, and J.O. Wolfe, The Role of Lubrication in the Propagation of Contact Fatigue, *J. Lubr. Technol.,* Vol 90 (No. 1), Jan 1968, p 89–100
- W.E. Littmann and C.A. Moyer, "Competitive Modes of Failure in Rolling Contact Fatigue," Preprint 620A, Society of Automotive Engineers, Warrendale, PA, Jan 1963
- W.E. Littmann and R.L. Widner, Propagation of Contact Fatigue from Surface and Subsurface Origins, *J. Basic Eng.,* Vol 88 (No. 3), Sept 1966, p 624–636
- W.E. Littmann, The Mechanism of Contact Fatigue, in *Interdisciplinary Approach to the Lubrication of Concentrated Contacts,* NASA SP-237, National Aeronautics and Space Administration, Washington, July 1969
- T.W. Morrison, T. Tallian, H.O. Walp, and G.W. Baile, The Effect of Material Variables on the Fatigue Life of AISI 52100 Steel Ball Bearings, *Trans. ASLE,* Vol 5 (No. 2), Nov 1962, p 347–364
- C.A. Moyer and H.R. Neifert, A First Order Solution for the Stress Concentration Present at the End of Roller Contact, *Trans. ASLE,* Vol 6, 1963, p 324
- C.A. Moyer, "The Use of Elastohydrodynamic Lubrication in Understanding Bearing Performance," Preprint 710733, Society of Automotive Engineers, Warrendale, PA, Sept 1971
- R. Pederson and S.L. Rice, Case Crushing of Case Carburized and Hardened Gears, *Trans. SAE,* Vol 68, 1960, p 187
- J.C. Skurka, Elastohydrodynamic Lubrication of Roller bearings, *J. Lubr. Technol.,* April 1970, p 281–291
- F.F. Simpson and W.J.J. Crump, Effect of Electrical Currents on the Life of Rolling-Contact bearings, in *Lubrication and Wear Convention,* Institution of Mechanical Engineers, 1963, p 299–304
- F.F. Simpson, Failure of Rolling Contact bearings, in *Lubrication and Wear, Second Convention,* Institution of Mechanical Engineers, Vol 179 (Part 3D), 1965, p 248
- J.O. Smith and C.K. Liu, Stresses Due to Tangential and Normal Loads on an Elastic Solid with Application to Some Contact Stress Problems, *J. Appl. Mech.,* June 1953, p 157–165
- K. Sugino, K. Miyamoto, M. Nagumo, and K. Aoki, Structural Alterations of Bearing Steels under Rolling Contact Fatigue, *Trans. Iron Steel Inst. Jpn.,* Vol 10 (No. 2), 1970, p 98–111
- T.E. Tallian, On Competing Failure Modes in Rolling Contact, *Trans. ASLE,* Vol 10, 1967, p 418–439
- T.E. Tallian and J.I. McCool, An Engineering Model of Spalling Fatigue Failure in Rolling Contact, II. The Surface Model, *Wear,* Vol 17, 1971, p 447–461
- N.M. Wickstrand, Depth of Permanent Indentations in Flat Plates Due to Loaded Cylindrical Rollers, *J. Lubr. Technol.,* Jan 1970
- R.L. Widner and J.O. Wolfe, "Analysis of Tapered Roller Bearing Damage," Technical Report C7.11.1, American Society for Metals, Oct 1967
- R.L. Widner, W.K. Dominik, and A.J. Jenkins, "High Performance Bearings," SAE Paper 780784, Society of Automotive Engineers, Warrendale, PA, Sept 1978
- D.C. Witte, Operating Torque of Tapered Roller Bearings, *Trans. ASLE,* Vol 16 (No. 1), Jan 1973, p 61–67
- H. Zantopulos, The Effect of Misalignment on the Fatigue Life of Tapered Roller Bearings, *J. Lubr. Technol.,* Vol 94 (No. 2), April 1972, p 181–187
- C.V. Zaretsky and J.W. Anderson, Effect of Materials—General Background, in *Interdisciplinary Approach to the Lubrication of Concentrated Contacts,* NASA SP-237, National Aeronautics and Space Administration, Washington, July 1969, p 379
- E.V. Zaretsky, W.J. Anderson, and R.J. Parker, *Effect of Nine Lubricants on Rolling-Contact Fatigue Life,* NASA TN D-1404, National Aeronautics and Space Administration, Washington, Oct 1962

Failures of Lifting Equipment

Frank L. Jamieson, Stelco Inc. (retired)

Introduction

LIFTING EQUIPMENT is used for raising, lowering, and transporting materials, parts, and equipment, generally within a limited area. The types of metal components used in lifting equipment include gears, shafts, drums and sheaves, brakes and brake wheels, couplings, bearings, wheels, electrical switchgear, chains, steel wire rope, and hooks. This article will primarily deal with many of these metal components of lifting equipment in the following three categories:

- Cranes and bridges, particularly those for outdoor and other low-temperature service
- Attachments used for direct lifting, such as hooks, chains, wire rope, slings, beams, bales and trunnions, and the members to which they are attached, such as lifting lugs or eyes
- Built-in members that are the items necessary for the operation of lifting equipment, such as shafts, gears, and drums

Most of the failures discussed are related to the more common and critical components of lifting equipment used in steel mills and similar industrial applications, but the problems encountered and the methods of analysis are the same as for lifting equipment used in other industries.

Failure Mechanisms. Failures of lifting equipment commonly result from fatigue, ductile fracture, brittle fracture, and wear. A part may fail from any one of these mechanisms or from a combination of two or more of them. A member may initially fracture by fatigue, with final fracture being of a ductile or brittle nature. Total brittle fracture of a part is most undesirable, because it is unpredictable and often has catastrophic results. Periodic inspection of a part that is susceptible to brittle fracture is often of little purpose. It is preferable for fracture of a component to begin by fatigue rather than suddenly in a brittle manner. Properly timed inspections will usually reveal a slowly propagating crack, and the equipment can be removed from service before failure occurs.

Wear is easily recognized; excessive wear can usually be corrected by changing the material or its processing. However, the complete elimination of wear of lifting equipment may require selection of a material that is subject to brittle fracture—an alternative that may well be unacceptable.

Failure Origins. Failures occur for a number of reasons, which may be related to operation, design, material selection, material quality, and manufacturing practices. The largest portion of lifting-equipment failures are of operational origin. Overloading of a lifting mechanism is a common practice and often leads to either ductile or brittle fracture or to fracture by fatigue from repeated overstressing. Figure 1 shows a steel wire rope that fractured in tension due to overloading. With a fatigue type of failure, the member will usually fail at a load well below its specified load limit; thus, a false sense of security exists when dealing with a member that is fracturing by fatigue. Periodic inspection of such parts is therefore very important. Items that show signs of excessive wear or abuse should be removed from service for replacement or repair. Figure 1 shows the end of a steel wire rope that failed in tension because of overloading. Excessive wear on a 16-mm ($\frac{5}{8}$-in.) diam 6 × 37 (6 strands of 37 wires each) fiber-core improved plow steel wire rope is described in Example in this article. Corrosion failures are also relatively frequent because of the environments in which lifting equipment is operated.

Fig. 1 End of a steel wire rope that failed in tension because of overloading. Necking at the ends of the wires indicates ductile fracture; no worn or abraded areas were found at the break.

Section changes, keyways, lubrication holes, rough-machining marks, and threaded sections are often points of initiation of cracks. Small fillets, sharp corners, grooves, and threads act as stress raisers from which fatigue cracks initiate and fractures propagate. Correct design of a threaded shank on a hook, for example, and the stress-relief groove at the base of a threaded shank are very important to failure prevention. Failures have also occurred where names, numbers, and other identification marks are die stamped or imprinted on highly stressed surfaces. Information on the effects of identification marks is provided in the articles “Fatigue Failures” and “Failures of Shafts” in this Volume.

Another common cause of lifting-equipment failures is improper material selection. The choice of a material is related to such aspects as its basic quality, composition, overall mechanical properties, notch toughness, corrosion resistance, weldability, and machinability. In most applications, many materials may be satisfactory for a particular part, but only a few materials will be optimum.

In hardenable steels, tempered martensite has greater fatigue resistance than mixed structures. From a practical and theoretical aspect, alloy steels are a better material selection than carbon steels because of the deeper-hardening characteristics of alloy steels.

The most encompassing cause of failures of lifting equipment is poor manufacturing, assembly, and maintenance practices. Inferior machining, defective welds, residual welding stresses, misalignment, and improper and insufficient lubrication are all common and critical causes of lifting-equipment failures. Metallurgically, the most common cause of failures is improper heat treatment. Quench cracks and residual heat-treat stresses contribute to such failures. The examples in this article illustrate many of these causes.

Investigation of Failures. The investigation of a failure can provide valuable information that can be used for design improvement, material selection, and other factors related to the efficient operation of lifting equipment. Much information can be obtained by visual examination. Close, careful visual inspection should always precede any technique requiring destruction or cutting of a failed part.

A frequent error in investigating a failure is to interpret it automatically on the basis of the most commonly known causes indicated from past experience. Although this cannot be avoided entirely, efforts should be made to be alert for other or unique details that may be contributory or related factors and that may significantly alter the interpretation.

In the investigation of a fatigue failure, the following should be considered: material, design of the entire machine as well as that of the failed part, fabrication practice (machining, welding, forging, or casting), heat treatment, types of loading to which the equipment is subjected, and operating environment, such as corrosive or high-temperature atmospheres.

Many failures may be the result of a combination of several faults or contributing factors. Correcting one fault may not be sufficient or may even have a strongly adverse effect in the presence of other uncorrected faults. It is important that the investigator sort out the contributory causes of failure from the associated but immaterial aspects of the failed part to determine the proper corrective measures. The techniques for conducting failure analyses are provided in the article “Conducting a Failure Examination” in this Volume.

Frank L. Jamieson, Stelco Inc. (retired)

Materials for Lifting Equipment

Recommendations given here are for materials commonly used in the manufacture of lifting equipment for the steel industry; other industries may use different materials as required by the application. The recommendations are general; thus, each situation must be considered individually. Transition temperatures of the steels used in the equipment should be considered in all cases, particularly for applications that involve temperatures below 20 °C (68 °F). For a discussion of ductile-to-brittle transition, see the article "Mechanisms and Appearances of Ductile and Brittle Fracture" in this Volume.

Chains, Chain Hooks, and Fittings. Low-carbon steel chains are in wide use and are generally acceptable. However, to ensure a greater margin of safety, alloy steel chains with a hardness of 302 to 352 HB are recommended.

Chain hooks are usually made of a grade of steel similar to that of the chain assembly, except in special cases. The design of chain hooks is often characteristic of the manufacturer rather than the application. Materials for chains for use in acid or other corrosive environments must be specified for the particular application.

Shafts. Misalignment of shafts resulting in failure is a common occurrence and is difficult to eliminate when dealing with large equipment. Frequently, 1040 or 1045 steel is specified for shafts. Failures in shafts made of these steels can often be minimized by heat treatment. In other cases, it may be necessary to make shafts from quenched-and-tempered alloy steels, such as 4140 or 8640, which have excellent fatigue properties. For severe applications and large section sizes, 4340 or a comparable high-alloy steel may be used with excellent results. More information is available in the article “Failures of Shafts” in this Volume.

Bridges and Cranes. The use of a steel having a high impact resistance at low temperatures is required for the prevention of brittle fracture of structures, especially welded ones, that are subjected to cold-temperature conditions in service. Use of a fine-grain, normalized, relatively low-carbon (0.20% C) steel will generally ensure adequate impact resistance and freedom from high hardness in the heat-affected zone (HAZ) of welds. Fully quenched-and-tempered low-carbon steels and similarly treated alloy steels are also used for such applications.

Miscellaneous Equipment. Lifting booms, coil hooks, lugs and rings, boxes and containers, trunnions, and miscellaneous attachments should be made of either normalized or quenched-and-tempered fully killed fine-grain steels. For equipment operated at temperatures below 20 °C (68 °F), the transition temperature of the material must also be considered; the material selected should have a transition temperature below the operating temperature. When the weight of these components is a limiting factor, plain carbon steels may not be suitable. A high-strength low-alloy steel can be used instead, the grade and fabrication depending on the requirements of the specific application.

Flame-cut parts and welded components, insofar as possible, should be stress relieved at 600 to 650 °C (1100 to 1200 °F); flame cutting may produce hard edges that are deleterious to machining and may act as stress raisers in a weldment. Nondestructive inspection of welded assemblies is highly recommended. In the fabrication of a member, the material should be oriented so that the rolling direction will be parallel to the direction of principal service stress.

Bolts, nuts, and washers should be made in accordance with ASTM A 325 (Ref 1) or other recognized specifications for applications in which they are used only as fasteners. Pins and bolts used as pivots should be of fully killed fine-grain steel of compositions that will satisfy processing and application requirements. Rivets should be made in accordance with ASTM A 502 (Ref 2). Particular designs or service conditions may require the use of heat-treated alloy steels.

References cited in this section

1. “Standard Specification for High-Strength Bolts for Structural Steel Joints,” A 325, *Annual Book of ASTM Standards,* ASTM, Philadelphia

2. “Standard Specification for Steel Structural Rivets,” A 502, *Annual Book of ASTM Standards*, ASTM, Philadelphia

Frank L. Jamieson, Stelco Inc. (retired)

Steel Wire Rope

Many factors must be considered when investigating a failure of steel wire rope used on lifting equipment. Environment is of obvious concern because wire rope may be subjected to corrosive atmospheres, such as water, acid, and various chemicals, or to elevated temperatures. Another important factor that must be considered is the flexibility of wire rope. When sheaves or drums are involved, and they almost always are, the diameter of the sheave or drum has a direct relationship to the flexibility of the wire rope. The degree of flexibility of a wire rope is determined by its construction. Figure 2 illustrates the component parts of a wire rope.

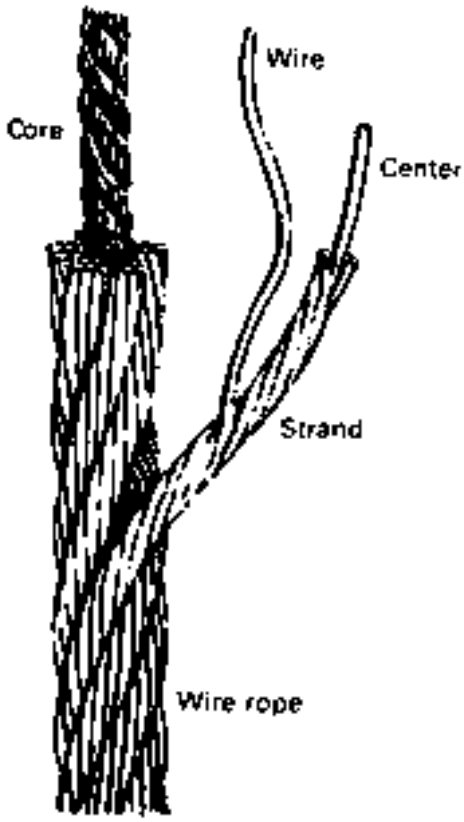


Fig. 2 Components of a steel wire rope. Source: Ref 3

Strength and Stretch. The ultimate breaking strength of a wire rope is by design less than the aggregate strength of all the wires and will vary, depending on the construction of rope and grade of wire used. All manufacturers of wire rope publish catalogs that list minimum breaking strengths for the various sizes, constructions, and grades of ropes. The proper design factor for a wire rope demands consideration of all loads. These loads should include (when applicable) acceleration, deceleration, rope speed, rope attachments, number and arrangement of sheaves and drums, conditions producing corrosion and abrasion, and length of rope.

Cold-drawn high-carbon steel wire, the type generally used in wire rope, has a modulus of elasticity of approximately 1.93 to 1.99 × 10^5 MPa (28,000 to 29,000 ksi). Modulus of elasticity is the measure of the degree to which the wire will stretch under increasing load. The amount of rope stretch caused by the relative movement of all the wires in their attempt to adjust their positions to a stable condition corresponding to the load imposed varies with the construction of the rope.

Generally, the more flexible ropes, which contain the greatest number of wires and have fiber cores, will stretch more than all-metal ropes with fewer wires and therefore less flexibility.

Stretch resulting from the movement of the wires is of two types: constructional and elastic. When a rope is first placed under load, the resulting slight rearrangement of the wires will cause a permanent lengthening, known as constructional stretch, and a recoverable lengthening, which is elastic stretch.

The constructional stretch, depending to some degree on the magnitude of the load imposed, amounts to approximately 0.5 to 0.75% of the length of the rope for six-strand ropes with fiber cores. For six-strand ropes with steel cores, the corresponding value is 0.25 to 0.5%; for eight-strand ropes with fiber cores, 0.75 to 1%.

The preceding remarks apply to bright ropes of a type intended to operate over sheaves and drums, and it is assumed that the ends of the rope are restricted against rotation, are free from corrosion, and have not deteriorated appreciably either internally or externally. Corrosion and other kinds of rope deterioration have a marked effect on stretch properties.

Sheaves. The size of a sheave is usually expressed in terms of its diameter measured at the base of the groove. This is its tread diameter. It can also be expressed as a ratio determined by dividing the tread diameter by the nominal rope diameter.

As sheave size is decreased, the stresses resulting from bending and the contact pressure between rope and sheave are increased. Higher bending stresses cause more rapid fatigue of the wires in the rope. Increased pressure also accelerates rope deterioration and, at the same time, increases sheave wear.

As sheave size is increased, rope-to-sheave pressure is decreased, and bending becomes easier. If bending alone were involved, an increase in rope life could obtained by increasing sheave size up to a limit of about 90 to 100 times the rope diameter for a 6 × 19 (6 strands of 19 wires each) classification rope. However, except on some shaft hoist installations, sheaves this large are seldom in actual use for two reasons. First, bending is seldom, if ever, the only factor involved. Large, expensive sheaves would be unnecessary if the abrasion or scrubbing encountered by the rope on some other part of the installation were the determining factor in its ultimate service life. Second, most machines would not be practical with such large sheaves, and many would never have been designed if sheaves of this size were mandatory.

In practice, many factors other than bending influence rope life, such as repeated stressing, abrasion, pounding, impact, vibration, twisting, speed, drum-winding abuse, corrosion, and lack of maintenance. For many applications, one or more of these factors affects rope life more than sheave size alone.

The optimum sheave size is determined by evaluating the factors affecting safe, economical operation. On high-speed mobile-type equipment and on equipment requiring low initial cost, it is customary to use smaller sheaves. The need for more frequent rope replacements is outweighed by the speed and flexibility of movement provided by the more compact equipment.

Although no definite minimum sheave size can be established for all types of installations, one important factor must not be overlooked. The heavy pressures between the wires at the contact points in the rope, combined with the high bending stresses resulting from operation over small sheaves, have a definite effect on rope deterioration. Under such adverse conditions, wire breakage often occurs at and between the points where the strands contact each other and contact the core of the rope. Broken wires in these sections are difficult, and sometimes impossible, to detect.

Therefore, for applications in which a high degree of safety is essential, sheaves should be liberal in size to provide better assurance that the wires will deteriorate progressively on the surface of the rope where they can be readily seen and evaluated. Extremely small sheaves should be limited to those types of equipment on which such usage is in line with acceptable practice.

Sheave size is such an important factor in wire-rope life that many laboratory bending or fatigue tests have been conducted to evaluate this relationship. Laboratory bending life can be determined; however, field service life can seldom be judged by these results. There are too many factors existing under actual field conditions to be simulated under controlled laboratory tests.

Laboratory tests have their value for research and development, particularly to the wire-rope engineer. It may also be important at times to show how bending alone, as performed in such tests, affects wire rope.

One series of tests was made on 16-mm ($\frac{5}{8}$-in.) diam wire-rope specimens in the 6 × 7, 6 × 19, 6 × 37, and 8 × 19 classifications. The sheaves ranged from 20 to 50 cm (8 to 20 in.) in diameter. The load approximated a design factor of five, and specimens were operated until they broke. The number of cycles of operation was the measure of bending life.

The laboratory tests showed that as the ratio of sheave diameter to rope diameter was reduced, there was a reduction in bending life, with the reduction in bending life showing the same trend for all specimens tested, and that the reduction in bending life was faster than the reduction of the sheave-to-rope diameter ratio.

Sheave Grooves. To allow wire rope to perform the maximum amount of useful work, it is essential that sheave grooves be of sufficient diameter to provide and maintain proper rope clearance. All wire rope, when new, is slightly oversize to allow for some pulling down while becoming adjusted to its normal working tensions.

Table 1 lists the recommended groove clearances for the usual sheave applications. In some cases, departures from these recommendations are advisable. For example, with some deflectors or equalizers, smaller clearances can be used, or with conditions involving large fleet angles, larger clearances may be necessary.

Table 1 Groove clearance in sheaves based on nominal wire-rope diameter

Nominal rope diameter		Groove clearance[(a)]	
mm	in.	mm	in.
6.4–7.9	$\frac{1}{4}-\frac{5}{16}$	0.4	$\frac{1}{64}$
9.5–19	$\frac{3}{8}-\frac{3}{4}$	0.8	$\frac{1}{32}$
20.6–28.6	$\frac{13}{16}-1\frac{1}{8}$	1.2	$\frac{3}{64}$
30–38	$1\frac{3}{16}-1\frac{1}{2}$	1.6	$\frac{1}{16}$
40–57	$1\frac{9}{16}-2\frac{1}{4}$	2.4	$\frac{3}{32}$
59–76	$2\frac{5}{16}-3$	3.2	$\frac{1}{8}$
<76	<3	4.0	$\frac{5}{32}$

(a) Recommended groove diameter equals nominal rope diameter plus groove clearance.

The diameter of a sheave groove does not remain constant. Its change is influenced by the rope and, to a great extent, by the sheave material and the pressures involved. Because a groove of the recommended size becomes worn to a smaller diameter by the time a rope is taken out of service, the question arises as to whether or not a sheave should be regrooved each time a new rope is installed. In theory, best service will be obtained when a new rope starts with all sheave grooves exactly proper for the conditions involved. In practice, it would be false economy in most installations to change sheaves with every rope replacement. In most cases, the economical procedure is to allow the new rope to take additional wear at first, even at the expense of some rope life. For most installations, the limit below which a sheave should be regrooved is a clearance value in excess of the nominal rope diameter equal to half the clearance listed in the second column of Table 1.

When a sheave is regrooved, it is generally necessary to dress the flanges and to cut the base of the groove to the proper diameter. If only the base is machined, the throat angle near the base may be left too small to provide proper clearance where the rope makes a fleet angle to one side or the other.

Rope Pressure. In addition to the bending stresses introduced by operation over sheaves, wire rope is subjected to radial pressure by its contact with the sheave. This pressure sets up shearing stresses in the wires, distorts the rope structure, and affects the rate of wear of the sheave grooves. Therefore, the magnitude of the pressure and the wear resistance of the sheave material should be considered when selecting the most suitable rope construction.

The radial pressure, in pounds per square inch of projected area of the rope, can be determined using:

$$P = \frac{2T}{Dd} \qquad \text{(Eq 1)}$$

where P is radial pressure (in pounds per square inch), T is tension in the rope (in pounds), D is tread diameter of the sheave (in inches), and d is diameter of the rope (in inches).

Just as is true for bending stresses, the magnitude of the stresses resulting from the radial pressure increases as the size of the sheave decreases. High bending stresses generally indicate the need for flexible rope constructions. However, the resulting relatively small-diameter wires have less ability to withstand heavy pressures than the larger wires in the less flexible rope constructions. Both factors should be considered in selecting the most suitable type of wire rope.

Furthermore, the pressure of the rope against the sheave tends to flatten the rope structure. This type of rope distortion can be controlled to a large extent by proper sheave-groove contour.

Contact of the rope with the sheave results in wear in the rope and the sheave groove. The rate of wear is influenced by the magnitude of the rope pressure against the sheave groove.

Sheave Material. If the sheave material has insufficient wear resistance, the grooves will wear to the diameter of the rope operating in it. When a new full-size rope is installed, the groove worn small by the previous rope will subject the new rope to unnecessary abrasion in its attempt to grind the groove to its own diameter. To avoid this condition, the sheave-groove material should be selected to resist the wear corresponding to the rope pressure that will be present.

Table 2 lists allowable radial pressures calculated from Eq 1 for three cast, annealed sheave materials. Heat treatment to improve wear resistance will permit greater pressures. The values listed in Table 2 are of necessity approximate, but do represent desirable limits for avoiding excessive groove wear. These values are for regular-lay wire rope, in which the wires are laid in the direction opposite to the twist of the strands in the rope. The values can be increased for lang-lay wire rope (wires and strands laid in the same direction).

Table 2 Allowable pressures on sheaves of three materials for various classifications of wire rope[(a)]

Rope classification	Allowable pressures on sheaves made of:					
	Cast iron		**Cast steel**		11-13Mn steel	
	Mpa	**psi**	**Mpa**	**psi**	**Mpa**	psi
6 × 7	2.1	300	3.8	550	10.3	**1500**
6 × 19	3.4	500	6.2	900	17.2	**2500**
6 × 37	4.1	600	7.4	1075	20.7	**3000**
8 × 19	4.1	600	7.4	1075	20.7	**3000**
Flattened strand	**5.5**	**800**	**10**	**1450**	**27.6**	**4000**

(a) Values are for regular-lay rope: for lang-lay rope, these values may be increased 15%, except for flattened-strand rope, which is normally lang lay

Drums. A common method of driving a wire rope is by a drum. One end of the rope is fastened to the drum, and the rope is wound and stored as the drum revolves. In some applications, both ends of the rope are attached to the drum.

Most drums are cylindrical with flanges at the ends. Flanges may not be necessary with grooved drums when the rope will never wind in more than one layer and when the rope cannot get out of the grooves. Flanges are necessary for multiple-layer winding and for cases where the rope might slip or wind off the face.

Drum size, groove contour, pressure, and drum material relate closely to the matters previously discussed on sheaves. The groove contour in grooved drums should be the same as in sheaves. The radial pressure between rope and drum on the first layer of a properly grooved drum is calculated by Eq 1. For these conditions, the same limitations apply to the materials listed in Table 2.

High pressures are exerted on the face of the drum with two or more layer winding. This has little influence on the choice of drum material because the additional pressure is applied after the first-layer wraps are fixed in position on the drum.

Higher pressures resulting from multiple-layer winding should be considered in the structural design of the drum. A drum for two-layer winding does not have to be twice as strong as that for single-layer winding, nor does one for three layers have to be three times as strong. When a second layer is wound on a drum, some tension is lost in the rope on the first layer. With the application of a third layer, additional tension is lost in the rope on the second layer. Studies have indicated that the combined rope tension with two layers is about $1\frac{3}{4}$ times the tension with one layer, with three layers about $2\frac{1}{8}$ times the tension with one and with four layers about $2\frac{1}{4}$ times the tension with one.

With plain-faced drums, the actual radial pressures are higher because the rope does not have support around part of its circumference. On drums of this type, wear to both rope and drum would be more severe than that on grooved drums, and rope deterioration, resulting from pressure, would be more pronounced.

The diameter of the drum should be influenced by considerations of safety and economy. For general conditions, the ratios for sheaves should be considered, but for some applications, departures are made from these values.

Where bending is the primary factor affecting the service life of the rope, the drum may be slightly smaller in diameter than the sheaves in the system, because in traveling in each direction the rope bends only once at the drum, whereas it bends twice at each sheave. In traveling over each sheave, the rope is bent the first time when it conforms to the curvature of the sheave and the second time when it straightens out and leaves the sheave. The number of bends is decreased on those machines in which the rope travel is not very great and the section of rope operating on and off the drum does not reach the sheaves. The reeving diagrams shown in Fig. 3 can be used to determine the number of bends in a wire rope when multiple-pulley blocks are used.

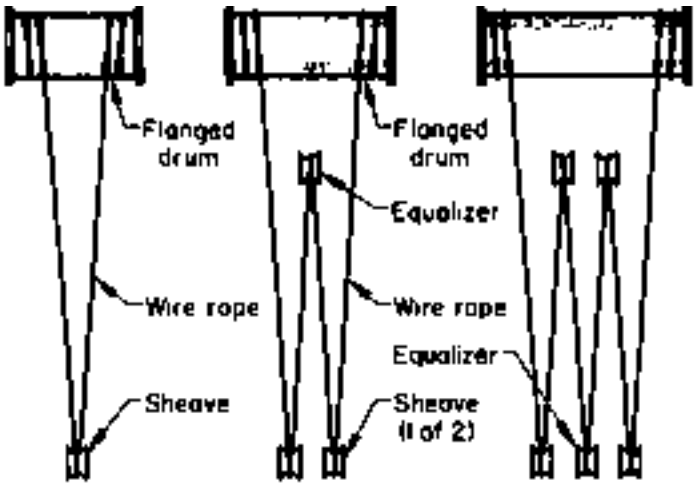


Fig. 3 Three reeving diagrams for double-drum hoists.

When wire rope is wound on a drum, it exhibits a slight rotation because of the spiral lay of the strands. Standing behind the drum and looking toward an oncoming overwind rope, rotation of a right-lay rope is toward the left, whereas rotation of a left-lay rope is toward the right. This rotation is extremely small and seldom of any significance.

However, with a plain-faced drum, in which the only other influence to the rope is winding on the first layer is the fleet angle, this slight rotation can sometimes be used to advantage in keeping the windings close and uniform. Rightlay rope is standard and the one most readily available from stock. Therefore, all machines or installations having plain-faced drums should be designed for use with right-lay rope.

With a plain-faced drum, a right-lay rope, and overwind reeving, the attachment should be made at the left flange. With a right-lay rope and underwind reeving, the attachment should be made at the right flange.

When grooved drums are used, there generally is sufficient control by the grooving to wind the rope properly whether it is right or left lay. With either an overwind or an underwind installation or a left or right flange attachment, the standard right-lay rope construction can be used. Only in special applications should a change to left-lay rope be considered. One instance would be when opposite rotation of the rope might help prevent open winding, or a piling up at the flange under adverse fleet-angle conditions.

Grooved drums offer greater control and uniformity of rope winding compared with plain-faced drums. Moreover, when several layers are involved, grooved drums also influence the winding of the second and subsequent layers, provided the change from each layer to the next is properly accomplished. Grooving also provides some degree of circumferential support for the rope, which is advantageous to drum and rope life.

Maximum support is provided by grooving that has a depth equal to half the diameter of the rope, if the contour is proper. However, it is seldom possible or advisable to make the grooves this deep. For some applications, deep grooves have another advantage: they help keep the rope winding properly where a swinging load or some other condition of abnormal displacement causes the rope to lead improperly to the drum.

Deep grooves can be a disadvantage by adding to rope abrasion and restricting freedom of movement as the rope enters the groove. In leading to the drum, there is only one point at which the lead is absolutely straight into the groove. On either side of this point, some fleet angle is encountered. This causes rope and groove wear if the rope is confined around too much of its circumference.

There is seldom any advantage in designing a groove to have more than 150° circumferential support at the base, and in most cases, it would be disadvantageous to do so. If single-layer winding is involved and if deep grooves are necessary, this can be best accomplished by spreading the wraps apart and providing higher ridges between grooves.

It is important that grooved drums be designed with the proper pitch or distance from center to center of grooves to allow ample but not excessive clearance between successive wraps of rope. This is essential to prevent crowding and scrubbing of the oncoming rope against the rope already on the drum. To provide proper conditions for multiple-layer winding, the pitch between grooves should only be enough to prevent rope contact when winding under the maximum angle of fleet.

Example 1: Bending-Fatigue Failure of a Steel Wire Hoisting Rope for a Stacker Crane. A 13-mm ($\frac{1}{2}$-in.) diam 18 × 7 fiber-core improved plow steel nonrotating wire rope broke after 14 months of service on a stacker crane. Previously, 13-mm ($\frac{1}{2}$-in.) diam 6 × 37 improved plow steel ropes with independent wire-rope cores had been used. These ropes were in service for 12 months. The change to an 18 × 7 rope was made because of difficulties caused by twisting of the 6 × 37 rope.

Investigation. Chemical analysis of the steel wire indicated a carbon content of 0.51 to 0.55%, as specified. The tensile strength of the individual wires averaged 1234 MPa (179 ksi), which was above average. The quality of the steel was satisfactory, with the crown wires showing very little wear.

The hoist arrangement for this crane consisted of one rope with each end attached to a separate drum; the rope wound around two 30-cm (12-in.) diam sheaves in the block and back up and around an equalizer sheave (Fig. 3b). The section of the rope that had been in contact with the sheaves was found by measurement checks. Reverse bending of the section of the rope normally subjected to this flexing revealed the presence of broken wire ends (Fig. 4), which indicated that the rope failed by fatigue. The minimum sheave diameter for a 13-mm ($\frac{1}{2}$-in.) diam 18 × 7 wire rope should be 43 cm (17 in.). Thus, the 30-cm (12-in.) diam sheaves were too small.

Fig. 4 13-mm ($\frac{1}{2}$-in.) diam 18 × 7 fiber-core improved plow steel nonrotating wire rope that failed in bending fatigue. The rope was operated over a sheave that was too small in diameter.

Conclusions. Failure of the wire rope occurred by bending fatigue. Continually running the same section of the rope over a sheave too small in diameter resulted in excessive bending stresses.

Corrective Measures. The sheave diameter could not be increased; therefore, the flexibility of a 6 × 37 rope was required. The 13-mm ($\frac{1}{2}$-in.) diam 18 × 7 rope was replaced by two 13-mm ($\frac{1}{2}$-in.) diam 6 × 37 steel-core ropes stranded side by side, one with left lay and the other with right lay. The twisting problem was eliminated by the use of the two counter-stranded cables.

Corrosion is another common cause of wire-rope failure. The corrosive atmospheres in which wire ropes operate are created by blast furnaces, cleaning tanks, plating tanks, and exposure to outdoor elements. Figure 5 shows corrosion failure of a steel wire rope that operated partially underwater.

Fig. 5 Steel wire rope with heavy corrosion and broken individual wires resulting from intermittent underwater service.

Shock loading of a wire rope can cause it to vibrate, producing high-frequency cyclic bending stresses in the rope. Vibration is most severe at the connection end of the rope, and can result in fatigue failure.

Example 2: Fatigue Failure of a Steel Wire Rope Resulting From Shock Loading. The wire rope on a cleaning-line crane broke while lifting a normal load of coils. This rope, which was specified for the application, was 11-mm ($\frac{7}{16}$-in.) diam 8 × 19 fiber-core rope of improved plow steel wire. Service life of the rope was 5 weeks; average expected life was 6 weeks. The rope was inspected weekly.

Investigation. Visual examination of a section of the wire rope adjacent to the fracture revealed several broken wires and fraying of the fiber core (Fig. 6a). The construction and mechanical properties of the rope were as specified.

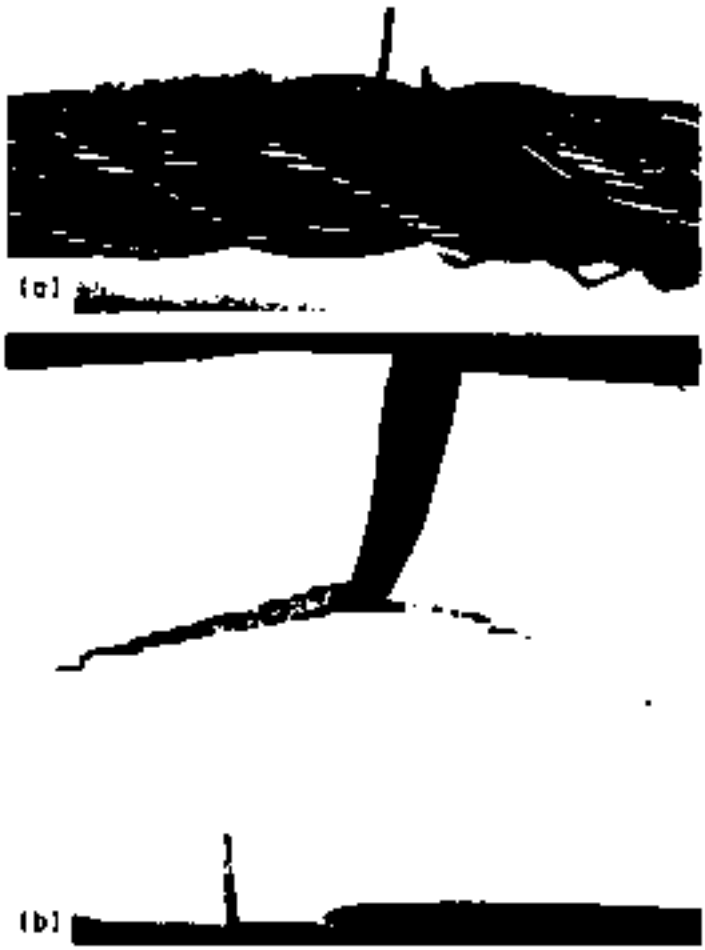


Fig. 6 Steel wire rope, used on a cleaning-line crane, that failed from fatigue resulting from vibration caused by shock loading. (a) Section of the wire rope adjacent to the fracture. Approximately $1\frac{1}{2}$×. (b) Unetched longitudinal section of a wire from the rope showing fatigue cracks originating from both sides. 75×

Metallurgical examination of several wires revealed a uniform cold-drawn microstructure with no evidence of severe abrasion or martensite. Microscopic examination of a longitudinal section of a wire revealed fatigue cracks originating from both sides of the wire (Fig. 6b). One crack changed direction and propagated parallel to the centerline of the wire. The diameter of the sheave on the bale, 27 cm ($10\frac{5}{8}$ in.), was slightly below that specified for the 11-mm ($\frac{7}{16}$-in.) diam rope.

Observation of the crane in operation revealed that in rolling the coils over the edge of the rinse tank after pickling the hook received a sudden shock load, which was transmitted to the rope, causing vibrations. The vibrations were most severe at the clamped end of the rope.

Conclusions. Failure of the rope was attributed to fatigue, resulting from vibration caused by shock loading.

Corrective Measures. Pitched roll plates were installed between the tanks where rolling of coils was required. The plates reduced the free fall of the coils and aided in rolling. Also, the diameter of the sheave was increased to 33 cm (13 in.).

Scrubbing of a wire rope against a foreign object can cause excessive heating of the crowns of the outside wires. Because these regions are small in comparison to the total area of the wire, they may be rapidly quenched to a martensitic structure by the adjacent and underlying cooler metal. Martensite has very little ductility and will readily crack under the slightest bend.

Martensite is more easily formed in rope made of the higher-carbon grades of steel wire. Two ropes were analyzed that had operated under similar conditions as main lines in a logging machine. The ropes were 3.2 cm ($1\frac{1}{4}$ in.) in diameter and of 6 × 19 independent wire rope core (IWRC) construction. One of the ropes had a carbon content of 0.685% and a manganese content of 0.607%; the other rope had a carbon content of 0.751% and a manganese content of 0.607%.

Both ropes developed martensite on the crowns of the outside wires. The martensite was caused by the friction and heat generated by the action of the rope sliding against the sides of the sheave. The outside wires were rapidly heated, then

rapidly quenched by the adjacent metal. The difference in carbon contents manifested itself in the layer of martensite formed being greater in the wires that contained the greater amount of carbon. Figure 7 shows a micrograph that illustrates the martensite layer developed by the wires having 0.685% carbon content. The ropes containing the 0.751% C wires displayed a deeper layer of martensite. The rope made of the lower-carbon wires had a life of about three times the other; however, this life was not considered satisfactory. The formation of martensite must be avoided, which means that the cause of failure was the abuse the ropes received and not necessarily the carbon content or the martensite formation.

Fig. 7 Transverse section through 2.6-mm (0.102-in.) diam steel wire. Light-etching surface layer (top) is untempered martensite; adjacent dark-etching zone is self-tempered martensite. The matrix was composed of deformed pearlite. Etched with 5% nital. 265×

Example 3: Fatigue Fracture of Individual Steel Wires in a Hoisting Rope. The wire rope on a crane in a scrapyard broke after 2 weeks of service under normal loading conditions. Expected life of the wire rope was 2 months. The rope was 16-mm ($\frac{5}{8}$-in.) diam 6 × 37 fiber-core improved plow steel. This type of rope is made of 0.71 to 0.75% C steel wires, with a tensile strength of 1696 to 1917 MPa (246 to 278 ksi).

The rope became damaged while it was attached to a chain for pulling jammed scrap from the baler. The chain broke during this operation, skipping the cable up onto the edge of the sheave. The rope then apparently seated itself in a kink on the flange of the sheave.

Investigation. As shown in Fig. 8(a), the rope was heavily abraded (scrubbed) and crushed. Also, several of the individual wires were broken.

Fig. 8 Wire rope, made of improved plow steel with a fiber core, that failed because of heavy abrasion and crushing under normal loading. (a) Crushed rope showing abraded wires and crown wear. 1.8×. (b) Nital-etched specimen showing martensite layer (top) and uniform, heavily drawn microstructure. 500×

Metallographic examination of several of the broken wires revealed a uniform cold-drawn microstructure, with patches of untempered martensite in regions of severe abrasion and crown wear.

Chemical analysis of the wire showed a carbon content of only 0.46%, whereas 0.71 to 0.75% C was specified. The tensile strength (1620 MPa, or 235 ksi) was below specification.

A hard layer of martensite (Fig. 8b) was formed on the wires as a result of abrasion. This surface layer, being very brittle, was susceptible to fatigue cracking while bending around the sheave.

Conclusions. The wire rope failed in fatigue because of the brittle layer of martensite on the wires. Both the carbon content and the tensile strength of the wires were below specifications.

Corrective Measures. Because of a record of poor service life on this wire rope, 6 × 19 rope was substituted because it withstands abrasion and drum crushing better than 6 × 37. Service life was doubled after this change in rope construction.

Example 4: Failure of Steel Wire Rope Because of Overheating. A 3.8-cm ($1\frac{1}{2}$-in.) diam 6 × 37 rope of improved plow steel wire broke in service during dumping of a ladle of hot slag.

Investigation. Examination of the rope showed a heavy blue oxide extending 0.6 to 0.9 m (2 to 3 ft) back from each side of the break. The broken ends of the rope showed tensile fractures. Microscopic examination of the wires adjacent to the break revealed that the steel had been recrystallized. The rope had thus been heated in excess of 700 °C (1300 °F). The tensile strength of the wires in the rope that broke was 896 MPa (130 ksi), whereas the specification required 1724 MPa (250 ksi).

Conclusions. Failure of the wire rope was attributed to overheating, which resulted in a 50% loss of tensile strength.

Recommendations. Wire ropes must not be subjected to extremes of heat, such as exposure to hot slag, except for brief, intermittent periods of time. Prolonged or continuous exposure will cause a rapid deterioration of service life.

Example 5: Fatigue Fracture of a Steel 8 × 19 Elevator Cable. Fracture occurred in one of six cables on a passenger elevator. The elevator continued to operate on the remaining five cables until routine inspection discovered the broken cable. The cable was made of 16-mm ($\frac{5}{8}$-in.) diam steel wire rope designated as 8 × 19 G Preformed Extra High Strength Special Traction Elevator Cable with fiber core. The cables had been in service for $1\frac{1}{2}$ years. The end of the wire rope was sealed into a conical shape (Fig. 9a) in a low-melting alloy. Fracture occurred at the shackle where the end of the cable was socketed.

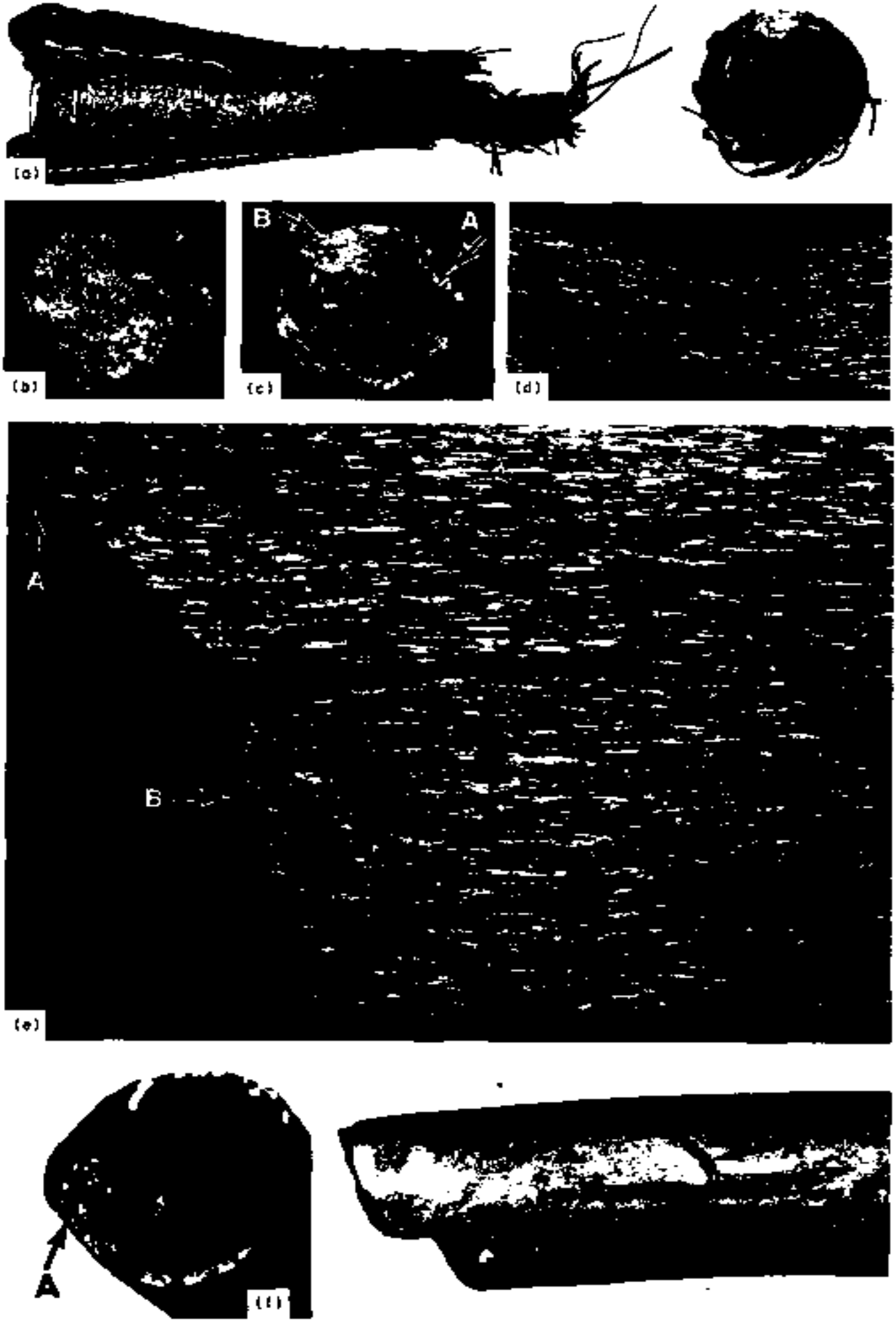

(a)
(b)
B
A
(c)
(d)
A
B
(e)
A
(f)

Fig. 9 Fatigue fracture of a steel 8 × 19 elevator cable. The fracture resulted from cyclic torsional and tensile stresses. (a) Conical shape at end of cable, and end of broken cable. (b) As-received 1.2-mm-diam wire. 25×. (c) Same wire after cleaning with a cold aqueous solution of 10% HCl. 25×. A indicates a nick in side of wire; B, a bright, smooth area containing fatigue marks. (d) Longitudinal section of 0.6-mm-diam wire etched in 2% nital showing necked region. 55×. (e) 2% nital-etched longitudinal section through 1.6-mm-diam wire showing cold working at A, flat-type fracture surface at B, and longitudinal cracks. 100×. (f) Fractured end (left) and longitudinal view (right) of 1-mm-diam wire. Both 34×. Arrows A indicate nick in a region showing cold working and wear in service; arrow at far right, a secondary crack origin.

The fractured end, a length of wire rope away from the fracture, and several mounted specimens of various wires from the rope were examined. Samples of wire received were 1.2-mm (0.0465-in.) diam heart wire, 1-mm (0.0405-in.) diam outer wire, and 0.6-mm (0.023-in.) diam inner wire.

Investigation. Close examination of the wire rope under a stereomicroscope revealed two general types of fracture:

- A flat-type fracture in the samples of larger-diameter 1.2- and 1-mm (0.0465- and 0.0405-in.) wires
- A cup-and-cone type of fracture in the samples of smaller-diameter 0.6-mm (0.023-in.) wire

Generally, the larger wires, which failed with a flat-type fracture, were rusted; the smaller wires were oxidized but were not as heavily rusted. Figure 9(b) illustrates the fractured ends typical of the larger wires as received; Fig. 9(c), after removal of the rust. Arrow A in Fig. 9(c) indicates what appeared to be a nick in the side of the wire; arrow B indicates a bright, smooth area containing beach marks radiating inward, which indicated fatigue cracking. Flat-type fractures were believed to result from cyclic stresses, with the major stress component being torsional.

The nature of ductile behavior in the 0.6-mm (0.023-in.) diam wires is illustrated in Fig. 9(d), which shows a reduction of area in a necked-down region under excessive tensile stress without fracture. The fracture surfaces of the smaller-diameter wires were slightly oxidized; however, some bright regions could be seen, which indicated that the smaller wires broke later than the larger-diameter wires.

Numerous specimens were examined microscopically on a longitudinal plane at the failed ends. Figure 9(e) shows the typical microscopic characteristics of the larger wires. The important features are the cleanness of the material, microstructure, and the nature of the failure. No evidence of material or processing defects could be associated with the failure. It appears that as transverse cracking occurs under cyclic torsional stresses, longitudinal cracking follows as the transverse cracking progresses, placing the maximum torsional fiber stress below the surface. Because of the fibrous nature of cold-drawn materials, longitudinal crack propagation would be expected under conditions of cyclic torsional loading with transverse crack propagation. Restriction of free movement of the socket-end in the shackle would be expected to promote fracture by increasing the magnitude of the torsional stress under cyclic loading.

Evidence of torsional stress was also observed in the 1-mm (0.0405-in.) diam wires. Figure 9(f) shows nicks and cold working on the surfaces of the wire in the region of fracture. This particular wire probably failed under greater tensile stress after the 1.2-mm (0.0465-in.) diam heart wires broke. Ductility of this wire is indicated by some necking and by cup-and-cone characteristics in the fracture surface. Most of the wires examined showed essentially flat fracture surfaces. The cleanness of the wire in the region of fracture is also illustrated in Fig. 9(f).

Conclusions. Mechanical damage to the surfaces of the wires was sufficient to cause fatigue cracking under the stresses encountered in service. The wire rope was composed of several sizes of wire; rust on the fracture surfaces of the larger-diameter wires indicated that they had failed first. The smaller-diameter wires behaved in a ductile manner under excessive loads before ultimate failure. The microstructure and cleanness of the material appeared normal for the application and could not be associated with the broken wires examined.

Stress-corrosion cracking (SCC) frequently occurs in terminals that have been roll swaged on ends of wire ropes for marine use. In the following example, rolling lines from the swaging operation acted as stress raisers. Cracks developed along these stress raisers, and corrosion by seawater occurred in the cracks.

Example 6: SCC of a Stainless Steel Wire-Rope Terminal. An eye terminal made of AISI type 303(Se) stainless steel that was roll swaged on the end of a 9.5-mm ($\frac{3}{8}$-in.) diam wire rope was found to have cracked extensively after 1 year of service. The terminal was sectioned, and specimens were mounted for metallographic examination.

Investigation. Examination of one specimen revealed that a hairline crack had initiated at the inner surface of the fitting. Holes in the region adjoining the crack and the rough texture of the crack surface (Fig. 10) indicated that a corrosive medium (presumably seawater) had entered the crack from the inner surface of the fitting and, coupled with the hairline

crack, developed crevice corrosion. High residual stresses in the swaged metal caused the crack to propagate toward the outer surface, followed closely by corrosion. The last 0.3 mm (0.012 in.) of metal thickness failed in pure tension, as evidenced by the smooth appearance of the crack surface in this region.

Fig. 10 Composite micrograph of a transverse section through a type 303(Se) stainless steel eye terminal for a wire rope showing corroded crack surface and final-fracture region. 75×

Swaging creates a few small cracks in the surface of the hole in a terminal, particularly when swaging onto stranded wire rope. Residual stresses plus load stress and corrosion frequently result in stress corrosion.
Conclusions. The terminal failed by SCC as a result of residual stresses from swaging, load stresses, and corrosion by seawater.
Recommendations. Rotary swaging or swaging in a punch press is recommended instead of roll swaging. Roll swaging is accomplished by nonsymmetrical radial metal deformation. The size of internal cracks is likely to be greater under these conditions than when swaging is accomplished in a more uniform symmetrical reduction, such as by rotary swaging or press swaging. Corrosion then combines with the internal stress to cause SCC.

Reference cited in this section

3. A.B. Dove, Ed., *Steel Wire Handbook,* Vol 3, The Wire Association, Inc., Branford, CT, 1972

Failures of Lifting Equipment

Frank L. Jamieson, Stelco Inc. (retired)

Chains

Chains made of resistance-welded plain low-carbon steel are in wide use and are generally acceptable. However, to ensure safety and to minimize chain failures, heat-treated alloy steel chains with a hardness range of 302 to 352 HB should be used. This hardness range is recommended to reduce the frequency of brittle fractures and to avoid premature wear and excessive stretching. However, it is much more desirable for a chain to stretch or to exhibit wear than to be so hard that it can fail by brittle fracture from a small fatigue crack, as often happens to chains with hardnesses of 375 HB and higher.
Fatigue cracks in chains are usually initiated at one of the following origins:

- A heavily cold-worked zone resulting from misuse of the chain
- A weld zone with entrapped inclusions or with inadequate weld penetration
- The portion of a weld at the inside surface of a link, which, because of its position, cannot be planed as smoothly as the outside surface

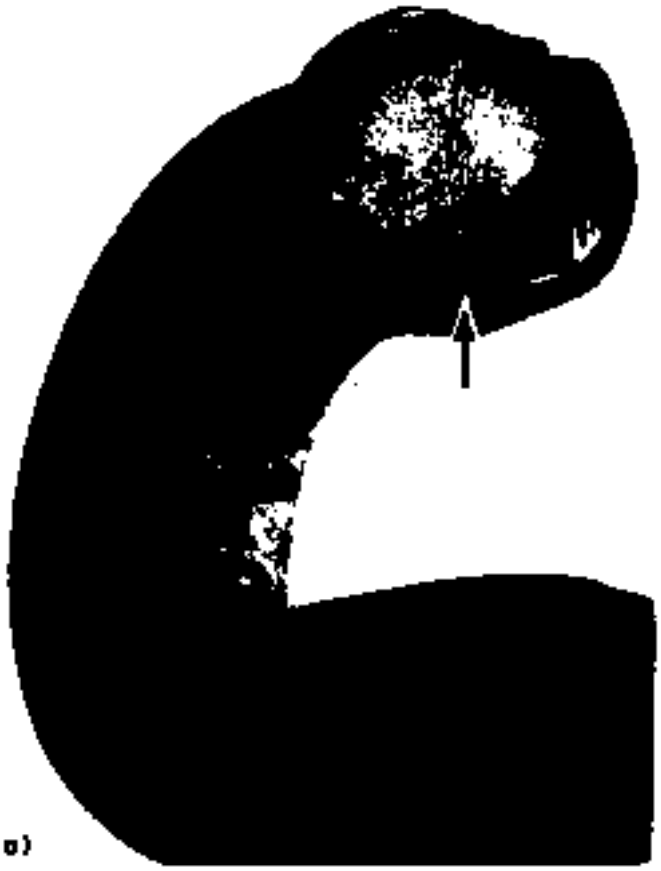
In the investigation of chain failures, the following possibilities must be considered in determining the cause of failure:

- Subjection of the chain to sudden impact loads or unbalanced loads
- Kinks, twists, or knots in the chain, which create severe stresses that can bend, weaken, and break chain links
- Deterioration of the chain by strain, usage, corrosion, or operation at elevated temperatures. The specified load limit should be reduced when alloy steel chains are used above 260 °C (500 °F). Chains should be replaced when wear has exceeded the specified limit of the manufacturer
- Stretching of the chain, which usually indicates overloading and which can be avoided by the use of proper chain size. Stretch is often expressed in terms of percentage of overall length. This is not recommended, because individual links may often be severely elongated; thus, only a small portion of the entire chain is stretched. To determine the degree of stretching, the chain should be inspected link by link
- Improper repair of the chain. Alloy steel chains should always be identified as such to ensure that any repairs—for example, replacement of damaged section or joiner links—are made properly. Alloy steel chains are usually returned to suppliers for repair and proof testing

Regular inspection of chains is imperative. Where many chains are in service, inspection is usually done visually. There are critical applications, however, in which magnetic-particle or liquid-penetrant inspection should be implemented. Inspection of individual links by these methods is a slow and tedious process and is carried out only when chain failure would cause injury to personnel and damage to equipment.

Example 7: Fracture of a 4615 Steel Chain Link Because of a Weld Defect. A resistance-welded chain link made from 16-mm ($\frac{5}{8}$-in.) diam 4615 steel broke while lowering a 9070-kg (10-ton) load of billets into a rail car. The acceptable load limit of the chain was approximately 13,600 kg (15 ton). The chain had been in service for 13 months. Because of the large number of chains in service and the fact that chains were left around billets in storage, no routine inspection was carried out.

Observations. The link broke at the weld. Beach marks, typical of fatigue, originated at the inside of the link (arrow, Fig. 11a). Metallographic examination of a section through the fracture surface revealed cracks (arrows A, Fig. 11b) in the weld zone that were up to 1.2 mm (0.0465 in.) deep. Examination of this region at a magnification of 65× showed that the cracks were filled with scale, indicating that they had formed during resistance welding of the link. The chain was made of fine-grain quenched-and-tempered 4615 steel with a hardness of 285 HB, as specified.


(a)

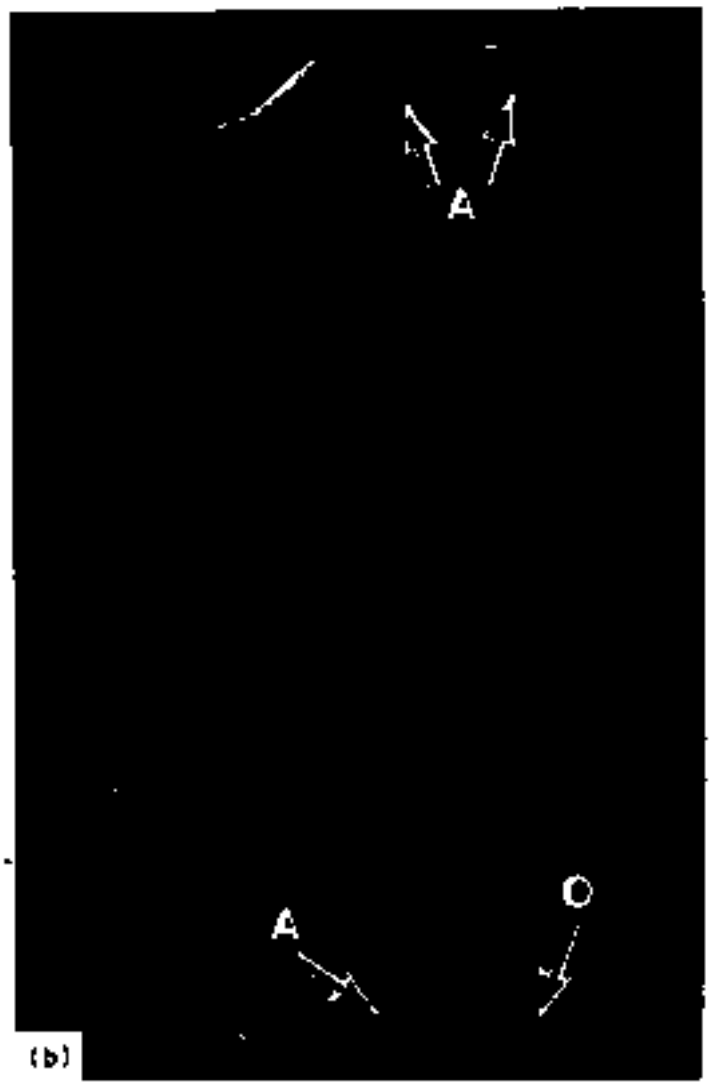

A
A
O
(b)

Fig. 11 Resistance-welded 4615 steel chain link that broke because of a weld defect. (a) A fracture surface of the chain link showing fatigue beach marks (arrow) progressing across the surface from the inside of the link. (b) Nital-etched longitudinal section through the link showing fracture origin (arrow O) and weld cracks (arrows A). Weld cracks at top were approximately 1.2 mm (0.0465 in.) deep.

Conclusions. Failure of the chain link was attributed to weld defects, which acted as stress raisers from which fatigue cracks originated.

Corrective Measures. Inspection revealed welding laps in all chains in service. All such chains were removed from service and replaced with defect-free chains. The chain manufacturer changed the welding method to ensure defect-free chain links.

Example 8: Brittle Fracture of Alloy Steel Chain Links Because of Excessive Hardness. Over a 1-year period, chain-link fractures occurred in many of several thousand 16-mm ($\frac{5}{8}$-in.) diam alloy steel sling chains used for handling billets. Several shipments of new chains had been acquired because of expansion of production facilities. No failures had occurred before delivery of the new chains.

Observations. The links broke at the weld, with the breaks originating at the inside of the link. The breaks were often undetected because the broken links opened only slightly at the fracture line. There had been no fracture-related damage to the links. Metallographic examination revealed no weld irregularities. All failures occurred in links having hardness values in the range of 375 to 444 HB.

Upon contacting the supplier, it was learned that the hardness level of the new chain links was 375 to 444 HB, a change from the previous hardness level of 302 to 375 HB. Hardness was increased to minimize wear, but it made the links notch sensitive, resulting in fractures that initiated at the butt-weld flash on the inside surfaces of the links. It is believed that most of the fractures occurred during the winter months; the low temperatures undoubtedly caused a further reduction in ductility.

Conclusions. The chain links failed in a brittle manner because the high hardness of the material made it notch sensitive, particularly at low temperatures.

Corrective Measures. All chains were visually inspected, and those with broken links were returned to the supplier for repair. All chains were retempered to a hardness of 302 to 375 HB.

All new alloy steel sling chains were subsequently ordered to a hardness of 302 to 352 HB. No failures of this type were reported since the chains were retempered and the hardness requirement of 302 to 352 HB was introduced.

Failures of Lifting Equipment

Frank L. Jamieson, Stelco Inc. (retired)

Hooks

Hooks are designed for individual applications to suit the type of load, the size and weight of the load, the intended service environment, and the degree of versatility required. Failure of hooks occurs frequently, and hooks are continually being altered to avoid future failures. Among these alterations are changes in method of fabrication, grade of steel, heat treatment, and design.

In general, the most important features required of a hook material are excellent fatigue properties and resistance to brittle fracture. It is desirable for a hook to exhibit wear rather than to fracture in a brittle manner because of the possible serious injury and damage that can result. Such a hook should be evaluated visually while in service, and if signs of excessive wear are found, the hook should be removed from service for repair or replacement. If worn in the saddle region, a hook may be ground smooth within the specified reduction limit (usually, a maximum of 20% of the cross-sectional depth), renormalized if originally in a normalized condition (or otherwise thermally treated, depending on prior treatment), and returned to service. If fatigue cracks are found by magnetic-particle inspection, it may be possible to eliminate these cracks by grinding within the specified limits; if not, the hook should be scrapped.

Welding on crane hooks is to be discouraged. When welding must be done to attach safety devices or to join laminated sections, it should be followed by a heat treatment that will eliminate any hard spots or residual stresses in the HAZ.

Frequently, hooks that have been in service for a considerable length of time are given a stress relief to eliminate the effects of cold working from repeated lifting contacts by chains or rings.

Materials for Crane Hooks. C-hooks, pin hooks, coil hooks, and sister hooks are usually forged from fully killed fine-grain carbon or alloy steels. Extremely large ladle hooks and coil hooks are made of laminated plate sections. Specifications governing forgings for industrial use are given in ASTM A 668 (Ref 4); these specifications may be augmented as required.

Low-carbon (1018, 1020, or 1025) and medium-carbon (1045) steels are widely used in the normalized or the quenched-and-tempered condition. Alloy steels used include the 8620, 4320, 4130, and 4140 grades, generally in the quenched-and-tempered condition. The low-carbon steels usually have greater resistance to crack propagation. Other alloy steels, or other materials, may be required in certain corrosive atmospheres or because of weight limitations placed on the hook.

Some hooks have been made of cast steel. Derrick hooks, for example, may be of quenched-and-tempered 4140 cast steel. This heat-treated steel has excellent fatigue properties. L-shape hooks, such as coil hooks, are usually constructed of low-alloy steel. Hooks that are torch cut from plate often fail because of cracks initiated in the torch-hardened zone. Such items should be thermally treated to eliminate this condition. A riveted, laminated L-shape hook is much more desirable. Depending on their ultimate service requirements, hooks are frequently magnetic particle inspected or ultrasonically tested during processing and at final inspection.

Design of C-Hook. The design of the hook shown in Fig. 12 has been developed from years of experience, and since its employment in a large steel complex, no failures of hooks of this design have occurred. The hook design may be adjusted to accommodate any load limit. The hook is forged from fully killed fine-grain 1020 steel and is normalized after forging. This steel exhibits excellent fatigue properties and will wear rather than fail in a brittle manner.

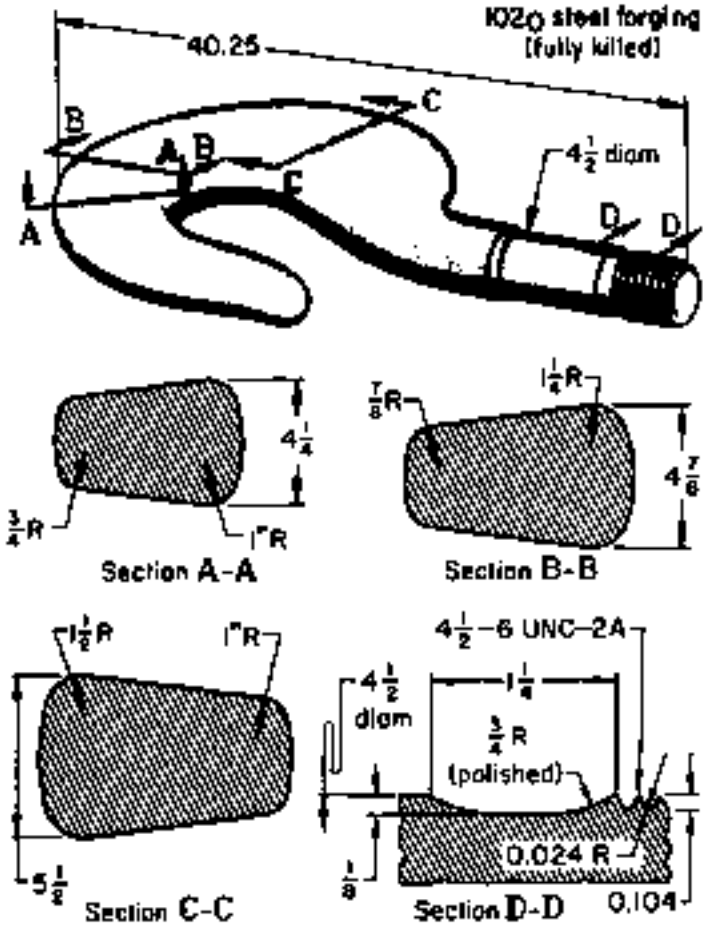


Fig. 12 Typical design of a 45,360-kg (50-ton) capacity 1020 steel C-hook with a stress-relief groove at end of threads and well-proportioned radii in body. Dimensions given in inches

Many hooks break at large changes in section or at poorly blended radiused section changes in the shank. The design shown in Fig. 12 avoids severe section changes. An important feature of this hook is the stress-relief groove at the end of the threads. If the nut is turned past the last thread—as it should be so as not to concentrate the stress at the bottom of the exposed thread groove—the stresses will be distributed over the length of the stress-relief groove, minimizing stress concentrations. The depth of a properly designed groove exceeds that of the thread roots. The thread design illustrated has also been found to be optimum. It is important, as with all hooks, to avoid rough-machining marks in the threads. Also, the nuts should be individually fitted to hooks.

Control and Testing. All hooks must be clearly identified, and accurate records must be kept of their characteristic features and location in the plant. Hooks should be inspected at least once a year, if not more often, depending on the predicted service life and the severity of the application. Magnetic-particle and liquid-penetrant inspection are commonly used for hooks. If feasible, new hooks should be checked by magnetic-particle inspection and should be spark tested to ensure that the correct grade of steel has been used.

Example 9: Fatigue Fracture of a 1020 Steel Crane Hook. The crane hook shown in Fig. 13 broke in the threaded shank while lifting a load of 9072 kg (10 ton). The crane was nominally rated at 13,000 kg (15 ton). Service life of the hook was not known. The nut, still threaded on the broken shank, and the hook were sent to the laboratory for failure analysis.

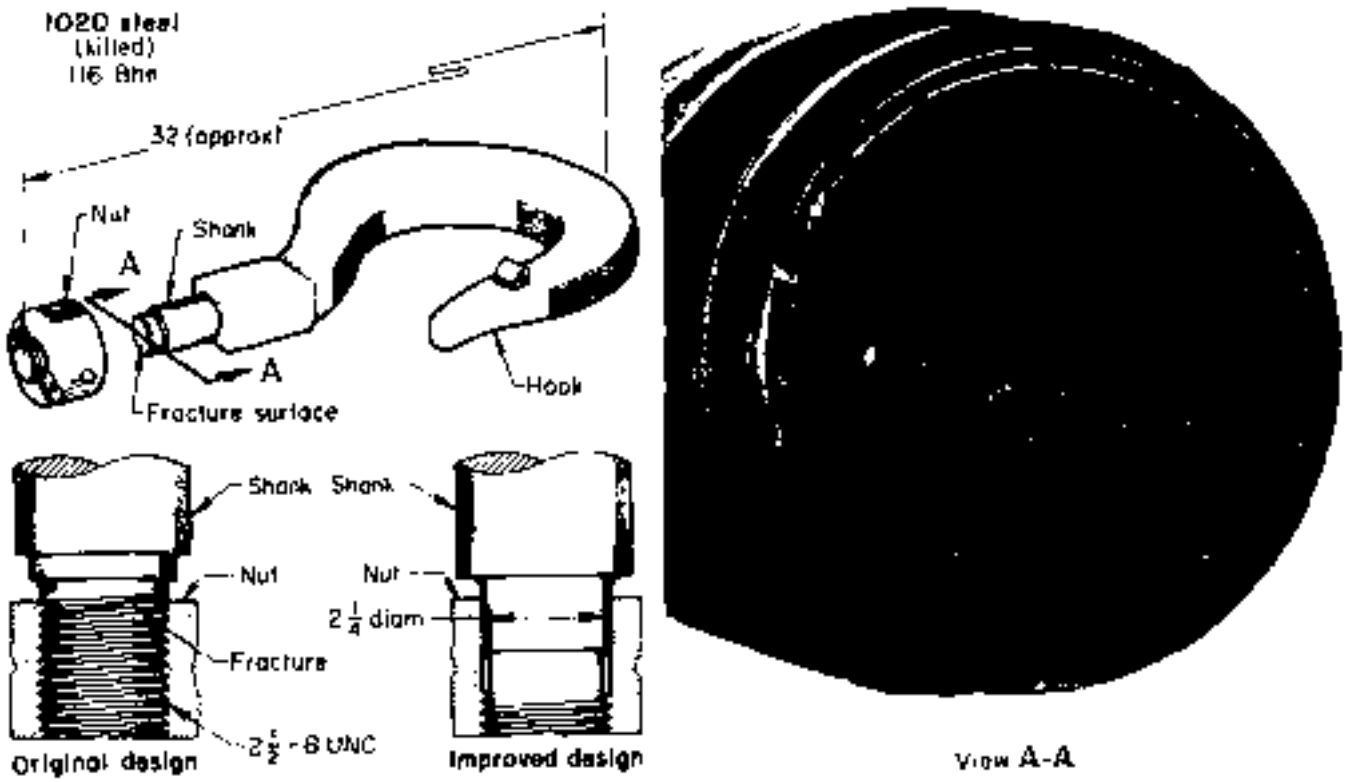


Fig. 13 13,600-kg (15-ton) 1020 steel crane hook that failed in fatigue. View of a fracture surface of the hook showing beach marks. Original and improved designs for the nut and the threaded end of the hook are also shown. Dimensions given in inches

Investigation. Chemical analysis of the metal in the hook showed that it was killed 1020 steel. The steel had a hardness of 116 HB and was judged to be satisfactory for the application.

Visual examination disclosed that fracture had occurred at the last thread on the shank. Rough-machining marks and chatter marks were evident on threads. The fourth thread from the fracture had a small crack at the root. Beach marks emanating from the thread-root locations on opposite sides of the fracture surface (View A-A, Fig. 13) identified these locations as the origins of the fracture.

Metallographic examination revealed a medium-coarse slightly acicular structure, indicating that the material was in the as-forged condition. Light segregation and nonmetallic inclusions were present, but the material was considered sound.

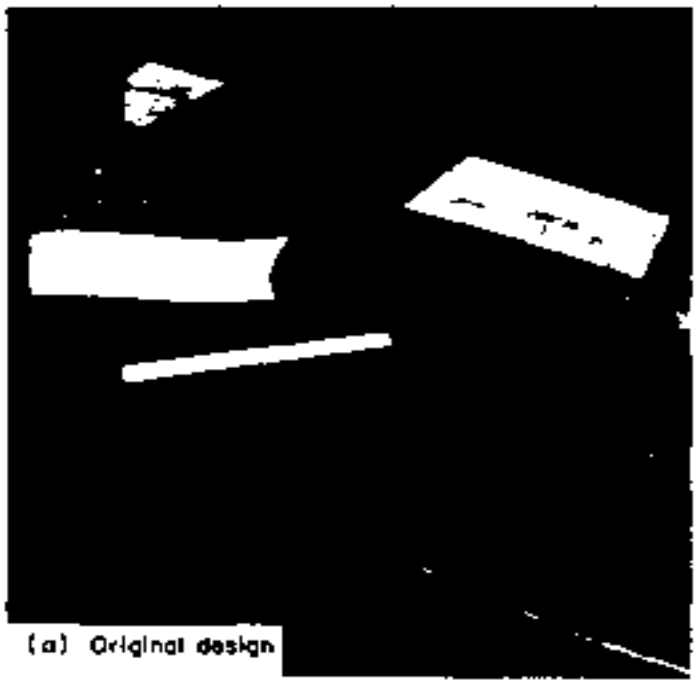
Conclusions. Fracture of the hook resulted from fatigue cracking that originated at stress concentrations in the root of the last thread. Life of the hook was shortened considerably because the material was in the as-forged condition, which resulted in low fatigue strength.

Corrective Measures. The crane hook was normalized after forging to produce the most desirable structure. A stress-relief groove with a diameter slightly smaller than the root diameter was placed at the end of the thread, and a large-radius fillet was machined at the change in diameters of the shank. Also, the nut was redesigned to extend beyond the threaded portion to relieve the stresses (Improved design, Fig. 13). Undercutting the last thread, using large-radius fillets, and ensuring that all crane hooks were properly normalized prevented additional failures for about 16 years.

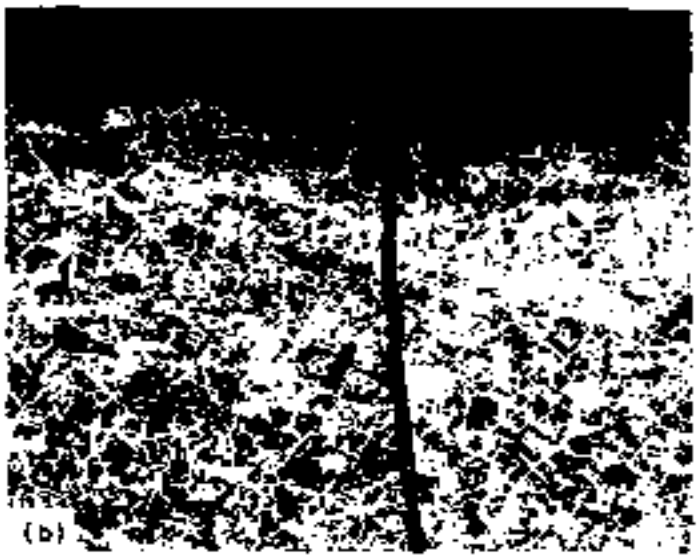
Example 10: Fatigue Fracture of a 1040 Steel Coil Hook. A 10,890-kg (12-ton) coil hook failed after 8 years of service while lifting a load of 13,600 kg (15 ton). The hook had been torch cut from 1040 steel plate.

Investigation. The inner surface of the hook exhibited the normal ironing (wear) marks. Visual examination of the fracture surface indicated that cracking had originated at the inside radius of the hook (Fig. 14a). Beach marks, typical of fatigue

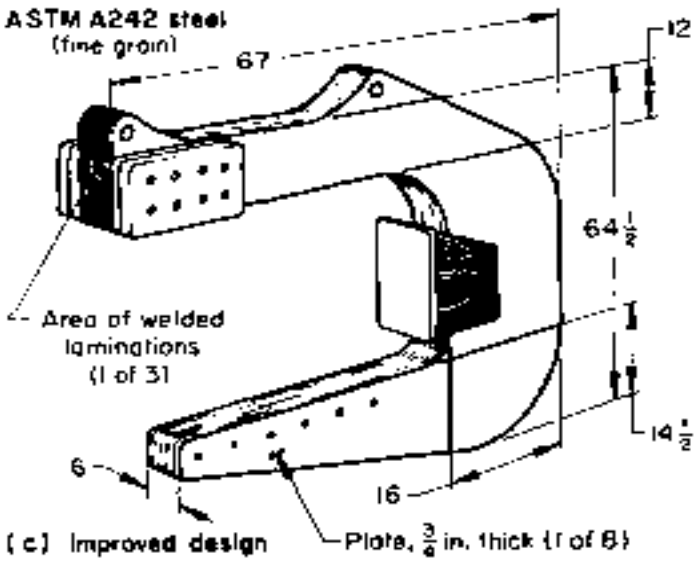
fracture, extended over approximately 20% of the fracture surface. The surface containing the beach marks was stained; the surface of the final-fracture region was bright.



(a) Original design



(b)



(c) Improved design

Fig. 14 Torch-cut 1040 steel coil hook that failed by fatigue. (a) Fracture region of the 10,890-kg (12-ton) hook. (b) Macrograph of a nital-etched section showing cracks propagating from the surface (top), which was hardened and embrittled during torch cutting. $7\frac{1}{2}\times$. (c) Replacement hook of a laminated design made of ASTM A242 fine-grain steel plate. Dimensions given in inches

Macroscopic examination of the torch-cut surfaces revealed numerous cracks. Figure 14(b) shows a macrograph of an etched section through the hook, illustrating cracks that were initiated in a hardened martensitic zone at the torch-cut surface. One crack is shown extending into the coarse pearlite structure beneath the martensitic zone.

Conclusions. The hook fractured by fatigue that originated at the brittle martensitic surface produced in torch cutting. The 25% overload contributed to the failure. The hook should have been normalized after torch cutting.

Corrective Measures. The coil hooks were remade to a laminated design (Fig. 14c); these hooks were flame cut from ASTM A242 fine-grain steel plate, then ground to remove the material damaged by flame cutting. The hooks were stress relieved at 620 °C (1150 °F) after welding but before riveting the wear pads. No failures were reported since adoption of the laminated design.

Example 11: Failure of a Forged Semikilled 1015 Steel Hook on a 13-mm ($\frac{1}{2}$-in.) Diam Chain Sling. A hook on a two-leg chain broke while lifting a 4990-kg (11,000-lb) load. The included angle between the two 13-mm ($\frac{1}{2}$-in.) diam chains was 60°. The service life of the sling was $2\frac{1}{2}$ years.

Investigation. The safe load limit for each leg of the sling was 2630 kg (5800 lb). The hook broke at the junction of the eye and shank. The diameter of the hook at this junction was approximately 22 mm ($\frac{7}{8}$ in.).

Visual examination of the fracture region revealed light intergranular oxidation at the surface on the side of the hook where cracking started. Approximately 50% of the fracture surface contained beach marks; the remainder contained cleavage facets.

Metallographic examination showed a medium-coarse acicular as-forged structure. The internal cleanness and soundness of the metal were satisfactory. Chemical analysis showed the metal to be semikilled 1015 steel.

Conclusions. Fatigue fracture was initiated at a region of intergranular oxidation that developed during forging. The acicular as-forged structure provided poor fatigue and impact properties, which contributed to failure of the hook.

Corrective Measures. The chain-sling hook was replaced with one made of normalized, fully killed, fine-grain 1020 steel.

Reference cited in this section

4. "Standard Specification for Steel Forgings, Carbon and Alloy, for General Industrial Use," A 668, *Annual Book of ASTM Standards*, ASTM, Philadelphia

Frank L. Jamieson, Stelco Inc. (retired)

Shafts

The most common or recurring initiation sites for fatigue fractures in shafts are section changes with insufficient fillet radii, roughly machined keyways and section changes, intersections of keyways with section changes, improperly located grease holes, and fretting corrosion from connecting attachments.

Shafts are usually subjected to bending or torsional loading or both. Careful visual examination of the fracture surfaces can provide the information necessary to identify the actual manner of loading. Bending stresses are largely the result of misalignment of bearings, fittings, or couplings and are usually difficult to eliminate completely. Although a particular

steel may be satisfactory for a certain application under normal conditions, misalignment will impose further stresses that result in the shaft being unsuitable for the application.

To provide a greater margin of safety, which is justified for lifting equipment, it is recommended that the shaft material be upgraded from the normally used hot-rolled medium-carbon steel to quenched-and-tempered carbon or alloy steel. Such materials, particularly alloy steels when properly heat treated, display superior fatigue properties and are much less susceptible to service failures. The danger and downtime resulting from a shaft failure make upgrading by heat treatment or a change in material economically feasible. It is presumed, of course, that any required improvement in design, such as increasing fillet radii or removing keyways from shoulders, would also be considered before or incorporated with any change in processing or material. For a more complete discussion of shaft failures, see the article “Failures of Shafts” in this Volume.

Example 12: Fatigue Fracture of a 1030 Steel Crane Shaft. A drum pinion shaft, part of the hoisting gear of an 18,140-kg (20-ton) capacity crane operating in a blooming mill, broke while the crane was lifting a 9070-kg (10-ton) load. Specifications indicated that the shaft was made of 1030 steel in the as-rolled condition.

Investigation. The end of the broken shaft containing the keyway was visually examined (Fig. 15). The keyway extended into a shoulder at a change in diameter; chatter marks, rough-machining marks, and sharp corner radii were visible in the keyway. Also, at each end of the keyway was a circular recess below the normal keyway surface—an outdated method of machining a keyway that is infrequently used. Both the bottoms and the sides of the recesses contained tool marks.

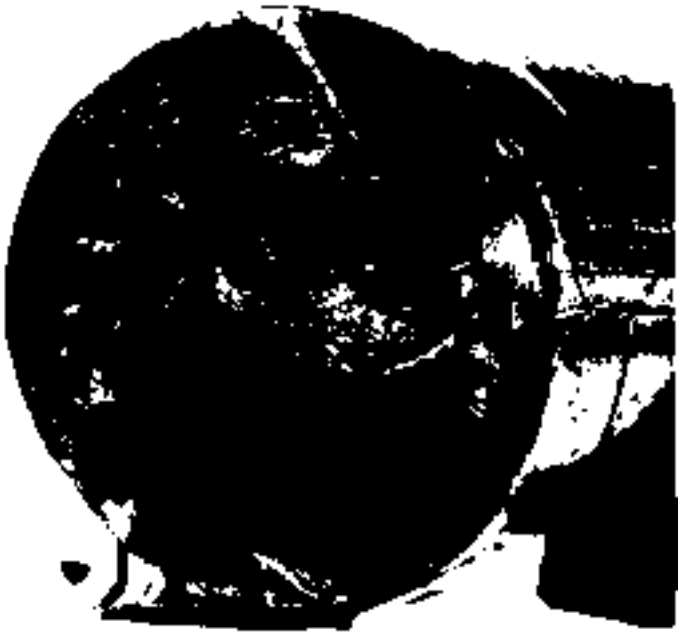


Fig. 15 Fatigue-fracture surface and keyway of a broken 1030 steel pinion shaft.

Examination of the fracture surface (Fig. 15) revealed that the origin of fracture was a sharp corner at the end of the keyway. Beach marks radiated from the origin over a large portion of the fracture surface. The final-fracture zone, a small ductile shear lip, was approximately 30° off-center, indicating low stress and rotational bending. The shaft material was 1030 steel, as specified, and had a slightly acicular fine-grain structure.

Conclusion. The shaft failed by fatigue fracture, which originated at a sharp corner at the end of a keyway at a change in section.

Corrective Measures. A replacement shaft was made of 4140 steel, quenched and tempered to a hardness of 286 to 319 HB. The keyway was moved away from the change in section and was machined with a 1.6-mm ($\frac{1}{16}$-in.) radius in the bottom corners. A larger-radius fillet was machined at the change in section.

Example 13: Fatigue Cracking of a 1040 Steel Main Hoist Shaft. The 14-cm ($5\frac{1}{2}$-in.) diam main hoist shaft of a mobile shovel was found to have multiple crack indications when ultrasonically inspected in the field. A crack indication located approximately 15 cm (6 in.) from the gear end extended around the shaft at a change in section adjacent to the bearing surface. The shaft was removed for further examination. It had been in service for 3 years. A previous shaft had failed after a short service life.

Investigation. Magnetic-particle inspection of the shaft showed a crack around the entire circumference at the change in section (Fig. 16). The fillet at the change of section where the crack was located was well polished and generally free of

rough-machining marks. However, the crack coincided with the junction of the fillet and the smaller diameter (10.8 cm, or $4\frac{1}{4}$ in.) at this change in section. A slight step, or break, in the continuity of the 8-mm ($\frac{5}{16}$-in.) radius fillet and some machining marks were noted at this junction. Microscopic examination of a section through the crack indication revealed a fine crack extending 2.5 mm (0.10 in.) from the surface, originating at the machining marks.

Fig. 16 Change in section in a 1040 steel main hoist shaft where a fatigue crack (arrow) was initiated at rough-machining marks and a break in a fillet.

Chemical analysis of the shaft indicated that it was made of 1040 steel. Hardness was 170 HB. Metallographic examination revealed a fine, uniform, normalized structure.

Conclusions. The shaft failed by fatigue; the step at the base of the fillet was the point of initiation of the fatigue crack. The shaft was underdesigned from a material standpoint; 1040 steel with a hardness of 170 HB was too low in fatigue strength.

Corrective Measures. Shaft material was changed to 4140 steel oil quenched and tempered to a hardness of 302 to 352 HB. All discontinuities related to fillet machining were also removed. With the above changes, no further failures were encountered.

Reversed Torsional Loading. Shafts used in lifting equipment are frequently subjected to reversed torsional loading. Insufficient fillets at the roots of spline teeth act as stress raisers, and fracture occurs frequently when the shaft is subjected to torsional loading.

Example 14: Fatigue Fracture of a 4140 Steel Cross-Travel Shaft. The horizontal cross-travel shaft on a derrick broke after 2 years of service. Specifications required the shaft to be made of 4140 steel quenched and to a hardness of 302 to 352 HB.

Investigation. Examination revealed that the shaft had fractured approximately 13 mm ($\frac{1}{2}$ in.) from the change in section between the splined end and the shaft proper. At this point, the cracks propagated in the longitudinal and transverse directions until failures occurred (Fig. 17a). A transverse section through the spline (Fig. 17b) showed that the longitudinal cracks had been initiated at the sharp corners at the roots of the spline teeth. The cracks visible in Fig. 17(b) were highlighted by techniques used in magnetic-particle inspection. Operation of the derrick had subjected the shaft to reversed torsional loading.

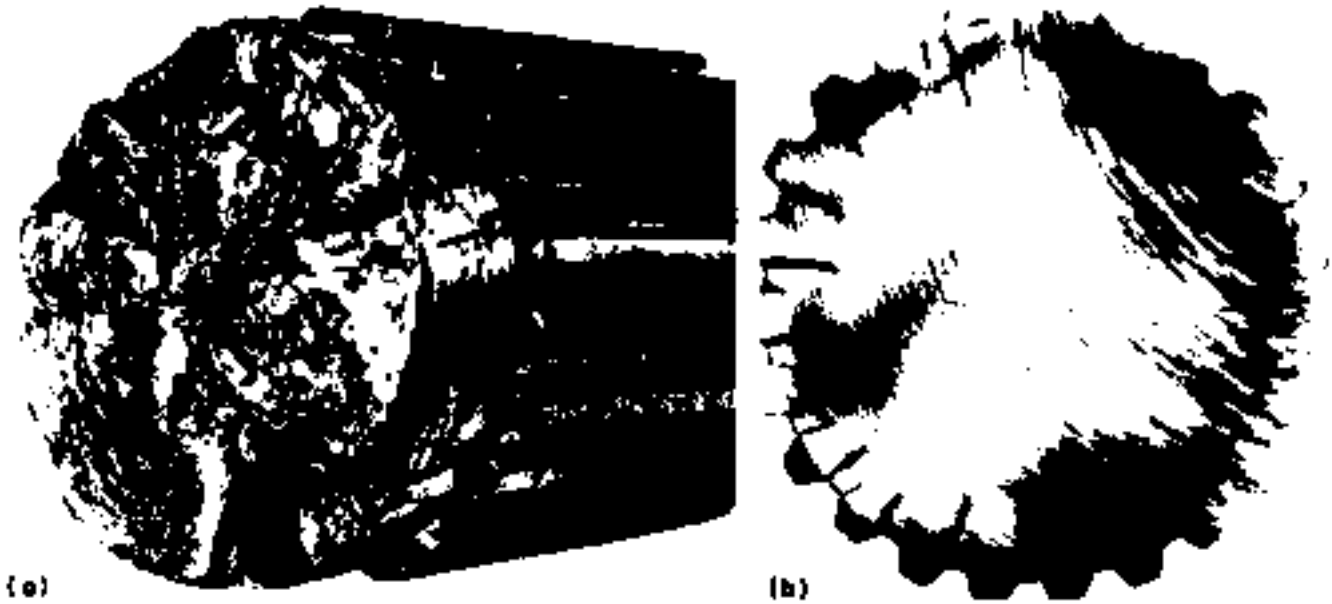


Fig. 17 4140 steel cross-travel shaft that failed in service. (a) Broken end of the shaft from a derrick showing the star-type fracture that results from reversed torsional loading. (b) Transverse section through the spline showing cracks initiated at sharp corners at the roots of the spline teeth

Conclusions. The shaft fractured in fatigue from reversed torsional loading; the sharp corners at the roots of the spline teeth acted as initiation sites for fatigue cracks.
Corrective Measures. The fillets at the roots of the spline teeth were increased in size and polished to minimize stress concentrations in these areas, and no further shaft failures occurred.

Frank L. Jamieson, Stelco Inc. (retired)

Cranes and Related Members

Failures resulting in the immobilization of cranes are most often related to shafts (which have been discussed previously), structural members (as related primarily to grade of steel and fabrication practice), crane wheels, and rail runways. Figure 18 shows the fracture surface of a 1055 steel crane wheel that failed by fatigue. The most common cause of failure is poor welding practice. Poor welding practice, such as a lack of preheat or postheat or use of the wrong filler metal, often results in hardening from the heat of welding, weld porosity, incomplete fusion, inadequate joint penetration, and weld cracks. Poor welding practice may occur not only during repair but also in initial fabrication or assembly.

Fig. 18 1055 steel wheel from a stripper crane that failed by fatigue. The wheel foiled after about 1 year of service in the 544,320-kg (600-ton) crane. In the center of the fracture is a fatigue zone showing beach marks that are concentric around the crack origin, which evidently was an internal flaw about 3.8 cm ($1\frac{1}{2}$ in.) beneath the wheel-tread surface. Chevron patterns are visible in the area of final fast fracture that surrounds the fatigue zone. 0.3 ×

The most common method of repairing structural members is welding. The part must be welded properly, or it may fail soon after being returned to service. Such faults as microcracks and macrocracks, mechanical gouges, and excessive wear may be corrected by welding. If a crack is present, it should be removed by grinding. Magnetic-particle or liquid-penetrant inspection should be done to ensure complete removal of the crack before repair welding. Generally, all parts should be preheated before welding to prevent hardening from the heat of welding. This is most critical in medium-carbon and high-carbon steels because of their sensitivity to hardening by welding heat. It is essential that the correct welding filler metal be used. After welding, low-carbon steel members may be air cooled if they have been preheated correctly. To prevent hardening from the heat of welding, the cooling rate of medium-carbon and high-carbon steel members should be controlled—for example, by wrapping them in asbestos.

The susceptibility of a crane to failure is also heightened by a poor choice of material. In general, structural members of a crane should be designed so as not to fail in a brittle manner. The members should be made from impact-resistant fine-grain steels to resist brittle fracture and to exhibit excellent fatigue properties. This becomes more critical in cold-temperature applications.

Crane members should be inspected periodically for fatigue cracks to ensure that a failure is not imminent. A standard specification covering all items related to cranes is difficult to establish because of the wide variety of associated items. Those specifications that do exist must be viewed as a minimal guideline.

Example 15: Fracture of a Steel Tram-Rail Assembly. Figure 19(a) shows a hoist carriage tram-rail assembly fabricated by shielded metal arc welding the leg of a large T-section 1020 steel beam to the leg of a smaller T-section 1050 steel rail. Several of the welds failed in one portion of the assembly. Cracks in weld regions of other portions of the assembly were found by magnetic-particle inspection.

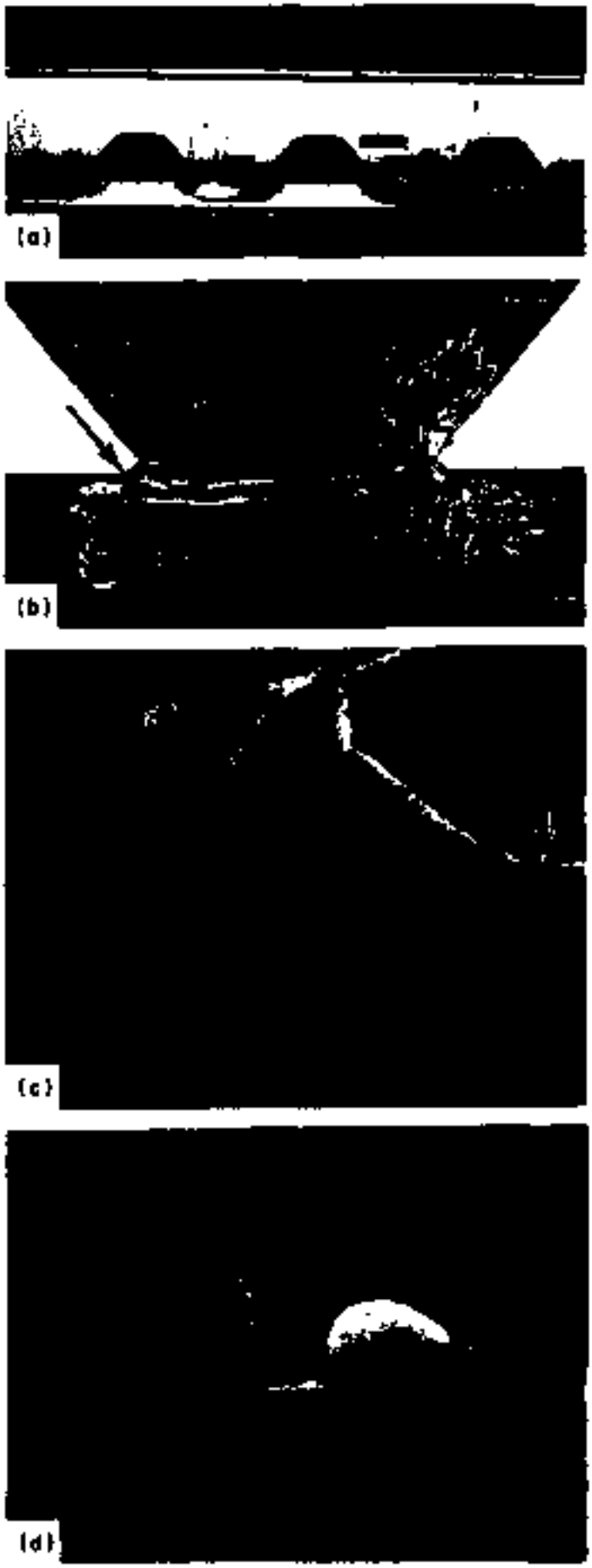


Fig. 19 Tram-rail assembly that fractured because of poor welding practices. (a) Section of tram rail as fabricated. T-section beam (1020 steel) is at top, T-section rail (1050 steel) is at bottom. (b) Enlarged view of welded area showing crack at toe of weld (arrow). (c)

Crack in rail initiated in the HAZ produced by feathering of the weld deposit. (d) Crack in rail initiated in the HAZ caused by weld spatter

Investigation. The weldment was inspected using the magnetic-particle method. This inspection revealed four weld cracks and several indefinite indications. All cracks were located at the toes of welds that joined the rail to the beam, such as the crack shown at the arrow in Fig. 19(b).

Metallographic examination of longitudinal sections through the welds revealed, in every instance, that cracks originated in HAZs in the rail section. A crack (arrow, Fig. 19c) was initiated in a localized HAZ produced by feathering the weld, which left a thin deposit of weld metal on the upper edge of the rail. Cracks were also found in HAZs resulting from weld spatter (Fig. 19d). Additional macrographs revealed cracks in welds and in HAZs resulting from arcing the electrode adjacent to the weld.

A relatively broad HAZ containing tempered and untempered martensite was present at each of the weld beads because of multiple-pass welding. The rail and beam both exhibited normal as-rolled steel structures with no irregularities. Chemical analysis revealed the rail to be made of 1050 steel and the beam of 1020 steel.

Hardness surveys revealed the following values: beam, 132 HB; rail, 255 HB; weld metal, 285 HB; tempered martensite, 321 HB; and untempered martensite, 578 HB.

Conclusions. The tram-rail assembly failed by fatigue cracking in HAZs. The size and shape of HAZs adjacent to welds indicated that the assembly had been neither preheated nor postheated during the welding operation. Welding workmanship was generally very poor. There was excessive feathering of weld metal and careless arcing during welding, which resulted in HAZs containing hard untempered martensite. Thus, vibration of the tram rail by movement of the hoist carriage on the rail could easily initiate and propagate fatigue cracking in the HAZs.

Corrective Measures. Welding procedures were improved, and the replacement rail assemblies were preheated and postheated.

Example 16: Brittle Fracture of a 1020 Steel Stop-Block Guide on a Crane Runway. A section broke from a stop-block guide (Fig. 20) on a crane runway and fell to the floor. A system consisting of wire ropes, pulleys, a counterweight, and a lever was used to raise the stop block from the crane runway to permit the crane to pass. When the crane was to be isolated within an area, the stop block was returned to the down position. Because the weight of the stop block and guide was about 30% greater than that of the counterweight, there was some impact against the rail each time the block was returned to the down position. One section of the fractured guide was sent to the laboratory for determination of the cause of failure.

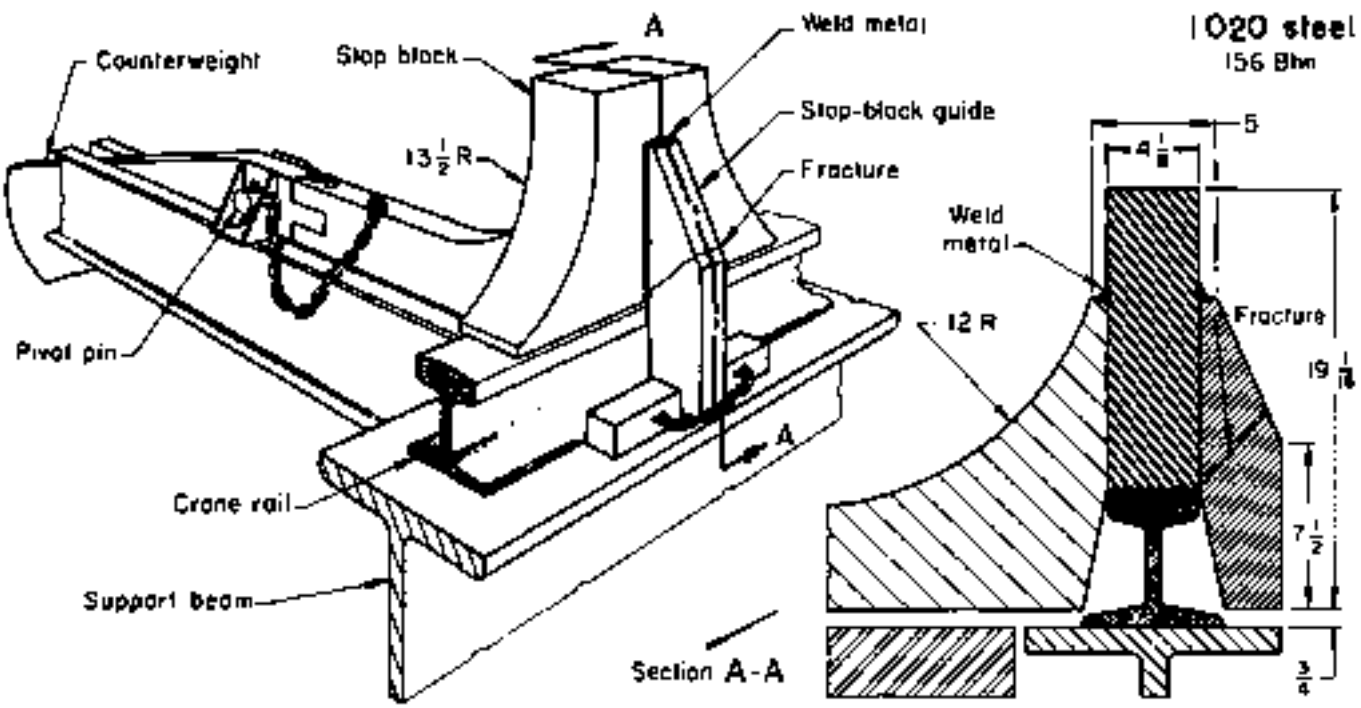


Fig. 20 Welded stop-block assembly for a crane runway showing stop-block guide that failed by brittle fracture. Dimensions given in inches

Investigation. Examination of the fracture surface disclosed a brittle crystalline-type break. The point of initiation was in a hardened heat-affected layer that developed during flame cutting and welding. Hardness of the heat-affected layer was 248 HB.

Chemical analysis showed the metal to be 1020 steel. The base metal (hardness, 156 HB) had a coarse as-rolled structure and a grain size of ASTM 00 to 4. The coarse grain size indicated that the weldment (stop block and guide) had not been normalized.

Conclusions. The stop-block guide failed in a brittle manner. Failure was initiated at a metallurgical and mechanical notch that was produced by flame cutting and welding.

Corrective Measures. New stop-block weldments were made from fully silicon-killed 1020 steel with a maximum grain size of ASTM 5. The weldments were normalized at 900 °C (1650 °F) after flame cutting and welding to improve microstructure and impact strength. All flame-cut surfaces were ground to remove notches. Stop-block weldments in service were checked for grain size and heat treatment and were normalized where required.

Example 17: Brittle Fracture of an Aluminum Alloy Lifting-Sling Member. The T-section cross member of the lifting sling shown in Fig. 21(a) broke in service while lifting a 966-kg (2130-lb) load. The L-section sling body and the cross member were made of aluminum alloy 5083 or 5086 and were joined by welding using aluminum alloy 4043 filler metal. Design load for the sling was 2722 kg (6000 lb).

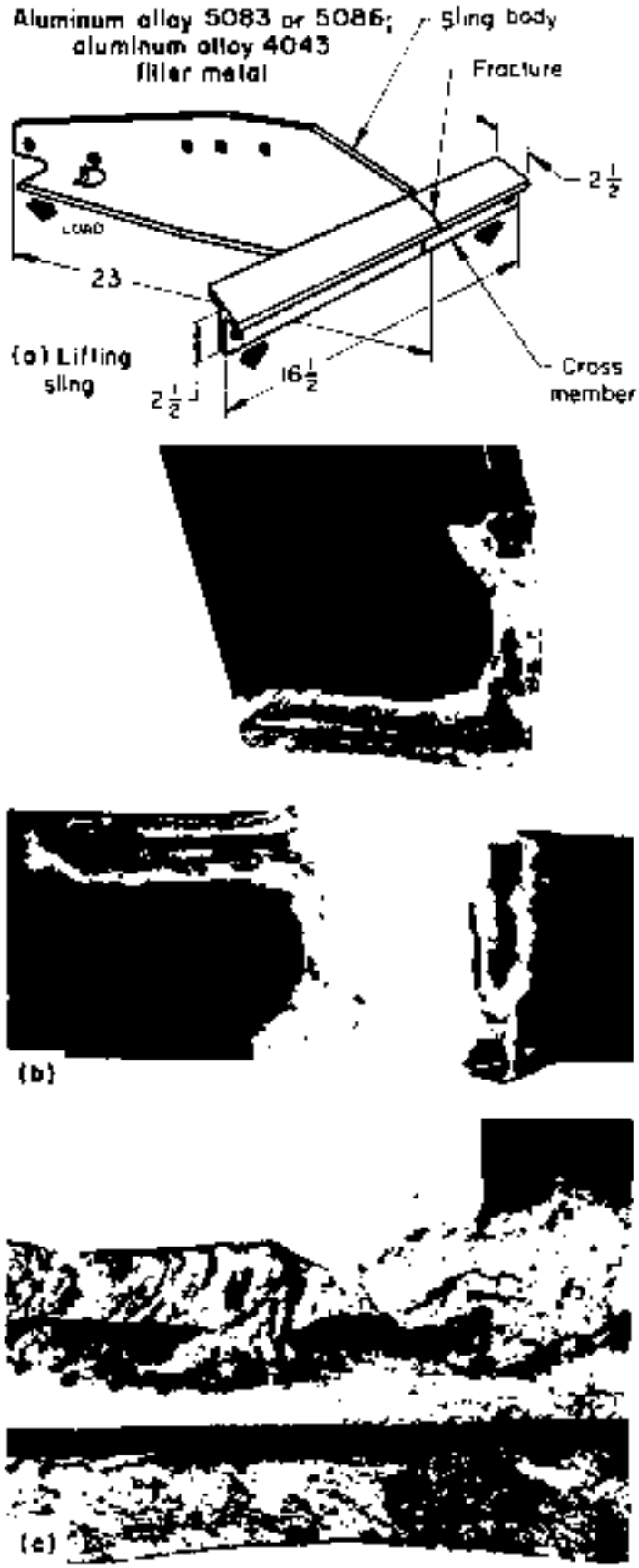


Fig. 21 Aluminum alloy lifting sling that fractured because of improper welding of the cross member to the sling body. (a) Configuration and dimensions (given in inches). (b) Fracture surfaces of the sling body (top) and the cross member (bottom). (c) Enlarged view of a fracture surface of the sling body showing weld defects

Investigation. Visual examination of the sling showed that fracture occurred at the weld joining the sling body and the cross member. Examination of the ends of the failed cross member showed that a rotational force had been applied on the cross member, causing it to fracture near the sling body. The fracture surfaces of the sling body and the cross member are shown in Fig. 21(b). Macrographic examination of the weld disclosed inadequate joint penetration and porosity (Fig. 21c).

Chemical analysis showed that the metal was within specifications for aluminum alloy 5083 or 5086. Analysis of the weld metal indicated that aluminum alloy 4043 filler metal had been used. The silicon content was lower and the magnesium and manganese contents were higher than normal for alloy 4043 filler metal. These differences were attributed to dilution of the weld metal by the base metal.

A similar sling was fabricated in the laboratory, set up in a test fixture, and subjected to three-point loading duplicating that intended for the sling that failed. The yield strength of the test sling (stress causing permanent deflection in the cross member) was 3538 kg (7800 lb), and the ultimate load was 7031 kg (15,500 lb).

The initial break in the test sling started at the welded joint between the sling body and the edge of the flange of the cross member and propagated through the body to failure. The cross member of the sling that failed in service sustained permanent deformation; however, the same member of the test sling exhibited no evidence of deformation.

Conclusions. The sling failed in a brittle fracture at the weld when the cross member was overloaded. Overloading was attributed to misalignment of the sling during loading, which concentrated the entire load on the cross member.

Recommendations. In welding of aluminum alloy 5083 or 5086, aluminum alloy 5183 or 5356 filler metal should be used to avoid brittle welds. Aluminum-silicon filler metals, such as alloy 4043, should not be used to weld aluminum alloys in which the magnesium content exceeds 3%, because an excess of Mg_2Si produces brittle welds.

Example 18: Fracture of a 1055 Steel Crane-Bridge Wheel. A bridge wheel on a crane fractured after 1 year of service. The wheel was forged from 1055 steel. The wheel fractured in the web between the hub and the rim.

Investigation. Visual examination of the fracture surface revealed a small area containing beach marks that originated in a heavily burned area on the web surface. Metallographic examination of a section taken through the region that contained the beach marks showed surface burning to a depth of approximately 0.8 mm (0.030 in.) (Fig. 22). The degree of decarburization and oxide dispersion that were visible indicated a forging defect.

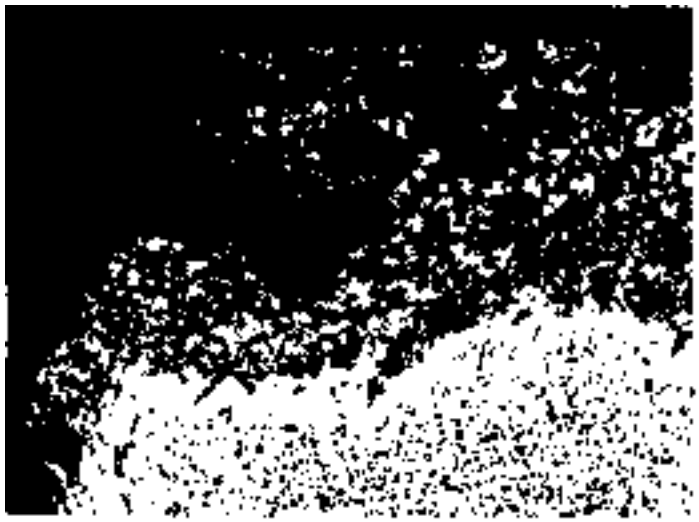


Fig. 22 Surface burning that initiated fracture in the web of a crane-bridge wheel forged from 1055 steel. Etched with 2% nital. Approximately 35×

Conclusions. Failure occurred because of surface burning during the forging operation. To prevent this type of failure, heating practice during forging should be more closely controlled to eliminate surface burning. If burning does occur, the burned region, if not too severely damaged, should be removed by machining.

Example 19: Fatigue Fracture of a Forged 1055 Steel Crane-Bridge Wheel. A bridge wheel from a 272,160-kg (300-ton) stripper crane failed after $1\frac{1}{2}$ years of service. The wheel was forged from 1055 steel, and the tread, hub faces, and hub bore were machined. The wheel fractured in the web near the rim.

Investigation. Macroscopic examination of the fracture surfaces revealed beach marks indicative of fatigue at ten locations. Fatigue cracking extended across about 15% of the surface before final fracture occurred. The dark areas on the fracture surface shown in Fig. 23(a) are the portions that failed by fatigue. The surface of the web was heavily scaled and decarburized.

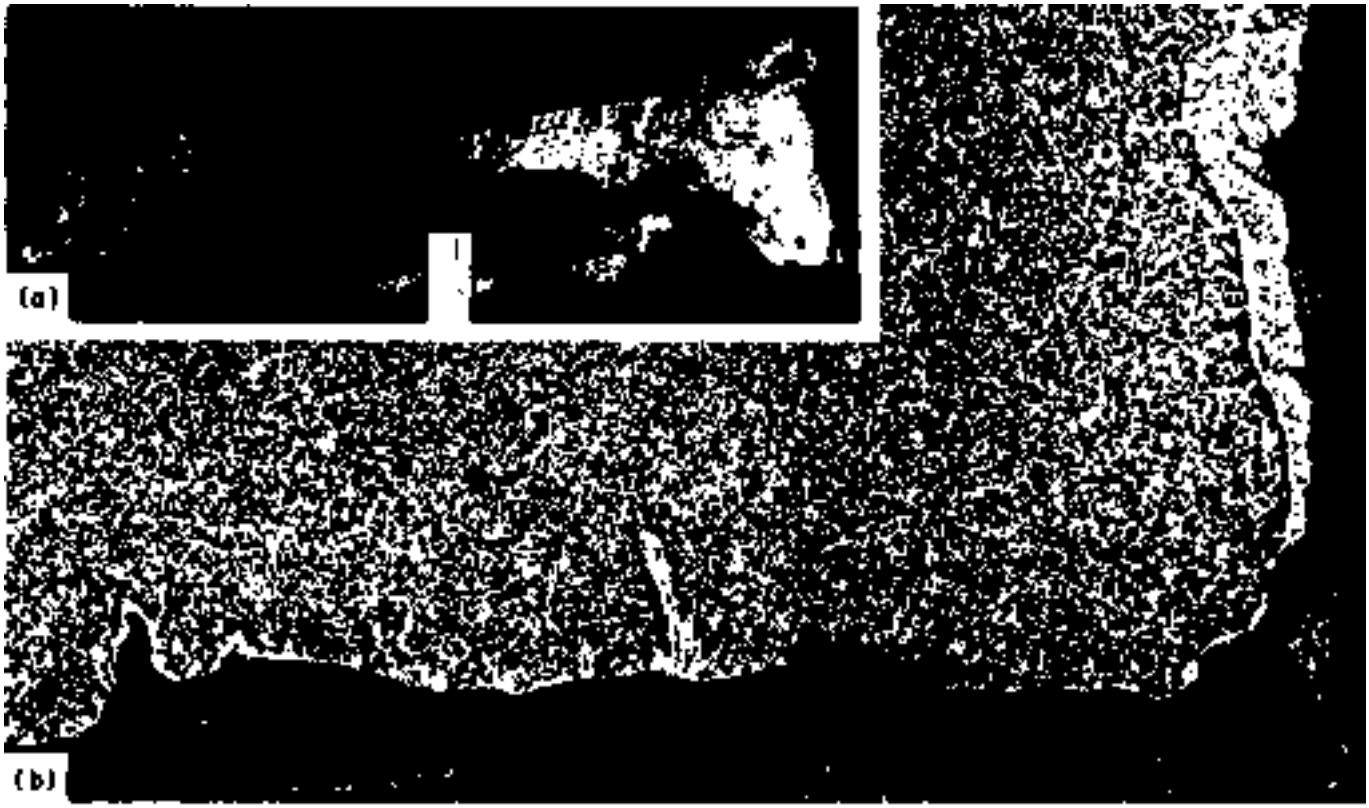


Fig. 23 1055 steel crane-bridge wheel that failed by fatigue. (a) Fracture surface of the crane-bridge wheel. Fatigue originated at forging defects. Dark areas are fatigue beach marks. (b) Micrograph of a nital-etched section through the fatigue origin showing a gross forging defect along the fracture surface (at right) and decarburized web surface (bottom). 35×

Chemical analysis showed that the material was equivalent to 1055 steel. Metallographic examination revealed that the wheel had been normalized and that the tread had been hardened to a depth of 5 mm (0.2 in.). The tread was tempered martensite; the core, fine pearlite.

Examination of a micrograph of a section through one of the fatigue origins showed a gross forging defect extending about 1.8 mm (0.072 in.) along the fracture surface (at right, Fig. 23b). Shallower forging defects and the decarburized surface were visible along the web surface (bottom, Fig. 23b).

Conclusions. Fatigue cracking of the wheel initiated at forging defects in the web.

Corrective Measures. Replacement wheels were machined all over and were magnetic particle inspected to detect any cracks that could act as stress raisers.

Failures of Lifting Equipment

Frank L. Jamieson, Stelco Inc. (retired)

References

1. "Standard Specification for High-Strength Bolts for Structural Steel Joints," A 325, *Annual Book of ASTM Standards*, ASTM, Philadelphia

2. "Standard Specification for Steel Structural Rivets," A 502, *Annual Book of ASTM Standards*, ASTM, Philadelphia

3. A.B. Dove, Ed., *Steel Wire Handbook*, Vol 3, The Wire Association, Inc., Branford, CT, 1972

4. “Standard Specification for Steel Forgings, Carbon and Alloy, for General Industrial Use,” A 668, *Annual Book of ASTM Standards,* ASTM, Philadelphia

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Introduction

THE PRIMARY FUNCTION of a fastener system is to transfer load. Many types of fasteners and fastening systems have been developed for specific requirements, such as higher strength, easier maintenance, better corrosion resistance, greater reliability at high or low temperatures, or lower material and manufacturing costs.

The selection and satisfactory use of a particular fastener are dictated by the design requirements and conditions under which the fastener will be used. Consideration must be given to the purpose of the fastener, the type and thickness of materials to be joined, the configuration and total thickness of the joint to be fastened, the operating environment of the installed fastener, and the type of loading to which the fastener will be subjected in service. A careful analysis of these requirements is necessary before a satisfactory fastener can be selected. The selection of the correct fastener or fastener system may simply involve satisfying a requirement for strength (static or fatigue) or for corrosion resistance. On the other hand, selection may be dictated by a complex system of specification and qualification controls. The extent and complexity of the system needed is usually dictated by the probable cost of a fastener failure.

Adequate testing is the most practical method of guarding against failure of a new fastener system for a critical application. The designer must not extrapolate existing data to a different size of the same fastener, because larger-diameter fasteners have significantly lower fatigue endurance limits than smaller-diameter fasteners made from the same material and using the same manufacturing techniques and joint system.

Types of Mechanical Fasteners. For descriptive purposes, mechanical fasteners are grouped into threaded fasteners, rivets, blind fasteners, pin fasteners, special-purpose fasteners, and fasteners for composites. Rivets, pin fasteners, and special-purpose fasteners are usually designed for permanent or semipermanent installation.

Threaded fasteners are considered to be any threaded part that, after assembly of the joint, may be removed without damage to the fastener or to the members being joined.

Rivets are permanent one-piece fasteners that are installed by mechanically upsetting one end.

Blind fasteners are usually multiple-piece devices that can be installed in a joint that is accessible from only one side. When a blind fastener is being installed, a self-contained mechanism, an explosive, or other device forms an upset on the inaccessible side.

Pin fasteners are one-piece fasteners, either solid or tubular, that are used in assemblies in which the load is primarily shear. A malleable collar is sometimes swaged or formed on the pin to secure the joint.

Special-purpose fasteners, many of which are proprietary, such as retaining rings, latches, slotted springs, and studs, are designed to allow easy, quick removal and replacement and show little or no deterioration with repeated use.

Mechanical fasteners for composites are frequently used in combination with adhesive bonding to improve the integrity of highly stressed joints. The usual bolts, pins, rivets, and blind fasteners are used for composites; however, the many problems encountered have stimulated the development and testing of numerous special-purpose fasteners and systems. Some of these problems are drilling of and installation damage to the composite, delamination of the composite material around the hole and pullout of the fastener under load, differences in expansion coefficients of the composite and the fastener, galvanic corrosion between the composite hole wall and the fastener, moisture and fuel leaks around the fastener, and fretting.

Load-carrying capacity has been significantly improved by increasing the head diameter of fasteners in order to distribute loads over a greater surface area. The bases of nuts and collars were similarly increased.

Fasteners for composites should fit the holes with no clearance in order to avoid fretting, but an interference fit may cause delamination. This problem has been alleviated to some extent by the development of several special sleeve/ fastener combinations.

The extreme temperature changes experienced by aircraft can cause differential thermal expansion and contraction of composites and their fasteners, thus altering clamp-up loads. Therefore, all new joint designs should be thoroughly tested to ensure satisfactory service. Although adhesives are generally used to join composites, mechanical fasteners are added to increase reliability when severe service conditions are involved. Table 1 compares fastener usage for several composite materials.

Table 1 Fasteners for advanced composites

Fastener type	Fastener material	Surface coating	Suggested application			
			Epoxy/graphite composite	Kevlar	Fiberglass	Honeycomb
Blind rivets[a]	5056 aluminum	None	Not recommended[h]	Excellent[h]	Excellent[h]	[e]
	Monel	None	Good[h]	Excellent[h]	Excellent[h]	[e]
	A-286	Passivated	Good[h]	Excellent[h]	Excellent[h]	[e]
Blind bolts[b]	A-286	Passivated	Excellent[h]	Excellent[h]	Excellent	[e]
	Alloy steel	Cadmium	Not recommended[h]	Excellent[h]	Excellent	[e]
Pull-type lockbolts	Titanium	None	Excellent[d]	Excellent[c]	Excellent[c]	Good or not recommended[f]
Stump-type lockbolts	Titanium	None	Excellent[d]	Excellent[c]	Excellent[c]	Good or not recommended[f]
Asp fasteners	Alloy steel	Cadmium/nickel	Good[g]	Excellent	Excellent	Excellent
Pull-type lockbolts	7075 Aluminum	Anodized	Not recommended	Excellent	Excellent	Not recommended

(a) Blind rivets with controlled shank expansion.
(b) Blind bolts are not shank expanding.
(c) Fasteners can be used with flanged titanium collars or standard aluminum collars.
(d) Use flanged titanium collar.
(e) Performance in honeycomb should be substantiated by installation testing.
(f) Depending on fastener design; check with manufacturer.
(g) Nickel-plated Asp only.
(h) Metallic structure on backside.
Source: Ref 1

Specifications are used to outline fastener requirements, to control the manufacturing process, and to establish functional or performance standards. Their goal is to ensure that fasteners will be interchangeable, dimensionally and functionally. The most common fastener specifications are product specifications that are set up to govern and define the quality and reliability of fasteners before they leave the manufacturer. These specifications determine what material to use; state the objectives for tensile strength, shear strength, and response to heat treatment; set requirements for environmental temperature or atmospheric exposure; and define methods for testing and evaluation. To ensure proper use of fasteners, additional specifications are necessary as a guide to proper design application and installation.

All ASTM and SAE specifications covering threaded fasteners require that the heads be marked for grade identification. Grade markings are a safety device that provides a positive check on selection, use, and inspection. The markings reduce the possibility of selecting and using a bolt of insufficient strength, which might lead to a failure and cause damage to equipment or injury of personnel. The failure analyst should be familiar with the different grade markings required by the specifications.

In an effort to standardize the reporting of fastener test failures, a failure-identification code has been established for reporting the failure of each fastener resulting from tests under MIL-STD-1312. Failures of threaded and blind fasteners and of fastened sheet covered by this identification code are illustrated in Fig. 1.

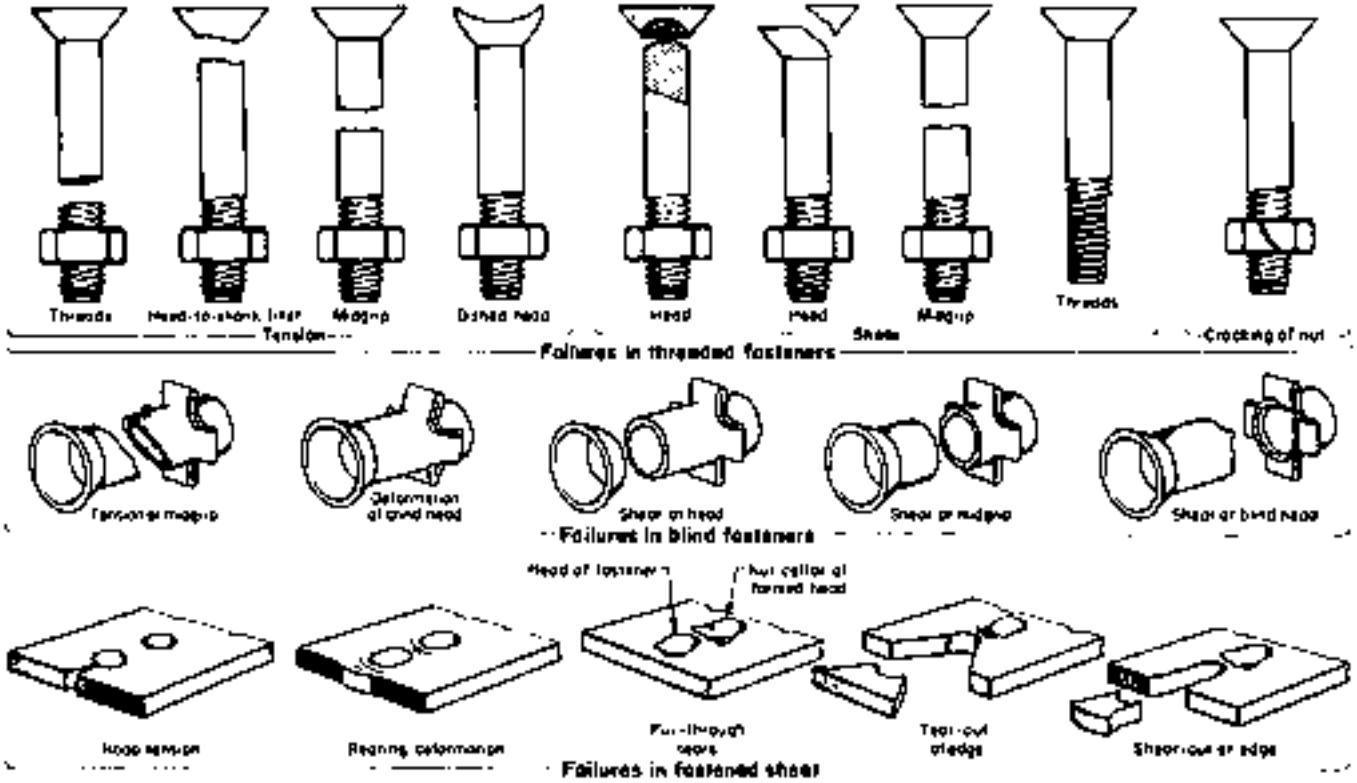


Fig. 1 Types of failures in threaded and blind fasteners and in fastened sheet. Source: Ref 2

References cited in this section

1. Mach. Des., 1985 Fastener and Joining Reference Issue, Vol 54 (No. 27), 1985, p 141

2. *Metallic Materials and Elements for Aerospace Vehicle Structures,* Vol II, Military Handbook, Figure 9.4.1.7(a)

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Failure Origins

The most common locations for fastener failure are in the head-to-shank fillet or, on threaded fasteners, through the first thread inside the nut or at the transition from the thread to the shank. Failure origins are often imperfections in the metal caused by segregation in the form of inclusions, such as sulfides and oxides in the ingot, or by folds, laps, or seams that have formed because of faulty working either in the semifinishing or finishing mills.

Seams are crevices in the surface of the metal that have been closed, but not welded, by working the metal. Seams seldom penetrate to the core of bar stock. Seams may cause additional cracking in hot forging and quench cracking during heat treatment. Oxides in a seam can be the result of exposure to elevated temperatures during processing. Seams are sometimes difficult to detect in an unused fastener, but they are readily apparent after a fastener has been subjected to installation and service stresses, as discussed below.

During periodic inspection of an airplane, a longitudinal crack was found in one of the wing-attachment bolts. The bolt was 22 mm ($\frac{7}{8}$ in.) in diameter and 89 mm ($3\frac{1}{2}$ in.) in length and conformed to National Aerospace Standard 634. Chemical analysis verified that the bolt material was 4340 steel, as called for in specifications, and hardness measurements indicated that it had been heat treated to a tensile strength of 1379 MPa (200 ksi).

As indicated by the arrows on the photograph of the bolt in Fig. 2(a), the crack extended almost the full length of the bolt shank, but it branched circumferentially at the head-to-shank fillet (arrows, Fig. 2b) and extended about halfway through

the bolt-head diameter (arrows, Fig. 2c). The bolt was sectioned to expose a portion of the crack surface, which was found to be heavily oxidized (Fig. 2d). Figure 2(e) shows a micrograph of a cross section transverse to the crack that illustrates the thickness and uniformity of the oxide coating. This heavy oxide coating and the longitudinal direction of the crack indicated that a seam was present in the bolt bar stock. Hot heading had disrupted the linear direction of the seam in the bolt-head areas so that the crack branched in a circumferential direction. Installation and service loads apparently opened the seam, which made it visually detectable.

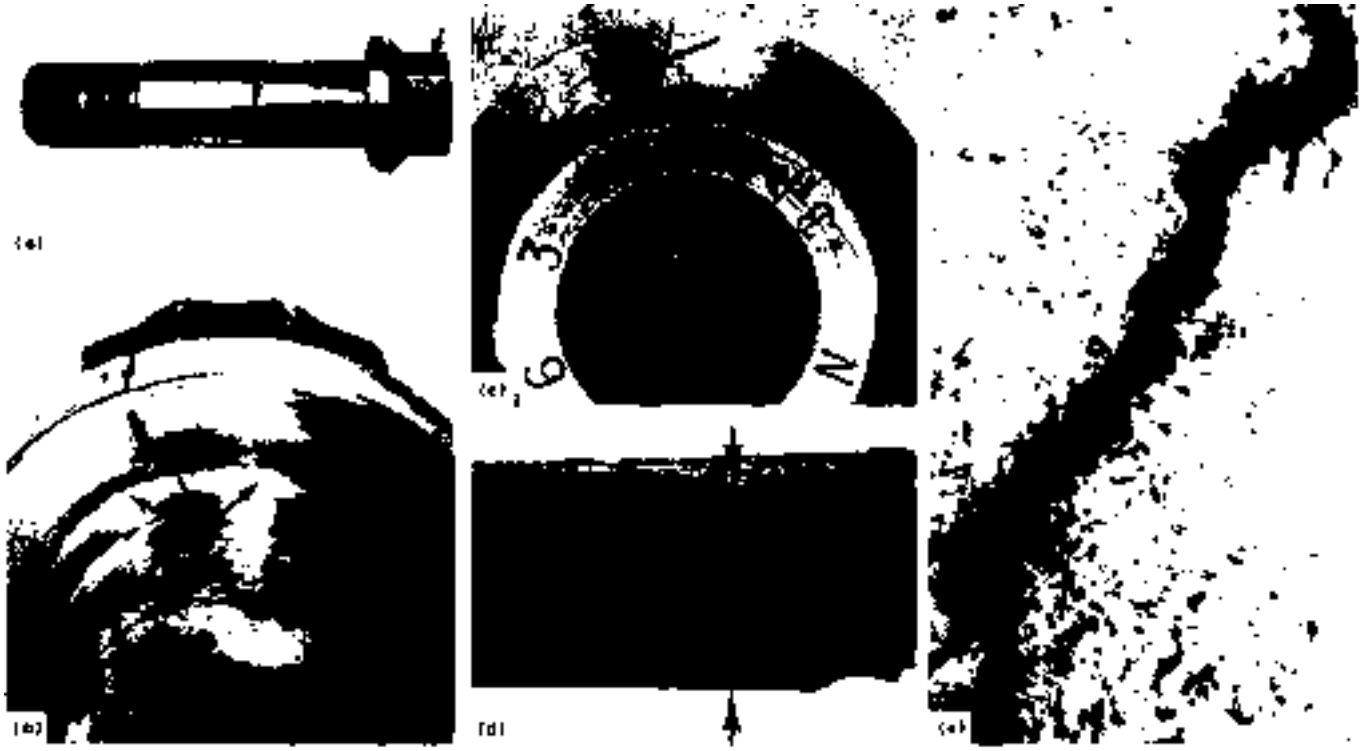


Fig. 2 4340 steel wing-attachment bolt that cracked along a seam. (a) Bolt showing crack (arrows) along entire length. (b) Branching cracks (arrows) at head-to-shank radius. (c) Head of bolt showing cracking (arrows) about halfway through bolt-head diameter. (d) Section through bolt showing heavy oxidation on crack surface (between arrows). (e) Section through crack showing thickness and uniformity of oxide coating. Etched with 2% nital. 1000×

Corrective action was not deemed necessary, because the periodic scheduled inspection of the aircraft was considered adequate to identify any faulty part for replacement and thus prevent failure of any safety-of-flight components.

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Causes of Fastener Failures

In a well-designed mechanically fastened joint, the fastener may be subjected to either static loading (overload) or dynamic fatigue loading. Static loading may be tension, shear, bending, or torsion—either singly or in combination. Dynamic forces may result from impact or from cyclic fatigue loading, including vibration. Pure shear failures are usually obtained only when the shear load is transmitted over a very short length of the member, as with rivets, screws, and bolts. When the ultimate shear stress is relatively low, a pure shear failure may result, but in general, a member subjected to a shear load fails under the action of the resulting combined stresses. In addition to fatigue and overload, other common causes of fastener failures include environmental effects, manufacturing discrepancies, and improper use or incorrect installation. The following illustrates a common manufacturing discrepancy.

During an assembly operation involving HL 22-8 fasteners, the fastener heads would sometimes crack (Fig. 3a and b). Aluminum alloy 7075-T73 nuts were being installed on type 7075-T6 fasteners with an installation torque range of 7.3 to 8.5 N · m (65 to 75 in. · lb). Dimpled fracture features indicated that the head cracks were caused by overload; however, chemical analyses, hardness tests, and microstructural examination showed that the fasteners were within specifications. Unused fasteners and nuts from the same batch were installed in a simulated joint in an effort to duplicate the head cracking. Overtorquing of the fasteners caused nut-thread stripping at torques of 10.2 to 11.3 N · m (90 to 100 in. · lb). Additional torque tests using stronger steel nuts stripped the fastener threads but did not produce any head cracks. However, when the fastener holes were drilled 2 to 4° off normal to the bearing surface, the head cracks shown in Fig. 3(c) were produced with torques of 7.9 to 10.2 N · in (70 to 90 in. · lb). The mandatory use of a drill guide to prevent off-line holes eliminated the problem.

A threaded fastener may be either a male or a female threaded cylinder—both of which are necessary as a two-element device for applying force, primarily to clamp two members together. Bolts, screws, and studs are male threaded cylinders; nuts or tapped holes are female threaded cylinders. For this discussion, bolts and nuts are considered representative of the two.

The primary force in the bolt is tension, which is set up by stretching the bolt during tightening, whereas the most important stress in the nut is the shear stress in the threads. If both the bolt and the nut had perfectly matched threads and both elements were inelastic, the torque load would be evenly distributed over all of the engaged threads. However, because the economy of mass production of fasteners requires that all fasteners be produced with practical dimensional tolerances and because all fastener materials are elastic, all threaded fastener devices experience nonuniform distribution of load over the threads. The bolt tension causes elongation, which results in more load on the threads near the bearing face of the nut, and the compression load of the nut is likely to have the same load-concentrating effect. To obtain a more uniform distribution of load along the threaded cylinder and thus minimize the chance of failure, a more ductile material must be provided for one of the elements. The usual combination is to provide material with high tensile strength for the bolt and material with good ductility for the nut (Ref 3).

In the theory of joint design, the ideal application can be represented as a situation in which there is a rigid structure and a flexible or elastic fastener. The bolt then behaves like a spring (Fig. 4a). When the bolt is preloaded or when the spring is stretched, a stress is induced in the bolt (and a clamping force in the structure) before any working load is encountered. As the working load is applied, the preloaded bolt does not encounter an additional load until the working load equals the preload in the bolt (Fig. 4b). At this point, the force between members is zero. As more load is applied, the bolt must stretch. Only beyond this point will any cyclic working load be transmitted to the bolt.

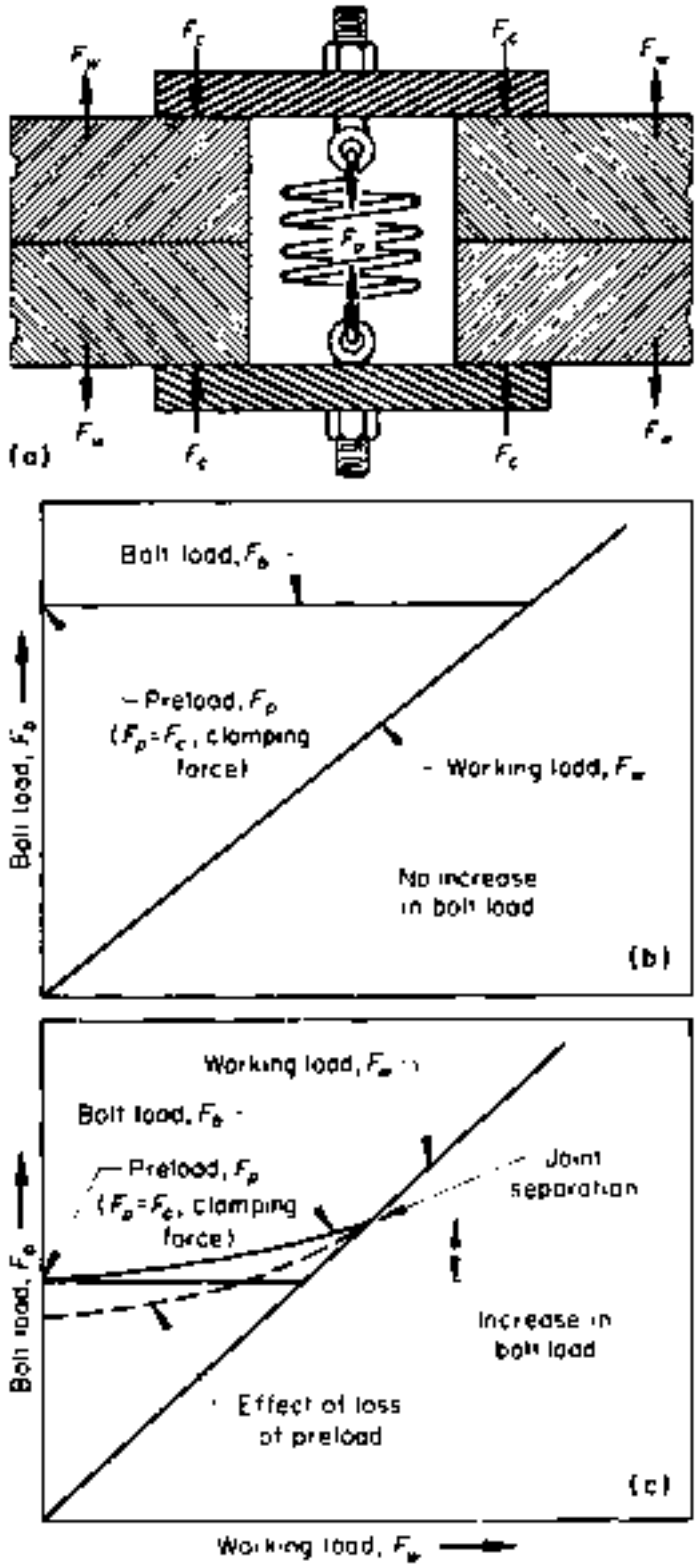


Fig. 4 Schematic showing the springlike effect of loading conditions on bolted joints. (a) Theoretical load condition for an elastic fastener and a rigid structure. (b) Ideal relationship of bolt load to working load with an elastic fastener and a rigid structure. (c) Actual relationship (both fastener and structure elastic) of bolt load to working load

For schematic representation of an actual condition, the structural material would also be represented by springs. The bolt would also be preloaded, but before application of the working load, the bolt encounters a slight increase in load because of the elasticity of the structural members. Actually, some of the additional load is encountered before the working load equals the preload. As the dynamic working load cycles, a fatigue condition is established.

When the bolt encounters an increase in load, being elastic, it stretches. If this stretch (strain) exceeds the elastic limit, the bolt yields plastically, taking a permanent set. The result is a loss in preload or clamping force (dashed curve, Fig. 4c). With a fluctuating load, this situation can cycle progressively, with continued loss of preload and possibly rapid fatigue failure (the effect of preload on steel automotive wheel studs and the fracture that resulted from the preload are described in Example in this article).

To eliminate fatigue problems that occur at room temperature, the designer should specify as high an initial preload as practical. The optimum fastener-torque values for applying specific loads to the joint have been determined for many high-strength fasteners. However, these values should be used with caution, because the tension produced by a selected torque value depends directly on the friction between the contacting threads. The use of an effective lubricant on the threads may result in overloading of the fastener, whereas the use of a less effective lubricant may result in a loose joint. With proper selection of materials, proper design of bolt-and-nut bearing surfaces, and the use of locking devices, the assumption is that the initial clamping force will be sustained during the life of the fastened joint. This assumption cannot be made in elevated-temperature design.

At elevated temperature, the induced bolt load will decrease with time as a result of creep, even if the elastic limits of the materials are not exceeded, and this can adversely affect fastener performance. Therefore, it is necessary to compensate for high-temperature conditions in advance when assembling the joint at room temperature. Failures that occur at elevated temperatures necessitate evaluation of the properties of the fastener material relative to the applied loads, operating temperature, and time under load at temperature.

Reference cited in this section

3. J.S. Davey, Why Nuts Should Have Their Own Specifications, *Fasteners,* Vol 13 (No. 3), Fall 1958

Fretting

Fretting results from a very small movement of one highly loaded surface over another. The slight movement prevents the formation of, or destroys, any protective oxide film. Thus, continually clean surfaces become available for oxidation. Also, because there is no oxide film to prevent true metal-to-metal contact, local bonding of microscopic areas can occur, ultimately resulting in dislodgment of extremely small metal particles from the surfaces. Because of their relatively small size, these metal particles are then rapidly oxidized by the ambient atmosphere.

Bolted machine parts that are subjected to vibration are the most susceptible to fretting. Fretting can be prevented by minimizing clearance in boltholes and by using high-strength fasteners properly tensioned to prevent relative motion of contacting surfaces.

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Corrosion in Threaded Fasteners

Corrosion is electrochemical in nature; therefore, an electrolyte must be present. Water is an excellent electrolyte, especially if it contains dissolved minerals. Temperature is also important because little reaction occurs below 5 °C (40 °F). Oxygen is required and is present in the atmosphere and, to a lesser extent, underwater and underground.

Types of corrosion include atmospheric corrosion, liquid-immersion corrosion, crevice corrosion, galvanic corrosion, stress-corrosion cracking, and hydrogen damage. Detailed information on these types of failures is available in the articles “Stress-Corrosion Cracking,” and “Hydrogen Damage and Embrittlement” in this Volume.

Atmospheric Corrosion. Atmospheric contaminants vary widely with location; therefore, rates of atmospheric corrosion vary widely. For ferrous materials, because of sulfur and chloride compounds, most severe corrosion generally occurs in highly industrialized and severe marine locations, with rural exposure being much less corrosive.

Liquid-Immersion Corrosion. Both fresh water and saltwater are corrosive. Saltwater is usually more severe than fresh water because saltwater is a better electrolyte and because the electrochemical current generated by inhomogeneities in the same metal or by different metals is greater. Also, the chloride ions present break down the protective oxide film and corrosion products. Completely submerged locations are not as corrosive as those at the waterline or splash zone, where more oxygen is present and where the surface is alternately wet and dry.

Crevice Corrosion. Every bolted connection offers the potential for crevice corrosion. Corrosion may be of two types, depending on the tightness of the joint or crevice opening. If there is a gap or opening, dirt can collect and retain moisture and thus enhance corrosion. On the other hand, in a correctly made tight joint, corrosion can result because of an oxygen deficiency or oxygen-starvation cell. The area of low oxygen becomes an anode, and the high-oxygen area becomes a cathode; thus, corrosion occurs. Fortunately, the ferrous metals most used for fasteners are not highly susceptible to the second type of crevice corrosion. Stainless steels and other metals that, in the presence of oxygen, form protective passive surface films are much more susceptible as the film breaks down and creates an active metal area.

Galvanic Corrosion. When dissimilar metals are in contact or are electrically connected together, galvanic corrosion results because of the electrical potential existing between the two metals. Current thus produced has a great effect on metal corrosion and must be considered whenever different metals are to be fastened together.

Galvanic corrosion particularly should be kept in mind at all times when designing structures using fasteners. Too often, errors are made—for example, using ferrous fasteners to join more noble metals such as copper alloys. In this case, galvanic corrosion and the size effect are involved, as described below.

For a structure to be well designed, the anode or corroding-metal area should be large relative to the cathode or protected metal in order to ensure reasonable anode life. With fasteners, this is sometimes difficult to accomplish because fasteners are always smaller than the structure they join. In one example of size effect, iron nails were used to fasten copper sheathing to the bottom of wooden ships to prevent marine growth. At the end of each voyage, most of this sheathing had fallen off because of the accelerated corrosion of the iron nails.

Galvanic corrosion can be avoided by using the same metal for both fastener and structure. Several alternatives exist when this is impractical or impossible: the structure may be painted, the fastener may be isolated by non-conducting materials or plated with a more noble metal, moisture may be prevented from contacting the couple, or fasteners may be made of a metal that will be protected by sacrifical corrosion of the much larger mass of the metal being fastened.

Stress-corrosion cracking (SCC) is an intergranular fracture mechanism that sometimes occurs in highly stressed fasteners after a period of time, and it is caused by a corrosive environment in conjunction with a sustained tensile stress above a threshold value. An adverse grain orientation increases susceptibility of some materials to stress corrosion. Consequently, SCC can be prevented by excluding the corrodent, by keeping the static tensile stress of the fastener below the critical level for the material and grain orientation involved, or by changing to a less susceptible material or material condition (for example, aluminum alloy 7075 is less susceptible to SCC in the T73 or T76 temper than in the T6 temper). In some cases, the corrodent can be the seemingly innocuous ambient atmosphere; in others, it can be unwittingly added, as in the following example.

Example 4: SCC of Stainless Steel Bolts. Stainless steel bolts broke after short-term exposure in boiler feed-pump applications. Specifications required that the bolts be made of a 12% Cr high-strength steel with a composition conforming to that of AISI type 410 stainless steel.

Several bolts from three different installations were sent to the laboratory for examination. One lot consisted of ten 100-mm (4-in.) long 1–8 UNC bolts, the second lot nine 100-mm (4-in.) long 1–8 UNC bolts, and the third lot five 70-mm (2.75-in.) long $\frac{5}{8}$-11 UNC bolts. An antiseizure compound consisting of metallic copper flakes in a silicone vehicle had been applied to the bolts during assembly.

All of the bolts had fractured in the threaded portion, but not all at the same thread. None of the 25-mm (1-in.) diam bolts had fractured under the head, but some of the 16-mm ($\frac{5}{8}$-in.) diam bolts had broken under the head and in the threads.

Visual Examination. The 25-mm, (1-in.) diam bolts exhibited external damage, which varied in degree, depending on the time that had been required to shut down the unit. No external damage was observed on the 16-mm ($\frac{5}{8}$-in.) diam bolts.

The first lot of 25-mm (1-in.) diam bolts exhibited flat fracture surfaces with chevron patterns. The second lot was similar in fracture-surface appearance; however, the fracture surfaces were badly stained by the corrosive boiler feedwater. The fracture surfaces of the 16-mm ($\frac{5}{8}$-in.) diam bolts exhibited stained brittle-fracture features in starlike patterns.

Examination of bolts in all three lots also disclosed a copper-colored residue adhering to the threads. This residue was from the antiseizure compound that had been applied during assembly.

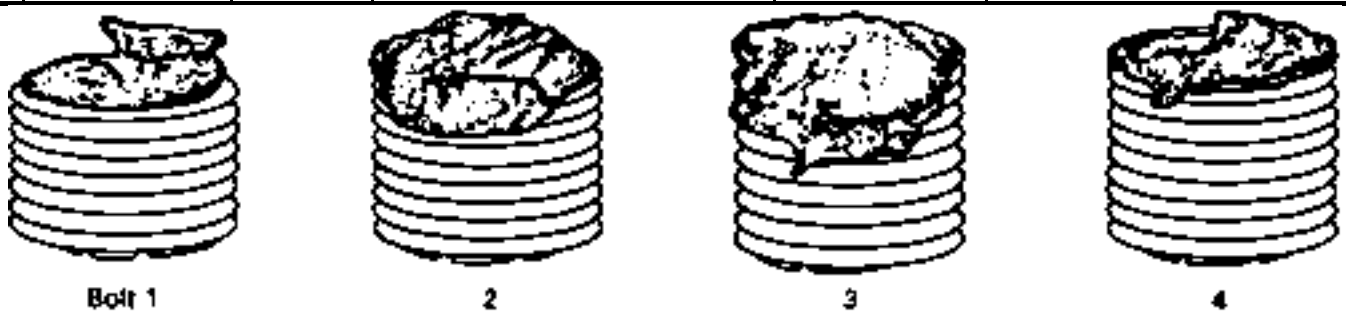
Accelerated Corrosion Testing. Laboratory corrosion tests were conducted on four bolts taken from stock. The purpose of the tests was to determine the effects of the antiseizure compound in acidified 3.5% NaCl solution on bolts at two hardness levels.

Two bolts, one with a hardness of 41 to 42 HRC and the other of 20 to 22 HRC, were coated with the antiseizure compound. Another bolt that had a hardness of 41 to 42 HRC and a bolt having a hardness of 18 to 20 HRC were left uncoated for the test. All four bolts were assembled in a static test fixture and tightened to a torque value equal to 90% of their yield strength and then intermittently dipped in an acidified solution of 3.5% NaCl. Intermittent dipping consisted of a 10-min total immersion and a 50-min air dry for 8 h per day, 5 days per week for 3 months.

After 3 months of testing, the coated bolt having a hardness of 41 to 42 HRC fractured completely across at the third thread from the head. The three other bolts had not fractured, but were tensile tested to destruction. The results of the test and appearance of the fracture surfaces are presented in Table 2. This test demonstrated the damaging effects of the metallic copper antiseizure compound to high-hardness 12% Cr steels in the presence of a corrosive medium.

Table 2 Results of stress-corrosion tests of 12% Cr steel bolts in a 3.5% NaCl solution to determine effect of antiseizure coating

Bolt	**Hardness, HRC**	**Surface**	**Appearance[a]**	**Tensile strength**		Remarks
				MPa	**ksi**	
1	41–42	Uncoated	Localized pitting, no cracks	1524	221	**No corrosion attack on fracture surface**
2	18–20	Uncoated	Localized pitting, no cracks	862	125	**No corrosion attack on fracture surface**
3	20–22	Coated[b]	Heavy pitting, no cracks	896	130	**No corrosion attack on fracture surface**
4	**41–42**	**Coated[b]**	**Bolt completely failed in test fixture**	**Not applicable**		**Heavy corrosion on fracture surface**



(a) After 3 months of intermittent dipping in a 3.5% NaCl solution.

(b) Coated with an antiseizure compound consisting of metallic copper flakes in a silicone vehicle

Microscopic examination showed that the microstructure of the metal in the bolts of lot 1 (Fig. 9a) contained tempered martensite. The metal in the two other lots (Fig. 9b) showed a two-phase structure consisting of tempered martensite, ferrite, and nonmetallic sulfide stringers. Ferrite content of the matrix was about 38%.

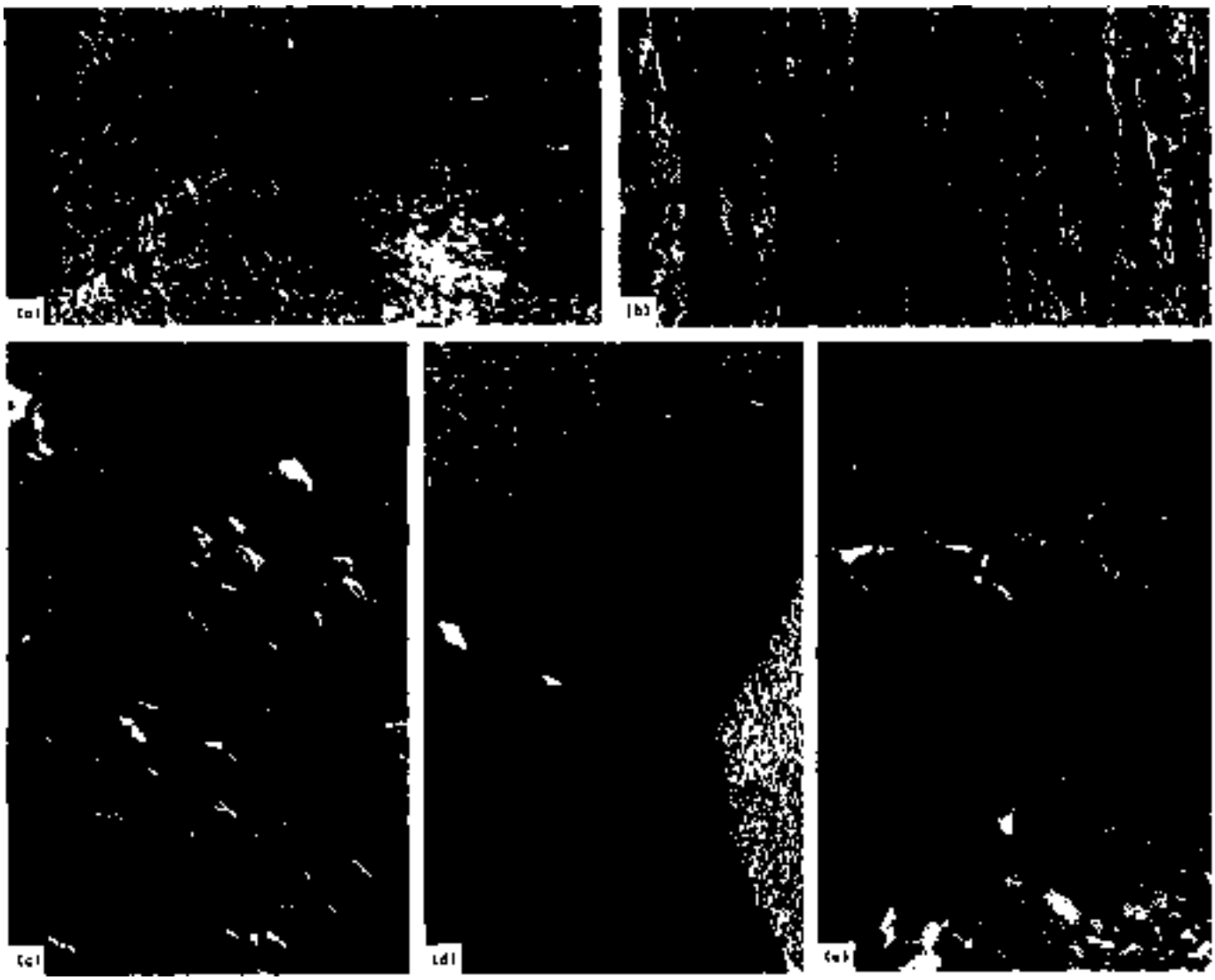


Fig. 9 Microstructures of stainless steel bolts that failed from SCC. (a) Branched intergranular cracking in a type 410 stainless steel bolt from lot 1 (see Example). Etched with picral plus HCl. 250×. (b) Microstructure of a type 416 stainless steel bolt typical of those in lots 2 and 3 (see Example). Same etchant as (a). 100×. (c) to (e) Typical fracture surfaces of the bolts showing carbide particles (c), intergranular fracture (d), and intergranular fracture with corrosive residue (e). Electron fractographs of unetched specimens. 6500×

The examination also showed that the threads on some of the bolts had been produced by cutting, and on others by rolling. Both types of threads had fractured on the bolts in service.

Longitudinal sections were taken through the fractured bolts and examined microscopically. Branched intergranular cracking at the thread root next to the fracture, similar in appearance to the crack shown in Fig. 9(a), was found in several of the bolts. Intergranular cracking and oxide entrapment were found under the bolt heads and at the fracture surfaces.

Fractographic examination by electron microscopy disclosed evidence of corrosion products on the fracture surfaces with localized pitting and second-phase carbide particles. The carbide particles (probably chromium carbide, $Cr_{23}C_6$) and evidence of intergranular fracture are shown in Fig. 9(c). The fractograph shown in Fig. 9(d) illustrates brittle intergranular fracture with corrosive residue on the surface. Large clusters of carbide particles and evidence of intergranular fracture are shown in Fig. 9(e). No fatigue indications were in evidence. The branching intergranular cracks shown in Fig. 9(a) and the intergranular fracture surface with corrosion products shown by the electron fractographs in Fig. 9(d) and (e) are evidence of stress corrosion.

Chemical analysis of the metal in a fractured bolt from each lot revealed the following compositions:

Element	Composition, %			
	Lot 1	**Lot 2**	**Lot 3**	Nominal[(a)]
C	0.140	0.058	0.061	**0.15**
Mn	0.560	0.370	0.410	**1.00**
P	0.010	0.017	0.018	**0.04**
S	0.016	0.251	0.241	**0.03**
Si	0.392	0.437	0.395	**1.00**
Cr	**12.09**	**12.54**	**12.14**	**11.5–13.5**

(a) Nominal composition for type 410 stainless steel (to which bolts were specified to conform). Percentages, except for chromium, are maximums.

Bolts from lot 1 conformed to specifications, but a deviation from prescribed chemical composition was found in lots 2 and 3, which also contained a sulfur addition to enhance machinability.

Conclusions. Fracture of the bolts was by intergranular stress corrosion. A metallic copper-containing antiseizure compound on the bolts in a corrosive medium set up an electro-chemical cell that produced trenchlike fissures or pits for fracture initiation. Because the bolts were not subjected to cyclic loading, fatigue or corrosion fatigue was not possible.

Corrective Measures. Bolts were required to conform to the specified chemical composition. The hardness range for the bolts was changed from 35 to 45 HRC to 18 to 24 HRC. Petroleum jelly was used as an antiseizure lubricant in place of the copper-containing compound. As a result of these changes, bolt life was increased to more than 3 years.

High static tensile stresses above the threshold for SCC can be minimized by introducing compressive stresses on the fastener surfaces where SCC is most likely to originate. Cold rolling and shot peening are two means of keeping static tensile stresses at the operating surface below the safe maximum. However, the beneficial effects of either of these will be negated if the fastener is subsequently heat treated, if the normal service temperature is above the tempering or aging temperature of the material, if the fastener is loaded above its yield point, or if general corrosion removes the compressively stressed layer.

Example 5: SCC of a T-Bolt. The T-bolt shown in Fig. 10(a) was part of the coupling for a bleed air duct of a jet engine on a transport plane. Specifications required that the 4.8-mm ($\frac{3}{16}$-in.) diam T-bolt be made of AISI type 431 stainless steel and heat treated to 44 HRC. The operating temperature of the duct was 425 to 540 °C (800 to 1000 °F), but that of the bolt was lower. The T-bolt broke after 3 years of service. The expected service life was equal to that of the aircraft.

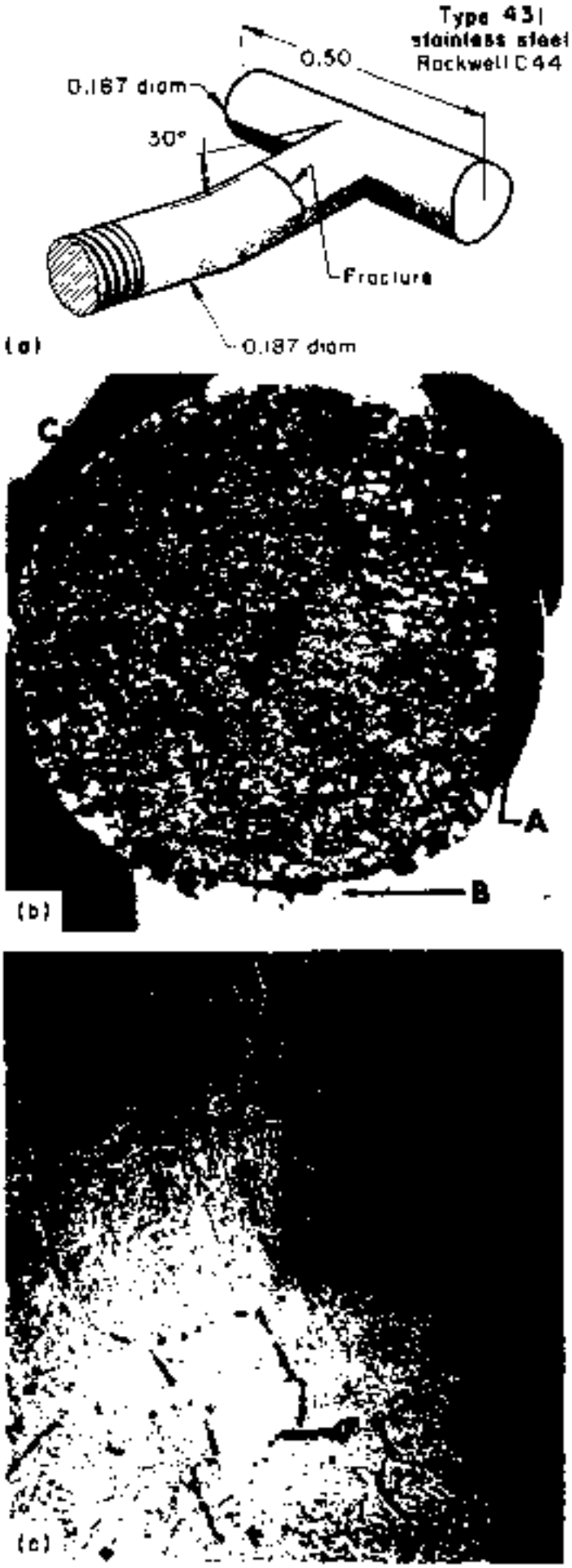

Type 431
stainless steel
Rockwell C44
0.50
0.187 diam
30°
Fracture
(a)
0.187 diam
C
A
B
(b)
(c)

Fig. 10 AISI type 431 stainless steel T-bolt that failed by SCC. (a) T-bolt showing location of fracture. Dimensions given in inches. (b) Fracture surface of the bolt showing shear lip (arrow A), fine-grain region (arrow B), and oxidized regions (arrows C). (c) Longitudinal section through the bolt showing severe intergranular carbide precipitation. Etched electrolytically with 10% NaCN. 540×

The bolt was made by cold upsetting bar stock to a T-shape, chasing the threads, bending to the specified curvature, then heat treating. Heat treatment consisted of heating to 1010 to 1040 °C (1850 to 1900 °F) and oil quenching, tempering at 275 to 415 °C (525 to 575 °F) for 2 h, and retempering for the same time and temperature. In the manufacture of similar bolts, the threads were rolled after heat treatment, and the bolt shank was shot peened.

Investigation. Visual examination of the fracture surface revealed intergranular cracking and evidence of brittle fracture. Macrographs were taken of the outer surface of the bolt and of the fracture surface. The outer surface contained circumferential cracks on the lower (tension) side of the T-bolt. The fracture surface (Fig. 10b) exhibited three distinct fracture areas—a shallow 45° shear-lip section along one side (arrow A, Fig. 10b), a thin region containing fine grains along part of the surface that could have been the result of stress corrosion or fatigue (arrow B, Fig. 10b), and regions that appeared to be oxidized as the result of a prior crack (arrows C, Fig. 10b). Hardness of the T-bolt was found to be 44 HRC, as specified.

Examination of micrographs of longitudinal sections through the T-bolt revealed the basic microstructure to be acicular martensite with no δ-ferrite present. However, there was evidence of severe intergranular carbide precipitation (Fig. 10c), which indicated that the material had been subjected to the sensitizing-temperature range after the bolt had been manufactured.

The micrographs also showed some intergranular surface and subsurface cracking near, but not at, the primary fracture. Although not confirmed, chlorine or chlorides in some form could have been present around the duct.

Conclusions. The bolt broke as a result of SCC. Thermal stresses were induced into the bolt by intermittent operation of the jet engine. Mechanical stresses were induced by tightening of the clamp around the duct, which in effect acted to straighten the bolt. The action of these stresses on the carbides that precipitated in the grain boundaries resulted in fracture of the bolt.

Corrective Measures. Because of the operating temperatures of the duct near the bolt, the material was changed to A-286, which is a heat-resisting alloy and is less susceptible to carbide precipitation. The bolt was strengthened by shot peening the shank and rolling the threads after heat treatment. Avoidance of heating in the sensitizing range would have been desirable, but difficult to ensure because of the application.

Example 6: SCC of a Die-Cast Zinc Alloy Nut. The two nuts shown in Fig. 11 (a) and (b) were used to secure the water-supply pipes to the threaded connections on hot-water and cold-water taps. The nut used on the cold-water tap (Fig. 11a) fractured about 1 week after installation.

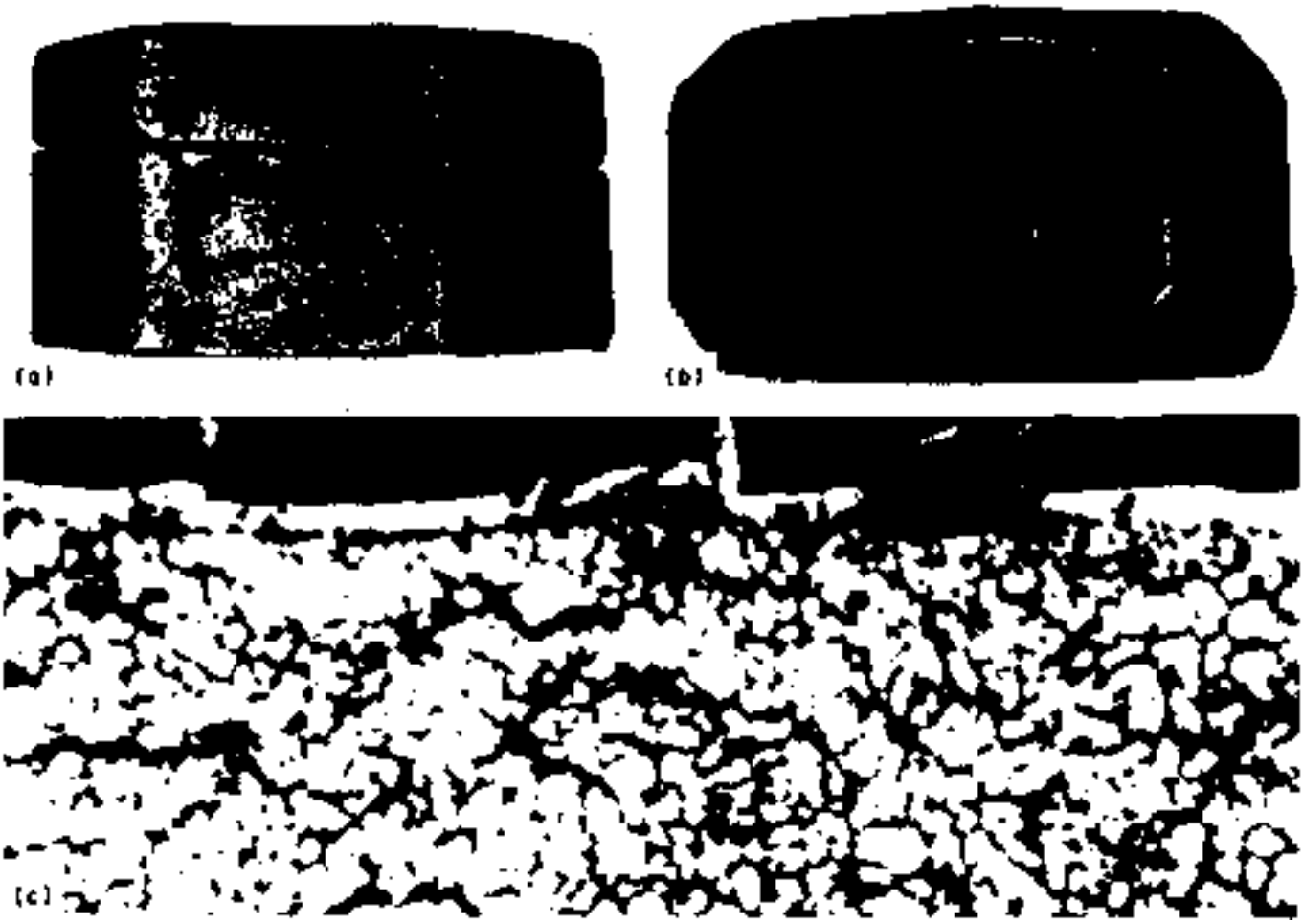


Fig. 11 Die-cast zinc alloy nuts from a water tap. (a) Nut for the cold-water tap that failed by SCC. (b) Mating nut for the hot-water top that shows only isolated areas of corrosion. (c) Unetched section showing metal in the cold-water tap after corrosion testing. 600×

The cold-water tap and fractured nut were submitted to the laboratory for examination to determine the cause of failure. The mating hot-water tap was forwarded for comparison.

Investigation. Visual examination of the two taps indicated that they were of similar design. As shown in Fig. 11(a) and (b), the two nuts on the threaded connections were of a different style, although they had the same size of thread. The two nuts did not exhibit any evidence of installation abuse, such as stripping of the threads. Various other marks and indentations were found on the flats of the nuts, particularly on the hot-water nut, such as would be made during normal installation.

Examination of the fracture surfaces of the cold-water nut did not reveal any obvious defects to account for the fracture, but there were indications of excessive porosity in the nut. The fracture had occurred through the root of the first thread that was adjacent to the flange of the tap.

Metallographic examination of the cold-water nut revealed evidence of considerable corrosive attack, mainly intergranular, extending into the material from the surface. This was present over the entire surface, including the threads, and extended into the material as deep as 0.25 mm (0.010 in.). Chemical spot tests indicated that the nut had been zinc plated.

Microscopic examination of the fracture surfaces of the cold-water nut indicated that considerable intergranular corrosion had also occurred at both sides of the fracture surface to a substantial depth, somewhat more than at the zinc-plated outer surface. Because this was a recent fracture, it would appear that corrosion on the fracture surface occurred before the failure rather than after.

Visual and microscopic examination of the hot-water nut also revealed some evidence of corrosion, but only in isolated areas mainly associated with the threads. This nut was plated with a copper-nickel coating. Corrosion was present only at breaks in the plating and then only to a limited extent.

Radiographic examination indicated that both nuts had excessive porosity in similar amounts and distribution. Metallographic examination revealed that both nuts had a cast structure and were of a similar basic composition of zinc-aluminum die-casting alloy.

Corrosion Testing. It appeared that the cold-water nut was highly susceptible to corrosive attack; the hot-water nut did not exhibit this characteristic to the same degree. Samples of nuts of both styles were subjected to a corrosion test designed to indicate whether a particular zinc-alloy casting was susceptible to intergranular attack in atmospheric service. The test consisted of exposing the nuts in air saturated with water vapor at a temperature of 90 °C (195 °F) for 10 days.

If the metal was attacked during this test, it indicated that there were deficiencies in magnesium content or that the content of the impurities (lead, cadmium, and tin) were above 0.004%, 0.003%, and 0.002%, respectively. In either case, fracture during the test would indicate that the parts are likely to fail in service.

After 4 days of testing, the cold-water nut was noted to be severely cracked and was therefore removed for further examination. Slight pressure between the fingers caused the nut to crumble into several fragments. Metallographic examination of the material in the cold-water nut indicated that it had completely disintegrated, as shown by the cracks and corrosion network in the micrograph in Fig. 11(c). The zinc plating on the outer surface appeared to be relatively unattacked.

The corrosion test for the hot-water nut was run for the full 10 days without any outward signs of attack. Metallographic examination revealed that there had been some additional corrosion, but not to as great an extent as in the cold-water nut.

Conclusions. The nut from the cold-water tap failed by SCC. Apparently, sufficient stress was developed in the nut to promote this type of failure by normal installation because there was no evidence of excessive tightening of the nut.

Corrosion testing of the nuts indicated that the fractured nut was highly susceptible to intergranular corrosion because of either a deficiency in magnesium content or excessive impurities, such as lead, tin, or cadmium. The primary cause of failure of the fractured nut was, therefore, an incorrect alloy composition.

Corrective Measure. This composition problem with zinc alloys was recognized many years ago, and particular attention has been directed toward ensuring that high-purity zinc is used. This corrective measure was reported to have resulted in virtual elimination of this type of defect.

Hydrogen Damage. Many fasteners require some kind of protective coating against corrosion. This need is sometimes satisfied by an electrodeposited coating, such as cadmium or zinc. An acid pickling bath is usually used to clean ferrous fasteners before plating. This pickling bath and the subsequent plating operation both supply nascent (atomic) hydrogen to the fastener surface, and some of the hydrogen is absorbed by the steel. In high-strength steel components that have been electroplated but have not been baked for expulsion of hydrogen, hydrogen diffuses to the area of the highest triaxial stress, such as at a notch. Figure 12 shows a cadmium-plated 8740 steel nut that failed by hydrogen embrittlement because the nuts were not baked after electroplating. After an incubation period, a crack initiates and, with time, propagates through the section. A characteristic of this phenomenon is that there is a stress value below which hydrogen-induced delayed failure will not occur. This is sometimes referred to as the static-fatigue limit.

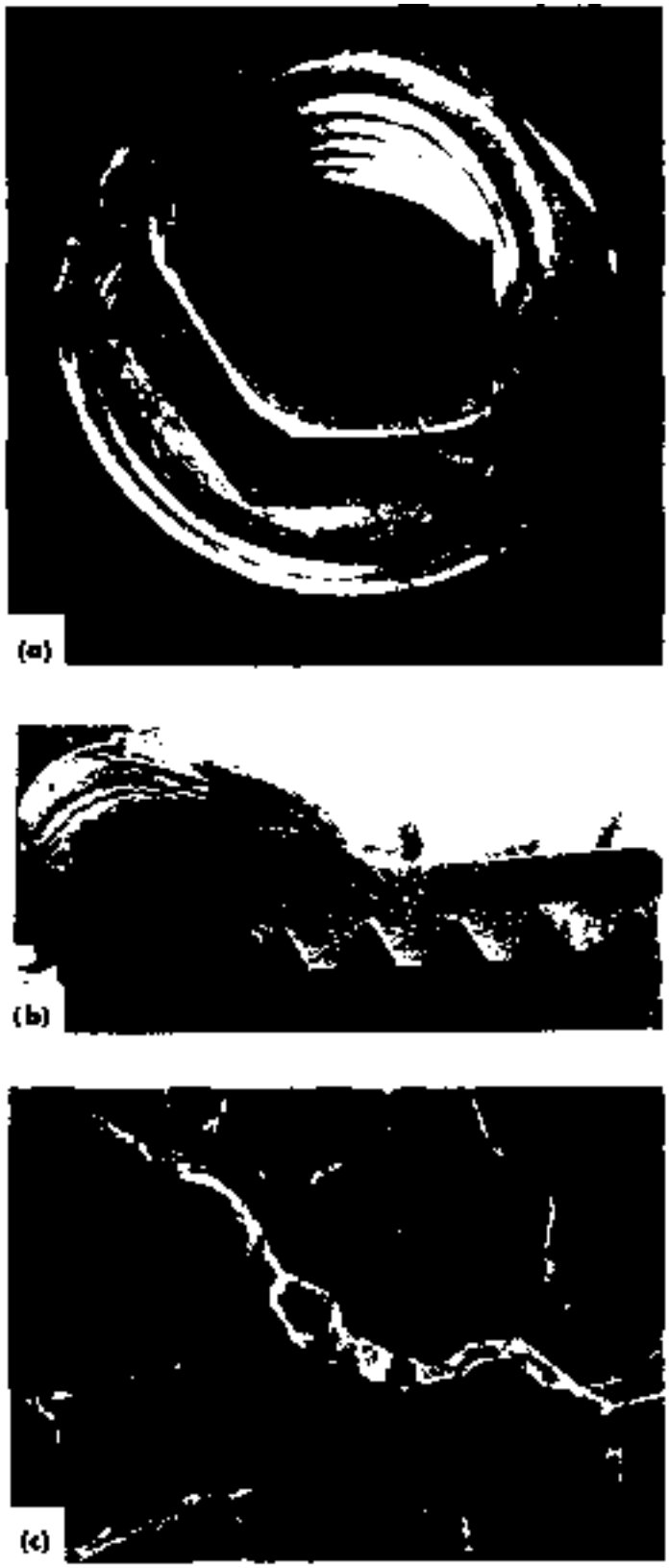


Fig. 12 Cadmium-plated 8740 steel aircraft-wing assembly nut that failed by hydrogen embrittlement. The nut was not baked after electroplating to release hydrogen. (a) Overall view. 5×. (b) Fracture surface. 9×. (c) Scanning electron micrograph of typical intergranular fracture shown in box in (b). 3600×

This difficulty, resulting from hydrogen charging during the pickling or electroplating process, can sometimes be eliminated without appreciable change in hardness by baking after plating at about 205 °C (400 °F) in air for a suitable time at temperature, depending on the size and finish of the fastener. In the following examples, cadmium-plated high-

strength steel bolts that had not been properly baked failed when normally stressed in service, and a case is described in which baking did not alleviate the hydrogen-embrittlement problem.

Example 7: Delayed Failure of an Electroplated Steel Bolt Because of Hydrogen Embrittlement. Cadmium-plated high-strength steel bolts were used to facilitate quick disassembly of a vehicle. One bolt was found fractured across the root of a thread after being torqued in place for 1 week. The bolts were made of 8735 steel heat treated to a tensile strength of 1241 to 1379 MPa (180 to 200 ksi) with a hardness of 39 to 43 HRC, followed by cadmium plating.

The bolt that failed and several that did not were sent to the laboratory for determination of the cause of failure. It was reported that the bolts that did not fail were stressed less than the specified torque level when attached to the assembly.

Investigation. It was assumed that the bolts had not been given a baking treatment for removal of hydrogen. Because there is no nondestructive test that will confirm such an assumption, two of the bolts were baked at 205 °C (400 °F) for 24 h, then a series of tests was performed on both the baked and the as-plated bolts.

First, short-term tensile tests were performed on one of the as-plated bolts and on one that had been baked. The two bolts exhibited similar notched tensile strengths (1448 MPa, or 210 ksi, for the as-plated bolt and 1510 MPa, or 219 ksi, for the baked bolt); therefore, this test did not disclose the presence of hydrogen. Consequently, the remaining bolts were tested under static loading.

The bolts were static loaded in tension on a creep frame. Delayed failure occurred in the as-plated bolts, but no failure occurred in the bolt that was baked to remove the hydrogen. Test results were as follows:

Condition	**Applied stress**		Time-to-failure, h
	Mpa	**ksi**	
As-plated	690	100	**5.1**
As-plated	517	75	**5.6**
As-plated	517	75	**1.1**
As-plated	345	50	**67**[(a)]
Baked[(b)]	**517**	**75**	**598**[(a)]

(a) No failure occurred; tests were concluded after time shown.

(b) Baked 24 h at 205 °C (400 °F)

These results indicated that the static-fatigue limit was between 345 and 517 MPa (50 and 75 ksi) for the as-plated (hydrogen-charged) bolts. The properly baked bolt did not fail when it was subjected to a torque stress of 517 MPa (75 ksi).

Conclusions. Failure of the bolts was the result of time-dependent hydrogen embrittlement. Had the remaining bolts been torqued to the normal stress levels, all would have failed within 2 weeks.

Corrective Measures. The bolts were baked, as specified by ASTM B 242, at 205 °C (400 °F) for 30 min. No further failures occurred. Baking for 30 min is the minimum baking time; however, baking times up to 24 h are recommended for greater safety.

Example 8: Hydrogen Embrittlement of Cadmium-Plated Alloy Steel Self-Retaining Bolts in a Throttle-Control Linkage. Two clevis-head self-retaining bolts used in the throttle-control linkage of a naval aircraft failed on the aircraft assembly line (a failed bolt is compared with an unused bolt in Fig. 13a). Specifications required the bolts to be heat treated to a hardness of 39 to 45 HRC, followed by cleaning, cadmium electroplating, and baking to minimize hydrogen embrittlement. The bolts broke at the junction of the head and shank. The nuts were, theoretically, installed fingertight.

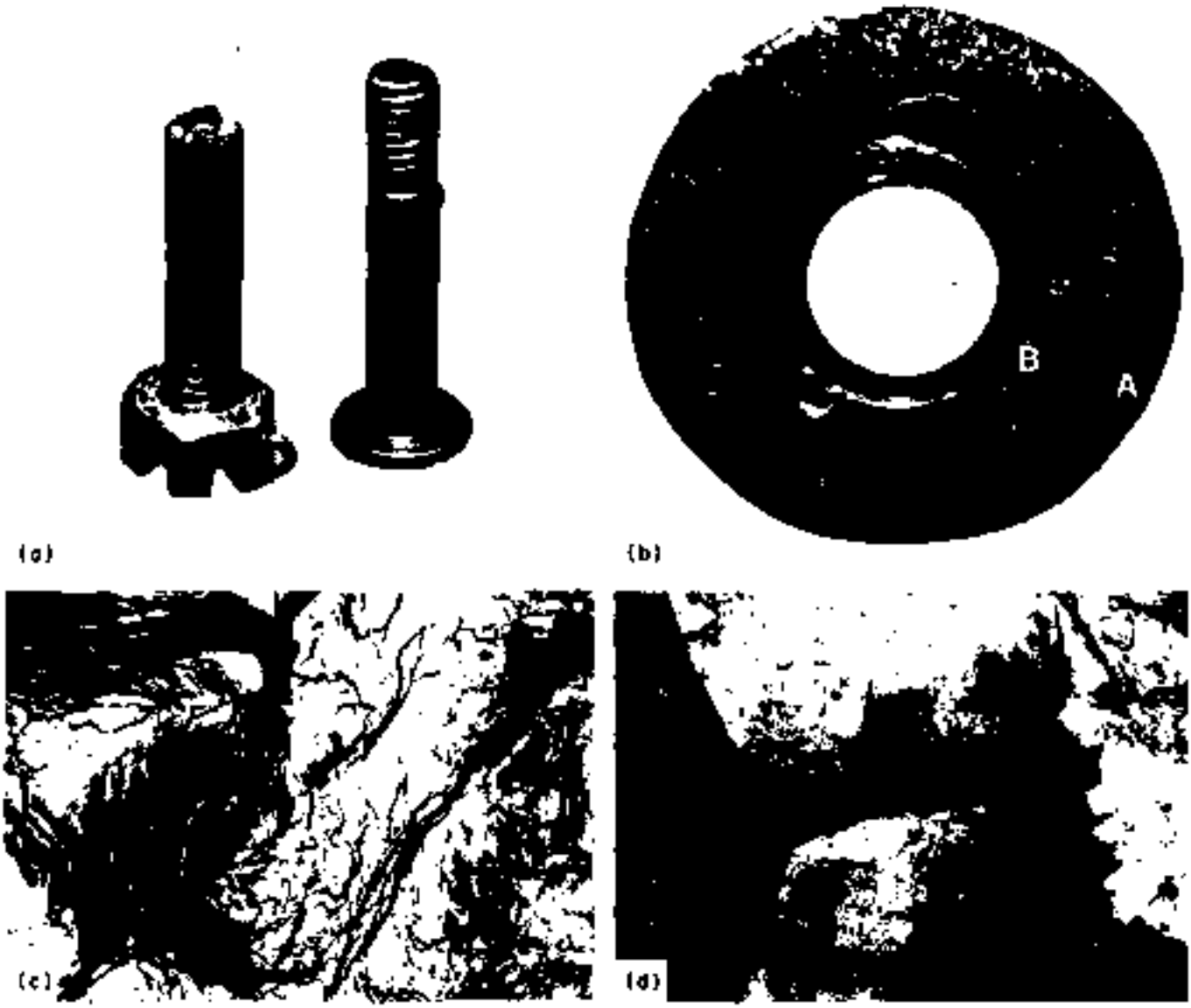


Fig. 13 Cadmium-plated alloy steel self-retaining bolts that fractured because of hydrogen damage. (a) Fractured and unused intact bolt. (b) Fractured bolt; brittle fracture surface is indicated by A and B. (c) and (d) Electron fractographs of surfaces A and B, respectively, showing brittle intergranular fractures. 5000×

Investigation. Inspection of the bolts and nuts showed excessive burrs in the threads at the cotter-pin holes and the nut slots, which made it necessary to use a wrench in assembly. This obviously had set up a tensile stress, which caused the bolts to fail. The fillet at the junction of the head and shank was undesirably sharp, but met the manufacturer's specification of 0. 13 to 0.25 mm (0.005 to 0.010 in.) for the fillet radius.

Macroscopic examination of the fracture surface (Fig. 13b) gave no evidence of corrosion, fatigue marks, or rapid ductile failure. Electron fractographs of the fracture surface (Fig. 13c and d) showed a typical brittle intergranular fracture, suggestive of hydrogen embrittlement. The microstructure was found to be sound and normal, and the hardness was within specifications.

Considerations of the event of failure before actual service and the possibility of hydrogen embrittlement that is inherent in cleaning and electroplating procedures led to the question of whether the bolts had been adequately baked after plating and whether postplating treatment had been performed as specified. To explore this possibility, bolts from the same lot as the failed bolts were tested in prolonged tension in three conditions: as-received, baked at 205 °C (400 °F) for 4 h, and baked at 205 °C (400 °F) for 24 h. The bolts in the first two conditions failed under a sustained tensile load, but the bolts baked for 24 h did not, indicating that baking by the manufacturer had either been omitted or had been insufficient.

Conclusions. The failure was attributed to hydrogen embrittlement (resulting from cleaning, cadmium plating, or both) that had not been satisfactorily alleviated by subsequent baking. The presence of the burrs on the threads prevented

assembly to fingertightness, and the consequent wrench torquing caused the actual fractures. The very small radius of the fillet between the bolt head and the shank undoubtedly accentuated the embrittling effect of the hydrogen.

Corrective Measures. The cleaning and cadmium-plating procedures were stipulated to be low-hydrogen in nature, and an adequate postplating baking treatment at 205 °C (400 °F), in conformity with ASTM B 242, was specified. A minimum radius for the head-to-shank fillet was specified at 0.25 mm (0.010 in.). All threads were required to be free of burrs. A 10-day sustained-load test was specified for a sample quantity of bolts from each lot.

Example 9: Hydrogen Embrittlement of Alloy Steel Fasteners (Ref 5). During an inspection of a structure 2 weeks after assembly, the heads of several cadmium-plated AISI 8740 steel fasteners were found to be completely separated from their respective shanks (Fig. 14). Scanning electron microscopy (SEM) examination of the fracture surfaces revealed a brittle, intergranular fracture mode (Fig. 15), indicating hydrogen embrittlement. An investigation was conducted to determine the extent of hydrogen embrittlement in the various lots of cadmium-plated 8740 steel fasteners.

Fig. 14 Cadmium-plated AISI 8740 alloy steel fasteners that failed by hydrogen embrittlement. See also Fig. 15.

Fig. 15 Scanning electron micrograph of fracture surface of fasteners shown in Fig. 14. 880×

Mechanical Testing. To simulate installation conditions, fasteners from several lots were torqued to 80% of rated ultimate tensile strength into a predrilled aluminum alloy plate. This high torque (greater than normal installation) was used to speed the failure rate of the fasteners.

After 335 h, a head was lifted with fingernail pressure from the shank of a fastener. Another was removed by inserting the blade of a pocket knife under the head. After 41 days, the test rig was disassembled. Nine more fasteners were found to be cracked in the head-to-shank radius. All of the failures were brittle in nature.

Stress-rupture tests. When torque tests were abandoned, all subsequently tested fasteners were installed in fixtures in which a constant load equivalent to 90% of the rated ultimate tensile strength could be maintained on the specimens. Stress-rupture testing was performed on lever-arm-type creep machines. Testing of any lot of pins was suspended when one pin failed (indicating hydrogen embrittlement) before the end of the 72-h test period. Four pins per lot were used as a representative sample. If none of the four pins failed during the 72 h, the lot was considered to be free of hydrogen embrittlement and was released for production. If a lot was embrittled, one of the fasteners being tested generally failed within the first 4 h.

Metallurgical examination showed that the microstructure of the steel base metal was tempered martensite, supposedly heat treated to an ultimate tensile strength range of 1240 to 1380 MPa (180 to 200 ksi). Testing revealed that the actual ultimate tensile strength of the fasteners varied from 1310 to 1480 MPa (190 to 215 ksi). The grain-flow pattern in the head-to-shank radius was satisfactory.

Condition of cadmium plating. The thickness of the cadmium plating on the fasteners varied from 7.9 to 11 μm (0.31 to 0.43 mils). The coating was applied from a plating bath to which brighteners were added. Bright, impervious coatings increase the possibility of hydrogen embrittlement in steels by forming a barrier to outward hydrogen diffusion from plated parts.

An accepted practice for removing hydrogen picked up during plating operations is to bake the cadmium-plated parts for 23 h at 190 ± 14 °C (375 ± 25 °F). This treatment was implemented, as well as baking fasteners

Walter J. Jensen, Lockheed-Georgia Company

Corrosion Protection

The most commonly used protective metal coatings for ferrous metal fasteners are zinc, cadmium, and aluminum. Tin, lead, copper, nickel, and chromium are also used, but only to a minor extent and for very special applications.

In many cases, however, fasteners are protected by some means other than metallic coatings. They are sometimes sheltered from moisture or covered with a material that prevents moisture from making contact, thus drastically reducing or eliminating corrosion. For fasteners exposed to the elements, painting is universally used. The low-alloy high-strength steel conforming to ASTM specification A 242 forms its own protective oxide surface film. This type of steel, although it initially corrodes at the same rate as plain carbon steel, soon exhibits a decreasing corrosion rate, and after a few years, continuation of corrosion is practically nonexistent. The oxide coating formed is fine textured, tightly adherent, and a barrier to moisture and oxygen, effectively preventing further corrosion. Plain carbon steel, on the other hand, forms a coarse-textured flaky oxide that does not prevent moisture or oxygen from reaching the underlying noncorroded steel base.

Steels conforming to ASTM A 242 are not recommended for exposure to highly concentrated industrial fumes or severe marine conditions, nor are they recommended for applications in which they will be buried or submerged. In these environments, the highly protective oxide does not form properly, and corrosion is similar to that for plain carbon steel.

Zinc Coating. Zinc is the coating material most widely used for protection of fasteners from corrosion. Hot dipping is the most often used method of application, followed by electroplating and, to a minor extent, mechanical plating.

Hot dipping, as the name implies, involves immersing parts in a molten bath of zinc. Hot-dip zinc coatings are sacrificial by electrochemical means. These coatings for fasteners meet ASTM specification A 394, which specifies an average coating weight of 1.25 oz/ft^2 (2.2-mil thickness). Zinc electroplating of fasteners is done primarily for appearance, where thread fit is critical, where corrosion is not expected to be severe, or where life expectancy is not great.

Specification ASTM B 633 for electrodeposited zinc coatings on steel specifies three coating thicknesses: GS, 25 μm (0.0010 in.); LS, 13 μm (0.0005 in.); and RS, 4 μm (0.00015 in.)—corresponding to coating weights of 0.59, 0.29, and 0.09 oz/ft^2, respectively. These electrodeposited coatings are often given supplemental phosphate or chromate coatings to develop a specific color and, to a minor extent, enhance corrosion resistance. Corrosion life of a zinc coating is proportional to the amount of zinc present; thus, the heaviest electrodeposited coating (GS) would have only about half the life of a hot-dip galvanized coating.

Mechanical (nonelectrolytic) barrel plating is another method of coating fasteners with zinc. Coating weight can be changed by varying the amount of zinc used and the duration of barrel rotation. Such coatings are quite uniform and have a satisfactory appearance.

Cadmium coatings are also applied to fasteners by an electroplating process similar to that used for zinc. These coatings meet ASTM specification A 165. As is true for zinc, cadmium corrosion life is proportional to the coating thickness. The main advantage of cadmium over zinc is in marine environments, where the corrosion life of cadmium is longer. Cadmium-plated steel fasteners also are used in aircraft in contact with aluminum because the galvanic characteristics of cadmium are more favorable than those of zinc. Chromate coatings are also used over cadmium coatings for the reasons given for zinc-plated fasteners.

Aluminum coating on fasteners offers the best protection of all coatings against atmospheric corrosion. Aluminum coating also gives excellent corrosion protection in seawater immersion and in high-temperature applications.

Aluminum coatings are applied by hot-dip methods at about 675 to 705 °C (1250 to 1300 °F). Aluminum alloy 1100 is usually used because of its general all-around corrosion resistance. As with any hot-dip coating, a metallurgical bond is formed that consists of an intermetallic alloy layer overlaid with a coating of pure bath material.

Aluminum coatings do not corrode uniformly, as do zinc and cadmium coatings, but rather by pitting. In some cases, these pits may extend entirely through the coating to the base metal; in others, only through the overlay to the intermetallic layer, Pits, which may occur in a part soon after exposure, sometimes discolor the coated surface, but cause little damage. The complex aluminum and iron oxide corrosion product seals the pits, and because the corrosion product is tightly adherent and impervious to attack, corrosion is usually limited. There is little tendency for corrosion to continue into the ferrous base, and there is none for undercutting and spalling of the coating.

Aluminum coatings will protect steel from scaling at temperatures up to about 540 °C (1000 °F); the aluminum coating remains substantially the same as when applied, and its life is exceptionally long. Above 650 °C (1200 °F), the aluminum coating diffuses into the steel to form a highly protective aluminum-iron alloy. This diffusing or alloying is time-temperature dependent; the higher the temperature, the faster the diffusion. However, scaling will not take place until all the aluminum is used up, which may take a thousand or more hours even at temperatures as high as 760 °C (1400 °F).

Prevention of galling at elevated temperatures is another characteristic of aluminum coatings. Stainless steel fasteners for use at 650 °C (1200 °F) have been aluminum coated just to prevent galling. Coated nuts can be removed with an ordinary wrench after many hours at these temperatures, which is impossible with uncoated nuts.

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Fastener Performance at Elevated Temperatures

The temperature, the environment, and the materials of the structure are normally fixed; therefore, a design objective is to select a bolt material that will give the desired clamping force at all critical points in the joint. To do this, it is necessary to balance the three time- and temperature-related factors (relaxation, thermal expansion, and modulus) with a fourth factor—the amount of initial tightening or clamping force.

Modulus of Elasticity. As temperature increases, the modulus of elasticity decreases; therefore, less load (or stress) is needed to impact a given amount of elongation (or strain) to a material than at lower temperatures. This means that a fastener stretched a certain amount at room temperature to develop preload will exert a lower clamping force at higher temperature. The effect of reducing the modulus is to reduce clamping force, whether or not the bolt and structure are of the same material, and is strictly a function of the bolt metal.

Coefficient of Expansion. With most materials, the size of the part increases as the temperature increases. In a joint, both the structure and the fastener increase in size with an increase in temperature. If the coefficient of expansion of the fastener exceeds that of the joined material, a predictable amount of clamping force will be lost as temperature increases. Conversely, if the coefficient of expansion of the joined material is greater, the bolt may be stressed beyond its yield or even fracture strength, or cyclic thermal stressing may lead to thermal fatigue failure. Thus, matching of materials in joint design can ensure sufficient clamping force at both room and elevated temperatures without overstressing the fastener.

Relaxation. At elevated temperatures, a material subjected to constant stress below its yield strength will flow plastically and permanently change dimensions in the direction of the stress. This phenomenon is called creep. In a joint at elevated temperature, a fastener with a fixed distance between the bearing surfaces of the male and female members will produce less and less clamping force with time. This characteristic is called relaxation. It differs from creep in that stress changes because the elongation or strain remains constant. Such elements as material, temperature, initial stress, manufacturing method, and design affect the rate of relaxation.

Relaxation is the most important of the three time- and temperature-related factors. It is also the most critical consideration in design of fasteners for service at elevated temperature. A bolted joint at 650 °C (1200 °F) can lose as much as 50% of preload. Failure to compensate for this could lead to fatigue failure or a loose joint, even though the bolt was properly tightened initially.

Other Considerations. Bolt preloading is not the only aspect of joint design that must be considered for elevated-temperature service. Other considerations include time, environment, serviceability, coatings, manufacturing methods, fastener design, and material stability. The element of time that is introduced with exposure to heat requires an assessment of the service life of the structure (or at least of the fastener) and of the amount of this life that will be spent at elevated temperature.

Galling and seizing are serious considerations if there is to be subsequent disassembly, which is usually necessary for threaded connections. Many of the materials with desirable elevated-temperature properties, such as the stainless steels, are inherently prone to galling during sliding contact of one member over another, even at room temperature. These materials require protective lubricating coatings. Additional problems of seizing of mating materials develop after exposure to temperature. Ease of disassembly requires that a suitable coating be used and that clearance sometimes be provided at the pitch diameter.

At moderate temperatures, where cadmium and zinc anticorrosion platings might normally be used, the phenomenon of stress alloying becomes an important consideration. Conventional cadmium plating, for example, is usable only to 230 °C (450 °F). At somewhat above that temperature, the cadmium is likely to melt and diffuse into the base material along the grain boundaries, causing cracking by liquid-metal embrittlement, which can lead to rapid failure. For corrosion protection of high-strength alloy steel fasteners used at temperatures between 230 and 480 °C (450 and 900 °F), special nickel-cadmium coatings such as that described in AMS 2416 are often used.

At extremely high temperatures, coatings must be applied to prevent oxidation of the base material, particularly with refractory metals. These must be selected on the basis of the environments. In static air, suitable coatings are available. However, at partial pressure or in dynamic air, there will be a significant reduction in service life.

Fastener-manufacturing methods can also influence bolt performance at elevated temperature. The actual design and shape of the threaded fastener are also important, particularly the root of the thread. A radiused thread root is a major consideration in room-temperature design, being a requisite for good fatigue performance. However, at elevated temperature, a generously radiused thread root is also beneficial in relaxation performance. Starting at an initial preload of 483 MPa (70 ksi), a Waspaloy stud with square thread roots lost a full 50% of its clamping force after 20 h, with the curve continuing downward, indicating a further loss. A similar stud made with a large-radiused root lost only 36% of preload after 35 h.

Selection of fastener material is perhaps the single most important consideration in elevated-temperature design. Most materials are effective over a limited temperature range. No fastener material is suitable for service at temperatures ranging from cryogenic to ultrahigh.

Standard high-strength bolts of medium-alloy steel have a usable temperature range up to about 230 °C (450 °F). The medium-alloy chromium-molybdenum-vanadium steel conforming to ASTM A 193, grade B16, is a commonly used bolt material in industrial turbine and engine applications to 480 °C (900 °F). An aircraft version of this steel, AMS 6304, is widely used in fasteners for jet engines.

For many industrial applications, corrosion- and heat-resistant stainless steels are usable to 425 °C (800 °F). Stabilized grades, such as types 321 and 347, can be used at slightly higher temperatures, at lower strength levels.

The 5% Cr tool steels, most notably H11, are used for fasteners having a tensile strength of 1517 to 1793 MPa (220 to 260 ksi). They retain excellent strength through 480 °C (900 °F).

From 480 to 650 °C (900 to 1200 °F), corrosion-resistant alloy A-286 is used. Alloy 718, with a room-temperature tensile strength of 1241 MPa (180 ksi), has some applications in this temperature range.

The nickel-base alloys René 41, Waspaloy, and alloy 718 can be used for most applications in the temperature range of 650 to 870 °C (1200 to 1600 °F). Between 870 and 1095 °C (1600 to 2000 °F) is a transition range in which the nickel- and cobalt-base alloys do not retain sufficient strength to give satisfactory service.

Mechanical fasteners are available for short-term exposure at ultrahigh temperature. Some refractory-metal fasteners are in service in carefully controlled applications at temperatures of 1095 to 1650 °C (2000 to 3000 °F).

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Liquid-Metal Embrittlement (LME)

Liquid-metal embrittlement is the brittle failure of a normally ductile metal when it is coated with a thin film of a liquid metal and subsequently stressed in tension. Embrittlement manifests itself as a reduction in fracture stress, strain, or both; failure occurs well below the yield stress of the solid. Even small amounts of liquid metal can cause catastrophic failure by LME. Fracture is usually intergranular, with little crack branching or striations to indicate slow crack propagation. More information on this subject is available in the article “Liquid-Metal and Solid-Metal Induced Embrittlement” in this Volume, and the article “Failures of Locomotive Axles” in this Volume gives examples of steel locomotive axles that failed by liquid-metal contamination by copper alloy bearing materials.

In one case of LME of mechanical fasteners, cadmium-plated 4140 steel nuts (hardness: 44 HRC) were inadvertently used on bolts for clamps used to join ducts that carried hot (500 °C, or 930 °F) air from the compressor of a military jet engine. The nuts were fragmented or severely cracked (Fig. 16a). Fracture surfaces were oxidized and tarnished (Fig. 16b), and fracture was brittle and intergranular (Fig. 16c). Failure occurred when the hot air melted the electroplated cadmium (melting point: 321 °C, or 610 °F), causing LME of the steel base metal.

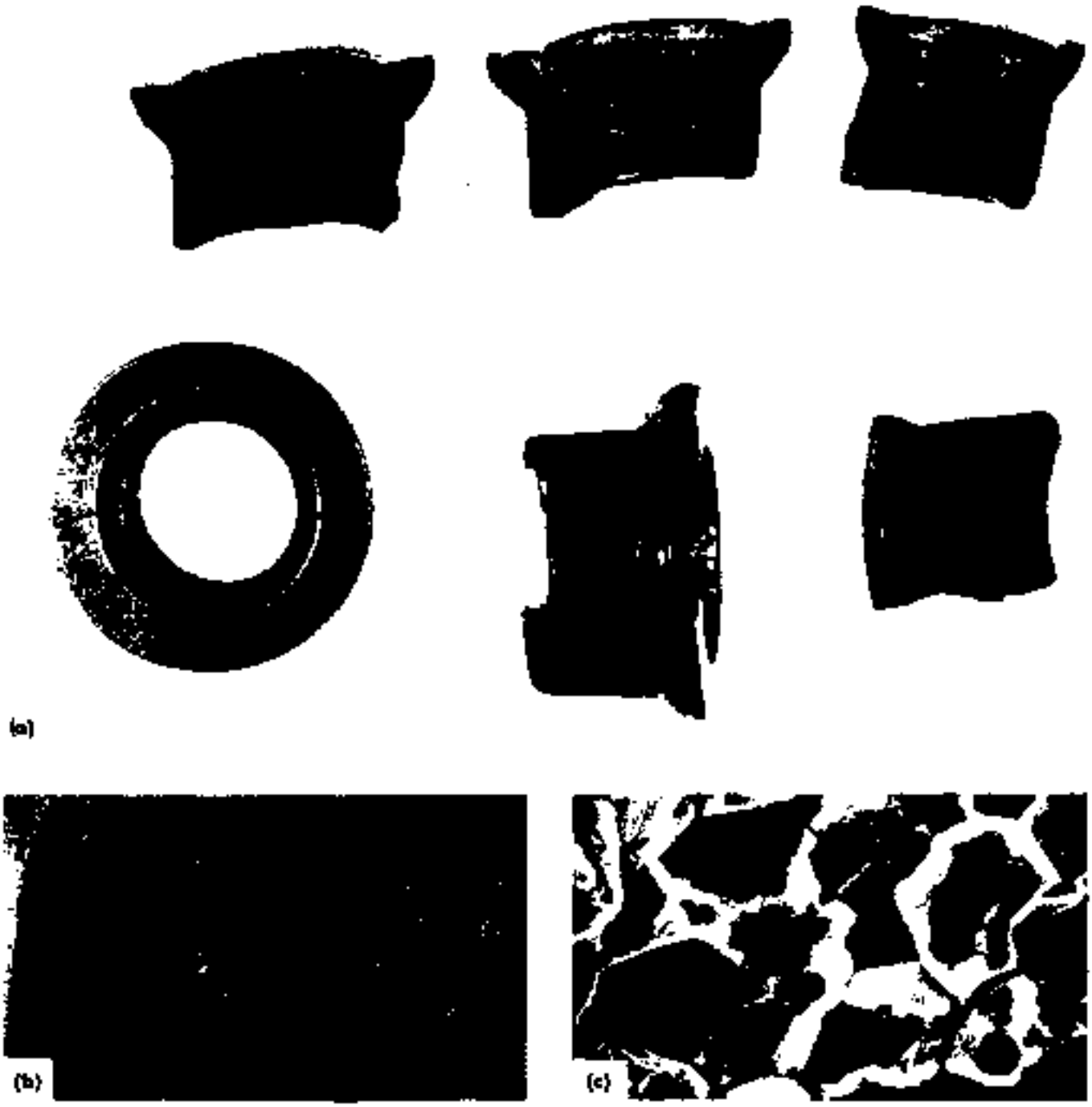


Fig. 16 Cadmium-plated 4140 steel nuts from a military jet engine that failed by LME. (a) Fragmented and cracked nuts. (b) Typical fracture surface. (c) Electron fractograph showing brittle intergranular fracture

---

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Rivets

Solid, tubular, and split rivets are among the most commonly used fasteners for assembled products. Rivets are used as mechanical fasteners because they have a favorable relative cost and weight, because they are available in a wide variety

of materials, head styles, and sizes, or because they have excellent hole-filling ability. To achieve these benefits, however, a product must be designed for riveting from the beginning, with consideration given to production as well as product design. The five basic types of rivets used are shown in Fig. 17.

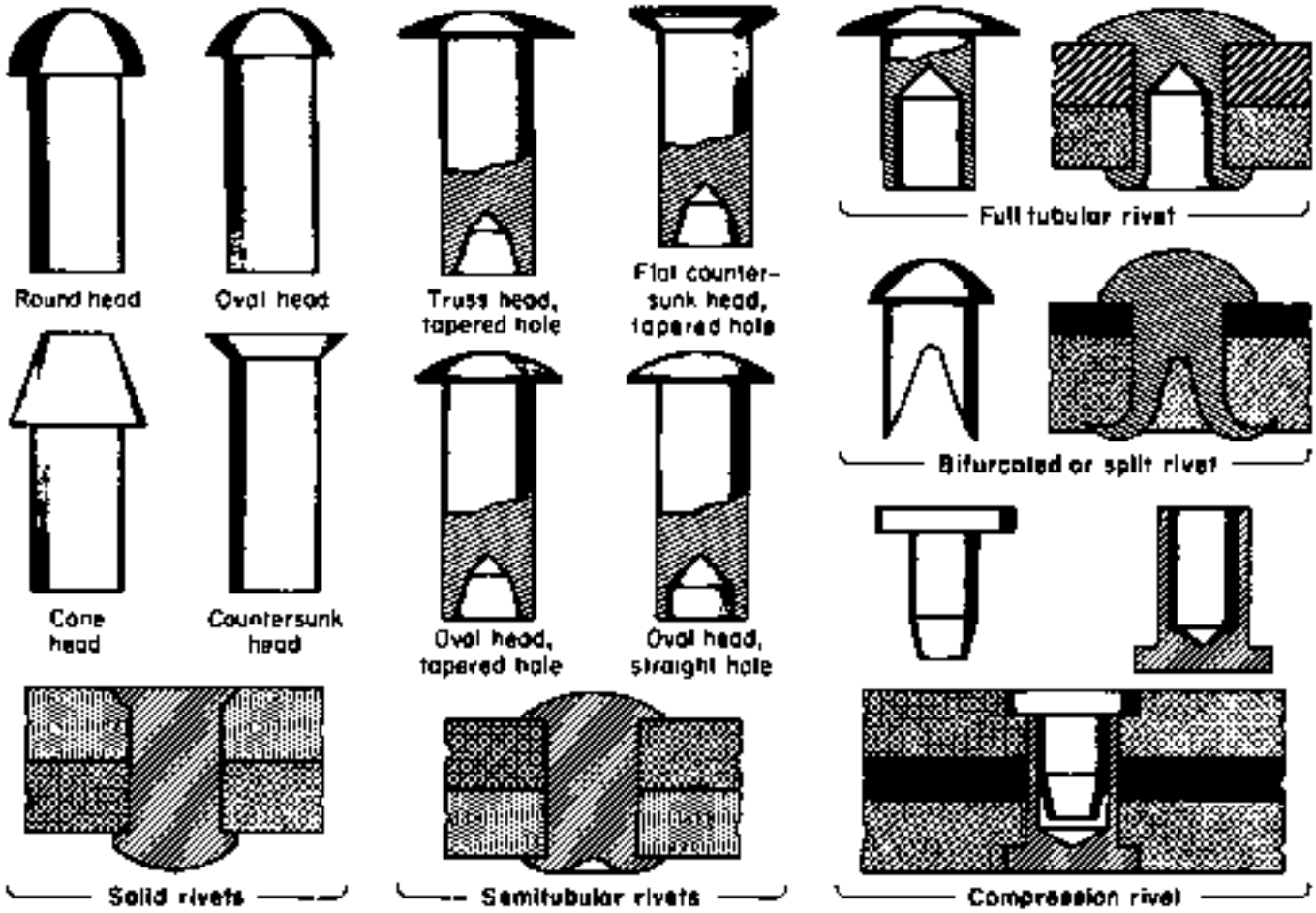


Fig. 17 Five basic types of rivets used to fasten assembled products.

Rivet Positioning. The location of the rivet in the assembled product influences both joint strength and clinching requirements. The important dimensions are edge distance and pitch distance.

Edge distance is the interval between the edge of the part and the centerline of the rivet and is controlled by the bearing strength of the material being joined. Maximum bearing strength is obtained with an edge distance equal to twice the diameter of the rivet shank. At less than this distance, the bearing strength of the joint decreases approximately directly with edge distance under some shear loadings. Even in the absence of shear load, edge distance of twice rivet-shank diameter should be specified to avoid buckling of the material while clinching. The recommended edge distance for plastic materials, either solid or laminates, is between two and three diameters, depending on the thickness and inherent strength of the material.

Pitch distance, which is the interval between centerlines of adjacent rivets, should be determined by the load to be carried across the joint and by the size of the fastener. Unnecessarily high stress concentrations in the riveted material and buckling at adjacent empty holes can result if the pitch distance is less than three times the diameter of the largest rivet in the assembly. The maximum pitch distance is limited by the tendency of the sheet to buckle between rivets. Multiple rows of rivets, sometimes staggered, are frequently necessary to provide the desired joint strength.

Types of Rivet Failures. Standard design analyses of riveted joints consider two primary types of failures: (1) shear in the shank of the rivet and (2) bearing or crushing failure of the metal at the point where the rivet bears against the joined members. The load per rivet at which each of these two types of failures may occur is separately calculated, and the lower value of the two calculations governs the design. However, the practical goal should be to design all riveted joints as bearing critical. This can be accomplished by selecting the correct material composition, condition, and thickness. For optimum performance, the rivet and the members being joined should have the same properties, but from the driving standpoint, it is often necessary to have a softer rivet.

In addition to the design considerations, rivets may also be susceptible to failure by many of the same failure mechanisms experienced by threaded fasteners. For example, rivets are especially susceptible to fatigue failure when used in the same

joint with threaded fasteners. This is because rivets fill the holes more completely than threaded fasteners and are thus subjected to more than their proportionate share of the load.

Shear in Shank. The number of planes on which a rivet may fail in shear depends on the type of joint that is riveted. Rivets are normally used in single or double shear. For design purposes, the load required to shear a rivet can be assumed to increase directly with the number of planes on which shear loading occurs; thus, in double shear, a rivet is approximately twice as strong as in single shear. However, if the members being joined are relatively thin and have significantly greater hardness than the rivets, the rivets will fail at a lower value of shear stress than that determined in tests with thicker material.

Machine-countersunk flush-head rivets are subject to failure of the sheet or to shear and tension failure of rivet heads. Riveted joints with thin outer sheets are critical because of the instability of the outer sheets, which can buckle around the rivet before full bearing stress is developed (Fig. 18).

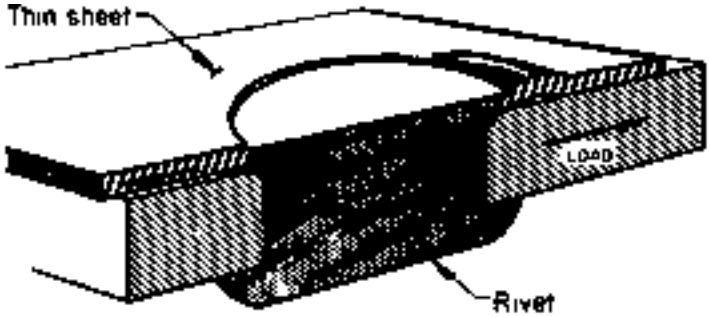


Fig. 18 Schematic of buckling failure of a thin sheet in a riveted joint. Countersinking the top sheet formed a sharp edge at the faying surface.

Bearing-Surface Failure. In riveted joints involving sheet of moderate thickness, the controlling factor is the strength of the rivet head, because induced shear and tension in the rivet head are more critical than shear through the shank. Determining factors are the magnitude of the induced loads in the rivet head, the deflection of the rivet head and sheet under load, and the bearing of the knife-edge of the countersunk hole in the top sheet of the rivet. Typical loading of a riveted joint is shown in Fig. 19(a); failure generally occurs as shown in Fig. 19(b). Because of these inherent limitations of flush-head rivets, protruding-head rivets are preferred. Flush-head rivets should be used only when a clean surface is required for appearance or for aerodynamic efficiency or when clearance for an adjacent structure is inadequate to accommodate protruding-head fasteners.

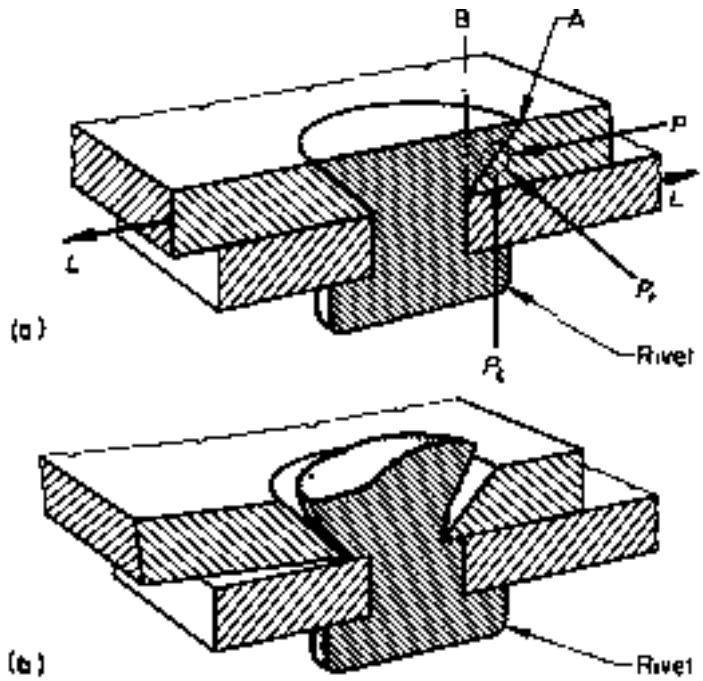


Fig. 19 Schematic of typical behavior of flush-head rivets. (a) Loading of rivet. A, bearing area of the upper sheet; B, bearing area of the lower sheet; *L*, load; *P*, shear component;

P_t, tension component; P_r, resultant of shear and tension components. (b) Shear and tension failure mode

Stress-Corrosion Cracking. In aluminum alloys containing more than 4% Mg, the introduction of plastic deformation followed by aging at a slightly elevated temperature, such as might occur in the tropics, can lead to SCC. A rivet of Alcan aluminum B54S-O, shown in cross section in Fig. 20(a), failed during a stress-corrosion test in the laboratory and illustrates this effect. The rivet was heated for 7 days at 100 °C (212 °F) before testing. The residual deformation from the heading operation, followed by the treatment at elevated temperature, resulted in failure from SCC. The path of the fracture along the grain boundaries is illustrated in Fig. 20(b). The specimen was etched in dilute phosphoric acid, which, on alloys of this type, outlines the grains in a material that is sensitive to stress corrosion. Electron microscope investigation has shown that this type of grain-boundary attack prevails when a fine, closely spaced precipitate is present at the grain boundary (Ref 6). If the precipitate is coarser and more widely spaced, as when treated at 230 °C (445 °F) for 6 h, small etch pits that resemble a string of pearls at the grain boundaries can be seen with the optical microscope. Such a structure is not susceptible to SCC.

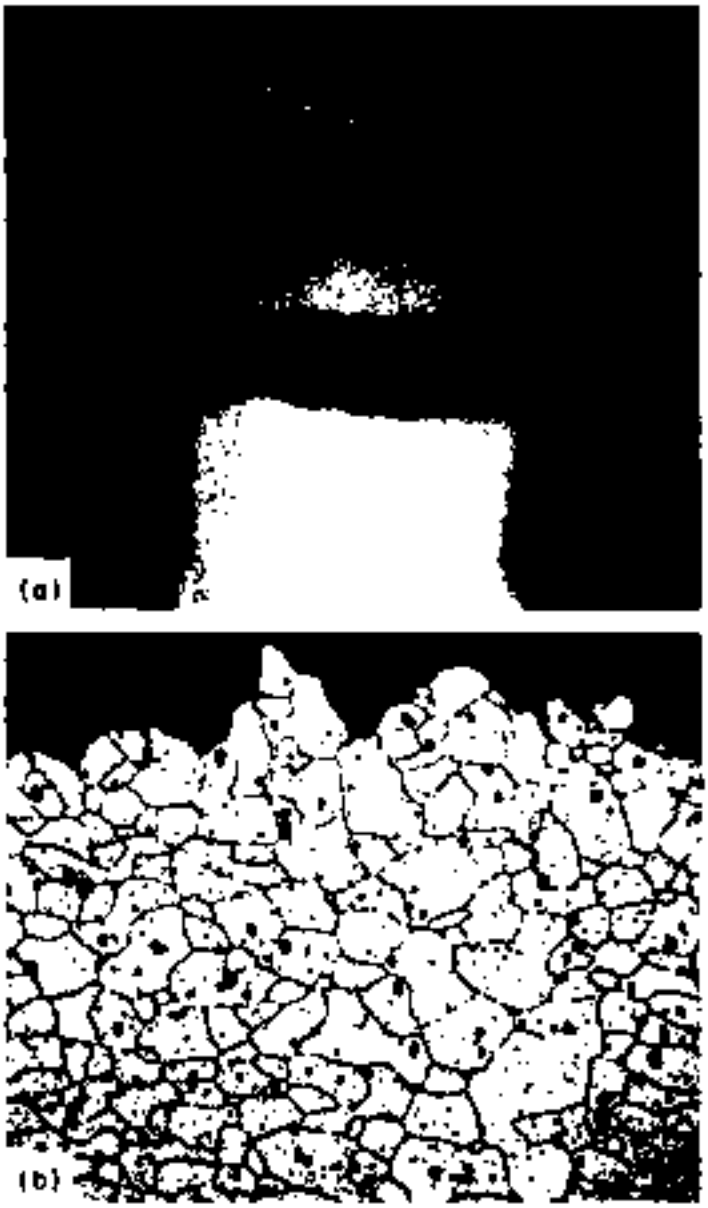


Fig. 20 Alcan aluminum alloy B54S-O rivet that failed by stress corrosion after being heated for 7 days at 100 °C (212 °F). (a) Section through rivet showing shape of fracture surface. 5×. (b) Micrograph showing an intergranular fracture path. 450×

Reference cited in this section

6. A.T. Thomas, Etching Characteristics and Grain Boundary Structures of Aged Aluminum-Magnesium Alloys, *J. Inst. Met.*, Vol 94, 1966

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Blind Fasteners

Blind fasteners are mechanical fasteners that can be installed in a joint that is accessible from one side only. When a blind fastener is set, a self-contained mechanical, chemical, or other device forms an upset on the inaccessible, or blind, end of the fastener and also expands the fastener shank, thereby securing the parts being joined.

Blind fasteners are classified according to the methods by which they are set, either pull-mandrel, threaded, drivepin, or chemically expanded. Rivets that conform to these classifications are illustrated in Fig. 21.

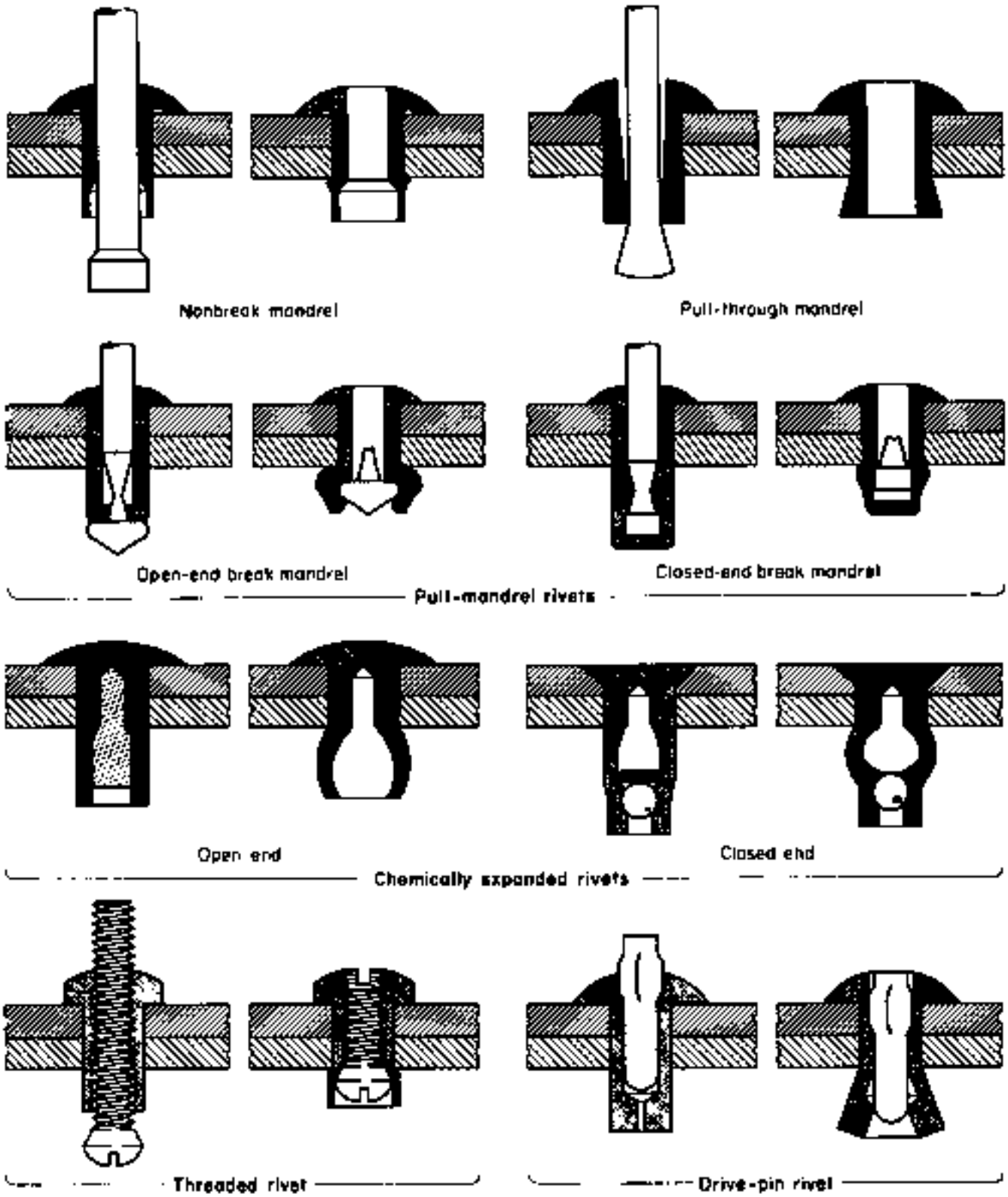


Fig. 21 Types of blind fasteners used in assembled components.

A blind-fastened joint is usually in compression or shear, both of which the fasteners can support somewhat better than tensile loading. The amount of loading that blind fasteners subjected to vibration can sustain is influenced by minimum hole clearance and such installation techniques as applying compression to the assembly with clamps before the fasteners are set.

Some blind fasteners can be used in material as thin as 0.5 mm (0.020 in.) if the heads are properly formed and if shank expansion is carefully controlled during setting. However, in joining sheets that are dissimilar in thickness, the best practice is to form the blind head against the thicker sheet. Also, if one of the components being joined is of a compressible material, rivets with extralarge-diameter heads should be used.

Tolerances applied to blind-fastener holes vary according to the type of rivet used. Ordinarily, it is more economical to select a fastener that clamps against the surfaces of the components being joined than to select one that depends only on hole filling.

Walter J. Jensen, Lockheed-Georgia Company

Pin Fasteners

Pins are effective fasteners where loading is primarily in shear. They can be divided into two groups: Semipermanent pins and quick-release pins.

Semipermanent Pins

Semipermanent pin fasteners require the application of pressure for installation or removal. The two basic types are machine pins and radial-locking pins.
General rules that apply to all types of Semipermanent pins are:

- Avoid conditions in which the direction of vibration is parallel to the axis of the pin
- Keep the shear plane of the pin a minimum distance of one pin diameter from the end of the pin
- Where the engaged length is at a minimum and appearance is not critical, allow pins to protrude at each end for maximum locking effect

Machine Pins. The four important types of machine pins are dowel pins, taper pins, clevis pins, and cotter pins. Machine pins are used either to retain parts in a fixed position or to preserve alignment. Under normal conditions, a properly fitted machine pin is subjected to shear only, and this occurs only at the interface of the surfaces of the two parts being fastened. Taper pins are commonly used to fasten parts that are taken apart frequently, where the constant driving-out of the pin would weaken the holes if a straight pin were used.
Example 10: Fatigue Fracture of a Taper Pin in a Clutch-Drive Assembly. The drive wheel on the clutch-drive support assembly shown in Fig. 22(a) was slightly loose and caused clutch failures in service after 6.8×10^6 cycles. After failure, removal of the taper pin holding the drive wheel on the shaft was difficult, indicating that the pin was tight in the assembly. The taper pin was made of 1141 steel, the shaft 1117 steel, and the drive wheel 52100 steel.

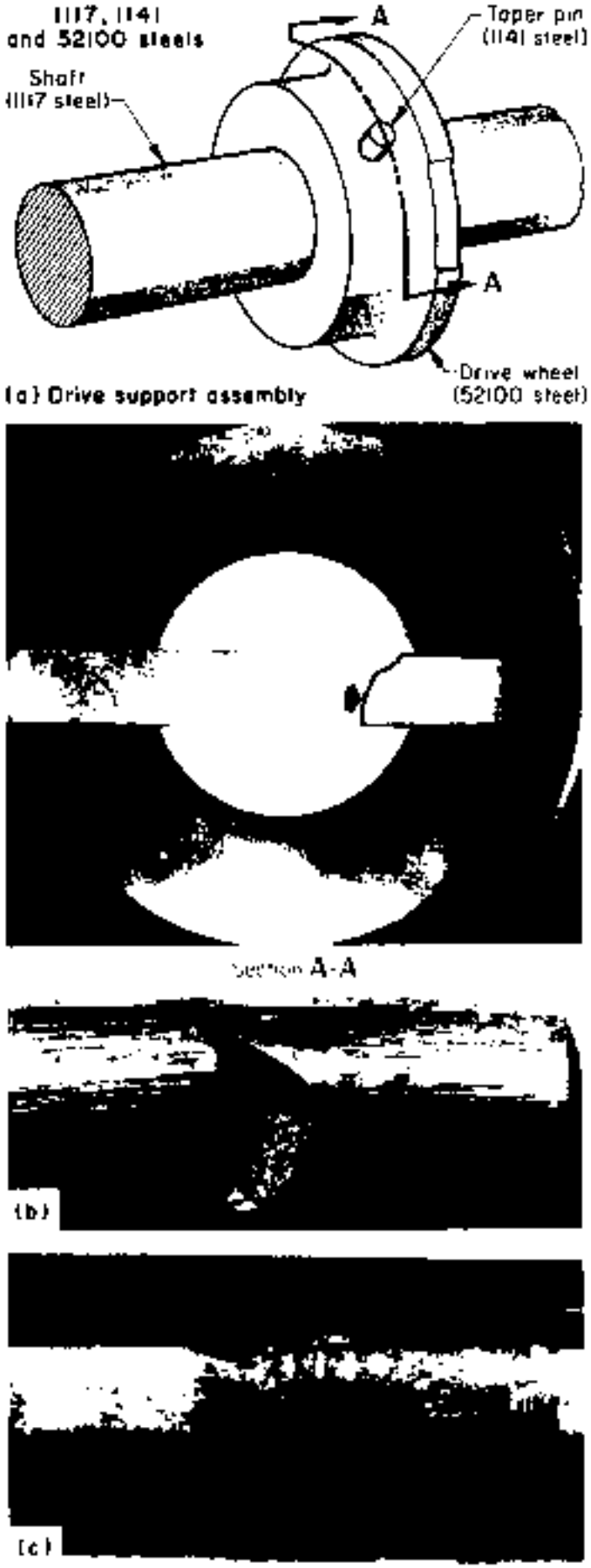


Fig. 22 Steel clutch-drive support assembly that failed in service after a taper pin fractured. The pin fractured because of a loose fit between components. (a) Drive support

assembly. Section A-A shows the break in the small end of the taper pin (arrow). (b) Fractured pin; note axial grinding marks on the surface. (c) Fretting on the portion of the shaft covered by the drive wheel

Investigation. Upon removal from the assembly, the taper pin was found to be broken at the small end (Section A-A, Fig. 22). In assembly, the fracture occurred near the interface of the shaft and the drive wheel. As shown in Fig. 22(b), the surface of the taper pin contained axial grinding marks, some of which were partly obliterated by movement of the components during operation, before and after the pin broke. Evidence of fretting was also found on the pin. Surface finish on the pin was 1 μm (45 μin.); specifications required 0.8 μm (32 μin.). The fracture surface of the pin contained ductile tears, fatigue striations, and shear lips.

Examination of the clutch shaft revealed that fretting had occurred on the portion of the shaft covered by the drive wheel (Fig. 22c). It did not appear that fretting alone had impaired machine operation.

During the investigation, a second clutch-drive support assembly was received at the laboratory. This second assembly exhibited a loose fit similar to the first. The assembly was sectioned transversely at the taper pin, and the pin was found to be fractured in the same area as the first. There was a tight line-to-line fit between the shaft and the taper pin. However, there was a maximum clearance of 0.08 mm (0.003 in.) between the drive wheel and the large end of the pin, and there was a lesser amount at the small end. Also, the clearance between the shaft and the drive wheel was 0.04 mm (0.0015 in.) (specifications permitted a maximum clearance of 0.025 mm, or 0.001 in.). Because of the excessive clearance between the shaft and the drive wheel, alternating stresses were transmitted to the pin.

Conclusions. Failure of the clutch-drive support assembly occurred as a result of fatigue fracture of a taper pin. A loose fit between the drive wheel and the shaft and between the drive wheel and the pin permitted movement that resulted in fatigue failure. Fretting of the pin and drive shaft was observed but did not appear to have contributed to the failure.

Recommendations. The assembly should be redesigned to include an interference fit between the shaft and the drive wheel. The drive wheel and the shaft should be taper reamed at assembly to ensure proper fit. In addition, receiving inspection should be more critical of the components and accept only those that meet specifications.

Example 11: SCC of 300M Steel Jackscrew Drive Pins. Both jackscrew drive pins on a landing-gear bogie failed suddenly when the other bogie on the same side of the airplane was kneeled for a tire change. The pins were smooth cylindrical tubes that fastened the top tubular ends of the jackscrew to the plane. The primary service stress on the pins was shear. The pins had nominal dimensions of 41-mm (1.6-in.) outside diameter, 25-mm (1-in.) inside diameter, and 178-mm (7-in.) length, and they had a smooth push fit with a hole tolerance of +0.04 to +0.1 mm (+0.0015 to +0.0040 in.). They were shot peened and chromium plated on the outside surface and were cadmium plated and painted with polyurethane on the inside surface. The pins were made of 300M steel heat treated to a tensile strength of 1931 to 2069 MPa (280 to 300 ksi). The top of the jackscrew was of 6150 steel heat treated to a tensile strength of 1241 to 1379 MPa (180 to 200 ksi).

Investigation. After the retrieved portions of the fractured pins had been reassembled (Fig. 23a and b), it was evident that they had sustained prior mechanical damage. Both ends of the pins were dented where the jackscrew had pressed into them, probably as a result of overdriving the jackscrew at the end of an unkneeling cycle. One end of the forward pin was noticeably bent (Fig. 23c). The chromium plating was missing from the dented areas, and these areas were heavily corroded.

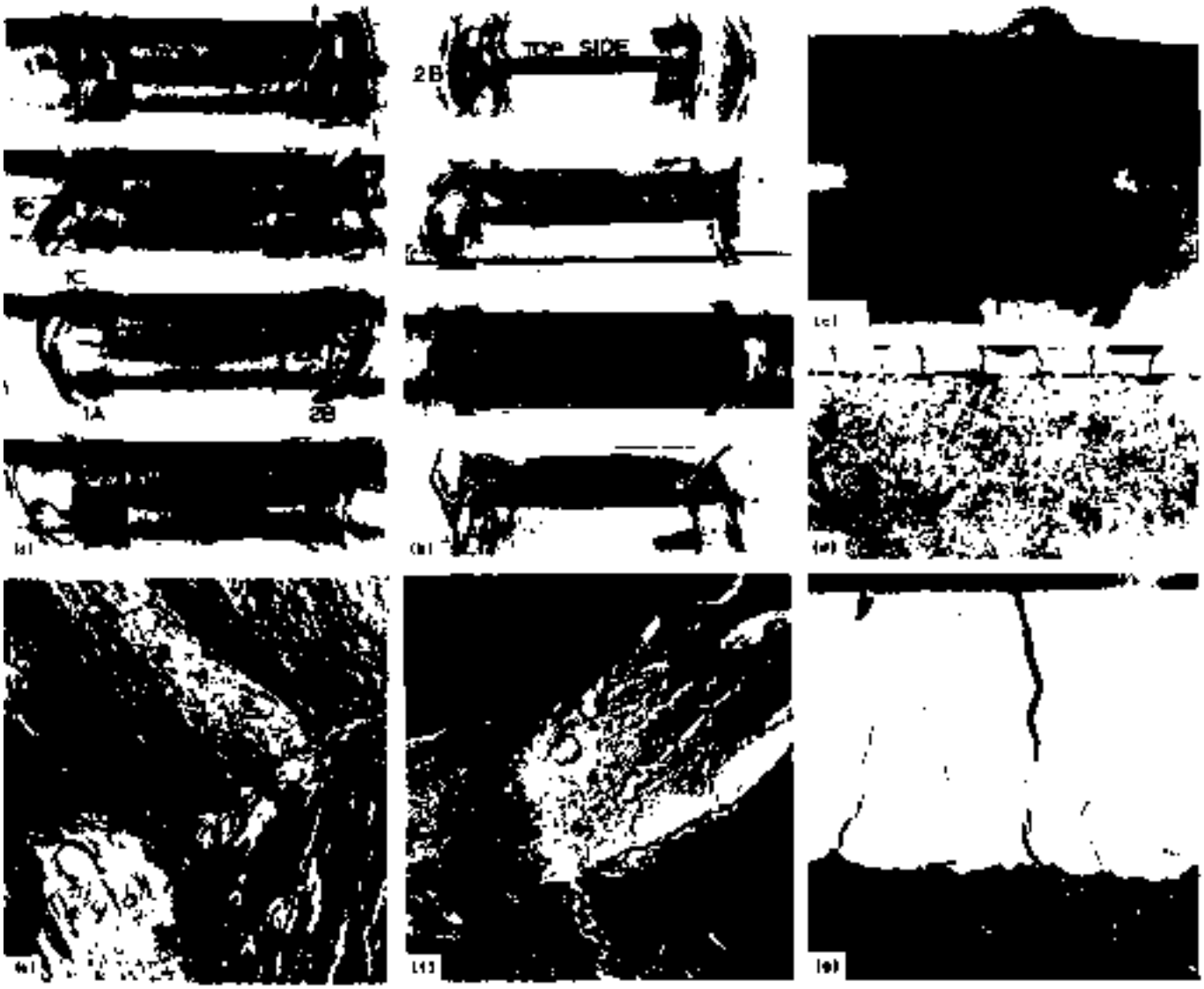


Fig. 23 300M steel jackscrew drive pins that failed by SCC. (a) Four views of aft-pin locations of individual origins (numbers), directions of fracture (arrows), and final-fracture regions (wavy lines). (b) Same as (a) except for forward pin. (c) Top surface of forward pin showing slight bend at left end. (d) Micrograph showing pits in base metal under cracks in chromium plating. Etched with nital. 100×. (e) Electron fractograph showing corroded intergranular fracture surface at origin 1 on forward pin (see b). 4000×. (f) Same as (e) except at 6500× showing origin 1A on aft pin. (g) Micrograph showing intergranular crack in base metal propagating from corrosion pit adjacent to crack. 400×

Chromium-carbon replicas of the areas of fracture origin showed a heavily corroded intergranular fracture mode (Fig. 23e and f). Subsequent metallographic examination revealed deep corrosion pits adjacent to the fracture origins and directly beneath cracks in the chromium plate (Fig. 23d). Intergranular cracks that had propagated from several of the pits under the cracks in the chromium plate are shown in Fig. 23(g).

The dimensions, surface finish, and composition of the pins were within the specified limits. Microhardness traverses and metallographic examination of sections through the pins confirmed that neither carburization nor decarburization had occurred on the surface of either pin. Alkaline chromate etching confirmed that no intergranular oxidation products were present. The microstructure was typical of tempered martensite and was satisfactory.

Conclusions. Overdriving the jackscrew dented the drive pins and caused cracking and chipping of the chromium plate. The unprotected metal rusted, and stress-corrosion cracks grew out of some of the rust pits. The terminal portion of the fracture resulted from overstress when the remaining uncracked cross section was not adequate to carry the load.

Corrective Measures. The jacking-control system was modified to prevent overdriving and subsequent damage to the drive pins. The pin material was changed from 300M steel to PH 13-8 Mo stainless steel, which is highly resistant to

rusting and SCC. The inside diameter of the pins was reduced to compensate for the lower strength of the PH 13-8 Mo steel. The protective plating and paint were eliminated because of the good corrosion resistance of the new material.

Radial-locking pins are of two basic styles: (1) solid, with grooved surfaces, and (2) hollow spring pins, which may be slotted or spiral wound. In assembly, radial forces produced by elastic action at the pin surface develop a friction-locking grip against the hole wall. Spring action at the pin surface also prevents loosening under shock and vibration.

Locking of the solid grooved pins is provided by parallel, longitudinal grooves uniformly spaced around the pin surface. When the pin is driven into a drilled hole corresponding in size to the nominal pin diameter, elastic deformation of the raised groove edges produces a force fit with the hole wall. Optimum results under average conditions are obtained with holes drilled the same size as the nominal pin diameter. Undersize holes should be avoided, because they can lead to deformation of the pin in assembly, damage to hole walls, and shearing-off of the raised groove edges.

Resilience of hollow cylinder walls under radial compression forces is the principle of spiral-wound and slotted tubular pins. Compressed when driven into the hole, the pins exert spring pressure against the hole wall along their entire engaged length to develop locking action. For maximum shear strength, slotted pins should be placed in assemblies so that the slots are in line with the direction of loading and 180° away from the point of application.

Quick-Release Pins

Quick-release pins are of two types: push-pull and positive locking. Both use some form of detent mechanism to provide a locking action for rapid manual assembly and disassembly.

Push-pull pins are made with a solid or hollow shank containing a detent assembly, such as a locking lug, button, or ball, which is backed by a resilient core, plug, or spring. These pins fasten parts under shear loading, ideally at right angles to the shank of the pin.

Positive-locking pins use a locking action that is independent of insertion and removal forces. These pins are primarily suited for shear-load applications. However, some tension loading can usually be tolerated without adversely affecting the function of these pins.

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Special-Purpose Fasteners

Special-purpose fasteners, many of which are proprietary, are frequently used in nonstructural applications to provide some unique feature, such as quick release, snap action, or cam action. These fasteners are subject to the same failure mechanisms experienced by threaded fasteners and rivets.

Spring clips are a type of special-purpose fastener that perform multiple functions, are generally self-retaining, and, for attachment, require only a flange, panel edge, or mounting hole. The spring-tension principle of fastening eliminates loosening by vibration, allows for design flexibility, compensates for tolerance buildup and misalignment, and minimizes assembly damage.

The basic material for spring-clip fasteners is steel with 0.50 to 0.80% carbon content. Fasteners are generally formed in the annealed stage, then hardened and tempered to 45 to 50 HRC. Zinc or cadmium electroplating is specified for certain applications. Other applications may use a zinc mechanical plating to eliminate hydrogen embrittlement.

Example 12: Failure of a Zinc-Electroplated 1060 Steel Fastener. The fastener shown in Fig. 24 was used to secure plastic fabric or webbing to the aluminum framework of outdoor furniture. Several clips were required on each piece of furniture, and these had to hold the fabric for several seasons of varying weather conditions.

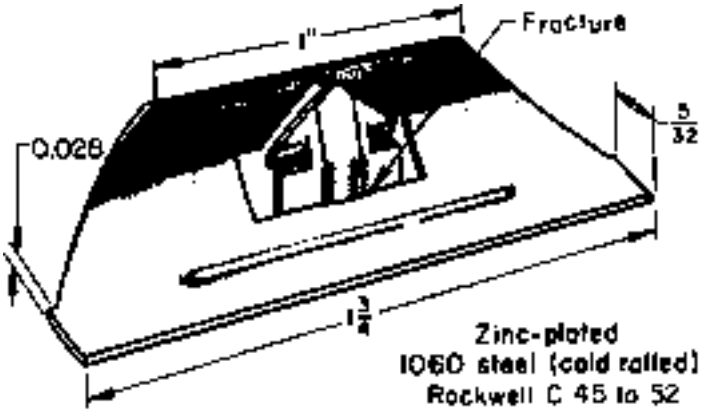


Fig. 24 Zinc-electroplated 1060 steel fastener that failed by hydrogen embrittlement. The part was used to secure fabric to lawn-furniture framework. Dimensions given in inches

The fasteners were made in high-production progressive dies from 0.7-mm (0.028-in.) thick cold-rolled 1060 steel. They were barrel finished to remove burrs and hardened and tempered to 45 to 52 HRC. The fasteners were electroplated with zinc and coated with clear zinc dichromate.

During attachment of the fabric to the furniture, approximately 30% of the fasteners cracked and fractured as they were compressed to clamp onto the framework prior to springback, which held the fabric in place. Fasteners on some of the furniture that was successfully assembled had fractured in warehouses, in retail stores, and in service, allowing the plastic fabric or webbing to come off the chair.

Discussion. The heat treatment of the fasteners consisted of austenitizing at 790 °C (1450 °F) for 10 min, quenching in oil at 80 °C (180 °F), rinsing, then tempering at 345 °C (650 °F) to obtain a microstructure of tempered martensite. After being acid cleaned, zinc electroplated to a thickness of 0.005 mm (0.0002 in.), and coated with a clear dichromate, the fasteners were baked 8 h at 165 °C (330 °F) to remove the nascent hydrogen. Fasteners treated in this manner were very brittle and had a limited capacity of being compressed without fracture. Low ductility at room temperature and delayed failure are characteristic of steel that has been damaged by hydrogen embrittlement.

Conclusions. The fasteners failed in a brittle manner as the result of hydrogen embrittlement. Hydrogen was induced into the surface of the steel when the hardened parts were acid cleaned before electroplating and during the electroplating operation. Baking at 165 °C (330 °F) for 8 h failed to remove all of the nascent hydrogen that had diffused interstitially into the steel.

Corrective Measures. The heat treatment was changed from austenitizing, quenching, and tempering to austenitizing at 790 °C (1450 °F) for 15 min, quenching in a potassium chloride salt bath at 345 °C (650 °F) for 20 min, air cooling, and rinsing. This heat treatment produced a more ductile and bainitic structure than the original heat treatment, as well as a hardness of 48 to 52 HRC. In addition, the method of plating the fastener with zinc was changed from electroplating to a mechanical deposition process.

These two changes eliminated the possibility of hydrogen embrittlement, because acid cleaning was not needed after austempering and before plating. Also, mechanical deposition of zinc coating does not produce hydrogen embrittlement in hardened steel.

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

Fastener Failures in Composites

The typical failure mechanism for statically loaded graphite composite joints is fastener pullout (Fig. 25a and b). The oversize fastener head is evident in Fig. 25(b); the collar flange, Fig. 25(a). Collars are used, rather than nuts, to avoid galling, which is frequently a problem when titanium nuts and bolts are installed, Figure 26(a) shows the typical pullout-type failure of a flush-head fastener in a composite joint. However, when this type of joint is fatigue tested, failure is usually first initiated by cocking of the fasteners, followed by fatigue failure of one or more of the fasteners (Fig. 26b). The fasteners in Fig. 26(b) were installed with close-fitting sleeves in the holes to improve fatigue life. After the sleeve is placed in the hole, the interference-fit fastener is pulled into the sleeve; this expands the sleeve to provide an interference

fit between it and the composite and reduces the possibility of delamination which can occur if interference-fit fasteners are forced into unsleeved holes. Non-destructive inspection of some type is needed to detect delamination or other internal damage without removal of the fasteners.

Fig. 25 Static tensile failures in carbon-graphite composite samples. Failure was by fastener pullout. (a) Single-lap shear specimen. (b) Double-lap shear specimen. Both $1\frac{1}{3}\times$

Fig. 26 Fatigue failure of fasteners in single-lop shear carbon-graphite composite joints. (a) Fastener pullout resulting from a static tensile load. (b) Fatigue failure of fasteners initiated by cocking of the fasteners. Both $1\frac{1}{3}\times$

Failures of Mechanical Fasteners

Walter J. Jensen, Lockheed-Georgia Company

References

1. Mach. Des., 1985 Fastener and Joining Reference Issue, Vol 54 (No. 27), 1985, p 141

2. *Metallic Materials and Elements for Aerospace Vehicle Structures,* Vol II, Military Handbook, Figure 9.4.1.7(a)

3. J.S. Davey, Why Nuts Should Have Their Own Specifications, *Fasteners,* Vol 13 (No. 3), Fall 1958

4. *SPS Bolts for the Aerospace Industry,* 1964–1965 ed., Standard Pressed Steel Company, Fort Washington, PA

5. E.L. Williams and D.M. Anderson, Hydrogen Embrittlement of AISI 8740 Alloy Steel Fasteners, *Mater. Perform.,* Vol 24 (No. 12), 1985, p 9–12

6. A.T. Thomas, Etching Characteristics and Grain Boundary Structures of Aged Aluminum-Magnesium Alloys, *J. Inst. Met.,* Vol 94, 1966

Failures of Springs

James H. Maker, Consultant, Associated Spring

Introduction

SPRINGS are made in many types, shapes, and sizes, ranging from delicate hairsprings for watches to massive buffer springs for railroad equipment. The vast majority of springs are made from one of several grades of carbon or alloy steel. However, to accommodate the broad spectrum of applications, springs have been made from a wide variety of metals and alloys, including stainless steels, heat-resisting alloys, copper alloys and titanium alloys.

The broad range of compositions, sizes, and types of springs is equaled by the required range of quality and characteristics. Quality can range from that required for low-cost ballpoint pens to the extremely high quality necessary for aerospace applications.

Ordinary springs, such as those used in noncorrosive environments and at temperatures not much higher than room temperature, are produced to high strength levels that often can be obtained entirely by severe coldworking (drawing or rolling). This approach to producing springs generally dictates that no metal be removed from the wire or strip during working, which allows any existing defects in the raw material to remain. Springs produced in this manner are adequate for many applications because minor surface discontinuities are of little or no consequence.

More rigorous applications are likely to require springs with defect-free surfaces. Such surfaces can be obtained by conditioning of billets or even by grinding of the wire or strip stock from which the springs are formed. These practices, however, significantly increase cost and decrease availability. For more detailed information on compositions, properties, methods of manufacture, and major applications for springs, see the article “Steel Springs” in *Properties and Selection: Irons, Steels, and High-Performance Alloys,*, Volume 1 of the*ASM Handbook.*

Failures of Springs

James H. Maker, Consultant, Associated Spring

Operating Conditions

Although most spring applications involve operating temperatures that are not far above or below room temperature, many applications require springs to operate over an appreciable temperature range. For instance, valve springs in internal-combustion engines must operate in frigid weather for start-up, then, within minutes, must function at the normal operating temperature of the spring chamber, which may approach 90 °C (195 °F). This temperature range, and even a range with a maximum temperature somewhat higher, can be accommodated by springs made of carbon or low-alloy steels (for examples of springs that failed because of excessive operating temperatures, see the article "Analysis of Distortion and Deformation" in this Volume). However, some springs are required to operate at much higher temperatures and thus must be made of heat-resisting materials. Such applications must be evaluated individually before a suitable selection of material can be made.

Corrosive Environments. Most springs are not subjected to environments more corrosive than humid air; there are, however many exceptions. As for elevated-temperature service, applications that involve corrosive environments must be evaluated individually before selection of materials for springs or of suitable protective coatings.

Common Failure Mechanisms

Fatigue is the most common mechanism of failure in springs, as demonstrated by the examples in this article. Any of the causes noted in the section "Common Causes of Failures" in this article may result in failure by fatigue; more than one cause is sometimes involved.

Relaxation is another common failure mechanism. This condition is caused by overstressing and often results from use of a grade of material that is marginal for the application. Relaxation from operation at excessively high temperatures is also a common failure mechanism.

Although some spring materials are quite notch sensitive and have limited ductility, brittle fracture is not a major problem in springs in the usual sense. Apparently, this is true because the section sizes are usually small, an extreme degree of working has been used (which creates a homogeneous structure), and care in minimizing stress concentrations has been applied.

One form of brittle fracture occasionally encountered results from hydrogen absorption in martensitic structures, which occurs in electroplating or other cleaning or finishing processes. Corrosion contributes to spring failure in much the same way as it contributes to failure of other highly stressed structures, although in springs it is most often connected with fatigue. Nevertheless, static stress-corrosion failures can occur at high stresses in the presence of electrolytes.

Examination of Spring Failures

For the most part, procedures that apply to analysis of failures in other parts also apply to analysis of spring failures (see the article "Conducting a Failure Examination" in this Volume). Loading and other service conditions as well as all details of the material should be determined, and a preliminary visual examination should be made to identify fracture-surface features, wear patterns, contact marks, surface deposits, and temperature colors. Fracture features often are damaged by clashing of the portions of the broken spring; therefore, it may be necessary to analyze several similar failures to arrive at the correct conclusion as to the cause of failure. For small springs, a stereoscopic bench microscope is essential because the entire cross section may be only as large as the origin of failure in many large sections.

Another important factor is lighting. Spot lighting is sometimes required, but because of reflection into and away from the objective lens, such lighting can be misleading. What is best for fractography is often not best for stereo viewing. A small ring-type fluorescent lamp around the objective is suitable for general use if supplemented by a more concentrated source of light. By exploration at all possible viewing angles with this lighting, features may be revealed that otherwise could escape detection.

Common Causes of Failures

Overstressing is a common cause of failure of springs. Settling and other types of distortion are generally caused by operation of springs at stresses that are higher than expected. It must be kept in mind, however, that the stresses a given spring can withstand are greatly affected by the operating environment. For example, helical springs made of 6150 steel provided failure-free service in fuel-injection pumps when the fuel oil being pumped was a normal low-sulfur grade, but several of these springs failed under identical stress conditions when the fuel oil contained substantial amounts of hydrogen sulfide. In addition, operating temperatures that are higher than those anticipated often result in failures of springs without changes in stress.

Design deficiencies, material defects, processing errors or deficiencies, and unusual operating conditions are common causes of spring failures. In most cases, these causes result in failure by fatigue.

In the remainder of this article, 19 examples of specific failures of springs are presented. Although most of the springs discussed in these examples failed by fatigue, the fatigue cracks were initiated by a wide variety of causes. In some of the examples, more than one cause contributed to failure.

Failures Caused by Errors in Design

Rules governing design of the more common types of springs are generally well understood and closely followed. Proper distribution of stresses and avoidance of sharp bends are among the principal requirements of a well-designed spring. Because the rules of design are generally understood and followed, the number of failures of springs that occur solely as the result of faulty design is comparatively small. However, there are occasional exceptions.

For example, one of a test set of titanium alloy Ti-13V-11Cr-3Al torsionally stressed compression springs failed by fatigue after 12×10^6 cycles at a maximum stress of 640 MPa (93 ksi). The surface of the fracture, which occurred one turn from the end, is shown in Fig. 1. Radial marks emanate from the fracture origin (arrow), which was in an area of contact with the end turn of the spring. In this area, small elongated gall marks were visible, indicating that there was sliding motion over a region about 0.1 mm (0.004 in.) in length. It was concluded that fretting in this area had lowered the fatigue strength. In this instance, and in many others, the proximity of one coil to another does not permit some areas to be shot peened properly, and this can contribute to a reduction in fatigue life.

Fig. 1 Fractograph of the surface of a fracture that occurred one turn from the end of a titanium alloy spring. The arrow indicates the point of contact with the end of the next coil. 10×

Failures Caused by Material Defects

Seams frequently contribute to failures of springs, usually by initiating fatigue cracks. Pipe, inclusions, and pits may also contribute to premature failures. The following three examples describe fatigue failures that resulted from material defects of these types. Additional examples of failures due to defects can be found in the section “Failures That Occur During High-Stress Fatigue” in this article (see Examples, , ,).

Example 1: Fatigue Failure of a Locomotive Suspension Spring that Initiated at a Seam. Hot-rolled hardened-and-tempered 5160 bars for suspension springs are purchased to a restricted seam-depth requirement. Outdoor exposure can result in rusting in the seam and on the fracture surface (Fig. 2a). This macrograph shows a smooth portion of the fracture that exhibits a step due to a seam, which can be seen on the surface.

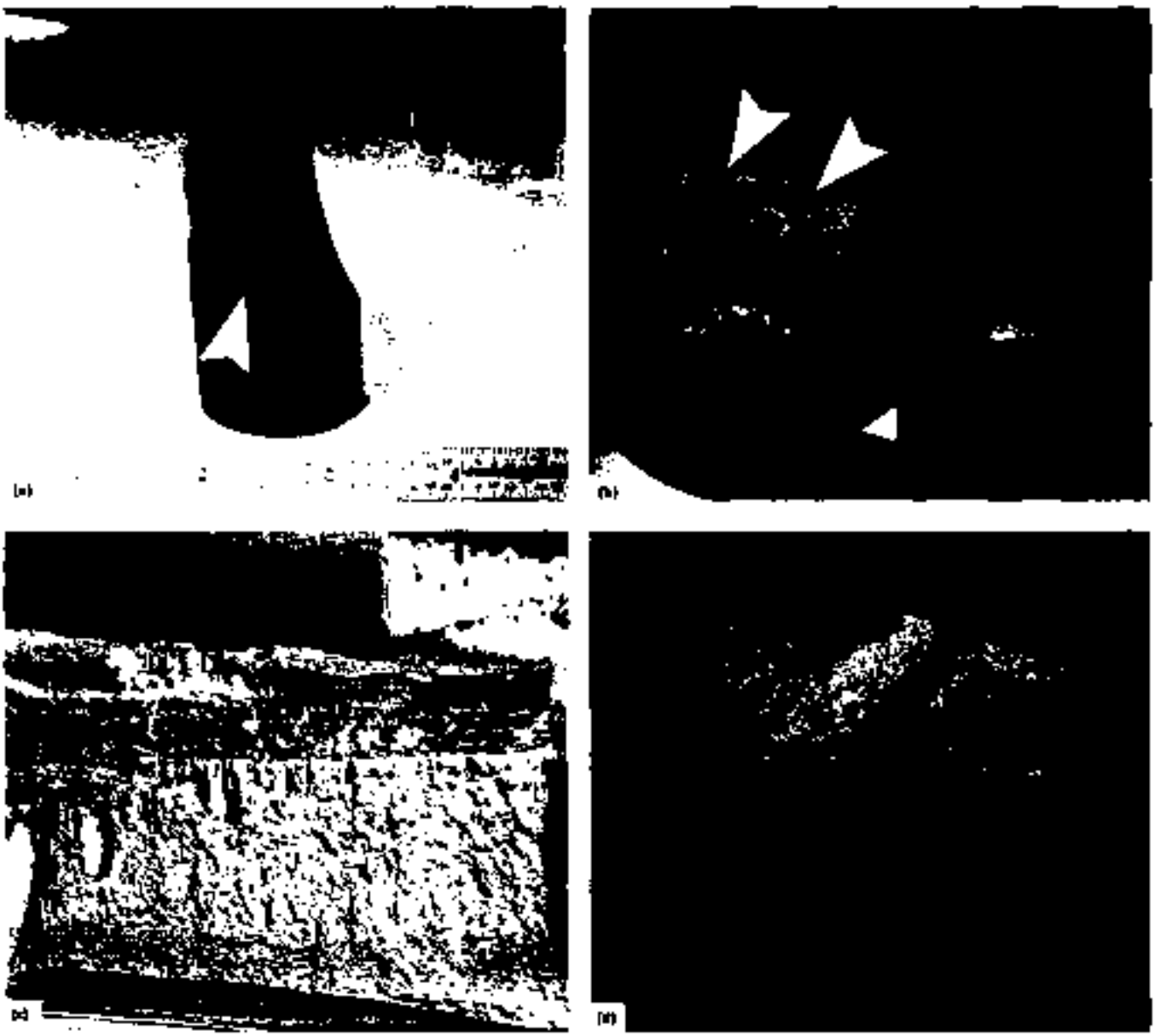


Fig. 2 Fatigue fracture of a locomotive spring with a bar diameter of 36 mm ($1\frac{7}{16}$ in.). (a) A step (shown by arrow) is visible at the fracture surface near the inner diameter of the spring. A seam is visible, extending from the step. 0.6×. (b) Higher magnification (1.7×) of thumb-nail position of the fracture. The step is in the center of a smaller semicircle indicated by arrows. Growth occurred principally to the right of the step until the area was approximately doubled, at which time rapid fracture (rough texture) began. (c) The seam wall (exposed by a specimen-fracture technique) is very dark in places because of scale. Below this is a mottled area exhibiting rust and scale. At the base of this area are dozens of fatigue origins with faces all parallel to the final fracture. At the bottom is the area of rapid, rough fracture. 14×. (d) Metollographic cross section of the seam. The light flecks near the surface are ferrite; the gray areas in the seam are unidentified nonmetallics, probably oxides. 100×

Investigation. Figure 2(b) shows a detailed view of the area around the step. The thumb nail looks off-center from the step, but there is a smaller thumb-nail shape that is concentric with the step and a second stage of growth, very slightly rougher, that spreads principally to the right of the step. The demarcation line is nearly invisible because of rust on the fracture. The rapid stage of failure, which begins at the edge of the thumb nail, is much rougher and exhibits rays that diverge approximately radially from it.

Figure 2(c) shows the seam wall at 15× from a section next to that shown in Fig. 2(b). This was accomplished by cutting the metal away beneath the seam with an abrasive cutoff disk oriented such that the cut was aligned with the seam. Cutting was continued until the wheel had penetrated to about 0.8 mm ($\frac{1}{32}$ in.) from the bottom of the seam. The section of the spring was then placed in a vise so that the abrasively cut walls could be pushed toward each other and the seam walls pulled apart by using the sound metal in between as a fulcrum. Under these conditions, the sound metal just under the seam is placed in tension, and it fractures in a ductile manner. The seam wall has two zones, the lower zone being mottled.

At the base of the seam wall are dozens of spear-head-shaped areas pointing away from the seam. These facets are all parallel to the final thumb-nail surface and are fatigue cracks. Although they extend laterally into the matrix metal, they are deeper than they are wide, showing that the fracture fronts did not develop at a uniform rate in all directions in their early stages. The orientation of these origins is normal to the direction of resultant tensile stress from torsional stressing of the spring material.

Figure 2(d) shows a metallographic view of a cross section of the seam. The light areas near the surface are predominantly ferrite resulting from decarburization. The dark areas in the seam are nonmetallics that were not identified.

Conclusion. The failure in this spring initiated at the base of a seam. Many fracture origins developed, which oriented normally to the direction of tensile stress at 45° to the wire axis, resulting from the torsional stressing of the spring material. As failure progressed, one or more of these origins developed into a semicircular fracture plane at the same 45° angle. This eventually caused complete fracture.

Example 2: Fatigue Fracture of Alloy Steel Valve Springs Because of Pipe. Two outer valve springs broke during production engine testing and were submitted for laboratory analysis. The springs were from a current production lot and had been made from air-melted 6150 pretempered steel wire. The springs were 50 mm (2 in.) in outside diameter and 64 mm ($2\frac{1}{2}$ in.) in free length, had five coils and squared-and-ground ends, and were made of 5.5-mm ($\frac{7}{32}$-in.) diam wire. Because both failures were similar, the analysis of only one will be discussed.

Investigation. The spring (Fig. 3a) broke approximately $1\frac{1}{2}$ turns from the end. Fracture was nucleated by an apparent longitudinal subsurface defect. Magnetic-particle inspection did not reveal any additional cracks or defects.

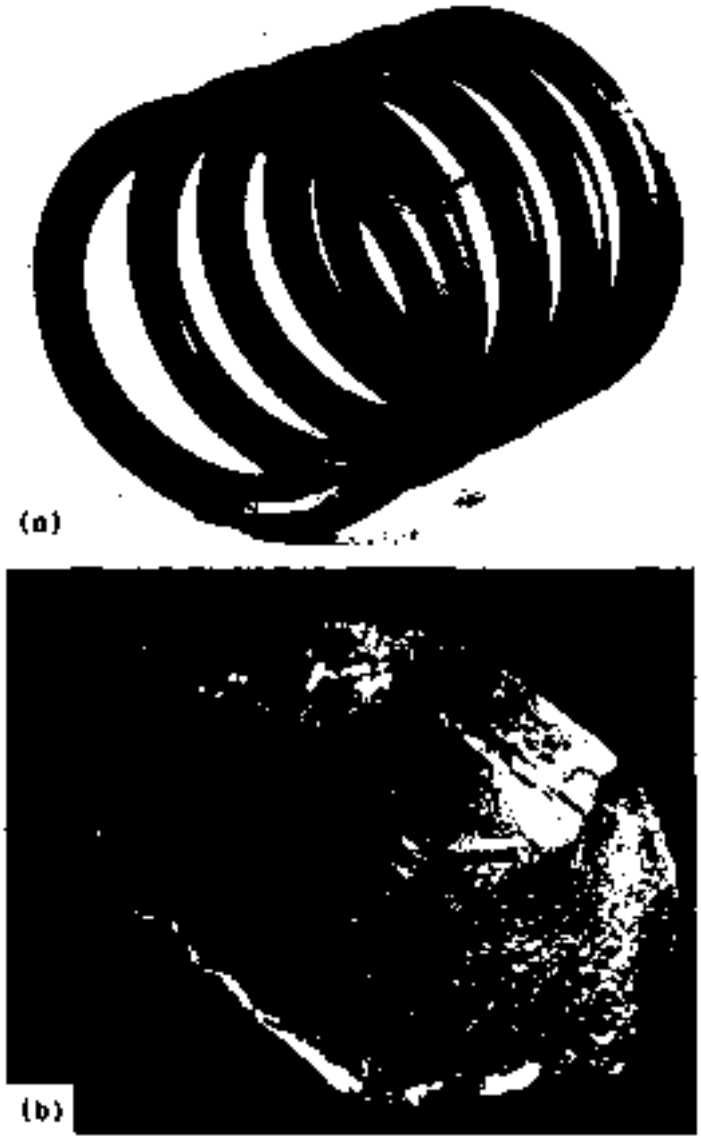


Fig. 3 Valve-spring failure due to residual shrinkage pipe. (a) Macrograph showing fracture as indicated by arrow. (b) Fracture surface; pipe is indicated by arrow.

Microscopic examination of transverse sections of the spring adjacent to the fracture surfaces revealed that the defect was a large pocket of nonmetallic inclusions at the origin of the fracture. The inclusions were alumina and silicate particles. Partial decarburization of the steel was evident at the periphery of the pocket of inclusions. The composition and appearance of the inclusions and the presence of the partial decarburization indicated that the inclusions could have been associated with ingot defects. The defect was 0.8 mm (0.03 in.) in diameter (max), 25 mm (1 in.) long, and 1.3 mm (0.05 in.) below the surface (Fig. 3b).

The steel had a hardness of 45 to 46 HRC, a microstructure of tempered martensite, and a grain size of ASTM 6 or 7, all of which were satisfactory for the application. The fracture surface contained beach marks typical of fatigue and was at a 45° angle to the wire axis, which indicated torsional fracture.

Conclusions. The spring fractured by fatigue; fatigue cracking was nucleated at a subsurface defect that was longitudinal to the wire axis. The stress-raising effect of the defect was responsible for the fracture.

Example 3: Fatigue Fracture of a Carbon Steel Pawl Spring That Originated at a Delamination at a Rivet Hole. The pawl spring shown in Fig. 4(a) was part of a selector switch used in telephone equipment. Three of these springs that broke and strip specimens of the raw material used to fabricate similar springs were examined to determine the cause of failure. The springs had been blanked from 0.4-mm (0.014-in.) thick tempered 1095 steel, then nickel plated.

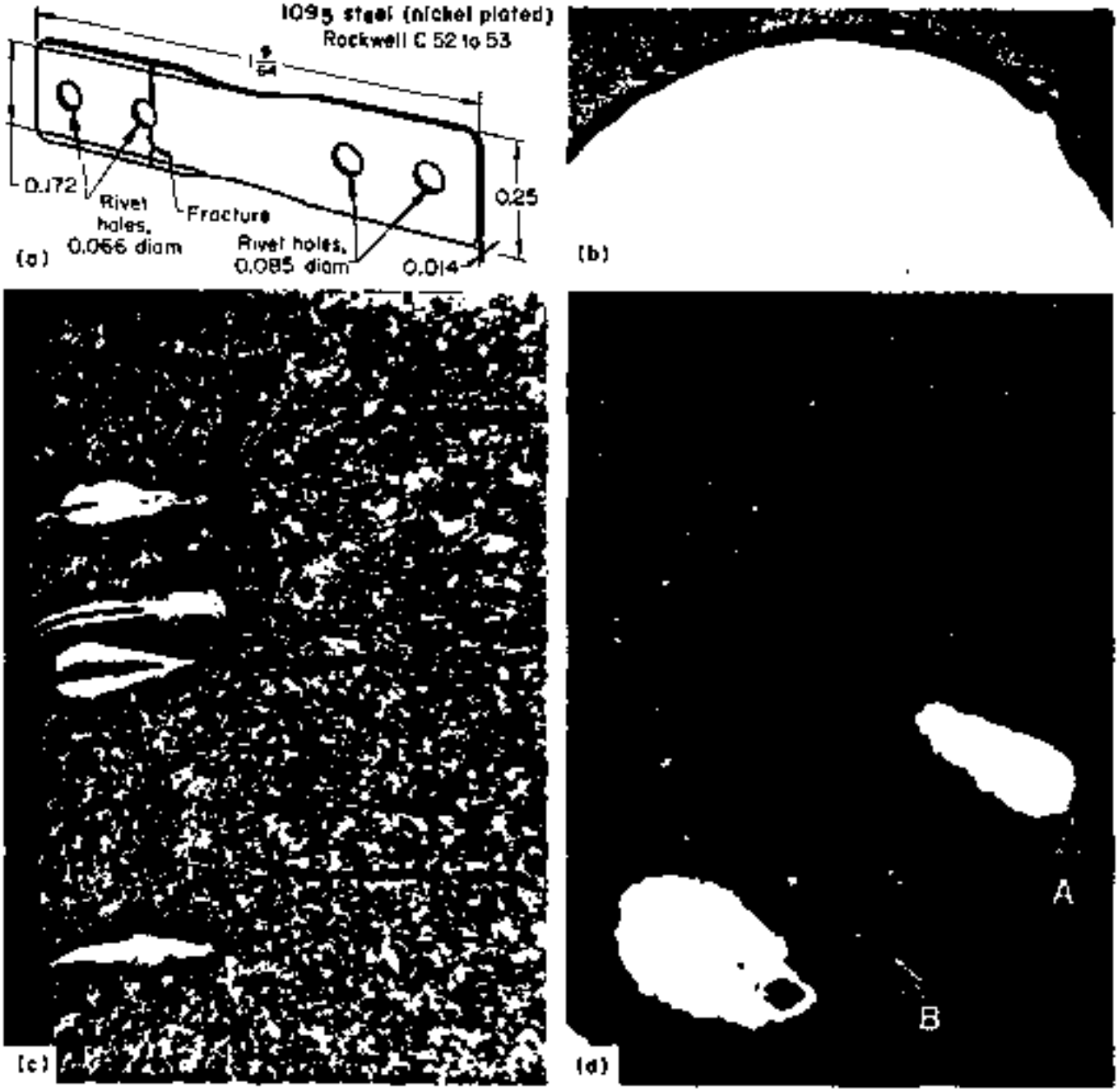


Fig. 4 Nickel plated 1095 steel pawl spring that fractured by fatigue. (a) Configuration and dimensions (given in inches) of the failed component. (b) Micrograph showing pits at edge of rivet hole. 45×. (c) Micrograph of area adjacent to rivet hole, showing delaminations (arrows) filled with nickel plating. 250×. (d) Micrograph showing delamination, at A, that initiated fracture; edge of rivet hole is at B. 250×

Investigation. Longitudinal specimens from the three broken springs were mounted and processed, and microscopic examination of these specimens revealed numerous pits around the rivet holes. These pits are readily visible in Fig. 4(b), a close-up view of one rivet hole, because their surfaces were covered with nickel plating that was not removed during grinding and etching of the specimen. The pits as well as other defects found at the rivet holes and at the bottom end of the spring were produced by a dull blanking-and-punching tool.

A longitudinal section through a rivet hole is shown in Fig. 4(c). This micrograph reveals delaminations that were formed at inclusion sites during punching of the rivet holes and that were filled with nickel during the plating operation. Each of the delaminations acted as a stress raiser, and one of them initiated the fracture (Fig. 4d).

Longitudinal specimens were sectioned from the raw material and bent slightly for mounting purposes. Examination of the stock revealed numerous long, narrow sulfide stringers, which probably were the main cause of delamination in this spring material.

The broken springs and the specimens of raw material had a hardness of 52 to 53 HRC, which was acceptable because specifications required a minimum hardness of 45 HRC. However, the hardness was higher than normal and may have contributed to delamination.

Conclusions. Fracture of the springs occurred by fatigue that originated at delaminations around the rivet holes. The fatigue was caused by stresses from blanking and rivet compression and by use of poor-quality spring material. Occasional failures have resulted from the stress pattern alone even though the material is satisfactory.

Failures of Springs

James H. Maker, Consultant, Associated Spring

Failures That Occur During High-Stress Fatigue

Figure 5 shows a modified 5160 (0.5% Cr and 0.1 to 0.2% V) hardened-and-tempered valve spring that failed due to fatigue after running 10 to 15 × 10^6 cycles at deflections considerably larger than those encountered in production engines. Depending on the size of the engine, typical loads and deflections are 70 to 110 kgf (150 to 250 lbf) and 13 mm (0.5 in.), respectively. The failure origin is at the inner diameter, and the fracture plane at that point is at an angle to the wire axis. It is not at 45° to this axis, as would be predicted for failure from torsional loading, but is much nearer the normal of the wire axis. Examples, , , , and in this article discuss valve-spring failures due to various causes that occurred during high-stress fatigue. Because all the springs are similar in appearance and composition to the valve spring shown in Fig. 5, no photographs of the individual springs will be shown; only scanning electron microscopy (SEM) fractographs of the failure origins will be given in this section. No specific data on stresses or life to failure will be given in the following examples. The failures are from a variety of automobile valve springs made on four continents and from modified valve springs run in laboratory tests.

Fig. 5 Valve spring that failed due to fatigue. Fractographs of similar valve springs are shown in Fig. 6, 7, 8, 9, 10, and 11. 0.8×

The methods for listing stresses differ. The dynamic stresses and range may be considerably greater than the calculated stresses due to harmonic vibrations excited by high engine speeds. If the material in a valve spring has a tensile strength of 1655 MPa (240 ksi), if there are no flaws in the surface deeper than a few tenths of a thousandth of an inch, and if the spring is made with a careful manufacturing process, there may still be a few field failures if the dynamic operation (torsional) stress range is of the order of 275 to 1035 MPa (40 to 150 ksi).

Example 4: Failure of Valve Spring Because of Grinding and Shot Peening Operations. The valve spring shown in Fig. 6 was made from ground rod. The transverse marks are remnants of the grinding and are apparent because the shot-peening pattern was light at this location. In this area a transverse crack grew from one such mark under the influence of local stress fields. The crack continued to grow deeper and wider until the loading became sufficient to reorient the crack front

to the plane normal to the major tensile axis. The transverse crack is somewhat galled because of the sliding motion of the crack surfaces on each other as the crack opened and closed with successive cycles.

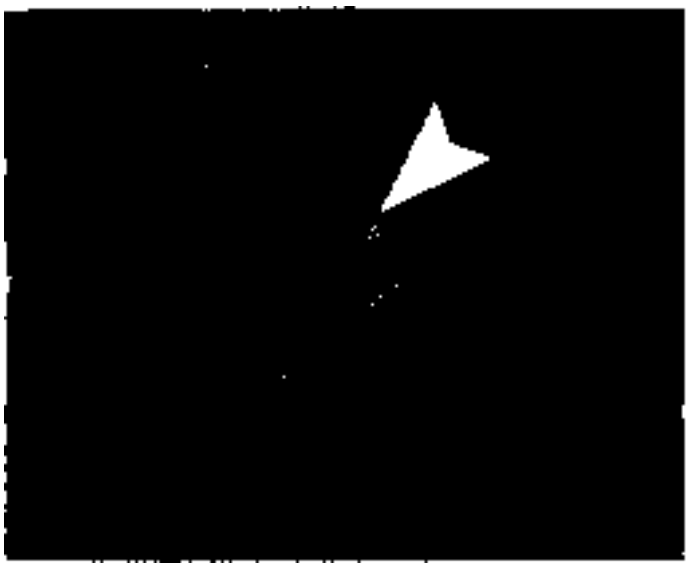


Fig. 6 Transverse failure origin in a valve spring made from ground rod. The transverse marks (arrow) are remnants of the grinding operation. 8×

Recommendations involved altering the shot-peening procedure to create adequate surface compression at all stressed points on the springs. The base of surface flaws should ideally be in the zone of high compressive stress from the peening. Example 5: Influence of Inclusions on Failure. Figure 7(a) shows two areas that are dark due to the electrostatic charging of the nonmetallics present. In Fig. 7(b), the inclusions are visible immediately below the surface (marked BB). The fracture in this zone consists of a number of triangular facets with the apexes pointing inward toward the second dark area (marked AA) with its inclusion. The triangles are much like the multiple origins discussed in Example in this article (see Fig. 2c), indicating multiple origins associated with the inclusion cluster. The second, deeper inclusion appears to be a secondary origin surrounded by a faint partial ring formation like a beach mark, with a ray pattern faintly visible. The inclusion at AA would probably not have caused a failure. Two of the small inclusions at the surface were less than 0.025 mm (0.001 in.) in length on the fracture, and because they are in the zone of maximum compression, either one might not have caused failure; however, the group created a stressed zone that fostered eventual fatigue failure.

Fig. 7 Spring failure originating at a cluster of inclusions. (a) Two adjacent dark areas (boxed zone) indicate presence of nonmetallics. 9×. (b) Two failure origins are located at BB, and one at AA. 43×

Figure 8 shows a failure origin formed at a nonmetallic inclusion located 0.2 mm (0.008 in.) below the wire surface. This is well below the zone of peak compression produced by shot peening; consequently, the protective effect is considerably lessened. The beach marks form nearly complete ellipses with a centroid near but not exactly at the inclusion. The first crack opening probably began at the boundary of the inclusion nearest the wire surface, progressing toward the centroid. A larger inclusion at the bottom of the fractograph took no part in developing the ellipse pattern. In this spring, the applied stress at this point would be about 20% lower than at the surface, and the metal would be essentially free of anything but a modest residual tension from peening. It was, therefore, not an appreciable factor in this fatigue failure.

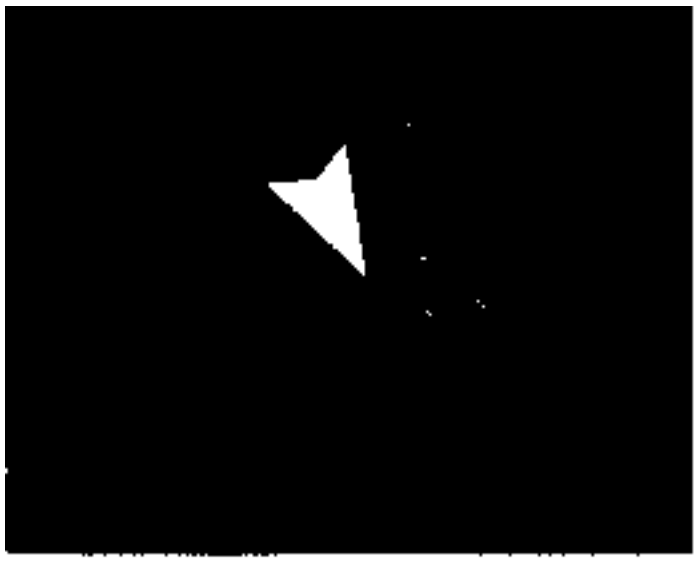


Fig. 8 Failure origin at an inclusion (arrow) located 0.2 mm (0.008 in.) below the spring surface. Note beach marks surrounding inclusion. 55×

Note also the smaller inclusion just below the surface at the left of the fractograph. The metal at this point shows a trace of postfailure damage, which did not break the hard inclusion. This nonmetallic had nothing to do with the failure, and may not have caused any difficulty, because it is in the zone of maximum compression from peening.

Example 6: Fracture at a Transverse Wire Defect. A failure originating at a transverse wire defect is shown as a band extending across the central part of Fig. 9. Above it is the shot-peened surface, and below is a zone of marks, each delineating the progress of the crack front after an increased number of cycles. The defect is about 0.13 mm (0.005 in.) deep and 0.64 mm (0.025 in.) long. This defect could have resulted from a scratch or gash incurred in handling the rod. Drawing causes the sides of such defects to be pushed into close contact.

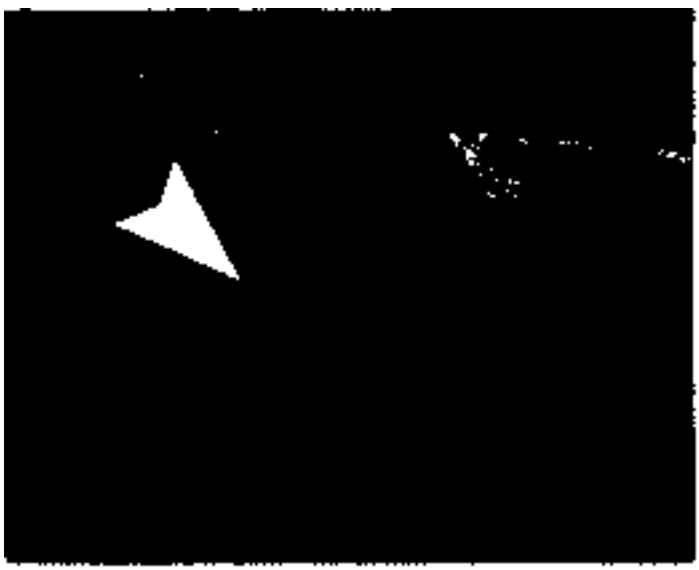


Fig. 9 Fracture at a wire defect. Beach marks are prominent beginning at the base of the flaw, which is indicated by the arrow. 39×

The striations below the defect indicate progressive crack development and are rarely seen in steel as hard as valve-spring wire. At the base of this area, several faceted areas developed, pointing away from the surface. A tensile-mode fracture plane then formed around all these fronts. This fills the bottom area of the fractograph.

Example 7: Spring Failure Due to a Surface Defect. Figure 10 shows light streaks arranged in a diagonal direction on the wire surface. The streaks are parallel to the wire axis. A darker depressed area is visible between the streaks and below the center of the fractograph. The light areas were raised and rubbed smooth by the other half of the spring. This occurred because the engine ran after failure of the spring. Careful examination of the depressed areas revealed distinct outlines that represent sharp corners in the depressions. The origin at one of the sharp corners is shown immediately above the center of the fractograph. The depressed area associated with this was in the other half of the spring that sustained more extensive postfailure damage. Apparently, a hard material (possibly mill scale) was impressed during drawing of the wire, but was broken out during peening, leaving the depressions with sharp-bottomed corners.

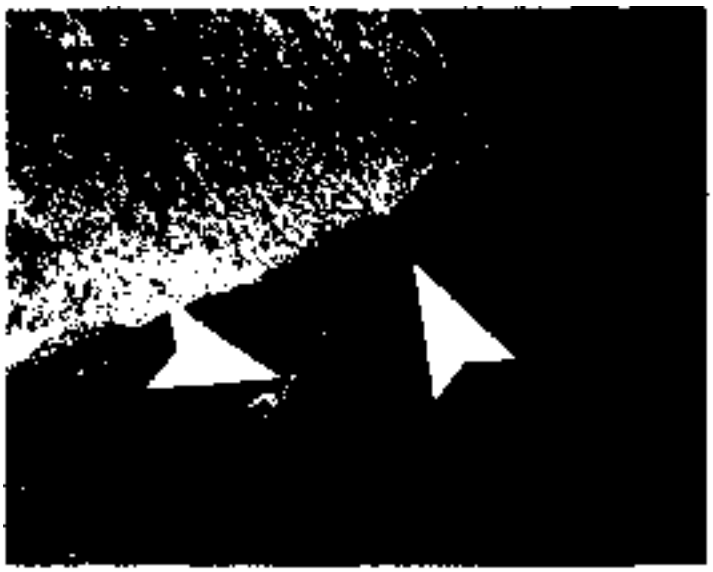


Fig. 10 Spring failure originating at a sharp-edged pitted area. Arrows indicate the location of the sharp-edged areas. 28×

Example 8: Spring Failures Originating at a Seam. The failure shown in Fig. 11(a) began at the seam that extended more than 0.05 mm (0.002 in.) below the wire surface. The fatigue-fracture front progressed downward from several origins. Each one of these fronts produces a crack that is triangular in outline and is without fine detail due to sliding of the opposing surfaces during the later stages of fracture. This occurs when the fracture plane changes to an angle with the wire axis in response to the torsional strain. These surfaces are seen in the lower part of Fig. 11(a).

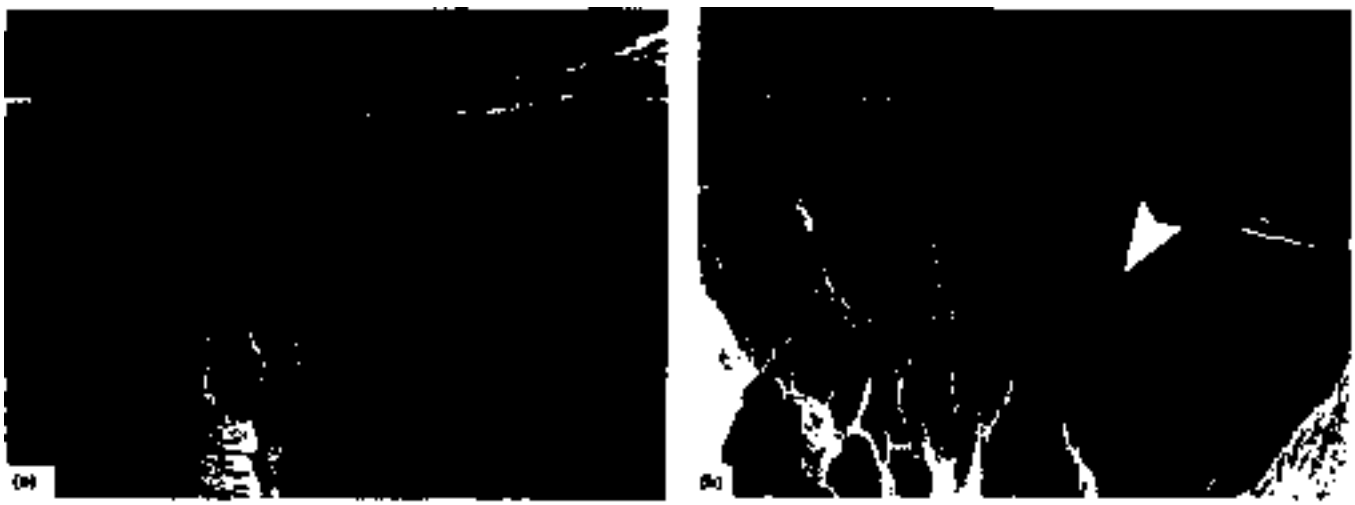


Fig. 11 Failures in wire springs. (a) Longitudinal failure originating at a seam. 45×. (b) Origin of failure at a very shallow seam. The arrow indicates the base of the seam. 115×

The failure shown in Fig. 11(b) has many of the characteristics of that shown in Fig. 11(a), except that the seam is scarcely deeper than the folding of the surface that results from shot peening. Observation of it requires close examination of the central portion of the fractograph. This spring operated at a very high net stress and failed at less than a million cycles.

James H. Maker, Consultant, Associated Spring

Failures Caused by Fabrication

Fabrication errors or deficiencies often result in spring failures. The most common of these are:

- Split wire, usually caused by overdrawing of the wire
- Tool marks
- Arc burns, sometimes occurring during welding and sometimes during plating; either can cause hard, brittle spots
- Hydrogen damage, usually resulting from improper pickling or plating practice
- Improper winding, which may allow formation of flats or scratches
- Improper heat treating
- Improper relation of spring index and hardness. A small index and high hardness may result in incipient cracks
- Unfavorable residual-stress pattern

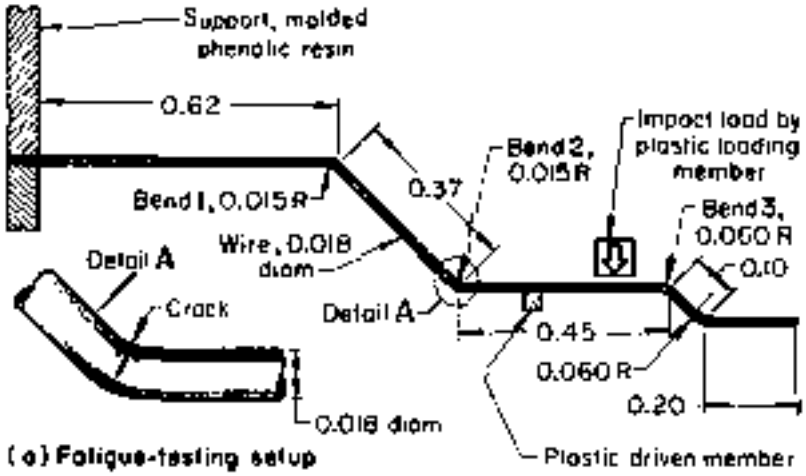
The examples that follow demonstrate how improper fabrication can result in failure.
Example 9: Failure of Carbon Steel Springs During Testing Because of Split Wire. The springs shown in Fig. 12 failed to comply with load-test requirements. These springs were formed from 3.8-mm (0.148-in.) diam cold-drawn carbon steel wire.

Fig. 12 Split wire in a 3.8-mm (0.148-in.) diam carbon steel spring (top). The spring at bottom appears to have a seam along its entire length, as indicated by the arrow.

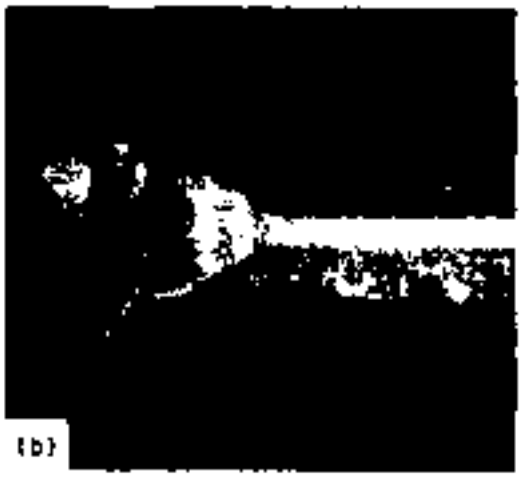
During load testing, most of the springs from a production lot supported a load of 90 kg (200 lb) without going solid. However, the springs shown in Fig. 12, which were from the same lot, supported only 60 to 70 kg (130 to 150 lb).
Investigation revealed split wire, as shown in the spring at top in Fig. 12. The spring at bottom in Fig. 12 is intact except for what appears to be a seam (arrow) that runs the entire length of the spring. The upper spring (originally similar to the lower) has been deliberately distorted to expose the split interfaces. The fracture shows a smooth heat-tinted longitudinal zone, which was the original split, and a rougher bright-appearing longitudinal zone, which resulted from tearing during testing and examination. The course of the defect can be followed for a full turn in Fig. 12.

Conclusion. The springs failed the load test because of split wire. This condition can vary from the extreme demonstrated here to occasional short lengths of one or two turns that have no visible or magnetically detectable opening to the surface and that cannot be detected until the spring is stretched out of shape. The cause of this condition is overdrawing, which results in intense internal strains, high circumferential surface tension, and decreased ductility.

Example 10: Fatigue Fracture of a Phosphor Bronze Spring Because of Tool Marks. Premature failure of a copper alloy C51000 (phosphor bronze, 5% A) spring occurred during life testing of several such springs. As shown in Fig. 13(a), the fracture occurred in bend 2, which had an inside radius of 0.38 mm (0.015 in.). The wire used for the springs was 0.46 mm (0.018 in.) in diameter and was in the spring-temper condition (tensile strength, 1000 MPa, or 145 ksi).



(a) Fatigue-testing setup



(b)

(c)

(d)

(e)

(f)

Fig. 13 Phosphor bronze (C51000) spring that failed prematurely during fatigue testing. Failure was due to the presence of a tool mark (indentation) at a bend. (a) Setup for fatigue testing, and detail of the spring showing location of crack at bend 2. (b) A broken end of the spring, 40×; the tool mark (indentation) is just to the right of fracture surface. (c) Spiral marks on spring surface. 450×. (d) and (e) Two areas of the fracture surface, 225×; flat along edge of surface in (d) is tool mark. (f) Micrograph of a longitudinal section through spring at bend 2, 145×; arrow points to crack that originated at surface at inside of bend

Test Procedure. These springs were formed, then assembled and tested as shown in Fig. 13(a) for up to 50×10^6 cycles. During testing, the springs were subjected to cyclic loading, mainly in the horizontal and vertical planes, with displacements of approximately 3.2 mm (0.127 in.) and 0.5 mm (0.020 in.), respectively. Some torsional loading also resulted. The vertical loading was applied by means of a typical cantilever arrangement. Near bend 3 the springs were subjected to impact loading by a plastic loading member, which moved a driven member approximately 0.5 mm (0.020 in.). This complex loading system developed very complicated stress patterns along the length of the springs.

The springs were examined by high-power microscopy at various intervals during testing to note the time at which cracks were initiated. After approximately 150,000 cycles, a crack was observed at the inside of bend 2 in one of the springs, as shown in detail A in Fig. 13(a).

Investigation. Microscopic examination of the surface of the fractured spring revealed an indentation at the inner surface of the bend, where fracture occurred (Fig. 13b). This indentation was presumed to have been made by the bending tool during forming. Spiral marks and other surface defects, such as those shown in Fig. 13(c), were observed on the surfaces of all the springs being tested. These spiral marks are similar to those that can be produced on springs during rotary straightening. The indentation and the other marks were, in effect, notches and would have some detrimental effect on fatigue life because of the reduced cross-sectional area in those zones.

Microscopic examination of the fracture surface (Fig. 13d) revealed two areas of prominent characteristics. One area was smooth, discolored, and rippled, indicating the gradual propagation of a crack from one or more origins. The indentation developed by the bending tool is also shown in Fig. 13(d). The remainder of the surface (Fig. 13e) had either a crystalline or a fibrous appearance, which indicated final fracture.

A longitudinal section was taken through bend 2. A micrograph of this section (Fig. 13f) revealed a crack that had originated at the surface at the inside bend and had propagated toward the outside of the bend. The crack had been initiated at a spiral mark where the fatigue limit had been decreased by a weak skin.

The small bend radius could create a condition that would result in straining at the bend zone and therefore render the section weak. It is difficult to determine to what extent the small bend radius contributed to spring fatigue, but this condition is recognized as an important factor in fatigue life.

The microstructure of the metal appeared normal for drawn wire and contained no non-metallic inclusions that would act as stress raisers.

Conclusions. The spring failed by fatigue that was initiated at a bending-tool indentation and at spiral rotary-straightener marks. The small bend radius at the fracture zone contributed to the failure.

Recommendations. The springs should be made of wire free from straightener marks, and the bending tool should be redesigned so as not to indent the wire. The small bend radius should be increased, particularly at bend 2, the area of maximum stress.

Example 11: Fatigue Fractures of Toggle-Switch Springs That Originated at Tool Marks. Several electrical toggle switches failed by fracture of the conical helical spring sealed within each switch enclosure. The springs were fabricated from 0.43-mm (0.017-in.) diam AISI type 302 stainless steel wire to the configuration shown in Fig. 14(a). In qualification testing of the springs at room temperature, fractures occurred after 11,000 to 30,000 switching cycles. Two broken springs and two unbroken springs were submitted for laboratory examination.

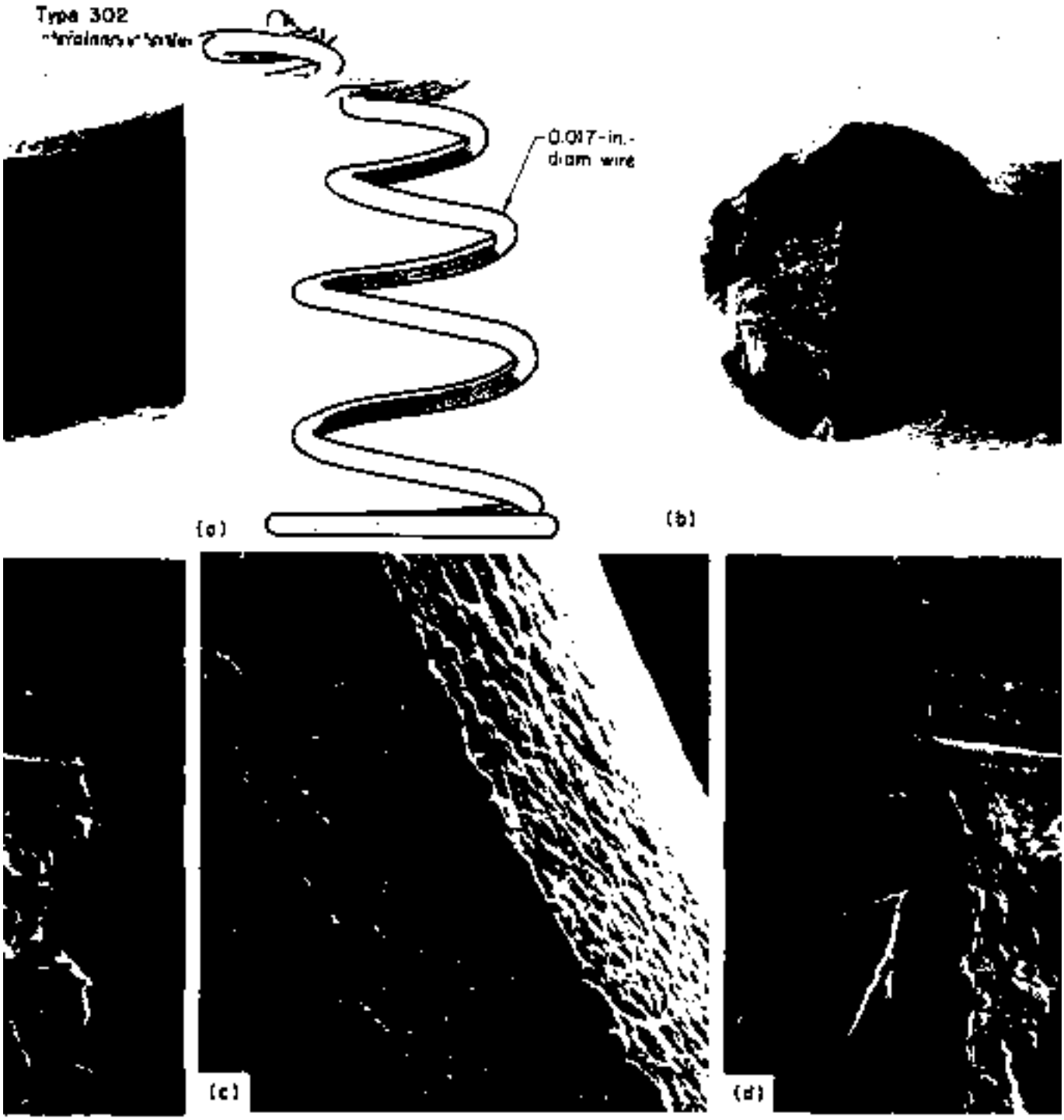


Fig. 14 Stainless steel toggle-switch spring that fractured by fatigue originating at a tool mark. (a) Configuration and dimensions (given in inches) of the spring. (b) Fracture surface, 85×; fracture origin (arrow) is at lower edge of tool mark. (c) SEM fractograph of fracture origin, 1000×. (d) SEM micrograph of surface of an unbroken spring, showing area around a tool mark, 300×

Investigation. Inspection of a fracture surface of one of the broken springs revealed a typical fatigue fracture that originated at a tool mark on the wire surface. This fracture surface is shown in Fig. 14(b); the fracture origin is indicated by the arrow at the fight edge of the fracture surface. Scanning electron microscope fractographs showed regions displaying beach marks around the fracture origin and parallel striations within the beach-mark regions. Outside the area containing beach marks the fracture surface was dimpled in a fashion characteristic of overload or rapid ductile fracture. Figure 14(c) is a fractograph of the fracture origin in Fig. 14(b). An appreciable amount of scale was observed on the fracture surfaces of both broken springs.

The inner surfaces of one unbroken spring and both broken springs had coarse textures and contained tool marks (Fig. 14d); the tool marks had been formed during the spring-winding operation. The other unbroken spring had a relatively smooth-textured surface and no tool marks. The microstructure of the wire in all four springs was normal. Chemical analysis of the material established that its composition was in conformity with specifications for type 302 stainless steel.
Conclusions. Fracture of the springs was caused by fatigue and was initiated at tool marks.
Corrective Measures. The spring-winding operation was altered to eliminate the tool marks. No further fatigue failures of the toggle-switch springs occurred.
Example 12: Fatigue Fracture of a Music-Wire Spring Caused by Poor Electroplating Practice. A cadmium-plated music-wire return spring that operated in a pneumatic cylinder was designed for infinite life at a maximum stress level of 620 MPa (90 ksi). Several such springs broke after only 240,000 cycles. No flaws were revealed by visual inspection.
Investigation. Figure 15(a) shows the surface of one spring at the origin of failure. The origin was at the outside surface of the coiled spring despite higher stress at the inside surface (due to the Wahl effect).

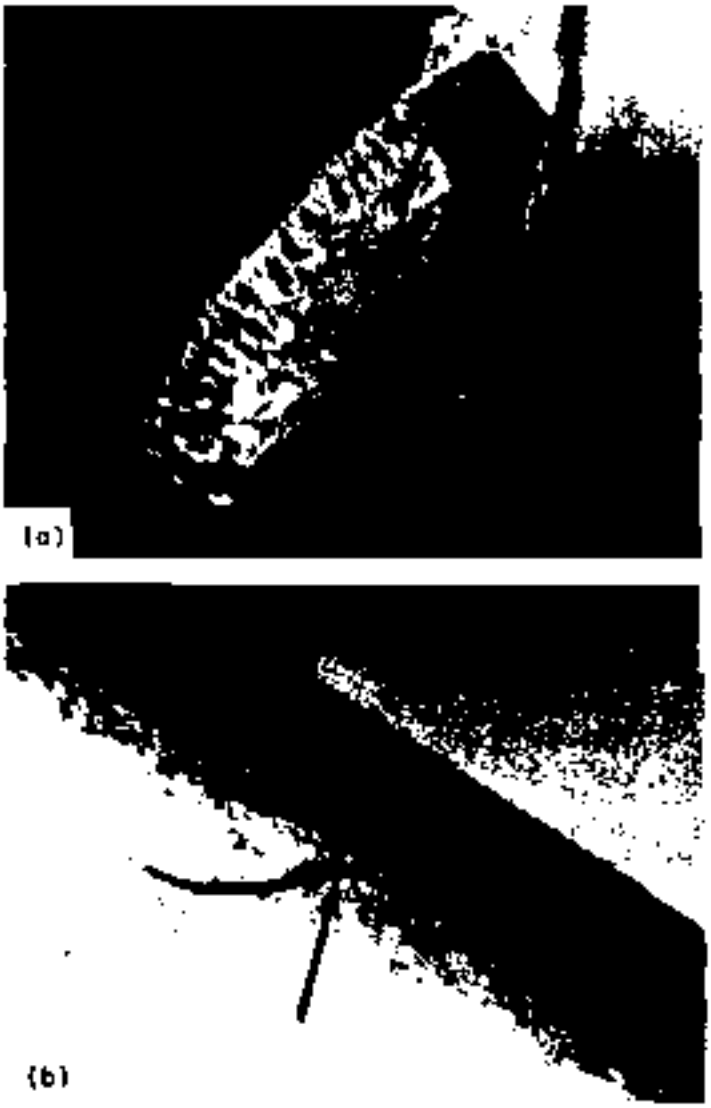


Fig. 15 Fractured music-wire spring. (a) Fracture surface at 9.2×; arrow indicates fracture origin. (b) Fracture surface, 9.2×; arrow indicates hard kernel that caused fracture

Figure 15(b) shows the fracture surface, which has a small kernel (arrow) at its center. This kernel appears as a circle on the surface and looks like a weld deposit. Stroking this kernel with a file showed it to be extremely hard, indicating that the kernel may have resulted from extreme localized overheating.
It was learned that these springs had been barrel electroplated after fabrication. It is well known that barrel loads that are too small or that tangle can result in intermittent contact with the dangler (suspended cathode contact) as the barrel rotates. This allows high local currents when the last contact is broken, resulting in an arc that causes local melting of the

metal being plated. Because the motion of the barrel rapidly moves the spring away from the dangler, the spot of molten metal caused by arcing is quenched instantly by the plating solution and by the mass of the cold metal of the spring.
Conclusion. Fatigue fracture of the spring resulted from a hard spot that was caused by arcing during plating.
Recommendations. Hard spots can be avoided either by resorting to rack plating, which is more expensive than barrel plating, or by using barrels that have fixed button contacts at many points instead of dangler-type contacts.
Example 13: Fatigue Fracture of a Carbon Steel Counterbalance Spring Caused by Hydrogen Damage. During fatigue testing, the power-type counterbalance spring shown in Fig. 16(a) fractured at the two locations indicated by the arrows. The spring had been formed from hardened-and-tempered carbon steel strip and subsequently subjected to a phosphating treatment. In the normal manufacture of these springs, winding to solid is performed to take out set (provide stable torque output) before phosphating. Although this operation creates residual tensile stresses in the interior surfaces of the coils, phosphating of this spring had not produced sufficient hydrogen absorption to cause embrittlement.

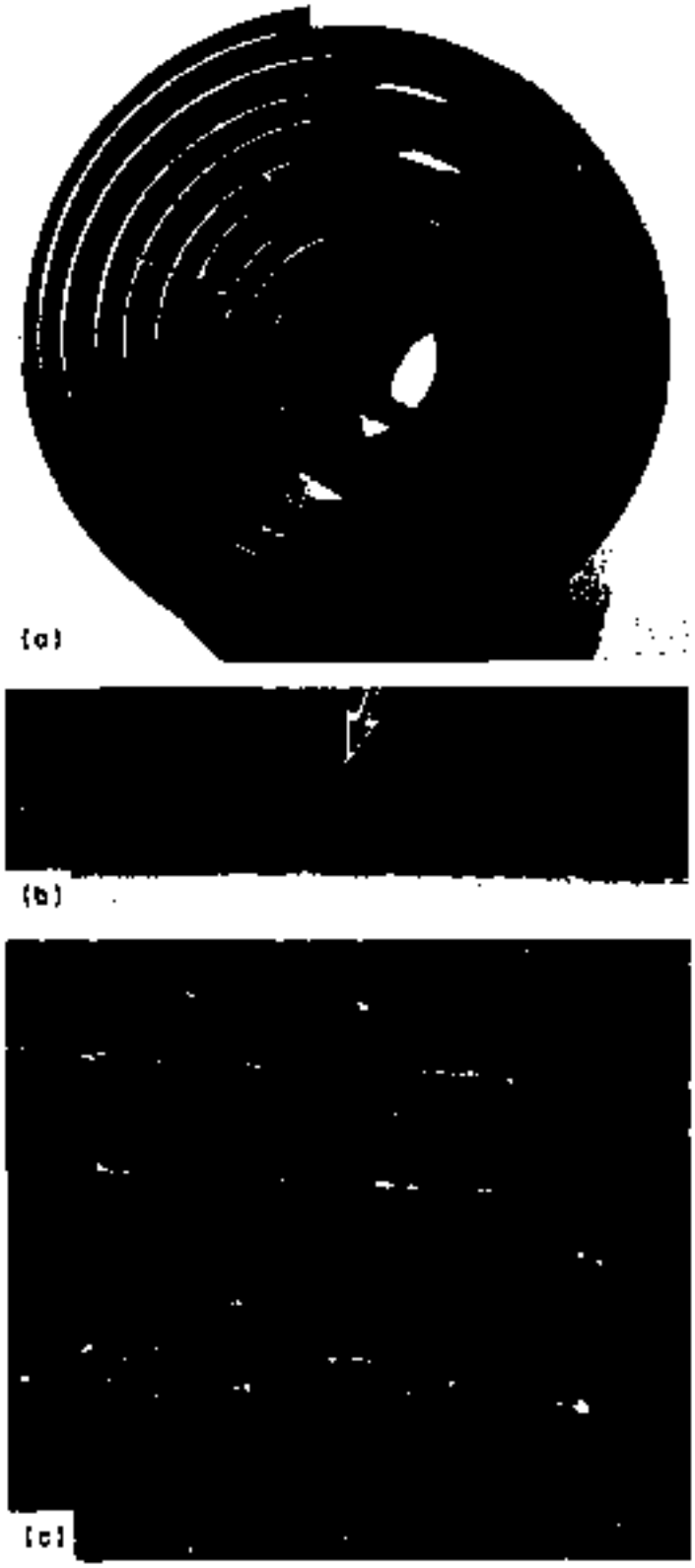


Fig. 16 Carbon steel counterbalance spring that failed during fatigue testing. (a) Macrograph showing fracture locations (arrows). $\frac{1}{3}$×. (b) Fracture surface showing dark band (arrow) that nucleated fracture. 6×. (c) Etch pits in surface. 100×

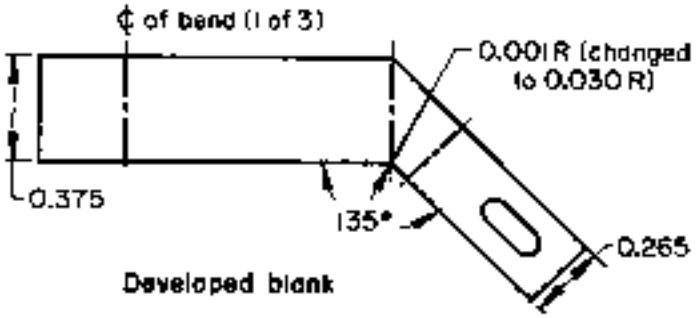
Investigation disclosed a dark band at the inside edge of the fracture surface, as indicated by the arrow in Fig. 16(b). This dark band was rust colored and extended through about 20% of the spring thickness. Examination of the cleaned surface revealed etch pits 0.05 to 0.1 mm (0.002 to 0.004 in.) deep (Fig. 16c). Such pits were never observed on properly phosphated springs; therefore, the spring that failed must have been subjected to an abnormal acid attack in pickling or phosphating. This attack resulted in considerable absorption of hydrogen by the metal.

Conclusion. The dark band in Fig. 16(b) could have been rust colored only by the chemical action of some aqueous medium, which means that the part cracked in a water-base solution—either the phosphating solution or the pickling solution.

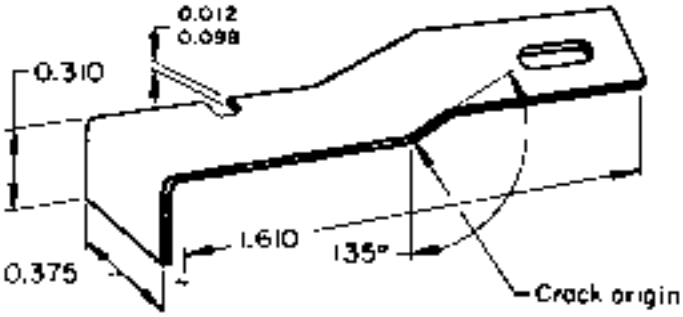
Cracks can result from vigorous pickling. The sum of the tensile stress from setout (which deforms the metal beyond the yield point) and the internal hydrogen pressure exceeds the strength of the metal. When the crack reaches the edge of the area of residual tensile stress, it stops. In normal practice, when hydrogen pickup is low, no fracture occurs. It was therefore suspected that the spring had been subjected to excessive acid pickling before phosphating and that this excessive pickling was responsible for hydrogen absorption by the spring. Normal phosphating treatments do not cause hydrogen embrittlement.

Example 14: Fatigue Fracture of a Carbon Steel Wiper Spring Because of Stress Concentration at a Sharp Corner. Parts such as the one shown in Fig. 17(a) were used as grease-wiper springs for cams. The springs were formed from stampings of 0.25-mm (0.010-in.) thick carbon spring steel (0.65 to 0.80% C) and were hardened to HR15N 84 to 86. The springs were fracturing at the 0.025-mm (0.001-in.) radius on the stamped 135° corner at a 90° bend after 5×10^6 cycles; normal life was 2.5×10^9 cycles.

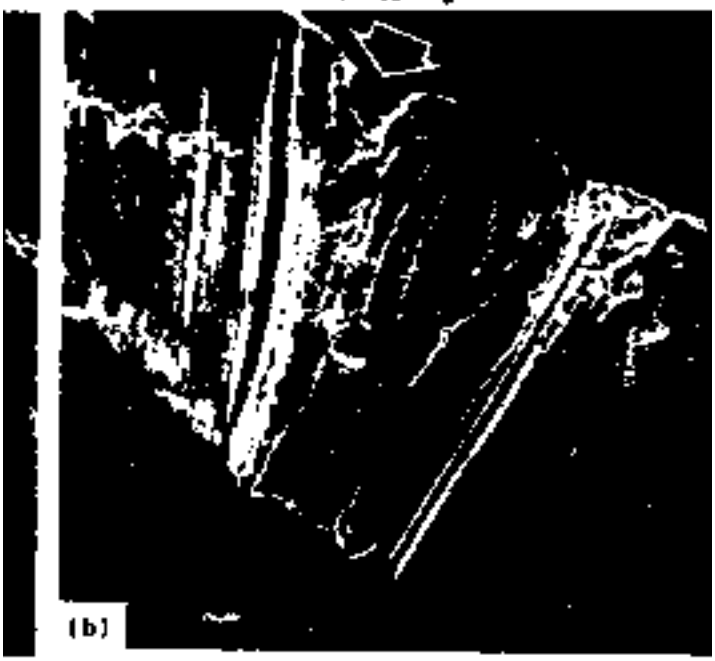
Carbon spring steel, 0.65 to 0.80% C
Rockwell 15-N 84 to 86



Developed blank



(a) Formed spring



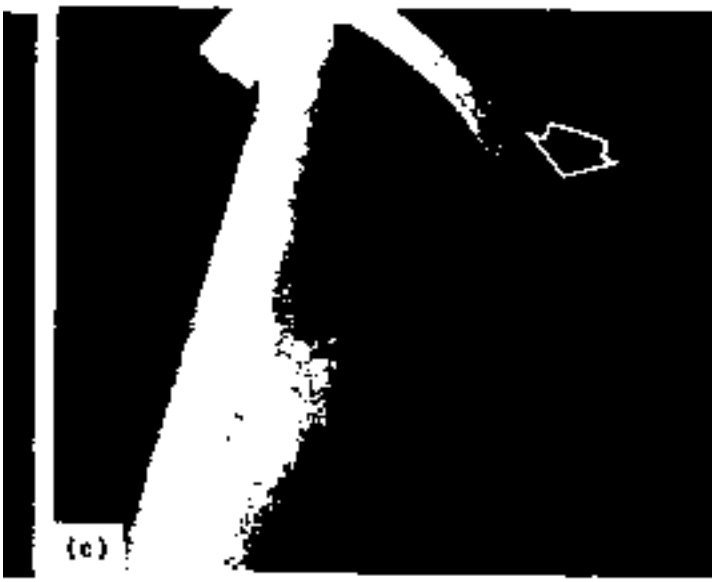


Fig. 17 Wiper spring that fractured at a small-radius corner of a stamped bend. (a) Configuration and dimensions (given in inches) of the spring. (b) SEM micrograph showing a forming crack (arrow) in the 135° corner on a new spring. 200×. (c) SEM micrograph showing a crack (arrow) that originated at the 135° corner in a used spring. 20×

Four broken springs were sent to the metallurgical laboratory to determine the cause of fracture. Three new springs from stock and six used springs from reconditioned machines were also sent to the laboratory for inspection and comparison. The grease-wiper springs that fractured were of an old design, but they had been made in a new die.

Investigation. Visual examination of all the springs in the laboratory disclosed tool marks 2 to 2.3 mm (0.080 to 0.090 in.) from the center of the stamped bend. The fracture surfaces were nearly parallel to these marks, but the fractures had not been initiated by them.

Examination of the fracture surface by SEM revealed fatigue striations originating from cracks at the 0.025-mm (0.001-in.) radius inside corner at the bend. Examination of a new part from stock and one from a reconditioned machine also showed cracks originating at the small radius inside corner. Fig. 17b). This micrograph also shows the condition of the cut edge of the stamping.

On the new parts, the cracks extended approximately 0.5 mm (0.020 in.) along the width. On the used parts, the cracks extended across approximately 75% of the width and completely through the stock thickness (Fig. 17c).

Stress calculations indicated the maximum stress at the bend, in stock of maximum thickness (0.3 mm, or 0.012 in.), to be 283 MPa (41 ksi) without considering the stress-concentration factor in the 135° corner. The minimum stress under the same conditions would be 83 MPa (12 ksi). Based on 0.3-mm (0.012-in.) thick stock and a maximum stress of 283 MPa (41 ksi), calculated stress as a function of the radius of the 135° corner was:

Radius		**Stress concentration factor**	Maximum stress	
mm	**in.**		**MPa**	ksi
0.025	0.001	3.0	841	**122**
0.38	0.015	1.82	510	**74**
0.64	0.025	1.56	434	**63**
0.76	0.030	1.52	427	**62**
0.89	**0.035**	**1.46**	**407**	**59**

The maximum allowable fluctuating stress for this material was 945 MPa (137 ksi). Therefore, the 841 MPa (122 ksi) maximum stress provides a very small factor of safety.

Conclusions. The wiper springs fractured in fatigue that originated at 0.025-mm (0.001-in.) inside radius at the corner of a stamped bend. The very small radius resulted in a high stress-concentration factor in the spring, which caused the maximum applied stress to approach the allowable limit. Tool marks at the bend were not considered major causes of fracture. Cyclic loading resulted from cam rotation.

Corrective Measures. The corner radius was increased to 0.76 mm (0.030 in.), and the tools were repolished to avoid tool marks. These changes increased fatigue life of the springs to an acceptable level.

Example 15: Fatigue Failure Caused by Weld Spatter. A medium-carbon helical spring was installed in a machine assembly that was welded into its final location. During welding, which was conducted several inches from the spring, no shield was used to prevent spatter from landing on the wire surface. Figure 18 shows an elongated drop of metal a short distance from the fracture and two tiny droplets between it and the fracture. Although no spatter was observed at the origin, visual observation revealed an oxide color difference on the surface at the origin. The other half of the fracture was not available, but it was assumed that a drop of molten metal landed at the origin. The degree of heating involved would lessen or eliminate the compressive stress produced by shot peening, and contraction of the surface in contact with the tightly adherent weld metal during cooling could generate other stresses. Adherence of the spatter drop could be affected by the opening and closing of the fatigue crack.

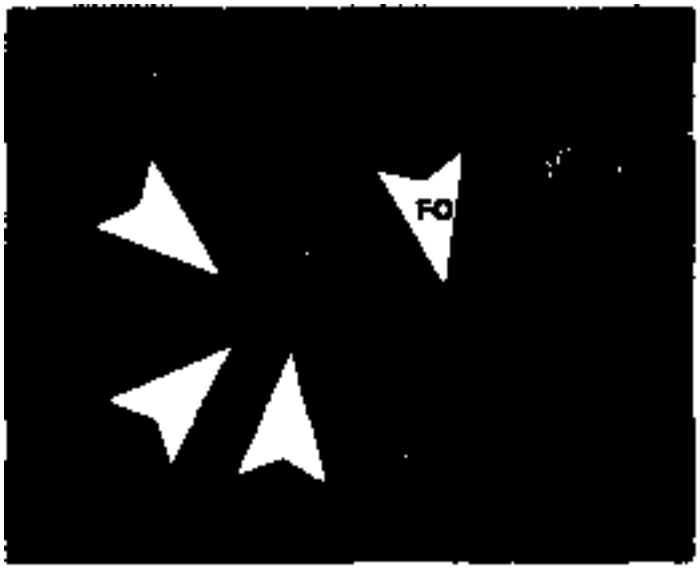


Fig. 18 Helical spring that failed by fatigue. Weld spatter (arrows) was believed to have caused the failure (see text). Arrow FO indicates fracture origin. 29×

Conclusion. No evidence of other reasons for failure could be found. It was therefore concluded that a weld spatter bead had resulted in the fatigue failure.

Example 16: Flat-Spring Failure Due to an Edge Defect. The face of a 6150 flat spring was under tensile stress. Although small flaws were visible on the surface (Fig. 19), the failure began at the dark spot on the edge, where roughness resulted from shearing during the blanking operation. On the fracture surface at the left, rays approximately 0.18 mm (0.007 in.) below the surface can be seen diverging from the origin.

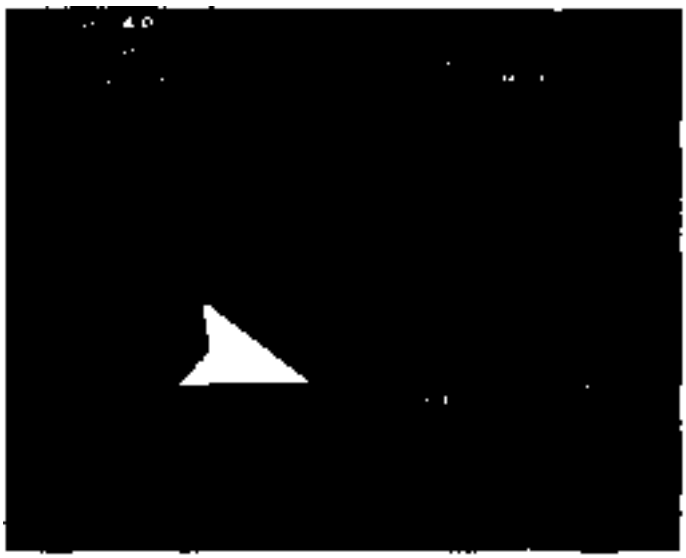


Fig. 19 Failure origin (arrow) on the edge of a flat spring. 58×

Failures Caused by Corrosion

Corrosive environments very often are principal contributors to failures of springs. Any metal is more likely to be attacked when under stress. Consequently, springs are extremely susceptible to stress-corrosion cracking.

Example 17: Stress-Corrosion Cracking of Inconel X-750 Springs. Springs such as that shown in Fig. 20(a) were used for tightening the interstage packing ring in a high-pressure turbine. After approximately seven years of operation, the turbine was opened for inspection, and several of the springs were found to be broken. Operation of the turbine was not impaired by the broken springs. The springs were made of 1.1-mm (0.045-in.) diam Inconel X-750 wire. The interstage packing rings were machined from centrifugally cast leaded nickel brass (German silver). The broken springs were submitted for laboratory analysis.

Fig. 20 Inconel X-750 spring that failed by stress-corrosion cracking. (a) Configuration and dimensions (given in inches) of the spring. (b) and (c) Unetched longitudinal sections showing intergranular cracking. 100× and 500×, respectively

There were chemical deposits in the turbine starting at about the fifth stage. These deposits were judged to have come from the water or boiler chemicals. There were no broken springs in the fifth or sixth stages. The temperatures attained in the turbine stages were:

- *Stage 1:* 462 °C (864 °F)
- *Stage 2:* 405 °C (761 °F)
- *Stage 3:* 372 °C (702 °F)
- *Stage 4:* 351 °C (664 °F)
- *Stage 5:* 330 °C (626 °F)
- *Stage 6:* 313 °C (595 °F)

Investigation. Many of the springs were broken into small pieces that had a white deposit on their rustlike surface scale. The white deposit was 100% water soluble and had a pH of 9.6 (slightly alkaline). Spectrographic examination showed the surface scale to contain a high amount of sodium, a small amount of tin, a high trace of zinc, and traces of lead and calcium.

Chemical analysis of the spring material showed that it was within specifications for Inconel X-750. The hardness values were within the specified range of 43 to 48 HRC.

Metallographic examination of a longitudinal section through the wire revealed intergranular cracks (Fig. 20b and c) about 1.3 mm (0.05 in.) in depth and oriented at an angle of 45° to the axis of the wire. In some places, these cracks were 90° apart on the wire circumference. A light-gray phase had penetrated the grains on the fracture surfaces. A fine lacey attack was also observed along the surface of the wire to a depth of about 0.013 mm (0.0005 in.). This general attack had the appearance of liquid-metal corrosion. When bent in the laboratory, the spring wires fractured in a brittle manner along a plane 45° to the wire axis, which is characteristic of brittle fracture from torsional loading.

High-nickel alloys are susceptible to embrittlement by some low-melting metals and alloys. Tin, lead, and zinc melt below the maximum service temperature of the turbine, which was 462 °C (864 °F). One Sn-Zn system had a eutectic melting point of less than 204 °C (400 °F). Such metals can cause liquid-metal embrittlement at temperatures above their melting points.

Sodium, tin, and zinc could have been present in a pigmented oil or grease used as a lubricant during spring winding and could have been left on the Inconel wire by the manufacturer. The white deposit on the surface of some of the springs was slightly alkaline and could have initiated stress-corrosion cracking by caustic penetration.

Conclusions. The springs fractured by intergranular stress-corrosion cracking during service. The cracks were promoted by the action of liquid zinc and tin in combination with static and torsional stresses on the spring wire. Stress-corrosion cracking by caustic action was an alternative but less likely mechanism of fracture.

Corrective Measures. The springs were cleaned thoroughly by the spring manufacturer before shipment to remove all contaminants and were cleaned again before installation in the turbine. Failure of the springs was thus reduced to a minimum.

Protective Coatings. Electroplated coatings of any of several metals or organic coatings, such as paint, are commonly used to protect springs that must operate in corrosive environments. These means of protection can be very effective, but either or both of two deficiencies are likely to prevail, particularly for helical springs that have closely spaced coils. The portions of springs that have the greatest need for protection (often about $1\frac{1}{2}$ to 2 coils from the ends) usually get the least protection because of the proximity of the coils to each other and are the areas from which protective coatings are most likely to be rubbed or broken away in operation.

Failures Caused by Operating Conditions

In some instances of spring failure, the cause is not evident, even after an exhaustive examination of material, processing, and design factors. When a thorough investigation in these areas does not reveal the cause, it is necessary to investigate operating conditions. Changes in operating conditions that are thought to be only minor deviations from normal sometimes prove to be the causes of spring failure.

Example 18: Fracture of a Landing-Gear Flat Spring. A flat spring for the main landing gear of a light aircraft broke after safe execution of a hard landing. The spring (Fig. 21a) was submitted for laboratory examination. No information was furnished concerning design, material, heat treatment, or length of service.

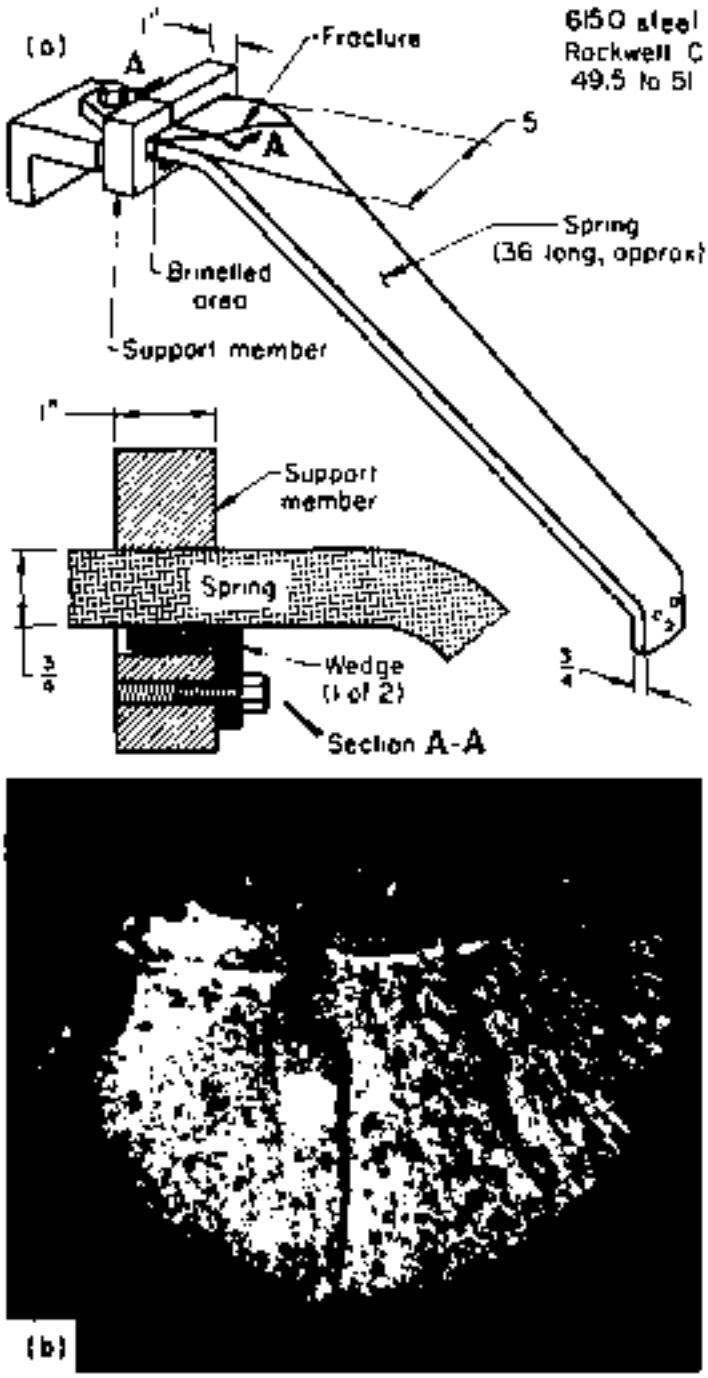


Fig. 21 Landing-gear spring, 6150 steel, that broke during a hard landing. (a) Configuration and dimensions (given in inches) of the spring. (b) Fractograph showing fatigue crack that initiated the brittle fracture. 7×

Investigation. Fracture occurred near the end of the spring that was inserted through a support member about 25 mm (1 in.) thick and attached to the fuselage by a single bolt. The required fit between the spring and the support member was obtained by inserting wedges beneath the spring at each side of the support. These wedges were bolted to the support member, and the degree of tightness of the assembly was controlled by the tension on the bolts. Details of the assembly are shown in section A-A in Fig. 21(a).

The spring had broken laterally at an angle of about 60° to its longitudinal axis (Fig. 21a). One end of the fracture corresponded to the point at which the forward edge of the spring contacted the support member. The other end of the fracture was in an unsupported area.

Brinelling (plastic flow and indentation due to excessive localized contact pressure) had occurred on the upper surface of the spring where the forward and rear edges of the spring contacted the support member. There was no evidence of brinelling on the bottom surface of the spring where it contacted the support member.

Visual examination of the fracture surface revealed that the spring had failed by mainly brittle fracture. Chevron marks indicated that fracture had started beneath the brinelled area at the forward edge of the upper surface of the spring. Figure 21(b), a light microscope fractograph of this corner, shows that the origin of the brittle fracture was in fact a small fatigue crack that had been present for a considerable period of time before final fracture occurred.

Microscopic examination of longitudinal and transverse sections near the fracture confirmed that the material had suffered extensive grain flow from cold working in the brinelled areas. The general microstructure was fine-grained tempered martensite of good quality. There was no surface decarburization.

Chemical analysis identified the spring material as 6150 steel. The hardness of the material was 49.5 to 51 HRC.

Discussion. Normal recurring vertical landing loads cause the top surface of the spring at the edge of the support to be stressed in compression. These probably were the major loads applied to the spring. On the other hand, side and drag loads produce tensile stresses at the top surface of the spring. When vertical, side, and drag loads are applied simultaneously during a landing, the resultant stress at the support is compressive. However, after the aircraft has touched down and the vertical load has decreased to the static load of the aircraft, the continuing application of side and drag loads during taxiing causes a tensile stress at the top surface of the spring.

Compressive stresses in the top surface of the spring at the support are generally localized at the edges of the support by the wedges and by any lack of straightness across the section that may have been introduced during fabrication or that may have occurred when the complicated loading pattern caused a torsional deflection in the region of failure. Also, it was considered probable that the assembly could have been more tightly compressed than it actually was. This would explain the brinelling observed near the forward and the rear edges of the top surface. The localized plastic deformation under compressive loading (brinelling) created high residual tensile stresses in the deformed region, reducing the capacity of the material to withstand fluctuating tensile stresses due to applied loads.

Ultimately, a crack started at the corner where the fluctuating tensile stress was highest under combined side and drag loads. This crack progressed by fatigue over a relatively long period of time (many landings). Final fracture occurred when the applied stress (primarily tensile stress due to side and drag loads) exceeded the critical fracture stress for the partially cracked spring. Under normal conditions, such a cracked part might last a long time, although complete fracture would occur eventually. An abnormal condition, such as a sharp change of direction on a runway or taxiway, rolling over the edge of a runway, or encountering a rut, could impose a sufficient amount of additional loading to cause final, fast fracture.

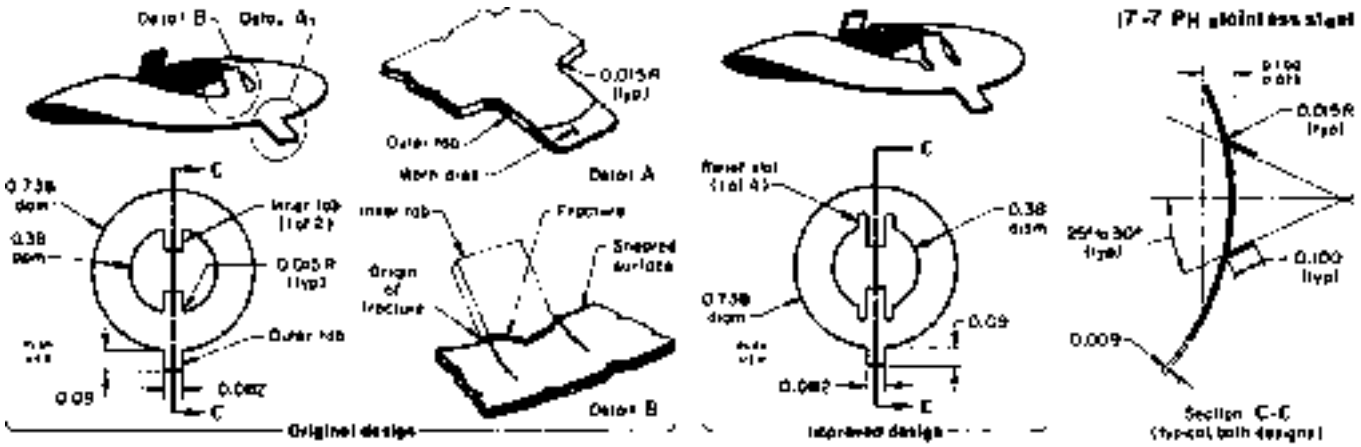
Conclusions. Fracture of the landing-gear spring was caused by a fatigue crack that resulted from excessive brinelling at the support point. The fatigue crack significantly lowered the resistance of the material to brittle fracture under critical side and drag loads.

Although the loading at the time of final fracture may have been unusual, a spring that did not contain a fatigue crack similar to the one found in the failed spring probably would not have broken. The fatigue crack had been present in the part for some time, and suitable inspection would have revealed it.

Recommendations included regular visual examinations to detect evidence of brinelling and wear at the support in aircraft with this configuration of landing-gear spring and means of attachment. If visual examination showed evidence of brinelling or wear, the spring should be magnetic-particle inspected for evidence of cracks. Liquid-penetrant inspection might be adequate if the amount of brinelling were minor, although care should be used in this type of inspection because brinelling can hide small cracks.

Cracked parts should be replaced. Brinelling itself need not be cause for retiring a part. The presence of a crack was considered necessary for failure to occur in the manner described in this example.

Example 19: Failure of a Retainer Spring Because of Cyclic Loading and Torsional Vibration. The part shown in the “original design” portion of Fig. 22 was a valve-seat retainer spring from a fuel control on an aircraft engine. This spring was found to be broken during disassembly of the fuel control for a routine overhaul after 3980 h of service. The two inner tabs had broken off but were not recovered. There had been no known malfunction of the fuel control.



the general area or location of failure, noting the proximity to any possible stress concentrations found when examining the part drawings.

Fracture surfaces should be examined visually to determine if there are indications of one or more fracture mechanisms and if there is an apparent crack origin. Surfaces of the component adjacent to the fracture surface should be examined for secondary cracks, pits, or imperfections. Photographs should be taken to record the condition of the pertinent parts before physical evidence is destroyed by subsequent examinations.

Nondestructive methods of inspection, such as ultrasonic inspection, can sometimes provide useful information. Such methods may reveal other cracks that have not progressed to rupture; these cracks may have other cracks or fracture surfaces that are not as badly damaged as the primary fracture and that can be diagnosed more readily.

Also, some machines and structures may have other shafts similar to the one that failed; these shafts may have the same service history as the failed shaft, and examination may reveal cracks that can provide useful information. Whenever possible, the analyst should also compare shafts that are from machines in service but have not failed.

Macroscopic Examination (from less than actual size to approximately 50×). Many characteristic marks on fracture surfaces, though visually identifiable, can be better distinguished with macroscopic examination. A magnifying glass or binocular microscope can be used to study the unique vestigial marks left on tensile-fracture surfaces in the form of fibrous, radial, and shear-lip zones from which the relative amounts of ductility and toughness possessed by the metal can be appraised. In the study of fatigue fractures, macroscopic examination of such features as beach marks, ratchet marks, fast-fracture zones, and crack-initiation sites, if present, may yield information relative to the kinds and magnitude of the stresses that caused failure.

Microscopic Examination (50 to 2000×.) Metallographic sections taken through fractures are used to classify fracture paths (transgranular or intergranular), to establish the mode of fracture (shear or cleavage), and to locate and identify crack-initiation sites. Plating before mounting can be used to preserve the edge or edges of the fracture surface. In addition, metallographic examination can reveal microstructure near the fracture surface, the grain size of the material, and the presence of undue segregation, inclusions, alloy concentrations, brittle grain-boundary phases, decarburization, and fabricating imperfections.

Scanning Electron Microscopy (SEM) and Transmission Electron Microscopy (TEM) Examination (10 to More Than 20,000×). The optical metallograph is limited for fracture studies by its restricted depth of field (the scanning electron microscope has a depth of field about 300 times that of the optical microscope). Thus, the electron microscope is better suited for fractographic work. With it, fractures may be classified by fracture path, fracture mechanism, and fracture features. There are two fracture paths: transgranular (or transcrystalline) and intergranular (or intercrystalline).

The fracture mechanisms and related microscopic features of transgranular fracture are microvoid coalescence (dimples, associated with ductile fracture), tearing (tear ridges), cleavage (river patterns, feather marks, Wallner lines, and cleavage tongues, all associated with brittle fracture), and fatigue (striations and tire tracks).

The fracture mechanism of intergranular fracture is grain-boundary separation. The causes may be the presence of grain-boundary phases, alloy-depleted boundaries, and environmental or mechanical factors such as stress-corrosion cracking, hydrogen damage, heat damage, and triaxial stress states.

Mechanical testing and chemical analysis occasionally pinpoint the cause of a failure as wrong material, improper heat treatment, or in-service changes in properties. Hardness testing and spectroscopic analysis should be performed as a matter of course. Impact tests, tensile tests, and other special mechanical tests may be performed if manufacturing procedures are in doubt or if other paths of investigation are not fruitful.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Stress Systems Acting on Shafts

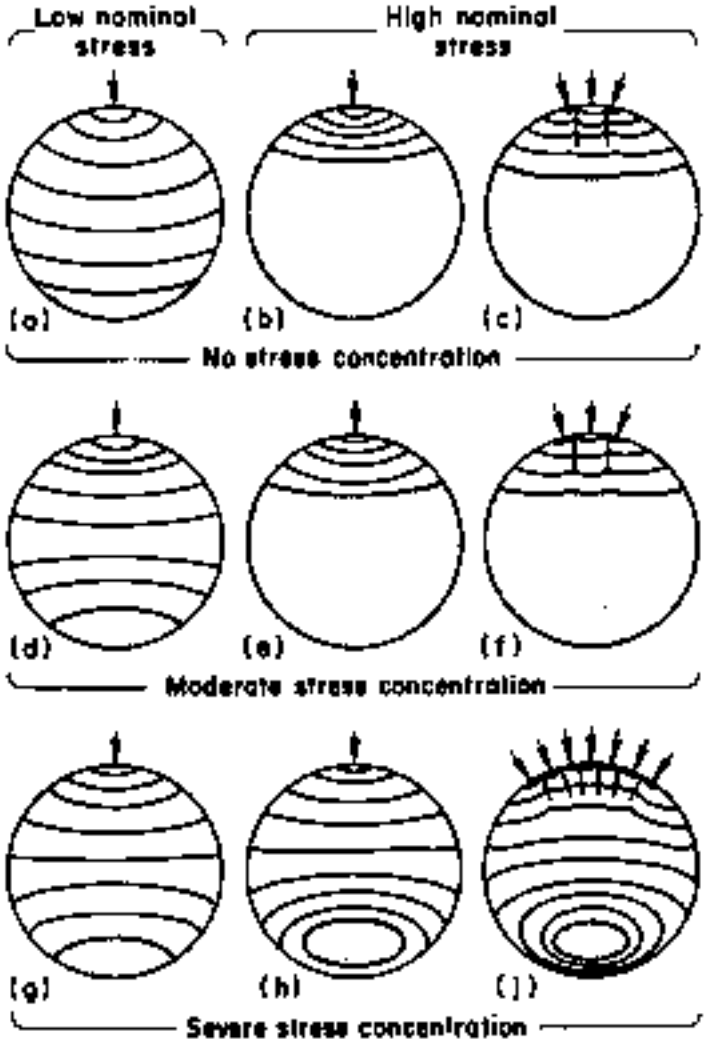


Fig. 2 Fatigue marks produced from single origins at low and high nominal stresses and from multiple origins at high nominal stresses. Fatigue marks are typical for a uniformly loaded shaft subjected to unidirectional bending. Arrows indicate crack origins; final-fracture zones are shaded.

Figure 2(c) shows a typical fatigue crack originating as several individual cracks that ultimately merged to form a single crack front. Such multiple origins are usually indicative of high nominal stress. Radial steps (ratchet marks) are present between crack origins.

Figures 2(d), (e), and (f) show typical fatigue beach marks that result when a change in section in a uniformly loaded shaft provides a moderate stress concentration. With a low nominal stress, the crack front changes from concave to convex before rapture (Fig. 2d). At higher nominal stresses, the crack front flattens and may not become convex before final fracture (Fig. 2e and f).

A change in section in a uniformly loaded shaft that produces a severe stress concentration will lead to a pattern of beach marks such as that shown in Fig. 2(g), (h), or (j). An example of a severe stress concentration is a small-radius fillet at the junction of a shoulder and a smaller-diameter portion of a shaft or at the bottom of a keyway. Such a fillet usually results in the contour of the fracture surface being convex with respect to the smaller-section side. The crack-front pattern shown in Fig. 2(g) was produced by a low nominal stress. The crack front in Fig. 2(h) developed more rapidly because of a higher stress in the peripheral zone. Multiple crack origins, high nominal stress, and unidirectional bending usually produce the beach-mark pattern shown in Fig. 2(j).

Example 1: Unidirectional-Bending Fatigue Failure of an A6 Tool Steel Shaft. The shaft shown in Fig. 3 was part of a clamping device on a tooling assembly used for bending 5.7-cm (2.25-in.) outer-diameter tubing on an 8.6-cm (3.375-in.) radius. The assembly contained two of these shafts, both of which failed simultaneously and were sent to the laboratory for examination. The maximum clamping force on the assembly was 54,430 kg (120,000 lb). The material specified for the shafts was a free-machining grade of A6 tool steel.

front usually swings around about 15° or more (Fig. 5a and c). A third difference arising from rotation is in the distribution of the initiation sites of a multiple-origin crack.

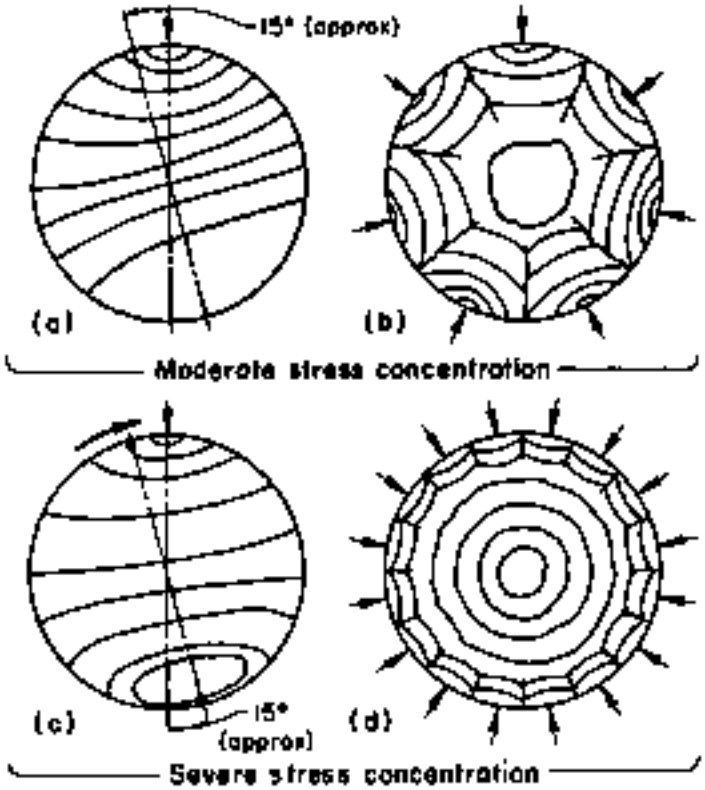


Fig. 5 Typical fatigue marks on the fracture surface of a uniformly loaded rotating shaft. Marks are produced from single and multiple origins (arrows) having moderate and severe stress concentration; shaded areas are final-fracture zones. Shaft rotation is clockwise.

In a nonrotating shaft subjected to unidirectional bending, the origins are located in the region of the maximum-tension zone (Fig. 2). In a nonrotating shaft subjected to reversed bending, the origins are diametrically opposite each other (Fig. 4). In rotary bending, however, every point on the shaft periphery is subjected to a tensile stress at each revolution; therefore, a crack may be initiated at any point on the periphery (Fig. 5b and d).

The crack surfaces are pressed together during the compressive component of the stress cycle, and mutual rubbing occurs. A common result of final fracture is that slight movement of one side of the crack relative to the other side frequently causes severe damage to the fracture surfaces and tends to obliterate many marks. However, although the high spots on one surface may rub the high spots on the other, the marks in the depressions are retained. Because the depressions are negative images of the damaged high spots on the opposing surface, they provide useful evidence; therefore, it is desirable to examine both parts of a cracked or fractured shaft.

The similarity in macroscopic appearance of fractures in shafts resulting from rotating-bending fatigue and from single overload torsional shear of a relatively ductile metal frequently results in misinterpretation. The fracture surface shown in Fig. 6(a) was the result of fatigue, as evidenced by the ratchet marks around the periphery and the pronounced beach marks. Under the low magnification of the fractograph shown in Fig. 6(b), beach marks are not visible, because they were obliterated by rubbing. The presence of ratchet marks around the periphery is also an indication of rotating-bending fatigue. However, the metal smearing apparent on the fracture surface and the twisting deformation of the shaft shown in Fig. 6(c) indicate torsional shear and would preclude mistaking this fracture for a fatigue fracture.

Beyond this stage, one crack usually takes the lead, and such inferences are no longer justified. If the shaft transmits a unidirectional torque but two cracks develop mutually at right angles, it can be presumed that the torque was of a reversing character. If a bending stress is applied to a shaft that is transmitting torque, the angle at which any fatigue crack develops will be modified. Therefore, if the angle differs significantly from 45° to the shaft axis, the presence of a bending stress is indicated.

Contact Fatigue

Contact fatigue occurs when components roll, or roll and slide, against each other under high contact pressure and cyclic loading. Pitting occurs after many repetitions of loading and is the result of metal fatigue from the imposed cyclic contact stresses. Factors that govern contact fatigue are contact stress, relative rolling/sliding, material properties, and metallurgical, physical, and chemical characteristics of the contacting surfaces, including the oil film that lubricates the surfaces.

The significant stress in rolling-contact fatigue is the maximum alternating shear stress that undergoes a reversal in direction during rolling. In pure rolling, as in antifriction bearings, this stress occurs slightly below the surface and can lead to the initiation of subsurface fatigue cracks. As these cracks propagate under the repeated loads, they reach the surface and produce cavities, or pits.

When sliding is superimposed on rolling, as in gear teeth, the tangential forces and thermal gradient caused by friction alter the magnitude and distribution of stresses in and below the contact area. The alternating shear stress is increased in magnitude and is moved nearer to the surface by friction resulting from the sliding action. Additional information is provided in the articles “Fatigue Failures” and “Failures of Rolling-Element Bearings” in this Volume.

Failures of Shafts

Revised by Donald J. Wulpi, Metallurgical Consultant

Wear

Wear of metal parts is commonly classified into either of two categories: abrasive wear or adhesive wear.

Abrasive wear, the undesired removal of material by a cutting mechanism, can reduce the size and destroy the proper shape of a shaft. The shaft may then fail by another means, such as by fracture, or may cease to perform its designed function. Foreign particles, such as sand, dirt and other debris, in the lubricant can cause wear of a shaft.

Example 3: Wear Failure of a Fuel-Pump Drive Shaft Caused by the Presence of Sand, Metallic Particles, and Vibration. The fuel pump in a turbine-powered aircraft failed, resulting in damage to the aircraft. The pump is shown in Fig. 8(a) and (b).

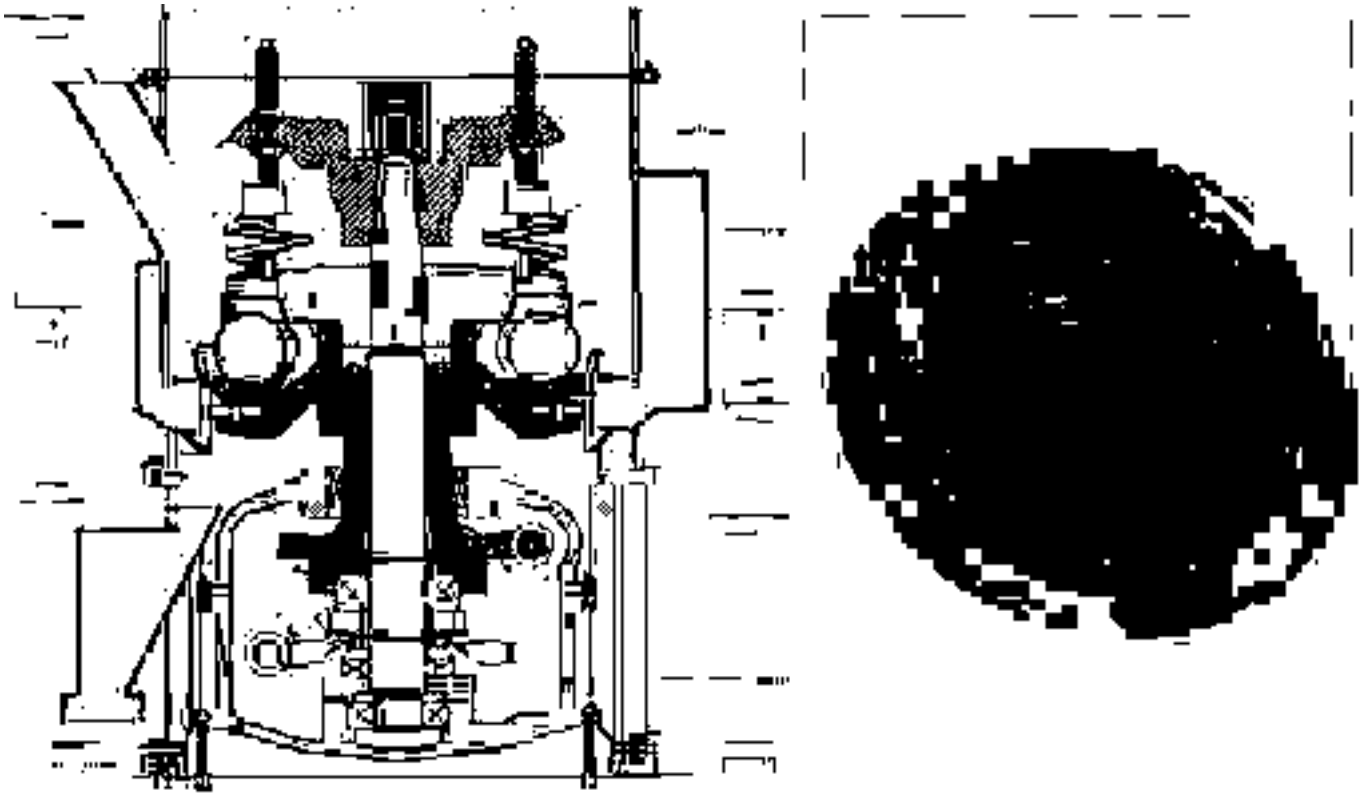


Fig. 11 6150 steel coal pulverizer shaft that failed by fatigue. (Left) Section through pulverizer showing the inner main shaft that fractured, repaired itself by friction welding, and fractured a second time. (Right) Photograph of the friction welded surface

The lower grinding ring of the pulverizer was attached to the outer main shaft (Fig. 11, left side). The upper grinding ring was suspended by springs from a spider that was attached to the main shaft. The weight of the upper grinding ring and the load imposed by the springs resulted in a total force of 8165 kg (18,000 lb) on the pulverizer balls. The shaft was made of 6150 steel normalized to a hardness of about 285 HB.

Investigation. Visual examination of the shaft revealed a circumferential crack in the main shaft just below the upper radial bearing at an abrupt change in shaft diameter. The shaft was set up in a lathe to machine out the crack for repair welding, and the smaller end (17.8 cm, or 7 in., in diameter) was found to be slightly eccentric with the remainder of the shaft. The crack did not disappear after a 6.3-mm (0.25-in.) deep cut, and the crack was opened by striking the small end of the shaft. The shaft broke about 1.3 cm (0.5 in.) from the crack. Examination of the fracture surface (Fig. 11, right side) revealed a previous fracture, almost perpendicular to the axis of the shaft, that resulted from torsional loading acting along a plane of maximum shear.

Although the shaft operated at the relatively low speed of 82 rpm, the weight of the upper portion of the shaft, the spider, and the upper ring caused the shaft to repair itself by friction welding in a band about 2.5 cm (1 in.) wide beginning at the periphery of the shaft. Thus, friction welding was confined to a small part of the 17.8-cm (7-in.) diam cross section. The center region of the fracture surface of the shaft (Fig. 11, right side) contained swirls of metal softened by frictional heat as a result of rotation of the lower part of the shaft. Welding of the plasticized outer ring of metal on the shaft and solidification of the swirls probably occurred while the pulverizer was shut down.

Examination of the machined surface revealed faint lines parallel to the visible crack. These faint lines generally followed bands containing large grains. The lines were thought to be fatigue cracks. The hardness of the shaft in the friction welded area was 38 to 49 HRC, and inward from the weld area, the hardness was 31 to 33 HRC.

Conclusion. The second fracture of the shaft was by fatigue, which resulted from an eccentric condition after the shaft was friction welded and from the inherent vibrations within the machine. Shafts in other coal pulverizers at the same plant fractured in the same region after similar service lives.

Corrective Measures. The shaft was repaired by welding a new 20.3-cm (8-in.) diam section to the older, longer section and machining to the required diameters. The fatigue cracks in the longer section were machined away before repair welding. Liquid-penetrant inspection was performed to ensure soundness of the metal in the older section. The repaired shaft was machined to provide a taper between the two diameters, rather than reproduce the sharp change in diameter in the original design. The taper section was carefully blended with the smaller diameter so that there were no sharp corners. The repaired coal-pulverizer shaft operated satisfactorily for more than five years. As shafts in other units failed, they were repaired by the same technique.

inches). (b) Fracture surface. (c) Hot trim marks. (d) Snag grinding marks. (e) Hot folds. (f) Section through a hot fold. Etched with 2% nital. 100×

The crankshafts were failing by transverse fracture of one cheek after one to three years of service. The expected life was 20 years of continuous service. One complete forging that had fractured (No. 1 in this Example) and a section containing the fractured cheek on the shorter shaft of another forging (No. 2) were sent to the laboratory so that the cause of failure could be determined.

Investigation. Visual examination of the fracture surfaces of both crankshafts revealed indications of fatigue failure; however, the origins were not readily visible (Fig. 17b). The surfaces had a clean fine-grain structure, but the edges were peened—evidently the result of damage after fracture.

The surfaces of the cheeks at the parting line contained rough grooves from hot trimming of the flash and from snag grinding (Fig. 17c and d). Forging 2 contained the most severe of such markings.

Longitudinal and transverse sections were prepared and etched with hydrochloric acid at a temperature of 71 to 77 °C (160 to 170 °F). The steel was of good quality and contained the normal amount of nonmetallic inclusions but no segregation or pipe.

Examination of an as-polished specimen revealed no intergranular oxidation. Etching the specimen with a 10% sulfuric acid and 10% nitric acid solution revealed no evidence of burning or overheating of the steel.

An area containing shallow surface folds was found on the outer face of one cheek of the throw on forging 1 (Fig. 17e). As shown in Fig. 17(f), the metal around one of the folds contained some ferrite, and the forged surface was slightly decarburized. Also, a fatigue crack had initiated in the fold and was propagating across the cheek. Examination of a section through the fracture surface disclosed cold working of the surface, which could have been the result of a rather extended period of fatigue cracking.

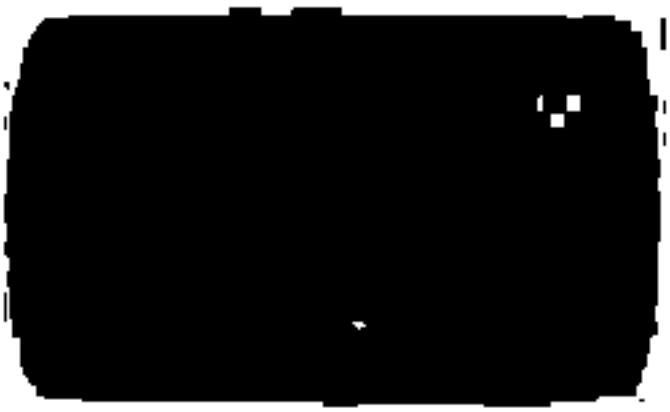
Chemical analysis of the two forgings found the metal to be 4140 steel, as specified. The hardness of the two forgings at the subsurface, midradius, and core ranged from 19 to 22 HRC. The specified hardness of the machined forging was unknown.

Tensile strength of the forgings ranged from 790 to 817 MPa (114.5 to 118.5 ksi), and yield strength at 0.2% offset was 580 to 620 MPa (84 to 90 ksi). Elongation in 3.8 cm (1.5 in.) was 20.6 to 22%, and reduction of area was 56.2 to 59.4%. These properties were representative of 4140 steel quenched and tempered to a hardness of 20 to 22 HRC. The general microstructure of the forgings was tempered bainite; the grain size was ASTM 6 to 8.

Conclusions. Fatigue cracking resulted in the transverse fracture of one cheek in each of the two crankshafts submitted for examination. In crankshaft 1, fatigue cracks were initiated at a shallow hot-work defect. A rough surface resulting from hot trimming or snag grinding of the forging flash was the point of initiation of fatigue cracks in crankshaft 2.

Corrective Measures. Before being machined, the forgings were normalized, hardened and tempered to 28 to 32 HRC to increase fatigue strength. The quenching procedure was changed to produce a more complete martensite transformation and to increase the ratio of yield strength to tensile strength. The surfaces were inspected by the magnetic-particle method, and shallow folds, notches, or extremely rough surfaces were removed by careful grinding.

Identification Marks. Excessive stresses may be introduced in shafts by stamped identification marks that indicate manufacturing date or lot number, steel heat number, size, or part number. The location of such a mark and the method by which it is made can be important. Identification marks should not be placed in areas of high bending or torsional stresses. For shafts, this often requires that they be located either on the end face or on an adjoining collar, but surface-finish requirements on the end of the shaft cannot be ignored if thrust loads are taken at that location. Stamping of marks with straight-line portions is the most likely to cause cracks, although characters with rounded contours also can cause cracking (Fig. 18). Stamping of metal shaft surfaces should be avoided because it is impossible to predict on which surface stamp marks will cause cracking in service (more information is provided in the article “Fatigue Failures” in this Volume.)



imperfections, for there is no neutral axis; any imperfection can be a stress concentrator and can be the origin of fatigue cracking if the stress is high with respect to the strength.

Example 12: Fatigue Cracking of a Forged 4337 Steel Master Connecting Rod Because of Nonmetallic Inclusions. Routine inspection of a reciprocating aircraft engine revealed cracks in the master connecting rod. Cracks were observed in the channel-shaped section consisting of the knuckle-pin flanges and the bearing-bore wall. The rods were forged from 4337 (AMS 6412) steel and heat treated to a specified hardness of 36 to 40 HRC.

Investigation. Visual examination revealed H-shaped cracks in the wall between the knuckle-pin flanges (Fig. 25a). The cracks originated as circumferential cracks, then propagated transversely into the bearing-bore wall. Magnetic-particle and x-ray inspection before sectioning did not detect any inclusions in the master rod.

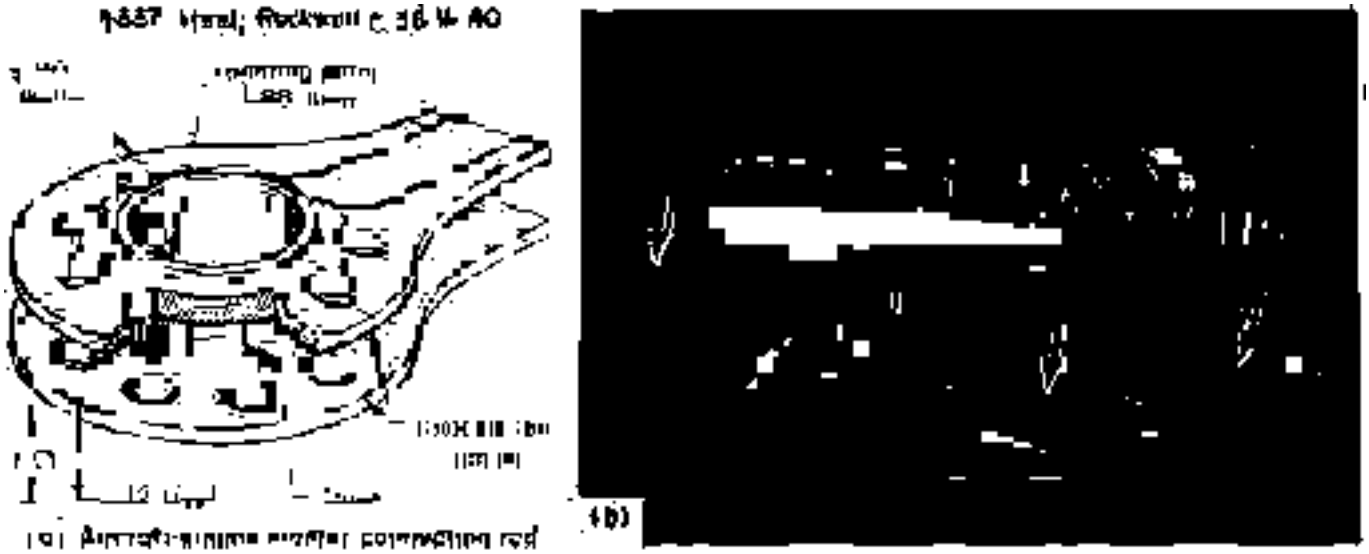


Fig. 25 Forged 4337 steel master connecting rod for a reciprocating aircraft engine that failed by fatigue cracking in the bore section between the flanges. (a) Configuration and dimensions (given in inches). (b) Fractograph showing inclusions (arrows) and fatigue beach marks

Macroscopic examination of one of the fracture surfaces revealed three large inclusions lying approximately parallel to the grain direction and fatigue beach marks around two of the inclusions. The inclusions and beach marks are shown in Fig. 25(b).

Microscopic examination of a section through the fracture origin showed large nonmetallic inclusions that consisted of heavy concentrations of aluminum oxide (Al_2O_3). These inclusions were of the type generally associated with ingot segregation patterns.

The hardness of the rods, 36 to 40 HRC, and the microstructure of the heat-treated alloy steel were satisfactory for the application.

A preliminary stress analysis indicated that the stresses in the area of cracking, under normal operating conditions, were relatively low compared with other areas of the rod, such as in the shank and the knuckle-pin straps.

Conclusions. The rod failed in fatigue in the bore wall between the knuckle-pin flanges. Fatigue was initiated by the stress-raising effect of large nonmetallic inclusions. The nonmetallic inclusions were not detected by routine magnetic-particle or x-ray inspection because of their orientation.

Recommendations. The forging vendors were notified that nonmetallic inclusions of a size in excess of that expected in aircraft-quality steel were found in the master connecting rods. Forging techniques that provided increased working of the material between the knuckle-pin flanges to break up the large nonmetallic inclusions were not successful. A nondestructive-testing procedure for detection of large nonmetallic inclusions was established.

Example 13: Fatigue Cracking of a 1040 Steel Crankshaft Because of Excessive Segregation of Nonmetallic Inclusions. The crankshaft in a reciprocating engine had been in operation for less than one year when the engine was shut down for repairs. Examination of the engine components disclosed that the crankshaft had suffered fairly severe cracking. The crankshaft was sent to the laboratory for a complete examination.

Investigation. The journals of the main and crankpin bearings were inspected by the magnetic-particle method. At least four of the main-bearing journals had three to six indications of discontinuities 1.5 to 9.5 mm (0.0625 to 0.375 in.) long. Another main-bearing journal had approximately 20 similar but generally shorter indications. One main-bearing journal had a crack along the fillet and almost entirely through the web (Fig. 26). Another main-bearing journal had a 1.3-cm

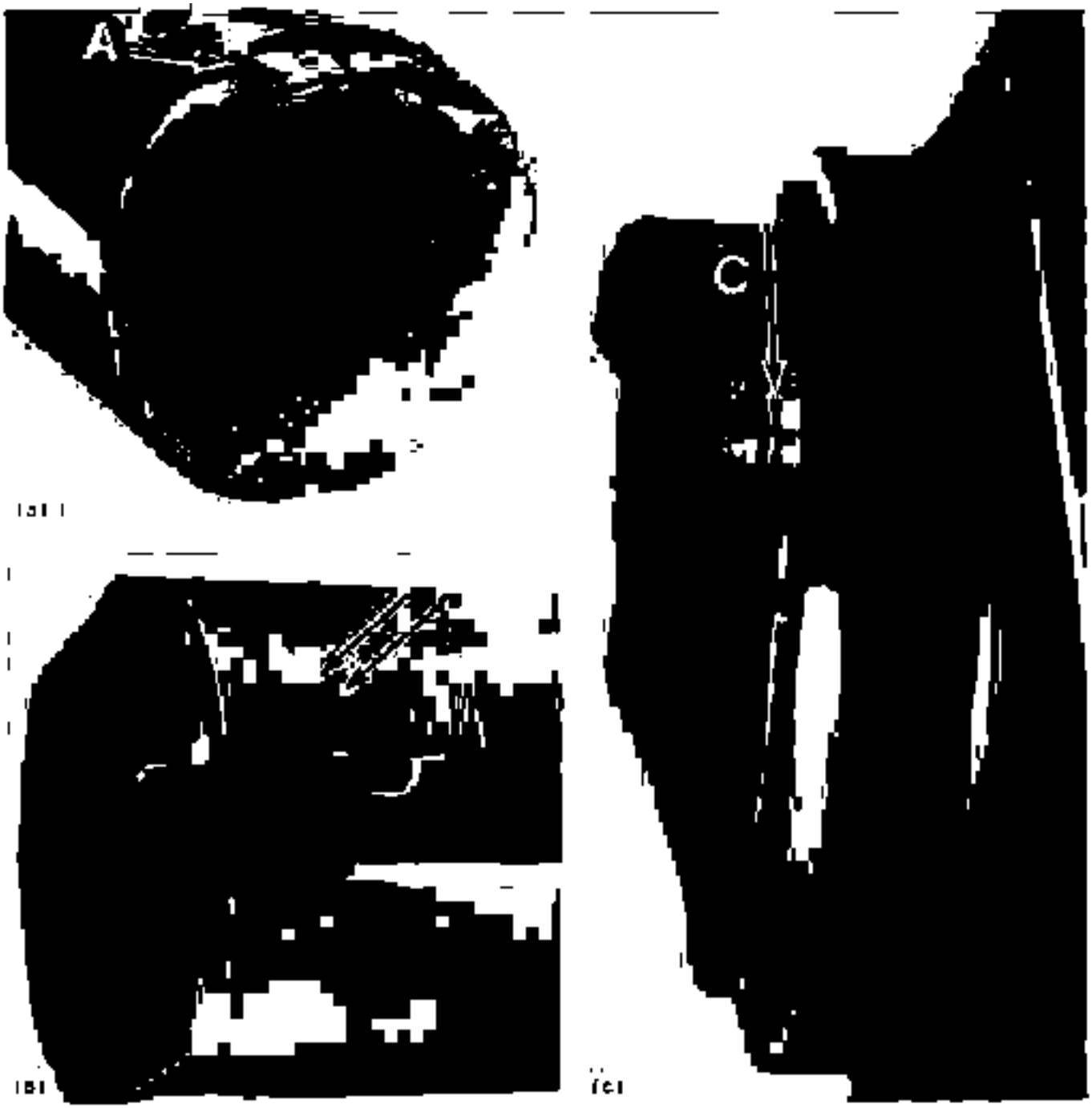


Fig. 28 Crankpin of a crankshaft that fractured in fatigue after the crankpin journals and main-bearing journals were rebuilt by metal spraying. (a) Fracture surface showing fracture origin (arrow A). (b) Machined grooves and punch marks (arrow B) on surface of crankpin. (c) Groove in crankpin, where cracking was initiated

It was found that adhesion of the sprayed-metal coating to the fractured crankpin was very poor and that the coating could be detached easily by tapping a chisel inserted between the coating and the pin. Portions of the sprayed coating were removed from all the pins and journals so that the surface preparation carried out before metal spraying could be examined. In each of two pins, a helical groove approximately 0.8 mm (0.03 in.) wide and deep, with a pitch of 10 per inch, had been cut in the surface (arrow B, Fig. 28b). The grooving extended over one-half of the pin periphery only; the remaining portion of the surface had been indiscriminately indented with a tooth punch (Fig. 28b).

From one pin, a large portion of the sprayed metal, extending the whole width of the crank-pin and covering one-third of its circumference, came away in one piece. Examination indicated that the tongues of metal that filled the grooves had broken away from the main layer at the level of the original pin surface. The average thickness of the sprayed layer was 0.8 mm (0.03 in.). The hardness of the crankpin was only 163 HB, indicating that it was readily machinable.

making machinery) or lubricated with available moisture (some water pumps) or a specifically applied lubricant. It is usually the case that the more severe the duty of bearings the more precisely they will be made and the more likely there will be a support (lubrication) system to ensure long life and to prevent catastrophic failure.

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

---

Bearing Materials

The lowest cost bearings are bushings of molded polymers (plastics) in which steel shafts rotate. The lowest cost polymers are the nylons and acetals. At low sliding speeds, they usually wear by abrasion due to dirt or the roughness of the surface of the shaft.

Polymers have two disadvantages—their low thermal conductivity and their tendency to cold flow and become distorted if loaded very heavily. If polymers are used at high speeds, the surfaces may melt, possibly even with some of the molten polymer being squeezed out. There are three solutions to the thermal problems. One is to operate the bearing with a low contact pressure, p, and at a low speed, V. A recommended pV limit is 0.2 MPa · m/s (Ref 1). Another is to lay a thin film of polymer on a metal backing to increase the heat-flow rate. A third is to fill the polymer with a material that has high thermal conductivity. The third solution to the thermal problem also offers resistance to cold flow. For these purposes, the polymer can be filled and reinforced with particles or fibers of glass, metal, and minerals. However, these fillers usually cause more wear of the shaft than the polymer does.

For low friction, the more expensive polytetrafluoroethylene (PTFE) is available. Polytetrafluoroethylene is sometimes added to acetals to reduce friction. Graphite and molybdenum disulfide (MoS_2) are also added; although they decrease friction, these solid lubricants usually do not prevent wear.

Another low-cost bushing material is porous bronze, which can be impregnated with lubricant. For a little higher cost, cast iron bearings can be used, particularly for large and flat surfaces.

Polymers, gray cast iron (with large amounts of graphite), and bronzes (see the section “Bronze Bearing Materials” in this article) are selected as bearing materials because of their ability to survive even though they may be made to poor tolerances or may be operated with occasional lubricant starvation or without maintenance. Poor treatment of these bearings will often appear as adhesive grooving and galling. Where these materials are run against a harder counter-surface, there may be metal transfer to that surface.

Almost all materials, such as carbon, ceramics, white cast iron, and metal alloys of all kinds, can be used as bearings. These materials may be chosen primarily for corrosion resistance, for prevention of sparking, for resistance to high volumes of abrasives, or simply because they are available. High-performance bearings are usually made with considerable care and are comparatively more expensive.

High-Performance Bearings. The soft metals, in various combinations, are used in high-performance bearings, and they are virtually always lubricated. The soft metals are sometimes used in thicknesses up to 6 mm (0.24 in.), but to increase fatigue resistance, the soft metals are coated or plated on a steel (or other stronger metal) backing in layers as thin as 25 μm.

The major groups of soft metal bearing materials are:

- *Tin-base or lead-base babbitts or white metals:* These materials are good for embedding hard contaminant particles and for resistance to galling, but fatigue quickly. Lead alloys also corrode readily
- *Copper-lead alloys:* These are superior to the white metals for corrosion resistance in many applications
- *Aluminum alloys that contain tin and lead:* The alloying elements do not mix, thus forming small solid droplets in the aluminum that smear over the aluminum surface at times of inadequate lubrication. The aluminum alloys are superior in corrosion resistance

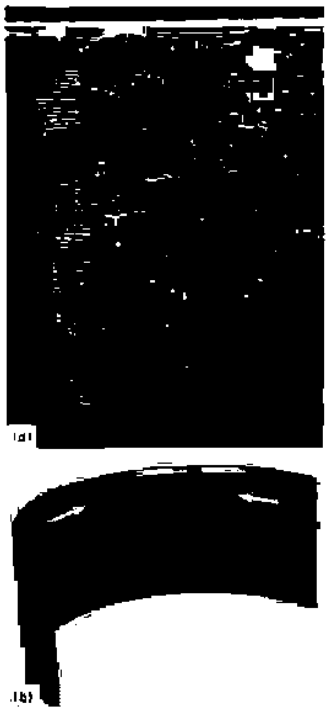


Fig. 4 Sliding bearing surfaces that failed by fatigue. (a) Wormhole appearance in a lead-base babbitt bearing. (b) Flaking (between arrows) of overlay in a trimetal bearing

The fatigue limit of engine bearings is a major topic. The performance of such bearings in service is very difficult to quantify, but there are practical guidelines for selecting material types based on applied stress. Table 1 lists such information.

Table 1 Peak stresses for normal fatigue life of various bearing materials used in automotive applications[(a)]

Bearing metal	Peak stress (max)	
	MPa	psi
Solid-metal bearings		
Aluminum alloy SAE 770-T101 4.1 mm (0.162 in.) thick	34.5	5000
Bimetal bearings		
Tin babbitt 0.6 mm (0.022 in.) thick on steel	8.3	1200
Lead babbitt 0.6 mm (0.022 in.) thick on steel	9.7	1400
Tin babbitt 0.1 mm (0.004 in.) thick on steel	15.2	2200
Lead babbitt 0.1 mm (0.004 in.) thick on steel	16.5	2400
Copper-lead 0.6 mm (0.022 in.) thick on steel	34.5	5000
Aluminum alloy SAE 780 0.8 mm (0.03 in.) thick on steel	41.4	6000
Trimetal bearings		
Lead-tin 0.025 mm (0.001 in.) thick on copper-lead 0.6 mm (0.022 in.) thick on steel	51.7	7500
Lead-tin 0.025 mm (0.001 in.) thick on silver 0.3 mm (0.013 in.) thick on steel	51.7	7500
Lead-tin 0,025 mm (0.001 in.) thick on aluminum alloy SAE 780 0.8 mm (0.03 in.) thick on steel	55.2	8000
Lead-tin-copper 0.4 mm (0.015 in.) thick on aluminum alloy SAE 781 0.4 mm (0.015 in.) thick on steel	69	10,000

(a) Fatigue strength of bearing materials was determined under laboratory conditions of excellent alignment, excellent oil flow, a high degree of cleanness, ideal clearance, and normal operating temperature of 80 to 105 °C (180 to 220 °F). Normal fatigue life of bearings is considered to be 161,000 km (100,000 miles) in gasoline automobile engines and 402,000 km (250,000 miles) in diesel truck engines.

Electrical Erosion. Some bearings, particularly those in steam turbines, may have an electrical potential applied across them, usually inadvertently. When the gap spacing in the bearing is small and the applied potential is high enough, sparks will pass through the oil film and cause erosion. The bearing surfaces will not be reflective, but will appear frosted (Fig. 5).

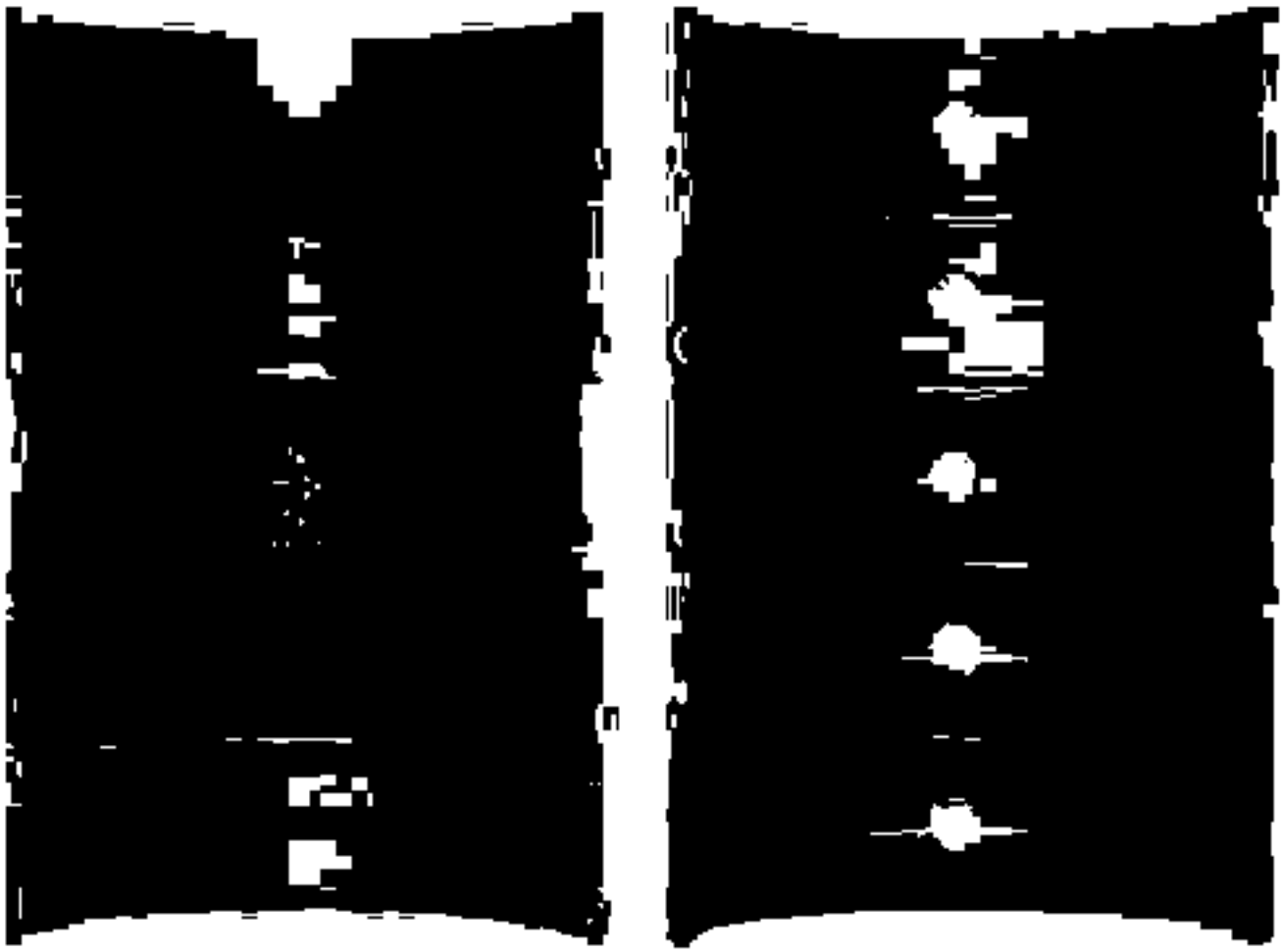


Fig. 13 Lower (left) and upper (right) halves of trimetal bearings for a diesel engine that failed because of distortion of the crankcase. The center three bearing halves sustained greater damage than the end halves.

Example 5: Bending-Fatigue Fracture of an Inertia-Welded Alloy Steel Pushrod Originating at a Sharp Internal Corner and a Decarburized Surface. The pushrod shown in Fig. 12 fractured two weeks after it was installed in a mud pump operating at 160 rpm with a discharge pressure of 14.4 MPa (2.1 ksi). The pushrod was made by joining two pieces of bar stock by inertia welding. Each piece was rough bored or drilled to produce a wall thickness of about 1.3 cm (0.5 in.) at the surfaces to be inertia welded. The flange portion was made of 94B17 steel, and the shaft was made of 8620 steel. After welding, the rod was machined, carburized, and hardened to 60 HRC; the shaft was then chromium plated to within 1.3 cm (0.5 in.) of the flange.

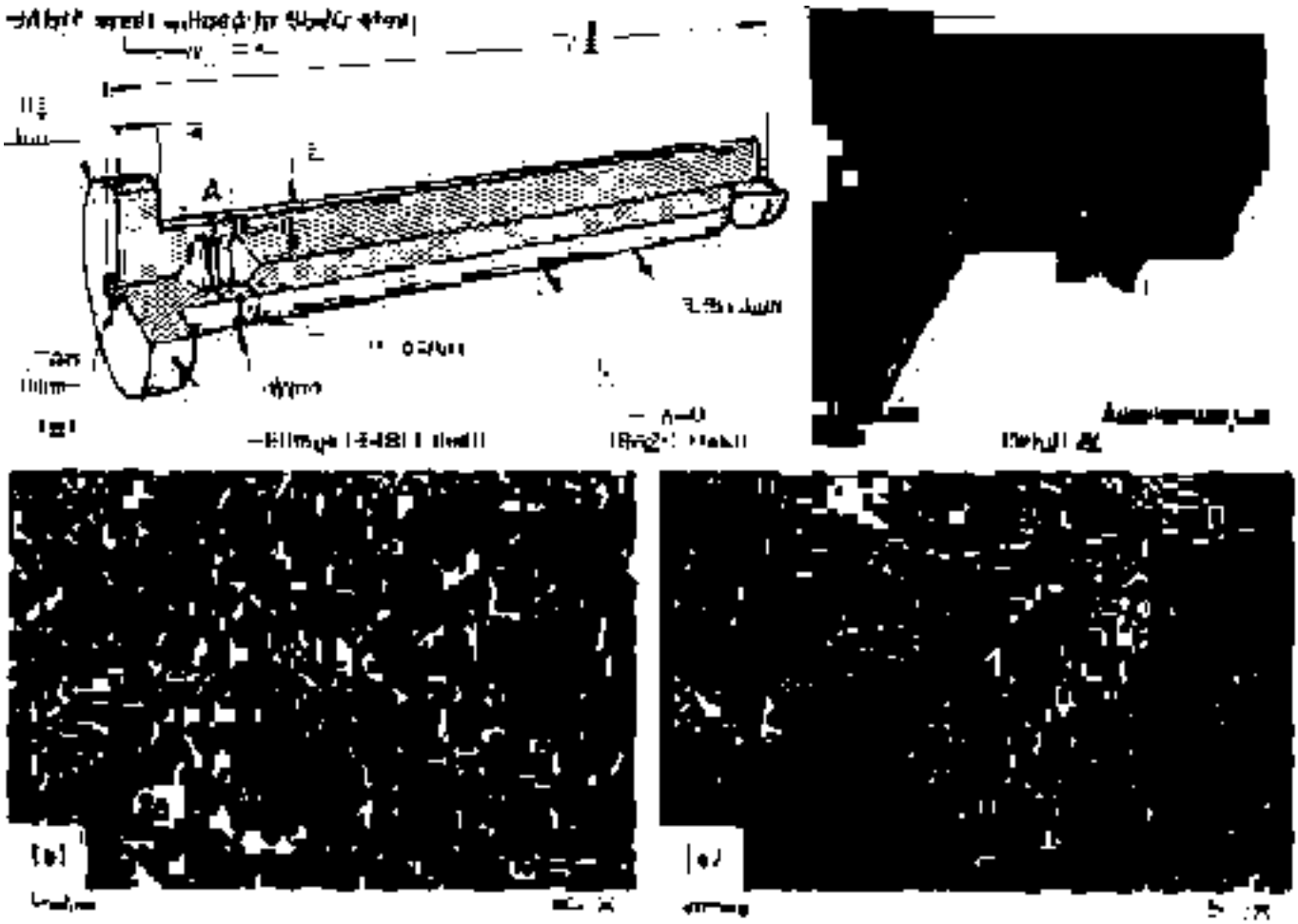


Fig. 12 Pushrod that fractured in bending fatigue after being fabricated by inertia welding. (a) Configuration and dimensions (given in inches). (b) and (c) Micrographs showing structure of decarburized inner surface and sound metal below the decarburized layer

Investigation. Visual examination of the pushrod disclosed that fracture had occurred in the shaft portion at the intersection of the 1.3-cm (0.5-in.) thick wall and the tapered surface at the bottom of the hole. This intersection was about 1.3 cm (0.5 in.) from the weld.

The fracture surface contained indications that a fatigue crack had initiated at the inner surface and progressed around the entire inner circumference, forming a narrow plateau, before propagating outward to final fracture. Operation of the pushrod after it had fractured obliterated some detail of the fatigue pattern on the fracture surface, but it still was evident that one-way bending stresses had influenced cracking. The amount of coarse texture on the fracture surface indicated that fatigue cracking had occurred for a short time.

Detail A in Fig. 12 shows a macrograph of a section through the weld and fracture region. A heavily decarburized layer was found on the inner surface of the flange portion, but was not detectable in the shaft portion. The inner surface had a poor finish, and there was a sharp corner at the intersection of the sidewall and bottom of the hole. The weld flash protruded from the inner wall, producing a small-radius fillet at the wall surface. Except for the metal extrusion, the weld zone was almost undetectable.

Fig. 5 Electrical wear on the surface of a thrust bearing from a steam turbine. The microscopic pitting is caused by sparking from electric current. Actual size

Fretting. When oscillatory sliding occurs with an amplitude of the order of 0.05 μm, fretting occurs. This may happen on the outside of the bearing backing, against the housing of the machine. The bearings may not have fit tightly, or the housing of the machine may have been too flexible, or the bearing may have been assembled with some dirt in it. A fretted surface is shown in Fig. 6.

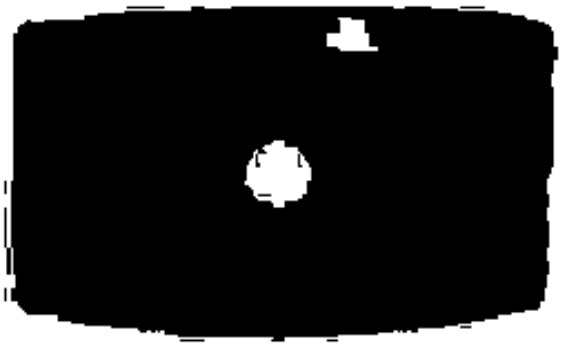


Fig. 6 Fretting on the outside surface of steel backing of a sliding bearing.

Corrosion. Several products of degradation of oil (alcohols, aldehydes, ketones, and acids), contaminants (such as water), and some antiwear additives in oil cause corrosion. A corrosive environment may selectively attack an alloy, removing only one element. In some environments, lead and zinc may be more readily dissolved than copper or aluminum, for example, leaving a surface such as that shown in Fig. 7. The remaining copper matrix is unsupported and can readily fatigue and fracture away.

Fig. 11 Denting damage to cone, cup, and rollers from debris entering the bearing from spalling in another component, through defective seals, or from improperly cleaned housings.

The first form of indentation results in rolling-contact fatigue, and the second causes bulk damage (damage outside the contact zone). The third form of indentation is a type of cold plastic flow. The well-defined indentations spaced at a uniform pitch generally result from balls or rollers having been inadvertently forced into the raceways and often arise from carelessness or the implementation of incorrect procedures in withdrawing or installing rings. In the most common situation, the force required to assemble the bearing is applied indirectly through the balls or rollers instead of directly to the ring itself. Careless handling of the machine in transit can also damage bearings, particularly where impact loading has been sustained. Similar, regularly spaced marks may be produced on the raceways of a bearing that has been forcibly assembled with the axis of one ring at a slight angle to the other. These effects (to which the term brinelling is sometimes applied) usually manifest themselves as noisy operation.

Brinelling, sometimes referred to as true brinelling or denting, is permanent deformation produced by excessive pressure or impact loading of a stationary bearing. Hollows or dents are produced by plastic flow of metal. Such indentations in raceways will be spaced in correspondence to ball or roller spacing, thus differentiating true brinelling from dents made by debris, such as dirt or chips, which usually are spaced randomly. The final grinding marks made during the manufacture of raceways can be seen, virtually undisturbed, within the true brinelled areas of Fig. 12. Improper mounting procedures, such as forcing of a tight-fitting or cocked outer ring into a housing, are common causes of true brinelling.

Generally, the more flexible ropes, which contain the greatest number of wires and have fiber cores, will stretch more than all-metal ropes with fewer wires and therefore less flexibility.

Stretch resulting from the movement of the wires is of two types: constructional and elastic. When a rope is first placed under load, the resulting slight rearrangement of the wires will cause a permanent lengthening, known as constructional stretch, and a recoverable lengthening, which is elastic stretch.

The constructional stretch, depending to some degree on the magnitude of the load imposed, amounts to approximately 0.5 to 0.75% of the length of the rope for six-strand ropes with fiber cores. For six-strand ropes with steel cores, the corresponding value is 0.25 to 0.5%; for eight-strand ropes with fiber cores, 0.75 to 1%.

The preceding remarks apply to bright ropes of a type intended to operate over sheaves and drums, and it is assumed that the ends of the rope are restricted against rotation, are free from corrosion, and have not deteriorated appreciably either internally or externally. Corrosion and other kinds of rope deterioration have a marked effect on stretch properties.

Sheaves. The size of a sheave is usually expressed in terms of its diameter measured at the base of the groove. This is its tread diameter. It can also be expressed as a ratio determined by dividing the tread diameter by the nominal rope diameter.

As sheave size is decreased, the stresses resulting from bending and the contact pressure between rope and sheave are increased. Higher bending stresses cause more rapid fatigue of the wires in the rope. Increased pressure also accelerates rope deterioration and, at the same time, increases sheave wear.

As sheave size is increased, rope-to-sheave pressure is decreased, and bending becomes easier. If bending alone were involved, an increase in rope life could obtained by increasing sheave size up to a limit of about 90 to 100 times the rope diameter for a 6 × 19 (6 strands of 19 wires each) classification rope. However, except on some shaft hoist installations, sheaves this large are seldom in actual use for two reasons. First, bending is seldom, if ever, the only factor involved. Large, expensive sheaves would be unnecessary if the abrasion or scrubbing encountered by the rope on some other part of the installation were the determining factor in its ultimate service life. Second, most machines would not be practical with such large sheaves, and many would never have been designed if sheaves of this size were mandatory.

In practice, many factors other than bending influence rope life, such as repeated stressing, abrasion, pounding, impact, vibration, twisting, speed, drum-winding abuse, corrosion, and lack of maintenance. For many applications, one or more of these factors affects rope life more than sheave size alone.

The optimum sheave size is determined by evaluating the factors affecting safe, economical operation. On high-speed mobile-type equipment and on equipment requiring low initial cost, it is customary to use smaller sheaves. The need for more frequent rope replacements is outweighed by the speed and flexibility of movement provided by the more compact equipment.

Although no definite minimum sheave size can be established for all types of installations, one important factor must not be overlooked. The heavy pressures between the wires at the contact points in the rope, combined with the high bending stresses resulting from operation over small sheaves, have a definite effect on rope deterioration. Under such adverse conditions, wire breakage often occurs at and between the points where the strands contact each other and contact the core of the rope. Broken wires in these sections are difficult, and sometimes impossible, to detect.

Therefore, for applications in which a high degree of safety is essential, sheaves should be liberal in size to provide better assurance that the wires will deteriorate progressively on the surface of the rope where they can be readily seen and evaluated. Extremely small sheaves should be limited to those types of equipment on which such usage is in line with acceptable practice.

Sheave size is such an important factor in wire-rope life that many laboratory bending or fatigue tests have been conducted to evaluate this relationship. Laboratory bending life can be determined; however, field service life can seldom be judged by these results. There are too many factors existing under actual field conditions to be simulated under controlled laboratory tests.

Laboratory tests have their value for research and development, particularly to the wire-rope engineer. It may also be important at times to show how bending alone, as performed in such tests, affects wire rope.

One series of tests was made on 16-mm ($\frac{5}{8}$-in.) diam wire-rope specimens in the 6 × 7, 6 × 19, 6 × 37, and 8 × 19 classifications. The sheaves ranged from 20 to 50 cm (8 to 20 in.) in diameter. The load approximated a design factor of five, and specimens were operated until they broke. The number of cycles of operation was the measure of bending life.

The laboratory tests showed that as the ratio of sheave diameter to rope diameter was reduced, there was a reduction in bending life, with the reduction in bending life showing the same trend for all specimens tested, and that the reduction in bending life was faster than the reduction of the sheave-to-rope diameter ratio.

Sheave Grooves. To allow wire rope to perform the maximum amount of useful work, it is essential that sheave grooves be of sufficient diameter to provide and maintain proper rope clearance. All wire rope, when new, is slightly oversize to allow for some pulling down while becoming adjusted to its normal working tensions.

Table 1 lists the recommended groove clearances for the usual sheave applications. In some cases, departures from these recommendations are advisable. For example, with some deflectors or equalizers, smaller clearances can be used, or with conditions involving large fleet angles, larger clearances may be necessary.

Fig. 21 Aluminum alloy lifting sling that fractured because of improper welding of the cross member to the sling body. (a) Configuration and dimensions (given in inches). (b) Fracture surfaces of the sling body (top) and the cross member (bottom). (c) Enlarged view of a fracture surface of the sling body showing weld defects

Babbitts are of two types: tin-base with 4 to 8% Cu and 4 to 8% Sb (SAE 11 and 12; ASTM B23, grades 1, 2, 3, and 11) and lead-base with a maximum of 10.7% Sn and 9.5 to 17.5% Sb (SAE 13–16; ASTM B23, grades 7, 8, 13, and 15). These babbitts generally cause the least damage to steel shafts when operated with inadequate lubrication or with contaminants. Tin-base babbitt is more corrosion resistant than lead-base babbitt and is easier to bond to a steel backing. Babbitts are inherently weak and strongly affected by high temperatures. For this reason, thin layers of babbitt are often applied to hacking materials, such as bronze, steel, or cast iron.

Copper-lead bearing alloys usually contain between 20 and 30% Pb, up to 3% Sn, and may contain up to 5.5% Ag and less than 0.1% Zn. Typical alloys are SAE 48, 49, 480, and 481 (CDA alloys C98200, C98400, C98600, and C98800, respectively). These alloys have all been used as main bearings and connecting-rod bearings of engines. However, in recent years, corrosion has been found to be a problem with these alloys. They are, therefore, usually covered with a thin layer of lead-tin, lead-tin-copper, or lead-indium alloy to reduce the corrosion of the substrate. In some instances, a nickel barrier about 1 μm thick may be placed between the substrate and the overlay to prevent diffusion of tin from the overlay to the copper in the substrate.

Bronze bearing materials are either leaded bronze, tin bronze, or aluminum bronze. Strength and high-temperature resistance increase in the order lead to tin to aluminum, but embeddability and scuff resistance decrease in that order.

The leaded bronzes contain from 5 to 25% Pb and up to 10% Sn. The high-lead alloy is soft and has a temperature limit of about 230 to 260 °C (445 to 500 °F). The addition of tin increases the fatigue resistance and hardness. Zinc in place of tin improves ductility but decreases strength and hardness. Some wrought leaded bearing bronzes are SAE alloys 791 to 794, 797, and 799. The alloy SAE 795 has especially good fatigue resistance. Some castable leaded bearing bronzes are SAE 40 (CDA C83600), SAE 660 (CDA C93200), SAE 66 (CDA C93500), SAE 64 (CDA C93700), CDA C94100, and CDA C94300.

Tin bronzes contain more than 20% Sn, with less than 10% Pb added to aid machinability. Small amounts of zinc, nickel, or both are sometimes added to increase strength.

Aluminum bronzes have iron, nickel, silicon, and manganese as the principal alloying elements. The hardness of these alloys may range to 240 HB. They have excellent shock resistance and will operate at comparatively high temperatures. They tend to scuff more readily than the leaded bronzes. Typical alloys are CDA C86200, C86300, C95200, C95300, C95400, and C95500.

Aluminum alloys other than bronzes contain about 7% Sn and 1% Cu, with 1 or 2% Si or Mg, plus some lead and cadmium. The copper is retained in solid solution in the aluminum, but the other alloys form soft-solid particles in the matrix. These alloys are usually relatively weak and require strong backing. They do not require an overlay for protection from corrosion; however, an overlay is often provided for start-up protection. Typical alloys are 850.0-T5, A850.0-T5, and B850.0-T5 (formerly 750-T5, A750-T5, and B750-T5).

Reference cited in this section

1. G.C. Pratt, Materials for Plain Bearings, *Int. Met. Rev.*, Vol 18 (No. 17), 1973

Failures of Sliding Bearings

Kenneth C. Ludema, Department of Mechanical Engineering, University of Michigan

Fluid-Film Lubrication

Fluid films can be provided in a bearing by surface-tension retention of a fluid in a gap, by pumping fluid into a contact region (hydrostatic lubrication), or by hydrodynamic action.

Surface Tension. If a drop of liquid is placed on a flat surface and then another flat surface is laid upon the wetted surface, some, but not all, of the liquid will be squeezed out. Surface tension, the same force that makes liquid rise in a very small diameter glass tube, makes complete exclusion of the liquid impossible. The amount that will be retained in the gap between two surfaces is related to the wettability of the liquid (lubricant) on the surface of interest. If a drop of the lubricant spreads out completely and spontaneously on the surface, that lubricant will also run out of the bearing

The striations below the defect indicate progressive crack development and are rarely seen in steel as hard as valve-spring wire. At the base of this area, several faceted areas developed, pointing away from the surface. A tensile-mode fracture plane then formed around all these fronts. This fills the bottom area of the fractograph.

Example 7: Spring Failure Due to a Surface Defect. Figure 10 shows light streaks arranged in a diagonal direction on the wire surface. The streaks are parallel to the wire axis. A darker depressed area is visible between the streaks and below the center of the fractograph. The light areas were raised and rubbed smooth by the other half of the spring. This occurred because the engine ran after failure of the spring. Careful examination of the depressed areas revealed distinct outlines that represent sharp corners in the depressions. The origin at one of the sharp corners is shown immediately above the center of the fractograph. The depressed area associated with this was in the other half of the spring that sustained more extensive postfailure damage. Apparently, a hard material (possibly mill scale) was impressed during drawing of the wire, but was broken out during peening, leaving the depressions with sharp-bottomed corners.

Fig. 10 Spring failure originating at a sharp-edged pitted area. Arrows indicate the location of the sharp-edged areas. 28×

Example 8: Spring Failures Originating at a Seam. The failure shown in Fig. 11(a) began at the seam that extended more than 0.05 mm (0.002 in.) below the wire surface. The fatigue-fracture front progressed downward from several origins. Each one of these fronts produces a crack that is triangular in outline and is without fine detail due to sliding of the opposing surfaces during the later stages of fracture. This occurs when the fracture plane changes to an angle with the wire axis in response to the torsional strain. These surfaces are seen in the lower part of Fig. 11(a).

Fig. 11 Failures in wire springs. (a) Longitudinal failure originating at a seam. 45×. (b) Origin of failure at a very shallow seam. The arrow indicates the base of the seam. 115×

temperatures. Figure 30 shows a graph of suggested minimum metal temperatures for avoiding external corrosion in economizers and air heaters; the temperatures are plotted as functions of the type of fuel and the amount of sulfur in the fuel.

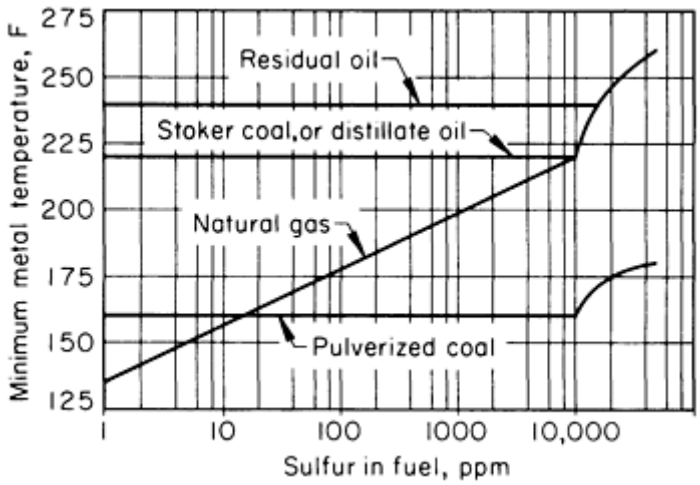


Fig. 30 Suggested minimum metal temperatures to avoid flue-gas corrosion in economizers and air heaters plotted as a function of as-fired sulfur contents of various fuels. Source: Ref 1

Low-alloy steels, particularly those containing copper, have been used successfully in economizers and air heaters that are prone to low-temperature corrosion. An additional advantage of these alloys is their good resistance to pitting caused by excess oxygen in boiler water.

The deposits that form as a result of low-temperature corrosion may contain corrosion products, fly ash, and the products of chemical reaction of the condensed acid with fly-ash constituents. Many of the compounds in these deposits are water soluble and can therefore be removed conveniently by washing the affected areas with water. However, the deposits sometimes become difficult to remove, especially when they are allowed to accumulate until they completely plug passages or when they contain insoluble compounds, such as calcium sulfate.

Low-temperature corrosion is encountered more often in oil-fired units than in coal-fired units, because the vanadium in oil-ash deposits is a catalyst for oxidation of SO_2 to SO_3 and because oil firing produces less ash than coal firing. In coal-fired units, a substantial portion of the sulfur oxides is absorbed by fly-ash deposits in the high-temperature regions, where the sulfur oxides participate in coal-ash corrosion. Furthermore, coal ash, which is composed chiefly of basic compounds, partly neutralizes the condensed acids when it deposits on moist surfaces in low-temperature regions. Oil ash, which is mainly acidic, is incapable of neutralizing condensed acids.

Control of Corrosion. Fire-side corrosion can rarely be eliminated altogether. Obviously, if alkali and sulfur could be removed from coal, or vanadium and sulfur from oil, prior to combustion, fire-side corrosion would largely disappear. However, although there are some processes for removing contaminants from these fuels, they are expensive and partly ineffective. Reduction of fire-side corrosion can be accomplished by one or more of the following methods, which are listed in approximately the descending order of their effectiveness:

- Fuel selection
- Combustion control
- Boiler design and construction
- Periodic ash removal
- Use of fuel additives

Frequently, several of these measures are required rather than just one.

Fuel Selection. Use of ashless or sulfur-free fuel is perhaps the most direct means of minimizing fire-side corrosion. Most natural and manufactured gaseous fuels are ashless, whereas most oils, coals, cokes, vegetation, residues, and waste gases have some ash content. Any type of fuel may contain sulfur; few coals, oils, tars, or burnable wastes are sulfur free, whereas gases are more likely to be sulfur free or to have low sulfur contents. Some fuels, such as the black liquor burned in kraft-mill recovery furnaces, have very substantial sulfur contents.

Some boilers are fitted with burners that can use more than one type of fuel; others are fitted with auxiliary burners to permit firing of secondary fuels (generally, main and auxiliary burners are not fired together). Most boilers are built to burn only one type of fuel; thus, selection of low-ash or low-sulfur fuel involves selection of a specific type of boiler as

deoxygenated water treated with caustic to a high pH, deoxygenated water treated with about 200 ppm of an oxygen scavenger, or treated deoxygenated water plus a cover of nitrogen gas called a nitrogen cap. Continuous recirculation of demineralized deoxygenated water is often used for protection against corrosion during lay-up of large central-station boilers.

To a certain extent, the volatile amines used to control corrosion of surfaces contacted by steam or condensate will also help to protect these surfaces during idle periods, but use of dry inert gas is preferable for this purpose.

Scaling

In most systems, the pH of the feedwater is maintained between 8.0 and 11.0 The optimum pH level, which is established by experience for each installation, is the level that keeps the amounts of iron and copper corrosion products in the boiler to a minimum. When water treatment is inadequate or control of pH is lax, there is a potential for excessive water-side corrosion in economizer and boiling-surface tubes. If thick corrosion scale forms, overheating failures can result.

However, the major source of water-side scale in boilers is not corrosion but dissolved solids in the feedwater. Compounds containing iron, calcium, magnesium, and sodium cations normally exist in most raw water; the usual anions are bicarbonate, carbonate, sulfate, and chloride radicals. Calcium and magnesium compounds (water hardness compounds) become less soluble in water as temperature is increased. These hardness compounds separate from solution at temperatures common in economizers and boilers and usually precipitate as tube-wall deposits, or hardness scale. As is true for corrosion scale, hardness scale is undesirable in regions of high heat transfer because it can cause rapid overheating of tube material. In addition, hardness scale may be somewhat porous. Dissolved solids in the boiler water, as well as strong alkalis sometimes used to treat feedwater, can be concentrated in the pores, which may lead to severe pitting corrosion of the tube surface beneath the hardness scale.

Scaling in Superheaters. Deposition of solids as scale in superheater tubing occurs primarily as the result of carryover of water droplets in wet steam. When carryover occurs, water droplets vaporize in the superheater, and any nonvolatile solids present in these droplets are deposited on heat-transfer surfaces. Deposits may not be as likely to cause rapid-overheating failures in superheater tubing as in boiling-surface tubing because of the lower heat flux in superheaters. However, because superheaters operate at higher metal temperatures than boilers, scaling may lead to stress-rupture failure or to loss of ductility because of such metallurgical changes as σ-phase formation or spheroidization of carbides.

Deposition in Turbines. Silica is the substance most commonly deposited in turbines, although other feedwater solids may also be found. At the higher boiling temperatures associated with high-pressure subcritical systems and with supercritical systems, a portion of the silica in the feedwater will volatilize and will be carried into the turbine along with the steam. As heat is extracted from the working fluid, steam temperature drops; volatile silica becomes nonvolatile and deposits out of the vapor stream onto turbine blades and surfaces of interstage passages. These deposits interfere with fluid flow in the turbine, resulting in a loss of power.

Silica, which is present in most raw water, is not removed by the same treatments used for control of water hardness; consequently, when silica deposition is excessive, special treatments must be used.

Removal of Water-Side Scale. Periodic removal of scale from internal surfaces in boilers and economizers is part of the normal maintenance of this equipment. Removal of scale can minimize unscheduled outages caused by tube rupture due to overheating. In addition, internal cleanness is an important factor in maintaining high thermal efficiency and low operating cost.

Although mechanical methods of scale removal can be used, chemical cleaning with a suitable inhibited acid offers the advantages of less downtime; lower cost; ability to clean otherwise inaccessible areas, such as sharp bends, irregular surfaces, crevices, and small-diameter tubes; and ability to clean internal surfaces without dismantling the unit. The particular acid and inhibitor that are appropriate for a specific situation depend on the composition and adherence of the deposits, the type of cleaning method (circulation or soaking), and the type of materials being cleaned—especially in applications that involve alloys with high chromium contents.

Chemical cleaning should be supervised by a competent chemist to ensure maximum effectiveness and safety and to ensure that cleaning solutions do not excessively corrode base metal under deposits. In chemical cleaning of nuclear steam generators, additional precautions must be taken in disposing of the solutions and in ensuring that personnel are not exposed to radioactive wastes from the cleaning process.

Chemical cleaning should always be followed by thorough flushing to remove loosened deposits and residual acid, after which the boiler should be filled with a neutralizing and passivating solution to prevent the clean metal surfaces from rusting. Chemical cleaning is not a substitute for adequate water treatment. Corrosion and scale deposition can lead to severe deterioration of tube materials, and the chemical-cleaning process itself removes a thin layer of base metal along with the deposits.

Fire-Side Corrosion

Except for most gaseous fuels, combustion of fossil fuels produces solid, liquid, and gaseous compounds that can be corrosive to structural components and heat-transfer surfaces. In addition, deposits of solid and liquid residues in gas passages can alter the heat-transfer characteristics of the system, with potentially severe effects on system efficiency and tube-wall temperatures.

Residues from the combustion process, called ash, normally constitute 6 to 20% of bituminous coals, but may run as high as 30%. The composition of coal ash varies widely, but is composed chiefly of silicon, aluminum, iron, and calcium compounds, with smaller amounts of magnesium, titanium, sodium, and potassium compounds.

Wood, bagasse (crushed juiceless remains of sugar cane), and other vegetable wastes used as fuel in some industrial plants contain lower amounts of ash than coal does. In many respects, however, the compositions of these vegetable ashes resemble that of coal ash.

Fuel oils have ash contents that seldom exceed 0.2%. Even so, corrosion and fouling of oil-fired boilers can be particularly troublesome because of the nature of oil-ash deposits. The main contaminants in fuel oil are vanadium, sodium, and sulfur—elements that form a wide variety of compounds, many of which are extremely corrosive.

Regardless of the fuel, there are similarities in all rapid fire-side corrosion circumstances. A liquid phase forms in the ash deposit adjacent to the tube surface. Once a liquid phase forms, the protective oxide scale on the tube surface is dissolved, and rapid wastage follows. In coal-fired boilers, the liquid phase is a mixture of sodium and potassium iron trisulfate—($Na_3Fe(SO_4)_3$ and $K_3Fe(SO_4)_3$, respectively. Mixtures of these melt at as low as 555 °C (1030 °F). In oil-fueled boilers the liquid phase that forms is a mixture of vanadium pentoxide (V_2O_5), with either sodium oxide (Na_2O) or sodium sulfate (Na_2SO_4). Mixtures of these compounds have melting points below 540 °C (1000 °F).

Coal-Ash Corrosion. During combustion of coal, the minerals in the burning coal are exposed to high temperatures and to the strongly reducing effects of generated gases, such as carbon monoxide and hydrogen. Aluminum, iron, potassium, sodium, and sulfur compounds are partly decomposed, releasing volatile alkali compounds and sulfur oxides (predominantly SO_2, plus small amounts of SO_3). The remaining portion of the mineral matter reacts to form glassy particles known as fly ash.

Coal-ash corrosion starts with the deposition of fly ash on surfaces that operate predominantly at temperatures from 540 to 705 °C (1000 to 1300 °F), primarily surfaces of superheater and reheater tubes. These deposits may be loose and powdery, or may be sintered or slag-type masses that are more adherent. Over an extended period of time, volatile alkali and sulfur compounds condense on the fly ash and react with it to form complex alkali sulfates, such as $KAl(SO_4)_2$ and $Na_3Fe(SO_4)_3$, at the boundary between the metal and the deposit. The reactions that produce alkali sulfates are believed to depend in part on the catalytic oxidation of SO_2 to SO_3 in the outer layers of the fly-ash deposit. The exact chemical reactions between the tube metal and the ash deposit including the alkali iron trisulfates are not well defined. However, in every example studied, the ash and corrosion products contain both carbon and sulfides. The sulfides cannot exist in a strongly oxidizing environment. Thus, at least on the tube surface, the conditions are reducing. The following coal-ash corrosion characteristics are common:

- Rapid attack occurs at temperatures between the melting temperature of the sulfate mixture and the limit of thermal stability for the mixture
- Corrosion rate is a nonlinear function of metal temperature, being highest at temperatures from 675 to 730 °C (1250 to 1350 °F)
- Corrosion is almost always associated with sintered or glassy slag-type deposits
- The deposit consists of three distinct layers (Fig. 27). The porous, outermost layer comprises the bulk of the deposit and is composed essentially of the same compounds as those found in fly ash. The innermost layer (heavy black lines, Fig. 27) is a thin, glassy substance composed primarily of corrosion products of iron. The middle layer, called the white layer, is whitish or yellowish in color, is often fused, and is largely water-soluble, producing an acid solution. The fused layers contain sulfides as part of the corrosion product and can be easily detected by means of a sulfur print
- Coal-ash corrosion can occur with any bituminous coal, but is more likely when the coal contains more than 3.5% sulfur and 0.25% chlorine
- None of the common tube materials is immune to attack, although the 18-8 austenitic stainless steels corrode at slower rates than lower-alloy grades. Although they vary widely depending on environment and operating conditions, corrosion rates of 2.25Cr-1Mo steel are in the range of 0.4 to 1 mm (15 to 40 mils, or 0.015 to 0.040 in.) per year, while the 18-8 stainless steels similar to 304 corrode in the range of 0.08 to 0.1 mm (3 to 5 mils, or 0.003 to 0.005 in.) per year
- The overall appearance of the wasted tube surface is rough, with what is termed an alligator hide appearance

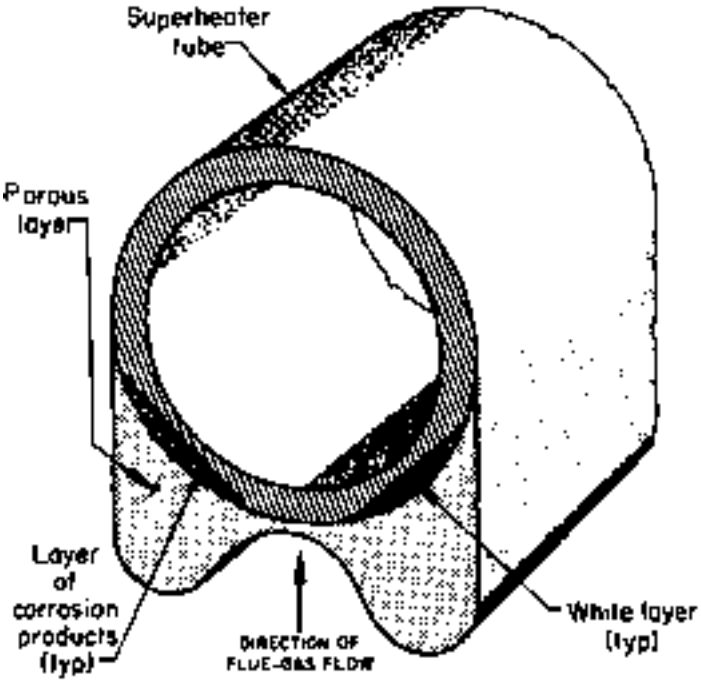


Fig. 27 Schematic illustration of the three layers formed during coal-ash corrosion of a superheater tube.

Because the layer next to the tube contains sulfides, the likely net corrosion reaction is:

$$2C + SO_2 + Fe = FeS + 2CO$$

It also seems probable that coal-ash corrosion of superheaters and reheaters can be reduced by better, more complete combustion of the coal. Because the corrosion occurs in the presence of carbon, removal of the carbon would seem to reduce the wastage.

Particles of fly-ash deposit on superheater and reheater tubes in a characteristic pattern (Fig. 27) in relation to the direction of flue-gas flow. The tube surfaces are corroded most heavily beneath the thickest portions of the deposit. When deposits are removed, shallow macropitting can be seen. Eventually, the tube wall becomes thinned to the point at which the material can no longer withstand the pressure within the tube, and the tube ruptures.

Example 7: Coal-Ash Corrosion of a Chromium-Molybdenum Steel Superheater Tube. The top tube of a horizontal superheater bank in the reheat furnace of a steam generator ruptured as shown in Fig. 28 after about $7\frac{1}{2}$ years of service (normal service life would have been 30 years). The tube consisted of 64-mm ($2\frac{1}{2}$-in.) diam × 5.6-mm (0.220-in.) wall thickness ferrite steel tubing welded to 64-mm ($2\frac{1}{2}$-in.) diam × 3.4-mm (0.134-in.) wall thickness austenitic stainless steel tubing. The rupture occurred in the ferritic steel tubing near the welded joint.

Fig. 28 2.25Cr-1Mo steel superheater tube that ruptured because of thinning by coal-ash corrosion.

The ferritic steel tubing, used in the low operating temperature portion of the tube, was made of 2.25Cr-1Mo steel (ASME SA-213, grade T-22); the austenitic steel tubing was made of type 321 stainless steel (ASME SA-213, grade TP321H).
The surface temperature of the tube was 620 to 695 °C (1150 to 1280 °F), which was higher than the operating temperature in use up to about 9 months earlier. At that time, the exit-steam temperature had been raised from about 520 °C (970 °F) to about 540 °C (1000 °F).

Investigation. The ferritic steel portion of the tube had split longitudinally for about 38 mm ($1\frac{1}{2}$ in.) along the top surface.

Heavy corrosion had occurred in the region of the rupture (portion at left of rupture, Fig. 28).
The top surface of the tube in the rupture area was free of deposits, but the sides and bottom were covered with a heavy accumulation of red and white deposits that had a barklike appearance (Fig. 28). The nearby stainless steel portion was free of deposits, but did show some pitting corrosion.
Spectrographic analysis showed that the ferritic tube steel contained 2.10% Cr and 1.10% Mo, both of which were within specified limits.
The white deposit was identified as sodium-potassium sulfate [$(Na,K)SO_4$], containing about 3% alkali acid sulfate [$(Na,K)HSO_4$]. The red deposit was found to be mainly Fe_2O_3 plus about 5% SiO_2, 2% Al_2O_3, and 0.5% Cr_2O_3. Attack of the steel by the alkali acid sulfate, which has a low melting point, thinned the tube wall and produced the red deposit. This type of corrosion is known to be severe on chromium-molybdenum steel at temperatures of about 595 to 705 °C (1100 to 1300 °F). Where the tube surface had been exposed to intermittent blasts of air from soot blowers, the corrosion products had been carried away. Although stainless steel is also vulnerable to this type of attack, the reaction is slower, and only a slight amount of tube-wall thinning was detected on the stainless steel portion of this tube.
The operating records showed two probable causes of the corrosion: the increase in operating temperature and the use of coal with an ash-fusion temperature 28 °C (50 °F) lower than normal during the week before the failure. Use of this coal could have caused severe slagging conditions in the boiler.
Sectioning of the tube through the region of failure showed that the attack by the alkali acid sulfate had significantly reduced tube-wall thickness. Metallographic examination revealed a structure that was largely normal except for marked spheroidization of the carbides, which indicated appreciable service time at temperatures close to the Ac_1 temperature.
Conclusions. Rupture of the tube was attributed to thinning of the wall by coal-ash corrosion to the point at which the steam pressure in the tube constituted an overload of the remaining wall. The attack was accelerated by the increase in operating temperature and may have been influenced by temporary use of an inferior grade of coal.
Corrective Measures. Adjacent tubes, which also showed evidence of wall thinning, were reinforced by pad welding. Type 304 stainless steel shields were welded to the stainless steel portions of the top reheater tubes and were held in place about the chromium-molybdenum steel portions of the tubes by steel bands. These shields were intended as a temporary expedient against coal-ash corrosion, not as a permanent solution. Because the shields would not be cooled by steam flow, their temperatures were expected to be about 705 °C (1300 °F), the maximum at which coal-ash corrosion would occur, and little corrosion was anticipated. The replacement for the ruptured tube was left bare to permit assessment of the degree of corrosion that would ensue after the improvements in coal quality and operating temperature were adopted.
The superheater operated for about 10 years with no further ruptures or additional repairs. At that time, the tubes were inspected, and those that were thinned the most were replaced; this amounted to about 10% of the superheater bank.
Furnace-wall corrosion is also a problem on occasion in coal-fired boilers. The first principal cause of furnace-wall corrosion is the reducing conditions. Just above the stoker or around wall burners, low-oxygen conditions may exist. Radiation from the flame can lead to high skin temperatures. Reduction of the protective oxide film on the steel tube exposes the metal and leads to quite rapid metal loss. Corrosion rates up to 6.4 mm (0.250 in., or 250 mils) per year are reported. Worst-case examples are steam leaks after 2000 h of operation in a stoker-fired unit. Corrosion morphology is smooth, fairly uniform wastage.
The second principal cause is low-melting ash constituents. Rapid wastage by components of the coal ash has also been frequently observed. Differential-scanning calorimeter measurements have measured ash melting points between 335 and 410 °C (635 and 770 °F), well within the operating temperatures of water walls. Chordal thermocouples have measured temperature spikes of 30 to 55 °C (50 to 100 °F) during soot-blower operation and during slag falls. Such rapid temperature excursions coupled with liquid-ash constituents lead to a thermal fatigue component in these failures. Circumferential grooves spaced along the length of the tube are characteristic of this corrosion attack (Fig. 29).

Fig. 29 Thermal fatigue plus liquid-ash corrosion on water walls leads to circumferential grooving. The cross section in an axial plane nearly parallel to the tube axis shows the deep fingerlike penetrations into the wall. Etched with nital. 210×. Courtesy of Riley Stoker Corp.

Oil-Ash Corrosion. During combustion of fuel oils, organic compounds (including those containing vanadium or sulfur) decompose and react with oxygen. The resulting volatile oxides are carried along in the flue gases. Sodium, which is usually present in the oil as a chloride, reacts with the sulfur oxides to form sulfates. Initially, vanadium pentoxide (V_2O_5) condenses as a semifluid slag on furnace walls, boiler tubes, and superheater tubes—in fact, virtually anywhere in the high-temperature region of the boiler. Sodium oxide reacts with the vanadium pentoxide to form complex compounds, especially vanadates ($n Na_2O \cdot V_2O_5$) and vanadylvanadates ($n Na_2O \cdot V_2O_4 \cdot m V_2O_5$). These complex compounds, some of which have melting temperatures as low as 249 °C (480 °F), foul and actively corrode tube surfaces.

Slag of equilibrium thickness (3.0 to 6.4 mm, or 0.12 to 0.25 in.) has developed in experimental furnaces within periods of time as short as 100 h. Slag insulates the tubes, resulting in an increase in the temperature of the slag, which in turn increases the rate of corrosion and promotes further deposition of ash. Thicker slag deposits generally lead to greater corrosion, because slag temperatures are higher and more of the corrodent is present to react with tube materials. However, higher slag temperatures also make the slag more fluid so that it will flow more readily on vertical surfaces. Consequently, slag generally builds up in corners and on horizontal surfaces, such as at the bases of water walls and around tube supports in the superheater.

Oil-ash corrosion affects all the common structural alloys. Even highly corrosion-resistant materials, such as 60Cr-40Ni and 50Cr-50Ni cast alloys, which are sometimes used for superheater-tube supports, are not immune. In addition, nonmetallic refractory materials used for furnace linings are attacked by vanadium slag; the mechanism of this attack appears to be a dissolution or stagging type of attack rather than the direct chemical attack that characteristically occurs with metals.

Low-Temperature Corrosion. In low-temperature zones of flue-gas passages, corrosion is caused chiefly by condensed water vapor containing dissolved SO_3 and CO_2. The dew point of sulfuric acid, which is the most active corrodent, ordinarily ranges from 120 to 150 °C (250 to 300 °F) for SO_3 concentrations of 15 to 30 ppm, which are common for coal-fired boilers. The dew point of the acidic vapors depends on the amount of moisture in the fuel and in the combustion air, the quantity of excess air, the amount of hydrogen in the fuel, and the amount of steam used for soot blowing—all of which influence the amount of water vapor in the flue gas.

Condensation of acidic vapors is most prevalent in air heaters, precipitators, stack coolers, flues and stacks, and near the inlets of economizers in units without feedwater heaters. Factors that increase the likelihood of acid condensation include low flue-gas flow, such as occurs during start-up and during periods of low-load operation, excessively low flue-gas exit temperatures during normal operation, too great an amount of excess air of high humidity, and very low atmospheric

temperatures. Figure 30 shows a graph of suggested minimum metal temperatures for avoiding external corrosion in economizers and air heaters; the temperatures are plotted as functions of the type of fuel and the amount of sulfur in the fuel.

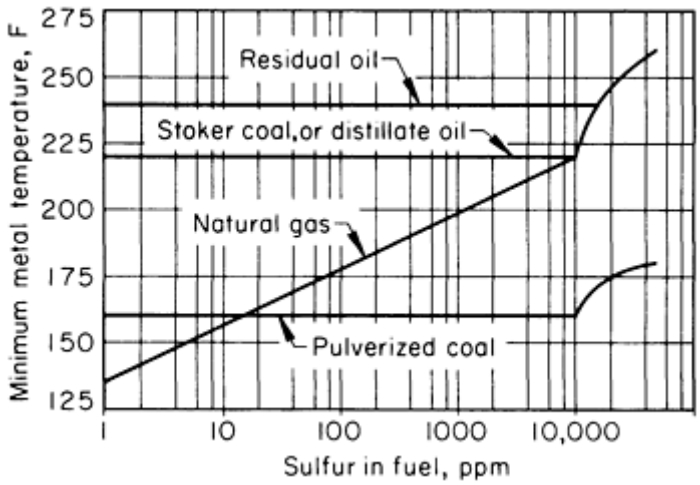


Fig. 30 Suggested minimum metal temperatures to avoid flue-gas corrosion in economizers and air heaters plotted as a function of as-fired sulfur contents of various fuels. Source: Ref 1

Low-alloy steels, particularly those containing copper, have been used successfully in economizers and air heaters that are prone to low-temperature corrosion. An additional advantage of these alloys is their good resistance to pitting caused by excess oxygen in boiler water.

The deposits that form as a result of low-temperature corrosion may contain corrosion products, fly ash, and the products of chemical reaction of the condensed acid with fly-ash constituents. Many of the compounds in these deposits are water soluble and can therefore be removed conveniently by washing the affected areas with water. However, the deposits sometimes become difficult to remove, especially when they are allowed to accumulate until they completely plug passages or when they contain insoluble compounds, such as calcium sulfate.

Low-temperature corrosion is encountered more often in oil-fired units than in coal-fired units, because the vanadium in oil-ash deposits is a catalyst for oxidation of SO_2 to SO_3 and because oil firing produces less ash than coal firing. In coal-fired units, a substantial portion of the sulfur oxides is absorbed by fly-ash deposits in the high-temperature regions, where the sulfur oxides participate in coal-ash corrosion. Furthermore, coal ash, which is composed chiefly of basic compounds, partly neutralizes the condensed acids when it deposits on moist surfaces in low-temperature regions. Oil ash, which is mainly acidic, is incapable of neutralizing condensed acids.

Control of Corrosion. Fire-side corrosion can rarely be eliminated altogether. Obviously, if alkali and sulfur could be removed from coal, or vanadium and sulfur from oil, prior to combustion, fire-side corrosion would largely disappear. However, although there are some processes for removing contaminants from these fuels, they are expensive and partly ineffective. Reduction of fire-side corrosion can be accomplished by one or more of the following methods, which are listed in approximately the descending order of their effectiveness:

- Fuel selection
- Combustion control
- Boiler design and construction
- Periodic ash removal
- Use of fuel additives

Frequently, several of these measures are required rather than just one.

Fuel Selection. Use of ashless or sulfur-free fuel is perhaps the most direct means of minimizing fire-side corrosion. Most natural and manufactured gaseous fuels are ashless, whereas most oils, coals, cokes, vegetation, residues, and waste gases have some ash content. Any type of fuel may contain sulfur; few coals, oils, tars, or burnable wastes are sulfur free, whereas gases are more likely to be sulfur free or to have low sulfur contents. Some fuels, such as the black liquor burned in kraft-mill recovery furnaces, have very substantial sulfur contents.

Some boilers are fitted with burners that can use more than one type of fuel; others are fitted with auxiliary burners to permit firing of secondary fuels (generally, main and auxiliary burners are not fired together). Most boilers are built to burn only one type of fuel; thus, selection of low-ash or low-sulfur fuel involves selection of a specific type of boiler as

well. The boiler operator should always be aware of the composition of the fuel being burned so that operating conditions can be adjusted to minimize ash deposition.

Use of a high-sulfur fuel sometimes cannot be avoided. In such instances, close control of operating conditions is mandatory. For example, in a fire-tube boiler that burned sulfur-contaminated refinery gas, leaks in the carbon steel tubes in the fifth pass caused the boiler to be taken out of service. Sulfuric acid corrosion on the internal surfaces of the tube was determined to have caused the leaks. Apparently, the operator had failed to purge the boiler by burning fuel gas with a low sulfur content for a sufficient length of time prior to several previous shutdowns, and acid had condensed in the tubes as the boiler cooled through the dew point of the acid in the flue gas. Corrosion of this type normally is not a problem, provided that clean fuel gas is burned long enough to expel all sulfur compounds from the tubes before the boiler is allowed to cool down.

Combustion control —that is, adjustment of firing rate, amount of excess air, air temperature, and amount of recirculated flue gas (applicable only to boilers equipped for gas tempering)—can be very effective in controlling the amount and composition of ash deposits. For example, in an experimental boiler burning high-vanadium high-sulfur residual oil, it was found that reducing the amount of excess air from 7% down to 1 to 2% reduced the rate of corrosion of 18-8 stainless steel superheater tubes by as much as 75%. This reduction in corrosion rate apparently occurred because the low amount of excess air did not permit oxidation of vanadium and sulfur to their highest states of oxidation and thus prevented formation of the most corrosive oil-ash compounds. Low-temperature corrosion was virtually eliminated by maintaining excess air at 1 to 2%. This effect of low excess air on low-temperature corrosion has been widely substantiated, but the results are not as conclusive for high-temperature corrosion.

Boiler design and construction, including tube size and spacing, furnace configuration, size and direction of flue-gas passages, and size and location of baffles or shields, can help control ash deposition. The ability to make changes in boiler design specifically for the purpose of corrosion control is somewhat limited by the need to maintain proper gas-temperature distribution in order to facilitate efficient generation of steam.

Periodic ash removal, although less effective in controlling fire-side corrosion than procedures that reduce ash deposition, is nevertheless an important aspect of boiler operation. Ash removal prevents development of conditions that alter the distribution of metal temperatures on heat-transfer surfaces, which can lead to such problems as failure by overheating. Almost all water-tube boilers are provided with soot blowers or air lances. Soot blowers and lances direct jets of air or steam against heat-transfer surfaces to remove fire-side deposits. The effectiveness of soot blowers in performing this task depends on the design and location of the soot blowers; the blowing medium and its temperature; the properties of the deposit, especially its friability and adherence and the frequency and duration of soot-blower operation. It is important to coordinate combustion control with the soot-blowing schedule to ensure that deposits are initially friable, or loosely adherent and powdery, and that they are removed before they are transformed into sintered masses.

Additives are used to control both coal-ash and oil-ash corrosion. Finely powdered magnesia, dolomite, and alumina are injected into furnace gases or blown onto heat-transfer surfaces through soot blowers. Additives are sometimes mixed with fuel oil before combustion. Additives promote the formation of ash deposits that are easily removed. Additives also reduce low-temperature corrosion by reacting with SO_3 to form innocuous sulfates. This reduces the SO_3 content of the flue gas, thus reducing the likelihood that acid will condense in low-temperature regions.

Dissimilar-Metal Welds

Weld failures are more completely covered in the article “Failures Related to Weldments” in this Volume. However, there are topics unique to the long-term elevated-temperature operation of the modern steam generator. One such topic is dissimilar-metal welds (DMW) between austenitic stainless steel and low-alloy ferritic steels, usually SA-213, grade T-22, material.

In the 1950s, when steam temperatures reached 540 °C (1000 °F), the need for stainless steels in the final stages of a superheater or reheater became apparent. Because stainless is required in only a portion of a superheater or reheater, the principal pressure-part tubing remained low-alloy ferritic steels. Problems of premature failure of welds between austenitic stainless steels and ferritic steels have plagued the boiler industry ever since. There are three root causes of these creep failures: carbon migration from the HAZ of the T-22 into the weld metal, expansion differences between the two varieties of steel, and the differences in corrosion resistance to flue gas leading to the formation of an oxide wedge on the outside diameter of the T-22 next to the weld.

Carbon Diffusion. When these DMWs are made with iron-base alloys of stainless steel, carbon will migrate from the T-22 into the weld metal. Particles of chromium-iron carbides will form along the weld interface in the stainless steel. The strength of T-22, especially the creep strength, is a function of carbon content. As the chromium-iron carbides form, the metal immediately adjacent to the carbides is weakened by the loss of carbon. The loss of carbon reduces the creep strength just where the temperature-induced strain is the largest. These two effects will lead to the formation of creep voids next to these carbide particles and ultimately to premature creep failure. To reduce the carbon migration and the formation of carbide particles, it is desirable to use a welding alloy that does not form a carbide as readily as stainless steel does. Such alloys are the nickel-base materials. Nickel does not form a carbide, and while these weldments do

contain chromium, the tendency to form carbides is substantially reduced. Carbon remains in the T-22, the creep properties of the HAZ remain high, the temperature-induced stresses are lower due to the compatibility of thermal coefficients, and so the formation of creep voids and failure is retarded.

Expansion Differences. The coefficients of thermal expansion for ferritic and austenitic steels differ by about 30% (Table 2). Because of these expansion differences, a 50-mm (2-in.) diam tube of stainless steel will be 0.1 mm (0.004 in.) larger in diameter at 540 °C (1000 °F) than a similar tube of T-22. This difference in expansion, the temperature-induced strain, must be accommodated at the weld.

Table 2 Coefficients of thermal expansion for ferritic and austenitic steels

Temperature range		**Ferritic steels**		**Austenitic steels**	
°C	**°F**	**$\times 10^{-6}$ m/m · K**	**$\times 10^{-6}$ in./in. · °F**	**$\times 10^{-6}$ m/m · K**	**$\times 10^{-6}$ in./in. · °F**
-18 to 315	0–600	11.5	6.4	15.5	8.6
-18 to 425	0–800	12.6	7.0	16.6	9.2
-18 to 540	0–1000	13.3	7.4	17.3	9.6
-18 to 650	0–1200	13.8	7.7	17.8	9.9

There are three likely combinations of expansion: (1) the weld metal may have the same coefficient of thermal expansion as the austenitic stainless, (2) in between the expansion of the two, and (3) the same as the ferritic steels. When the weldment has the same expansion as the stainless steel, the temperature strain must be absorbed at the interface between the weld deposit and the T-22. As noted above, the long-term creep strength of the T-22 is adversely affected by the loss of carbon. When the expansion coefficients are split (case 2, above), the temperature strains are reduced but not to zero. Potential premature creep failures may still occur, but at a reduced frequency and over longer service times. However, if the welding alloy has expansion similar to the ferritic steel, there is only limited temperature-induced strain on the T-22. The maximum temperature strain is transferred to the stainless interface. Both the weld and stainless have much stronger creep strengths, suffer no carbon loss, and are much better able to last. No DMW has failed on the stainless steel side of the weld.

In the past, the most popular welding electrode was an E-309, an austenitic stainless steel. The use of ferritic alloy welding rods, for example, E-9018, did not solve the problem. The creep voids and failures would still originate in the ferritic alloy—in this case, the weld metal next to the stainless steel. The current choice is a nickel-base welding electrode. These nickel-base alloys have thermal coefficients very close to ferritic steels of T-11 and T-22. Thus, the temperature-induced strains are nearly zero at the ferritic interface.

Corrosion. At the juncture of the weld and the T-22 on the outside-diameter surface, an oxide wedge forms whose volume is larger than the volume of metal from which it formed. Thus, the wedge acts as a stress raiser and creates its own stress as well. With time, this wedge penetrates the T-22 and may lead to failure. The cause of the formation of this oxide wedge is related to the difference in corrosion resistance between T-22 and the weld metal. The weld metal is more noble than the ferritic alloy. In effect, a galvanic cell is created that leads to rapid oxidation at this juncture. The use of nickel-base welding alloys will not prevent the formation of these external defects.

Stresses, regardless of origin, are additive. The life expectancy of DMWs may be improved by careful attention to stresses higher than the pressure stress. To minimize the system stresses on the DMWs:

- Place the welds close to a support; the support will often carry the load, not the weld
- Make the welds in a vertical run of the tube rather than in the horizontal run to reduce the nonuniform bending stresses
- Locate the DMWs in the penthouse, out of the convection pass. The temperature gradient through the tube adds yet another stress that may be removed entirely
- Fabricate the welds with as few welding defects as possible: no undercut; a smooth radius fillet; without backing rings, suck back, or lack of fusion at the root, and so on
- Use nickel-base welding alloys

All of the above considerations will not guarantee three decades of trouble-free operation, but will, on average, yield $2\frac{1}{2}$ times the life of stainless steel alloy weldments.

Once trouble begins in existing DMWs, the method of repair depends on whether the superheater or reheater is expected to last a few or many more years. If the condition is such that an overall general replacement is scheduled in the next few years, the oxide wedge can be ground out to a depth of about three-quarters of the tube-wall thickness and rewelded in place with a nickel-base alloy. If carefully done, this procedure will last several years and will give satisfactory results until the replacement is made. However, if more than 10 years of service are expected, the defective DMWs should be removed.

Although the previous discussion has focused on the pressure parts, principally tube-to-tube butt welds, no less important are the attachment welds. Spacers, supports, and alignment clips of stainless steel are routinely welded to T-11 and T-22. The preferred welding metals are also the nickel-base alloys, and for exactly the same reasons. However, extra care must be taken with the shape of the fillet; a smooth blending to the tube with no undercut is especially important. Full-penetration welds are usually not necessary or required, but it is better to have a continuous bead all around the attachment wherever possible. If an open gap is left, that is, welded only from the sides and not around the ends, flue gas and elevated temperature will fill the space with an oxide and fly ash. In time, this oxide will wedge open the attachment weld and cause failure. Because the load on a spacer is largest at the end, it is necessary to begin and finish the weld pass toward the center of the weld rather than at an end so that any crater cracks are away from the highest stress zones.

Reference cited in this section

1. D.N. French, *Metallurgical Failures in Fossil Fired Boilers,* John Wiley & Sons, 1983, p 155–165

Failures of Boilers and Related Equipment

Revised by David N. French, David N. French, Inc., Metallurgists

---

Failure by Fatigue

Fatigue is often considered the most common mechanism of service failure of mechanical devices. Basically, fatigue involves nucleation and propagation of brittle-appearing cracks under fluctuating (cyclic) loading at stresses that have peak values less than the tensile strength of the material. Macroscopic and microscopic features of fatigue fractures are discussed in the articles “Fatigue Failures” and “Failure Modes and Effects Analysis” in this Volume.

Steam equipment may fail by fatigue if mechanical service stresses are fluctuating or vibratory or if thermal cycles or thermal gradients impose sufficiently high peak stresses. Vibratory loads in steam equipment are fairly common—for example, as a result of a slight imbalance in rotating equipment, rapidly fluctuating fluid pressures (water hammer), vibration transmitted through settings or mountings, or turbulent fluid flow, especially at high flow rates. All but a few of the various components in a steam plant are subjected to pressure and thermal cycling during start-up and shutdown. Transient thermal gradients of a lesser degree may occur during operation, especially during changes in power output.

In contrast to most mechanical stresses, thermal stresses are often difficult to avoid and, if severe, can lead to fracture in relatively few cycles. Cyclic strain is more important than cyclic stress in thermal fatigue, because nonuniform or differential thermal expansion is the usual source of the cyclic loading. Frequently, extended periods of high-temperature operation ensue between thermal cycles, and thermally activated processes, such as stress relaxation, stress oxidation, and second-phase precipitation, may alter material properties between successive cycles, thus changing resistance to thermal fatigue and to other failure mechanisms.

Identification of Fatigue Fracture. On a fracture surface, features that indicate fatigue as the fracture mechanism include beach marks, multiple origins, ratchet marks, fatigue striations, and smooth, rubbed regions. Of these five types of features, striations are the most reliable indicators of fatigue.

Example 8: Fatigue Failure of Carbon-Molybdenum Steel Boiler Tubes Caused by Vibration. Tubes in a marine boiler on a new ship failed after brief service lives. Circumferential brittle cracking occurred in the carbon-molybdenum steel tubes near the points where the tubes were attached to the steam drum.

Investigation. Visual, macroscopic, and metallographic examination did not disclose the reason for the failures. There was no evidence of overheating, excessive corrosion, or metallurgical damage caused by improper fabrication. However, when the fracture surfaces were examined by electron microscopy at high magnification, fatigue striations were found.

Conclusion. It was concluded that the fatigue failures were caused by vibrations resulting from normal steam flow at high steam demand. The tubes were supported too rigidly near the steam drum, resulting in concentration of vibratory strain in the regions of failure.

Recommendations. The method of supporting the tubes was changed to reduce the amount of restraint and thus the strain concentration, and no further failures occurred. In similar circumstances, it might be necessary to lengthen or shorten tubes (or pipes) or to reduce operating load in order to eliminate fatigue caused by vibrations resulting from normal fluid flow.

Mechanism of Fatigue Fracture. Fatigue results from reversed plastic strain in metallic crystals. If plastic straining is confined to microscopic or submicroscopic regions in an otherwise elastically stressed component, it is likely that a single crack will occur, originating at the point of maximum local stress and minimum local strength in the entire structure. Microstructural discontinuities, such as inclusions, grain boundaries, and intersections of slip planes with free surfaces,

are the usual microscopic sites of crack initiation. Cracking occurs only after many cycles of loading (high-cycle fatigue), and usually there is no macroscopic evidence of plastic flow. On the other hand, when plastic straining is more extensive, there is a greater likelihood that cracks will initiate at many discontinuities after fewer cycles of loading (low-cycle fatigue) and that there will be macroscopic evidence of plastic flow.

Fatigue-crack initiation most often occurs at structural features that concentrate macroscopic stress, such as notches, fillets, holes, joints, changes in section thickness, and material discontinuities. In addition, service-induced damage, such as dents, gouges, and corrosion pits, frequently raise local stress sufficiently to initiate fatigue cracks.

Example 9: Fatigue Failure of a Carbon Steel Water-Wall Tube Because of an Undercut at a Welded Joint. In a new stationary boiler, furnace water-wall tubes were welded to the top of a dust bin for rigid support (Fig. 31a). The tubes measured 64 mm ($2\frac{1}{2}$ in.) outside diameter × 3.2 mm (0.125 in.) wall thickness and were made of carbon steel to ASME SA-226 specifications. Shortly after start-up, one of the tubes broke at the welded joint.

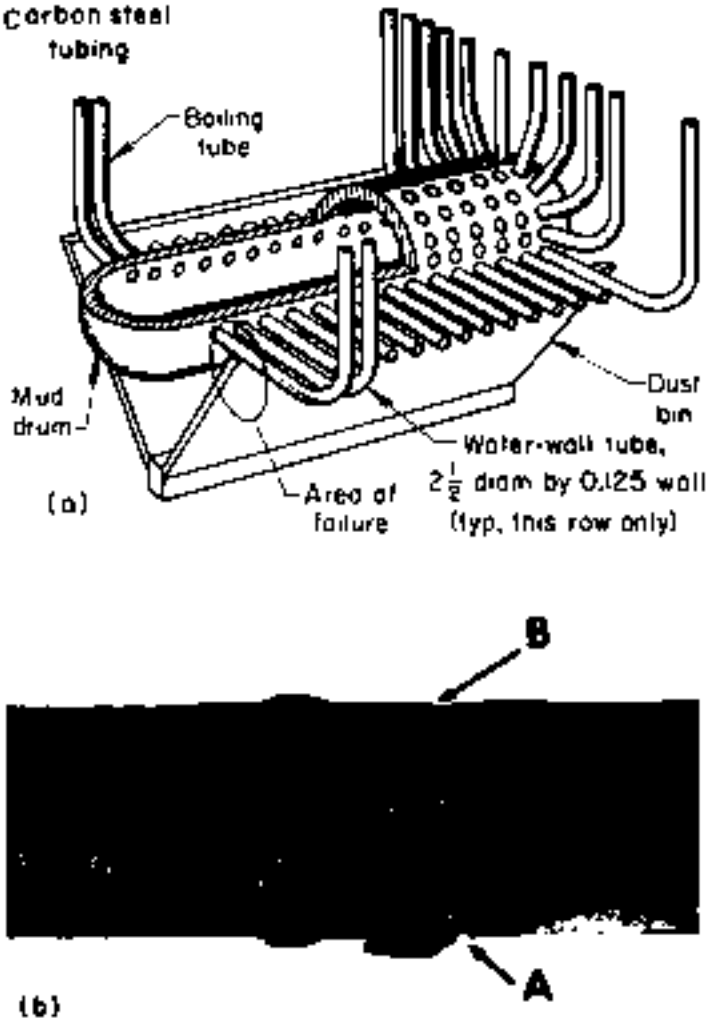


Fig. 31 Stationary boiler in which a carbon steel water-wall tube failed by fatigue fracture at the weld joining the tube to a dust bin. (a) Illustration of a portion of the boiler showing location of failure. Dimensions given in inches. (b) Photograph of fractured tube; fatigue crack is at arrow A, and ductile fracture is at arrow B.

Investigation. The portion of the fracture surface adjacent to the attachment weld around the bottom half of the tube had a flat, brittle appearance. Over the remainder of the tube wall, the fracture was ductile, with a noticeable shear lip and necking. Beach marks were evident on the brittle portion of the fracture surface, indicating that fatigue was the fracture mechanism. The origin was identified as a small crevicelike undercut at the edge of the weld bead. In Fig. 31(b), arrow A indicates the origin; arrow B, local necking opposite the origin.

Conclusion. Cracking began by fatigue from a severe stress raiser. As the fracture progressed out of the region of high restraint where the tube was welded to the dust bin, the fracture changed from fatigue to ductile overload. It was

concluded that the crevicelike undercut was the primary cause of the fracture and that the source of the necessary fluctuating stress was tube vibration inherent in boiler operation.

Recommendations. The remaining water-wall tubes in the row were inspected by the magnetic-particle method; several other tubes were found to have small fatigue cracks adjacent to the welds attaching the tubes to the dust bin. The broken tube was replaced, cracks in other tubes were repaired by welding, and the attachment welds were blended smoothly into the tube surfaces for the entire length of the dust bin. No further failures occurred after the severe undercut was eliminated.

Low-Cycle Fatigue. Pressure and thermal cycling during start-up and shutdown are the most common causes of low-cycle fatigue failures. Design factors that concentrate strain influence low-cycle fatigue as well as high-cycle fatigue; however, the effect is not quite the same. When service stresses are high enough to be conducive to low-cycle fatigue, the presence of a severe stress raiser will often cause complete fracture to occur on the first load cycle. Mild stress raisers usually do not lead to fracture on the first cycle, but only induce localized yielding. The resultant plastic flow may redistribute the stress, but usually not enough to avoid further yielding on successive cycles.

In piping exposed to high temperatures, for example, restraint produces bending moments (and strain) in the piping system when it expands at start-up; strain reversal occurs at shutdown. Elbows and U-bends frequently undergo larger bending moments than straight sections. The mild stress concentration associated with these bends can, under certain conditions, induce plastic strain in the bends, resulting in cracking after a sufficient number of thermal cycles. The thermal expansion of the entire system is important in determining the magnitude of strain concentration and consequently the number of cycles to failure. Lack of flexibility is a frequent cause of low-cycle fatigue in piping systems. Major redesign of the system is sometimes required to remove or reduce restraint, thus allowing the system to accommodate thermal strains by deflecting elastically over a greater portion of the system than was possible in the original design.

Example 10: Low-Cycle Thermal Fatigue Failure of a Type 304 Stainless Steel Tee Fitting. Several failures occurred in 64-mm ($2\frac{1}{2}$-in.) schedule 80 type 304 stainless steel (ASME SA-312, grade TP304) piping in a steam-plant heat-exchanger system near tee fittings at which cool water returning from the heat exchanger was combined with hot water from a bypass (Original design, Fig. 32). During operation, various portions of the piping were subjected to temperatures ranging from 29 to 288 °C (85 to 550 °F).

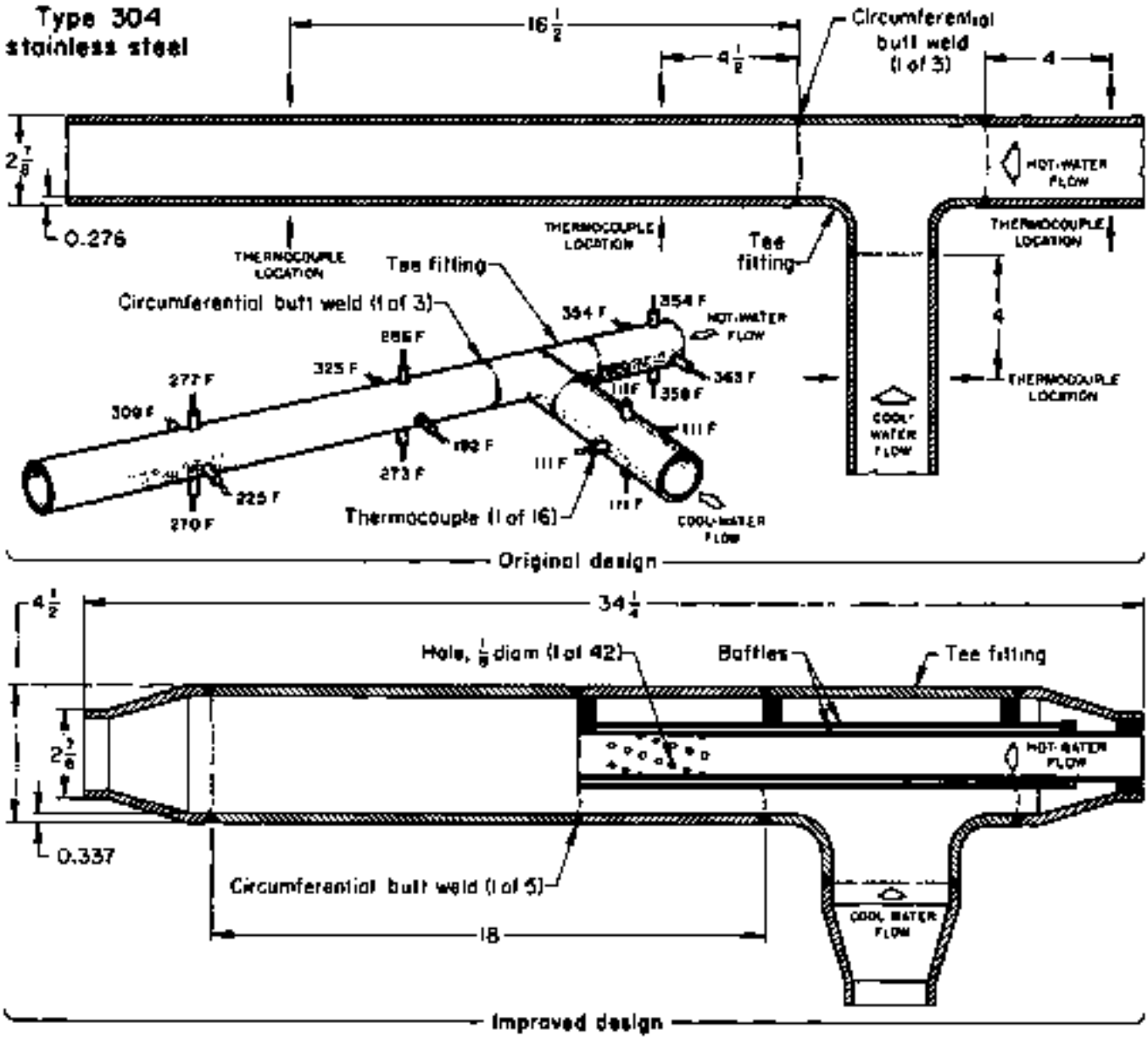


Fig. 32 Type 304 stainless steel tee fitting that failed by low-cycle thermal fatigue. Top: original design. Inset shows the locations of thermocouples used in analyzing thermal gradients and the typical temperatures at each thermocouple location. Bottom: the analysis resulted in an improved design. Dimensions given in inches

Investigation. Visual and metallographic examination showed that each of the failures consisted of cracking in and/or close to the circumferential butt weld joining the tee fitting to the downstream pipe leg, where the hot bypass water mixed with the cool return water. The cracking was transgranular and was located within the weld, adjacent to the weld, or in the base metal of either the pipe or the tee. Several cracks resulted in leakage. In one instance, the welds were radiographed before any leak was observed, and a long crack was found in the weld between the tee fitting and the downstream pipe leg. It was established that this crack had developed in service and had not been an original welding defect. In another instance, a crack occurred in a pipe segment about 100 mm (4 in.) from the nearest weld. This crack was outside the HAZ of the weld; however, additional cracks were found in the tee fitting and in the welds.

Because propagation of the cracks was entirely transgranular, thermal fatigue was a more likely cause of failure than SCC, in which some intergranular cracking would be expected.

To investigate thermal stresses, an array of thermocouples was installed in two tee fittings in the arrangement shown in Fig. 32. The temperatures at different points along the length and around the circumference of the piping are also shown for one of the two instrumented fittings; temperatures were similar for the other fitting. It was concluded that circumferential temperature gradients, in combination with inadequate flexibility in the piping system as a whole, had caused all of the failures.

Corrective Measures. To alleviate the thermal-stress pattern, the tee fitting was redesigned (Improved design, Fig. 32). A larger tee fitting, made of 100-mm (4-in.) schedule 80S type 316H stainless steel pipe (ASME SA-312, grade TP316H), was used, and a 455-mm (18-in.) long extension of 100-mm (4-in.) schedule 80 type 316H pipe was added to the downstream leg. Two concentric tubular baffles, both made of type 316H stainless steel, were placed in the run direction of the interior of the tee fitting. The inner baffle (455 mm, or 18 in., long × 38 mm, or $1\frac{1}{2}$ in., outside diameter × 2.1 mm, or 0.085 in., wall thickness, and perforated at the downstream end) carried the full flow of the hot water from the bypass. The outer baffle (400 mm, or 16 in., long × 50 mm, or 2 in., outside diameter × 2.1 mm, or 0.085 in., wall thickness), which was open at each end, served to insulate the inner baffle from the cool water entering the fitting from the heat exchanger and provided a region in which the cool water could form an annulus to receive the core of hot bypass water entering the downstream extension beyond the baffles. Mixing in that region was possible without generation of damaging circumferential thermal gradients.

Thermal fatigue may result in either low-cycle or high-cycle failure. Low-cycle thermal fatigue is usually associated with large plastic strains and is most often caused by large changes in temperature or large differences in thermal expansion between two structural members. If a change in temperature is especially severe and if it occurs rapidly, it is called thermal shock; cracks due to thermal shock generally initiate in ten thermal cycles or less—sometimes in only one thermal cycle when the material is sufficiently brittle.

High-cycle thermal fatigue frequently results from intermittent wetting of a hot surface by liquid having a considerably lower temperature. In such instances, the wetted surface contracts rapidly, whereas the metal below the surface does not; this produces large biaxial tensile stresses in the wetted surface. As the water absorbs heat and evaporates, the temperature of the surface returns to its previous value, and surface stresses are relaxed. After a sufficient number of thermal cycles, a crazed pattern of cracks appears on the surface (Fig. 33). Eventually, one or more of the surface thermal fatigue cracks may propagate completely through the section by fatigue or by another mechanism, as discussed in the section “Multiple-Mode Failure” in this article.

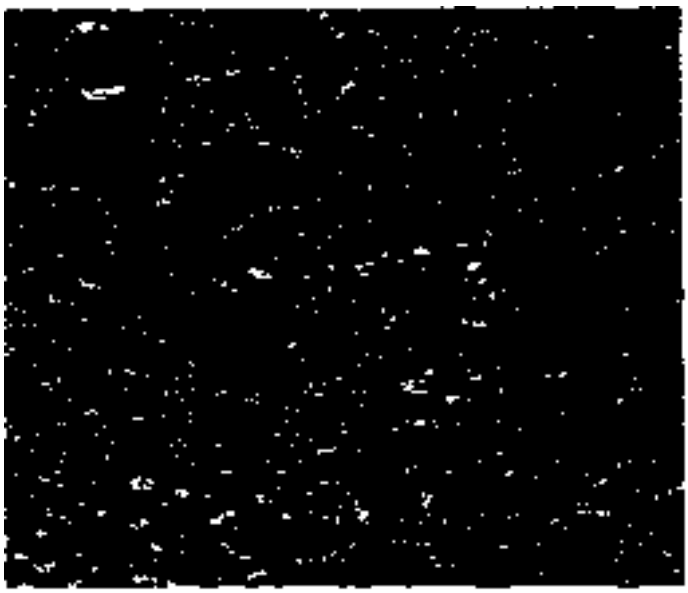


Fig. 33 Crazed pattern of thermal fatigue cracking on the outer surface of a stainless steel tube. See also Fig. 37. Approximately 4×

Cracking that is caused by thermal fatigue sometimes resembles cracking caused by corrosion fatigue or stress corrosion. However, in contrast to corrosion fatigue and stress corrosion, which can be retarded by use of inhibitors, coatings, or galvanic protection, thermal fatigue will proceed regardless of what is done to alter the environment or to coat the metal to prevent direct contact with the environment. Thermal fatigue can be prevented only by eliminating excessive strain caused by thermal cycling.

Ferritic steels are generally considered to be more resistant to thermal fatigue than austenitic steels. This has been confirmed in several instances in which parts made of ferritic steels have withstood service conditions under which austenitic steel parts have failed. One such instance involved two flanged 18Cr-12Ni austenitic stainless steel valves that were bolted to 5Cr-0.5Mo ferritic steel piping in a petroleum plant. The valves were alternately exposed to atmospheric temperature and to an elevated temperature between 595 and 620 °C (1100 and 1150 °F). Each valve was subjected to $1\frac{1}{2}$ thermal cycles per day, that is, three cycles every 2 days; about 8 h elapsed between the start of the heating portion of the

cycle and the start of the cooling portion, and conversely. After 10 days of operation, cracks were found radiating from the bore into the flange of one valve; the other valve failed shortly after. The mating 5Cr-0.5Mo ferritic steel piping gave excellent service, as did replacement valves made of 9Cr-1Mo ferritic steel.

The superiority of ferritic alloy steel over several wrought austenitic stainless steels in applications involving thermal cycling is a result of differences in thermal conductivity and coefficient of thermal expansion between the two types of steel. The higher thermal conductivity of terrific alloy steel results in a lower temperature gradient between the surface and the interior when the surface of a cool part is heated or when the surface of a hot part is cooled. This lower temperature gradient results in less thermal stress in the surface layer. In addition, because of the lower coefficient of expansion of ferritic steel, a given temperature gradient produces less expansion, and thus less thermal stress, in ferritic alloy steel than in wrought austenitic stainless steel. The combination of these two factors can result in appreciably less thermal stress and longer life for terrific steel under a given set of cyclic thermal conditions.

Corrosion fatigue is discussed in detail in the article “High-Temperature Corrosion-Related Failures” in this Volume. Strictly speaking, few service failures occur by pure fatigue. Most fatigue failures actually occur by corrosion fatigue, because most parts are exposed simultaneously to cyclic stress and to a medium that is somewhat aggressive toward the material; in this strict sense, even air can be considered an aggressive medium. In a more practical sense, however, because fatigue properties are normally determined in air, corrosion fatigue can be thought of as fatigue that takes place in an environment in which the combined action of corrosion and cyclic loading results in more rapid crack initiation and/or propagation than would occur in air. The microscopic mechanism of corrosion fatigue appears to be related to the galvanic or preferential dissolution processes that are believed to be responsible for SCC. In fact, all alloys that are susceptible to stress corrosion in a given medium appear to be susceptible to rapid corrosion fatigue in the same medium. However, corrosion fatigue has been identified as a failure mechanism in alloys that are considered immune to stress corrosion (but not immune to ordinary corrosion in the same environment).

The length of time necessary for a crack to be initiated and propagated by corrosion fatigue can be increased—sometimes substantially—by reducing or eliminating either the fluctuating stress or the corrosiveness of the environment. In addition to reduction of stress or stress concentration, the following measures have been used successfully to combat corrosion fatigue: (1) use of corrosion inhibitors, (2) cathodic protection, (3) shielding of the surface of a part from contact with the environment by seals, by paint, plastic or other nonmetallic coatings, or by plating with corrosion-resistant metals, and (4) use of a more corrosion-resistant material. The last method should be used with care, however, because some of the ordinarily corrosion-resistant materials can be more susceptible to failure by corrosion fatigue in certain environments than the materials they replace. Coatings should be used to combat corrosion fatigue only when local strains are uniform and low. The protection afforded by coatings, either metallic or nonmetallic, is only skin deep, and local breaches of this protective skin expose the metal underlying the coating to the same conditions as those that ordinarily lead to failure of uncoated parts.

Failures of Boilers and Related Equipment

Revised by David N. French, David N. French, Inc., Metallurgists

---

Failures Caused by Erosion

Erosion involves impact of large numbers of small solid or liquid particles against a surface, or it is caused by the collapse of gas-filled bubbles in a cavitating liquid. Erosion by solid particles is a form of abrasive wear, whereas liquid-impingement erosion—for example, as caused by the water droplets in wet steam—is more like cavitation erosion.

The microscopic mechanisms of material removal by erosion are discussed in the articles “Fundamentals of Wear Failures” and “Liquid-Impact Erosion” in this Volume. Failures caused by these mechanisms can be readily recognized by visual examination.

Abrasive erosion of screen tubes or superheater tubes results from impact by particles of fly ash entrained in the flue gases. Erosion is enhanced by high flow velocities; thus, partial fouling of gas passages in tube bands by deposition of fly ash can lead to erosion by forcing the flue gases to flow through smaller passages at higher velocity. This effect, sometimes called laning, exposes tube surfaces to a greater probability of impact by particles having higher kinetic energy, thus increasing the rate of damage. Erosion by fly ash causes polishing, flat spots, wall thinning, and eventual tube rupture.

Fly-ash erosion can be controlled by coating tube surfaces with refractory cements or other hard wear-resistant materials, although this reduces the heat-transfer capability of the surface. An alternative method is the channeling of gas flow away from critical areas with baffles.

Liquid-impingement erosion occurs chiefly in components that are subjected to high-velocity flow of wet steam. Among the components most susceptible to liquid-impingement erosion are low-pressure turbine blades, low-temperature steam piping, and condenser or other heat-exchanger tubes that are subjected to direct impingement by wet steam. Corrosion may or may not occur simultaneously with liquid-impingement erosion.

Liquid-impingement erosion in tubing or piping is most likely when fluid velocities exceed 2.1 m/s (7 ft/s). Damage occurs first at locations where direction of flow changes, such as elbows or U-bends. Large-radius bends are less susceptible to such damage; however, use of erosion-resistant materials, such as austenitic stainless steel, is often more effective. Erosion of heat-exchanger tubes caused by impingement of wet steam can sometimes be eliminated by redirecting flow with baffles.

Cavitation erosion occurs in regions of a system where a combination of temperature and flow velocity causes growth and subsequent collapse of large quantities of vapor-filled bubbles in a flowing stream of liquid. Damage occurs at the points of bubble collapse. In steam plants, the components most often damaged by cavitation erosion are feed lines, pump casings, and pump impellers. Figure 34 shows the typical appearance of a part extensively damaged by cavitation erosion; the surface is deeply and irregularly pitted.

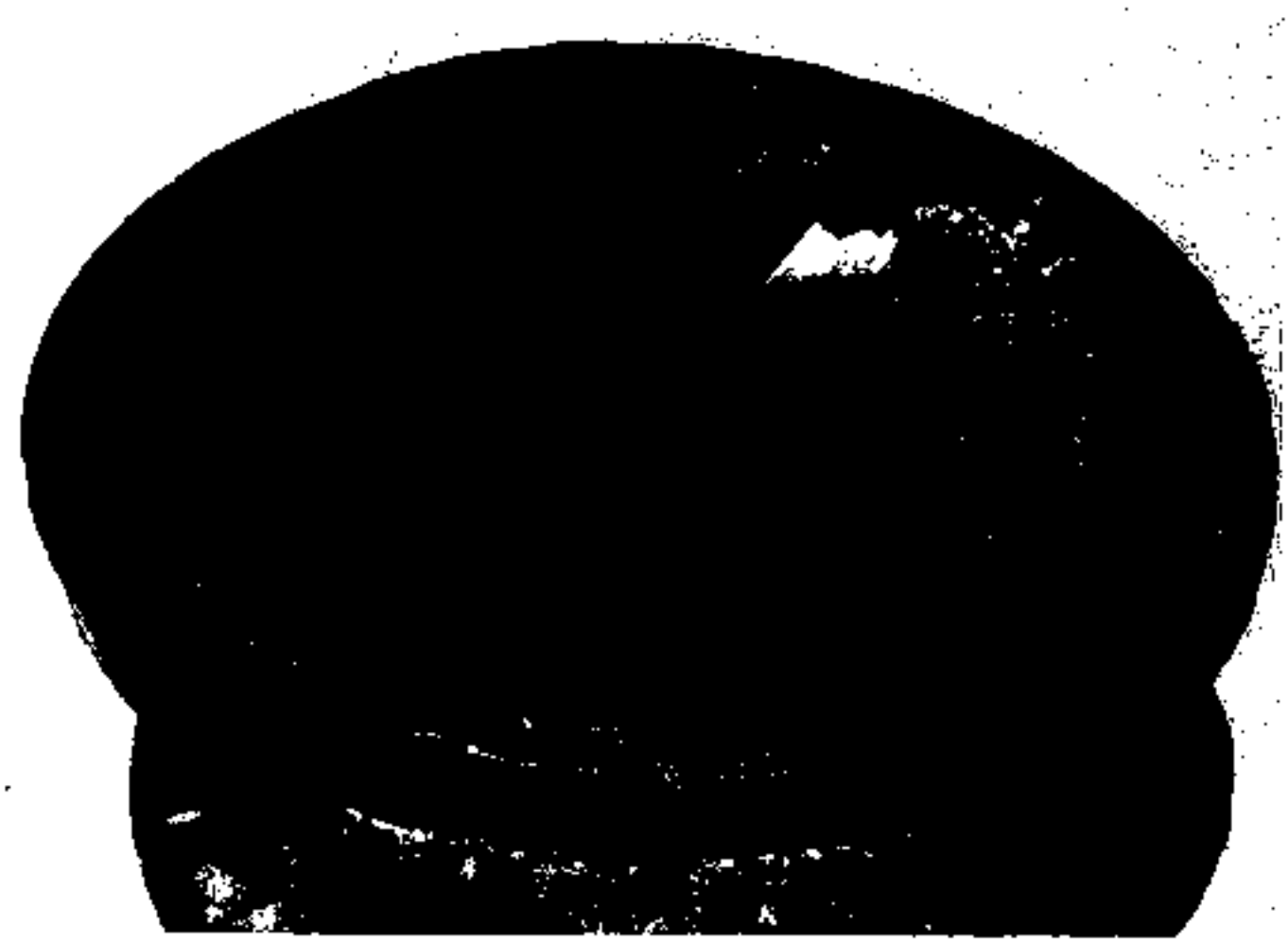


Fig. 34 Cast iron suction bell, 455 mm (18 in.) in diameter, from a low-pressure general service water pump that failed by cavitation erosion after about 5 years of service. Note the deeply pitted surface and the irregular shape of the erosion pattern, both of which are typical characteristics of caviation damage.

Cavitation erosion is most effectively avoided by reduction of either flow velocity or temperature (reduction of flow velocity is preferable). Other measures, including selection of material and the use of erosion-resistant welded overlays, are discussed in the article “Liquid Impact Erosion” in this Volume.

Revised by David N. French, David N. French, Inc., Metallurgists

Failure by SCC

Stress-corrosion cracking occurs in specific alloys exposed to specific environments, as discussed in detail in the article "Stress-Corrosion Cracking" in this Volume. Stress-corrosion cracking is a type of failure defined as fracture of a structural member under the combined effects of static tensile stress and a corrosive environment in circumstances where, if either the tensile stress or the corrosive environment were absent, fracture would not occur.

Stress-corrosion cracking can occur in boiler and superheater tubes, piping, valves, turbine casings, and other parts, especially where feedwater or condensate can collect. Concentration of caustic in portions of systems that use conventional or coordinated phosphate methods of feedwater treatment is probably the most common cause of stress-corrosion (caustic) cracking in steam equipment. Aqueous solutions containing chloride ions can cause SCC of austenitic stainless steel (Example in this article), and solutions containing ammonia can cause SCC of copper alloys. Stress oxidation in high-temperature regions can also occur.

Many cracks that result from stress corrosion are tight and difficult to detect visually (Fig. 35); they are particularly difficult to detect on scaled surfaces. Liquid-penetrant examination is useful in locating tight stress-corrosion cracks, but may not reveal all such cracks. Because the substances found in liquid penetrants are similar to some of the substances that cause SCC, use of a liquid penetrant to detect a crack can interfere with the identification of fracture-surface deposits. Consequently, liquid penetrants should not be used without due consideration of their undesirable side effects on the overall analysis.

Fig. 35 Inside surface of an austenitic stainless steel superheater tube showing a tight crack caused by stress corrosion. Arrows indicate ends of crack.

Microscopic examination is required to confirm SCC as the mechanism of failure. Stress-corrosion cracks may be transgranular or intergranular, depending on the alloy and the environment, but are usually extensively branched, becoming more branched as cracking progresses.

Although service conditions may produce static stress that exceeds the threshold for SCC, residual stress is more often responsible. Residual stress may result from metallurgical transformations or from uneven heating or cooling, especially in weldments. Hot- or cold-forming processes can introduce residual stress of high magnitude, as can assembly processes, such as press fitting, bolting, or riveting. Incidental damage, such as bending or hammering during assembly, is less frequently a source of residual stress.

Example 11: SCC of a Type 304 Stainless Steel Pipe Caused by Residual Welding Stresses. A 150-mm (6-in.) schedule 80S type 304 stainless steel pipe (170 mm, or $6\frac{5}{8}$ in., outside diameter × 11 mm, or 0.432 in., wall thickness), which had served as an equalizer line in the primary loop of a pressurized-water reactor, was found to contain several circumferential cracks 50 to 100 mm (2 to 4 in.) long. Two of these cracks, which had penetrated the pipe wall, were responsible for leaks detected in a hydrostatic test performed during a general inspection after 7 years of service.

Investigation. The general on-site inspection had included careful scrutiny of all pipe welds by both visual and ultrasonic examination. Following discovery of the two leaks, the entire line of 150-mm (6-in.) pipe, which contained 16 welds, was carefully scanned. Five additional defects were discovered, all circumferential and all in HAZs adjacent to welds. In contrast, scans of larger-diameter pipes in the system (up to 560 mm, or 22 in., in diameter) disclosed no such defects.

Samples of the 150-mm (6-in.) pipe were submitted to three laboratories for independent examination. Inspection in all three laboratories disclosed that all the defects were circumferential intergranular cracks that had originated at the inside surface of the pipe and that were typical of stress-corrosion attack. A majority of the cracks had occurred in HAZs adjacent to circumferential welds. Some cracks had penetrated the entire pipe wall; others had reached a depth of two-thirds of the wall thickness. In general, the cracks were 50 to 100 mm (2 to 4 in.) long. Branches of the cracks that had approached weld deposits had halted without invading the structure (austenite plus δ-ferrite) of the type 308 stainless steel welds. All HAZs examined contained networks of precipitated carbides at the grain boundaries, which revealed that welding had sensitized the 304 stainless steel in those local areas.

Additional cracks, also intergranular and circumferential and originating at the inner surface of the pipe, were discovered at locations remote from any welds. These cracks penetrated to a depth of only one-sixth to one-fourth of the wall thickness through a solution-annealed structure showing evidence of some cold work at the inner surface.

The water in this heat exchanger was of sufficient purity that corrosion had not been anticipated. The water conditions were as follows:

- *Temperature*: 285 °C (545 °F)
- *Pressure*: 7 MPa (1000 psi)
- *pH*: 6.5 to 7.5
- *Chloride content*: <0.1 ppm
- *Oxygen content*: 0.2 to 0.3 ppm
- *Electrical conductivity*: <0.4 μmho/cm

Analysis of the pipe showed the chemical composition to be entirely normal for type 304 stainless steel. The material was certified by the producer as conforming to ASME SA-376, grade TP304. All available data indicated that the as-supplied pipe had been of acceptable quality.

Several types of stress-corrosion tests were conducted. In one group of tests, samples of 150-mm (6-in.) type 304 stainless steel pipe were welded together using type 308 stainless steel filler metal. Two different welding procedures were used—one consisting of three passes with a high heat input and the other consisting of ten passes with a low heat input. The welded samples were stressed and exposed to water containing 100 ppm dissolved oxygen at 285 °C (545 °F) and 7 MPa (1000 psi). Samples that were welded with the high heat input procedure, which promotes carbide precipitation and residual stress, cracked after 168 h of exposure, whereas samples that were welded with the low heat input procedure remained crack free. These results were considered significant, because service reports indicated that the cracked pipe had been welded using a high heat input and a small number of passes and that the larger-diameter piping in the primary loop, which did not crack in service, had been welded with the use of a low heat input, multiple-pass procedure.

Conclusions. The intergranular SCC of the 150-mm (6-in.) schedule 805 type 304 stainless steel pipe exposed to high-temperature high-pressure high-purity water in 7 years of heat-exchanger service was believed to have been caused by the following factors:

- *Stress*: The welding procedure, which employed few passes and high heat input, undoubtedly generated unacceptably high levels of residual stress in the HAZs. The cold working of the internal surface in a sizing operation also set up residual stresses in areas apart from the welds. Both conditions were conducive to SCC
- *Sensitization*: The high heat input welding caused precipitation of chromium carbides at the grain boundaries in the HAZs. This rendered the steel sensitive to intergranular attack, which, combined with the residual stress, afforded a completely normal setting for stress-corrosion cracking
- *Environment*: Although the level of dissolved oxygen in the water was quite low, it was considered possible that, on the basis of prolonged exposure, oxygen in the range of 0.2 to 1.0 ppm could have provided the necessary ion concentration

Corrective Measures. All replacement pipe sections were installed using low heat input multiple-pass welding procedures. When stress corrosion is identified as the mechanism of a failure, the proper corrective action can be either a reduction in

stress (or stress concentration) or an alteration of the environment. When the cause of failure is inadvertent concentration of a corrodent or inadvertent exposure of the part to a corrosive foreign substance, such as use of a steam line to feed concentrated caustic during chemical cleaning, the preventive measure may be no more complicated than exclusion of the corrodent from that region of the system. In many instances, however, reduction of the level of residual stress is the most effective means of minimizing or preventing SCC.

Example 12: SCC of Type 316 Stainless Steel Tubing. A steam-condensate line made of type 316 stainless steel tubing that measured 19 mm (0.75 in.) outside diameter × 1.7 mm (0.065 in.) wall thickness had been in service for 5 to 6 years when leakage occurred. The line was buried, with no wrapping, in a sandy, caustic soil. The line carried steam condensate at 120 °C (250 °F) with a 2-h heat-up/cool-down cycle. No chemical treatment had been given to either the condensate or the boiler water. Two sections of the tubing, with a combined length of approximately 330 mm (13 in.), were submitted for laboratory analysis of the cause of the leakage.

Investigation. The outside surface of the tubing exhibited no signs of corrosion; however, on one of the sections there was a transverse crack along the edge of a 6.4-mm ($\frac{1}{4}$-in.) diam pitted area. Two adjacent areas were raised, resembling blisters, and also had transverse cracks along their edges. The cracks were irregular and had penetrated completely through the tube wall. The pitted area was believed to be a spall, and the blisters incipient spalls, resulting from SCC that originated at the inner surface of the tube. The cracks, the spall, and the blisters did not appear to have been related to any effect of the external environment.

The inside surface of the tubing was covered with a brown, powdery scale, was slightly rough, and contained small pits. Some cracks were found that had originated at pits on the inner surface but had not completely penetrated the tubing wall (Fig. 36). The cracks were transgranular and branching, which is typical of chloride-induced SCC of austenitic stainless steel.

Fig. 36 Section through type 316 stainless steel tubing that failed by SCC because of exposure to chloride-contaminated steam condensate. Micrograph shows a small transgranular crack that originated at a corrosion pit on the inside surface of the tubing and only partly penetrated the tubing wall. Etched with aqua regia. 250×

To check for chlorides, the inside of the tubing was rinsed with distilled water, and the rinse water was collected in a clean beaker. A few drops of silver nitrate solution were added to the rinse water, which clouded slightly because of the formation of insoluble silver chloride. This test indicated that there was a substantial concentration of chlorides inside the tubing.

Conclusion. The tubing failed by chloride-induced SCC. Chlorides in the steam condensate also caused corrosion and pitting of the inner surface of the tubing. Stress was produced when the tubing was bent during installation.

Recommendations. Water treatment should be provided to remove chlorides from the system. Continuous flow should be maintained throughout the entire tubing system to prevent concentration of chlorides. No chloride-containing water should be permitted to remain in the system during shutdown periods. Bending of tubing during installation should be avoided to reduce residual stress.

Revised by David N. French, David N. French, Inc., Metallurgists

Multiple-Mode Failure

Many failures in steam systems involve more than one failure process—one that initiates the failure and another (or others) that eventually destroys the part. For example, fatigue failures that cause tubes or pipes to leak may start at corrosion pits. A failure that involved a combination of thermal fatigue and stress-rupture cracking is described below.

The tube rupture shown in Fig. 37(a) has the macroscopic characteristics of a thick-lip rupture caused by overheating. However, as shown in Fig. 37(b), stress-rupture cracking was confined to the inner half of the tube wall. There was no microstructural evidence of overheating.

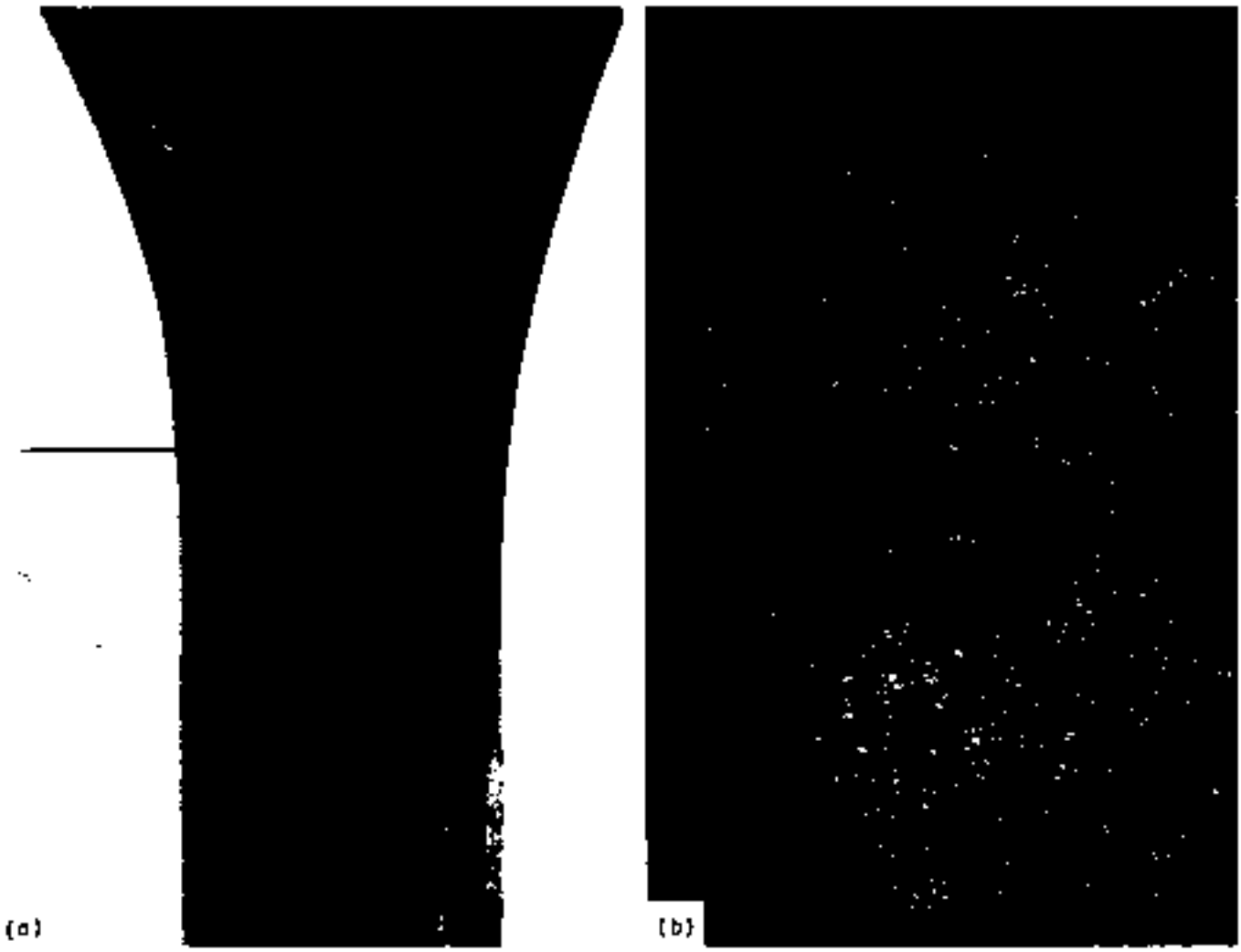


Fig. 37 Stainless steel superheater tube that failed by thermal fatigue and stress rupture. (a) Photograph of the tube showing thick-lip rupture. (b) Macrograph of a section taken transverse to a fracture surface of the tube showing that thermal fatigue cracking started at the outside surface (top) and was followed by stress-rupture cracking in the inner half of the tube wall. 10×. Thermal fatigue also produced cracking on the outside surface of the tube in the crazed pattern shown in Fig. 33

The outside surface of this 54-mm ($2\frac{1}{8}$-in.) diam stainless steel superheater tube exhibited extensive crazing by thermal fatigue, as shown in Fig. 33. Apparently, cyclic thermal stress initiated cracking at the fire-side surface. Pressure stresses

assisted in defining the preferential direction of fatigue-crack propagation (parallel to the tube axis and transverse through the tube wall), and stress-rupture cracking became the ultimate fracture process in the weakened tube.

The effects of internal and external scaling on tube ruptures caused by overheating are discussed in the section “Scaling” in this article. In tube and other high-temperature structural components, thinning caused by corrosion or erosion can also lead to overheating-type failures. Corrosion or erosion, particularly on the fire-side surfaces of screen tubes and superheater tubes, may reduce the tube-wall thickness to the extent that the tube can no longer sustain the pressure stress without creep. Continued thinning will accelerate the onset of tertiary creep, ultimately leading to stress-rupture fracture. Because localized loss of material is responsible for increasing the local tube-wall stress, overheating-type ruptures often occur with little or no swelling. Internal grain separation will be evident in the microstructure only in the thinned region, and macroscopic examination of the fire-side surface is usually sufficient to reveal the extent of external attack. Localized overheating is not necessary for tube rupture to occur. Consequently, a lack of microstructural evidence of overheating would be a positive indicator that corrosion or erosion was a factor. Conversely, the presence of such evidence does not rule out corrosion or erosion as a cause of failure.

Revised by David N. French, David N. French, Inc., Metallurgists

References

1. D.N. French, *Metallurgical Failures in Fossil Fired Boilers,* John Wiley & Sons, 1983, p 155–165
2. W.H. McAdams, *Heat Transmission,* McGraw Hill, 1952, p 184–189
3. *Principles of Industrial Water Treatment,* Drew Chemical Co., Boonton, NJ, 1979, p 194
4. I.M. Rehn, W.R. Apblett, Jr., and J. Stringer, Controlling Steamside Oxide Exfoliation in Utility Boiler Superheaters and Reheaters, *Mater. Perform.,* June 1981, p 27–31
5. *Steam,* 38th ed., Babcock & Wilcox Co., New York, 1972
6. *High Temp. Bull.,* No. 5, Babcock & Wilcox Co., Tubular Products Divison, Beaver Falls, PA
7. H.A. Grabowski and H.A. Klein, Corrosion and Hydrogen Damage in High Pressure Boilers, in *Proceedings of the 2nd Annual Educational Forum on Corrosion,* National Association of Corrosion Engineers, Houston, Sept 1964
8. R.W. Bailey, *Engineering,* Vol 129, 1930, p 265
9. R.W. Bailey, *Engineering,* Vol 133, 1932, p 261
10. W.L. Hemingway, *The Study of Graphitization,* Edwards Valve Co., 1952

Failures of Boilers and Related Equipment

Revised by David N. French, David N. French, Inc., Metallurgists

Selected References

- *Combustion-Fossil Power Systems,* Combustion Engineering, Inc., Windsor, CT, 1981

- O. Devereux, A. McEvily, and R. Staehle, Ed., *Corrosion Fatigue: Chemistry, Mechanics and Microstructure,* National Association of Corrosion Engineers, Houston, 1972
- B.E. Gatewood, *Thermal Stresses,* McGraw-Hill, 1957
- S.S. Manson, *Thermal Stress and Low-Cycle Fatigue,* McGraw-Hill, 1966
- *Manual for Investigation and Correction of Boiler Tube Failures,* EPRI CS-3945, Electric Power Research Institute, Palo Alto, CA, April 1985
- C.C. Osgood, *Fatigue Design,* John Wiley & Sons, 1970
- *Proceedings of the International Conference on Fatigue of Metals,* Institution of Mechanical Engineers, London, 1956
- *Proceedings of the Joint International Conference on Creep,* Institution of Mechanical Engineers, London, 1963
- P.D. Stevens-Guille, *Steam Generator Tube Failures: World Experience in Water-Cooled Nuclear Power Reactors During 1972,* Chalk River Nuclear Laboratories, Ontario, March 1974
- H. Thielsch, *Defects and Failures in Pressure Vessels and Piping,* Reinhold, 1965
- H. Thielsch, Why High Temperature Piping Fails, *Met. Eng. Q.,* Vol 10 (No. 1), 1970, p 7–11
- H.H. Uhlig, *Corrosion and Corrosion Control,* 2nd ed., John Wiley & Sons, 1971

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Introduction

HEAT EXCHANGERS are generally used to transfer heat from combustion gases, steam, or water to gases, vapors, or liquids of various types. Heat exchangers are of tube, plate, or sheet construction. Tubular heat exchangers are generally used for large fluid systems, whereas heat exchangers of plate or sheet construction are often preferred for smaller fluid streams. Heat exchangers are usually separate units, but can sometimes be incorporated as components in larger vessels.

In selecting materials for heat exchangers, corrosion resistance, strength, heat conduction, and cost must be considered. The demand for corrosion resistance is particularly difficult to meet because the material may be exposed to corrosive attack by various mediums. Therefore, damage to heat exchangers is often difficult to avoid.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Characteristics of Tubing

The primary function of tubes in a heat exchanger is to transfer heat from the shell side of the unit to the fluid on the inside of the tube, or conversely. Therefore, thermal conductivity, wall thickness, and resistance to scaling are extremely important. In most cases, tensile strength and yield strength of the tubes are not significant factors; loading, because it is opposed by the internal pressure, is so low that tubes with a high allowable stress are not required.

Stiffness and resistance to denting are properties of tubing that are important in the manufacture of heat exchangers. Stiffness of a tube is proportional to its wall thickness, diameter, and modulus of elasticity. Resistance to denting is related to the same factors plus the yield strength of the material.

Extended-surface (finned) tubing is sometimes used to provide the proper thermal conductance for effective transfer of heat from gas to liquid or from liquid to liquid. Fins on much of the tubing used in air coolers and hydrogen coolers are applied by soldering, brazing, or resistance welding; thus, the ability of the tube material to accept these processes is important. Oil coolers often use integral-finned (extruded) tubing. This type of tubing requires a material with a significant amount of ductility to ensure that the extended surface can be satisfactorily extruded.

Corrosion Resistance. To meet corrosion requirements, tubing must be resistant to general corrosion, stress-corrosion cracking (SCC), selective leaching (for example, dezincification of brass), and oxygen-cell attack in whatever environments are encountered before service, in service, and during periods in which the equipment is not operating.

The environment to which tubes will be exposed is often difficult to predict. Many heat exchangers operate with river water, seawater, brackish water, or recirculated cooling-tower water in the tubes. The compositions of these waters used for cooling can change over a period of time, through seasonal variations, or from an unpredictable accident. Long lead time in power-plant erection can have a considerable effect on the tubes. Heat exchangers may be completed and sent to the field years before the remainder of the plant is completed and put into operation. Under such circumstances, the heat exchangers should be adequately protected from the corrosion that might occur before start-up. Wet lay-up using a suitably buffered or inhibited solution can be used to retard corrosion. Another method is to keep the heat exchanger dry and, if necessary, fill it with an inert gas. Desiccants or vapor-phase corrosion inhibitors can also be added, and the heat exchanger can then be sealed until start-up.

After start-up, consideration must be given to the many unscheduled shutdowns that generally affect the early life of any new plant. Suitable precautions must be taken to protect equipment temporarily out of service if damage is to be avoided during downtime.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Examination of Failed Parts

Full background information on a part is important for proper failure analysis. The information should describe the circumstances of failure, operation, or other pertinent details. Inadequate information on the circumstances surrounding the failure often results when the investigation is not conducted at the failure site and samples are sent to the laboratory for examination. A list of the type of background information useful for a proper analysis of failures in heat exchangers is given in Table 1.

Table 1 Data needed for analysis of failures in heat exchangers

- Process name and function
- Unit name and function
- Material size, grade, specification
- Sketches of unit, failures or other particulars
    - Environment (inside and out):
        - Temperatures, max and min
        - Substances contacting tube:
            - Common name
            - Chemical composition
            - Concentration
            - Impurities
            - pH
            - Velocity
            - Aeration
            - Contact with other metals

- Particulars of failure:
  - Type
  - Location
  - Circumstances at time of failure
  - Service life
- History:
  - Service behavior of other materials
  - Unusual conditions before failure
  - Fabricating operations, sequence
  - Maintenance operations, sequence
  - Heat-treating operations, sequence

Sample Selection. The selection of samples that are representative of the failure is important. Samples should be taken not only from the failure area but also from areas remote from the site of the failure so that comparisons can be made. Care should be taken in selecting and removing samples from heat exchangers. Samples that have been burned or battered in removal or have been modified during the preliminary examination do not adequately illustrate the failure and cannot be considered representative. Removal of rolled tubes from tube sheets requires particular care.

Visual Examination. Much information concerning a failure can be obtained from a careful visual examination. Dimensional changes reflect swelling or thinning. The pattern, extent, or nature of corrosion or cracking is often an important clue to the cause of failure (consequently, the examination may be hindered by a sample that is too small). Discoloration or rusting on stainless steel may be a sign of iron contamination. Heavy deposits on a tube, such as scale or corrosion product, could be caused by overheating.

Visual examination will indicate the direction of further work; the detailed investigation should be planned at this stage. Care must be taken so that evidence is not inadvertently altered or destroyed. Provision should be made for preserving some of the original samples to provide material for a fresh start if an answer to the failure is not found. Sketches and photographs prepared at this time are helpful for later reference in interpreting results.

Microscopic Examination. For metallurgical investigation, microstructural details provide much information about thermal history, operating temperatures, chemical environment, manner of attack, and cracking.

Microstructural examination of a section through the fracture surface is an aid in determining if the fracture is transgranular or intergranular. Also, variations in material between the failure area and a remote region can reveal specific local microstructural effects, such as decarburization, alloy depletion, graphitization, or precipitation of embrittling constituents at the failure area. Quality of soldered or brazed joints, depth of penetration of welds, or occurrence of gas porosity can be revealed by proper sectioning. A critical combination of strain and subsequent heating can occasionally cause excessive grain growth. This has occurred in copper tubing, aluminum alloy tubing, or copper-clad steel tubing used for medium-pressure duty.

Materials may be identified by microstructural examination. For example, various grades of carbon steel pipe or tubing may be distinguished according to the deoxidation practice and resulting microstructure. Similarly, seamless and welded pipe or tubing may be differentiated. A gross error in material usage, such as using low-carbon instead of medium-carbon steel, can be determined quickly by microstructural inspection. In low-carbon steels, rimmed steels can be distinguished from killed steels by their larger-grain rimmed zone at the surface and concentration of carbon and impurities at the center of the section. The presence or absence of inclusions can also be checked.

In failures involving a weld, careful evaluation of the heat-affected zone (HAZ) is necessary. In carbon and alloy steels, HAZs can be abnormally hard. In austenitic stainless steels, HAZs can be sites for precipitated intergranular carbides; the resulting alloy-depleted zones surrounding the precipitates are potential origins for intergranular corrosion or cracking of the weldment.

Chemical Analysis. Information on corrosive agents is usually obtained by analysis of corrosion products or metal surfaces. For example, silver nitrate tests may be performed on residues clinging to a pitted stainless steel surface to confirm the presence of chloride ions—a common cause of pitting. Wet chemical methods are often supplemented by the use of x-ray diffraction or electron probe x-ray microanalysis, using wavelength- or energy-dispersive x-ray spectrometers for identification of compounds, which in turn may identify corrosives. These techniques are often used to analyze for the presence of contaminants on crack surfaces or at the roots of pits. More information on these analytical techniques is available in Volume 10 of *ASM Handbook*, formerly 9th Edition *Metals Handbook.*

Mechanical Tests. Tensile or stress-rupture tests can be used to determine if the metal had mechanical properties suitable for the intended service. Hardness tests can be used to check whether heat treatment was correct or uniform or if hardening occurred from cold working, overheating, carburization, or phase changes in service. Impact tests indicate brittle tendencies.

Revised by Robert J. Franco, Exxon Company, U.S.A.

Causes of Failures in Heat Exchangers

Failures in heat exchangers are commonly associated with methods of manufacturing pipe and tubing, handling methods during fabrication, testing methods in the shop and in the field, and the total environment to which the unit is exposed after fabrication—including conditions during shipment, storage, start-up, normal operation, and shutdown.

Fabricators and suppliers of pipe or tubing strive to minimize or eliminate latent surface or subsurface imperfections produced during manufacture. Nondestructive inspection, such as by eddy-current, ultrasonic, air-under-water, and hydrostatic tests, is used to detect these imperfections. Each test has its limitations. Hydrostatic tests and air tests are useful for detecting discontinuities extending completely through the tube wall, but they are of little use in detecting discontinuities that extend only partly through the wall. In these cases, ultrasonic and eddy-current methods are more useful. These test methods allow the sensitivity of the test instrument to be set at a level consistent with the type and size of imperfection being sought.

Where an imperfection extends through the wall but is very small and tight, the tube may not leak in a hydrostatic test, but it may leak in subsequent service. Air tests under water can be used to reveal imperfections of this type.

Fortunately, relatively few imperfections produced during manufacture escape detection and lead to failures in service. The combination of eddy-current and hydrostatic testing, whether conducted by the fabricator or tube supplier, effectively minimizes service failures resulting from imperfections produced during manufacture.

Secondary manufacturing techniques result in more service failures than do surface and subsurface imperfections produced during tubing manufacture. The methods used to draw, clean, heat treat, and straighten the tubes may be quite significant to service performance, but are often overlooked in specifications and inspection procedures. Drawing, heat treating, and straightening operations determine the level of residual hoop and bending stresses induced in the tube. By varying the severity of these operations, it is possible to produce tubes with very low residual stresses or with very high residual stresses that are near the yield strength of the metal. Low residual stresses are particularly important when tube materials must have maximum resistance to SCC in service. The residual stresses on the inner and outer surfaces of the tube and through the tube wall must be low; residual stresses at the surfaces should be compressive.

Cleaning methods, control of heat-treating atmospheres, and pickling operations affect the condition of the oxides on the inner surface of the tube. The oxide should be very thin and flexible and should completely cover the inner surface of the tubing because fractures in the oxide layer create anodic and cathodic sites that can lead to pitting corrosion. If the surface oxide scale is to be removed, it should be completely removed in order not to establish anodic and cathodic sites, which are starting points for corrosion during service. The inner surfaces of some tubes are abrasive blasted to remove the oxides produced during heat treatment and thus eliminate anodic and cathodic sites. The strip stock from which welded tubing is made can be inspected immediately before forming and welding and can be rejected if it does not meet quality requirements.

Alignment and drilling of tube sheets and tube-support plates to reasonable tolerances will ease tube insertion and will minimize stresses induced during the fabrication of heat exchangers. Overrolling the tube past the tube sheet may introduce residual stresses. Tubes rarely fail within the tube sheet, but failure behind the tube sheet because of overrolling is relatively common. The tube-expander cage must be properly located and the rolling process must be controlled to minimize overrolling.

Effects of Inspection Procedures. All heat-exchanger tubes purchased to ASTM specifications are hydrostatically tested at the mill; the purchaser sometimes specifies that a nondestructive test also be used. Hydrostatic testing is also performed after the heat exchanger has been fabricated. One advantage of not hydrostatic testing at the tube mill is that tubes can be kept dry during shipment and storage. The presence of corrosives in the water used in hydrostatic testing, inadequate drying after hydrostatic testing, and improper boxing of the tubes at the tube mill can lead to corrosion. Most tube mills are aware of the problems associated with hydrostatic testing and drying and take steps to ensure that the test water is free of corrosives and that tubes are dry before boxing.

Some copper alloys are very sensitive to stagnant-water conditions, especially if the water contains biological material or chemicals that can generate or decompose into ammonia. Hydrostatic-testing procedures that permit water to become stagnant in the tubes can result in serious pitting and SCC.

Failure to remove shop-hydrostatic-test water that contained contaminants has caused tubing failures to occur before start-up. The water used for hydrostatic testing of units in the field is also important, as is the method used to drain and dry the units after field hydrostatic testing. The following example describes pitting in ferritic stainless steel heat-exchanger tubes caused by the presence of chlorides in the hydrostatic test water.

Example 1: Pitting of Stainless Steel Heat-Exchanger Tubes Due to Chloride Ions in Flush Water. Many tube bundles in a heat exchanger fabricated from AISI type 410 stainless steel experienced leakage during hydrostatic testing. The tubes had not yet been in service, and no metallurgical deficiencies that could have caused the failure were found.

Investigation. Figure 1 shows a typical example of the pitting observed on the inside surfaces of the tubes. A cross section of a pit (Fig. 2) revealed undercutting characteristic of chloride-ion pitting. X-ray spectrometry of two of the pits revealed evidence of chlorides.

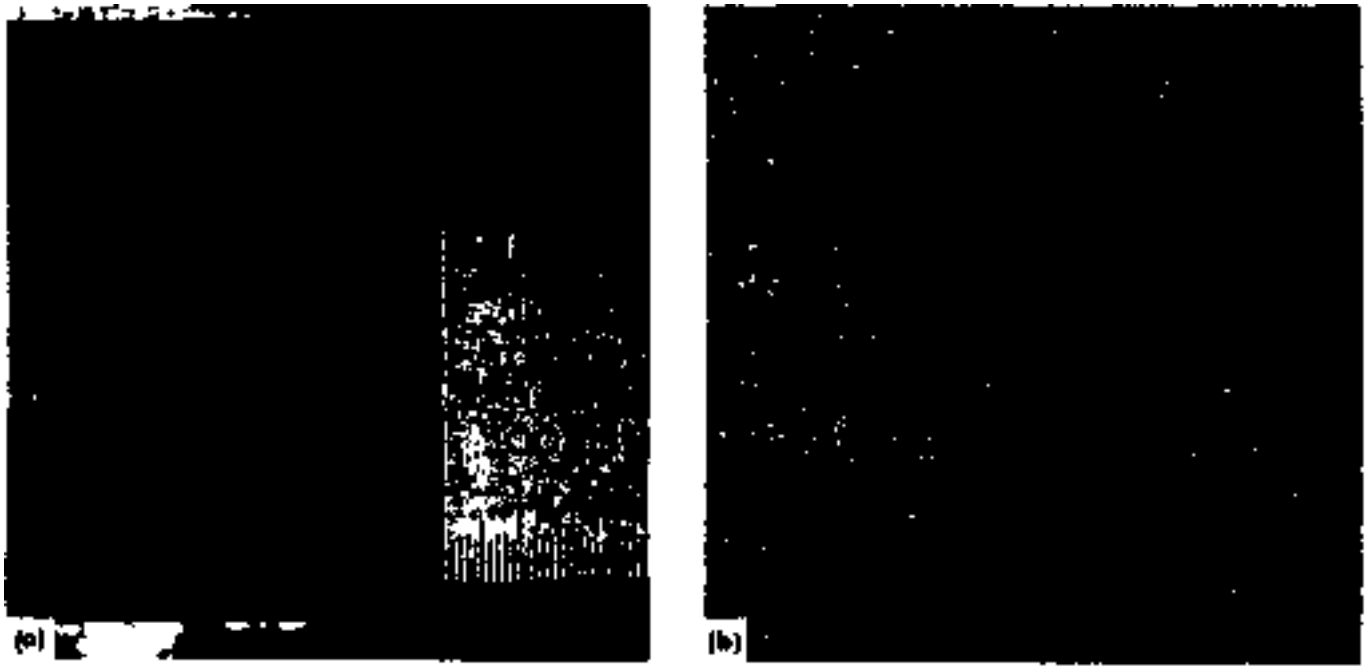


Fig. 1 Pitted inside-diameter surface of AISI type 410 stainless steel tube. (a) Typical example of pitting. Approximately $2\frac{1}{2}\times$. (b) Enlargement of pit shown in (a). Approximately 50×

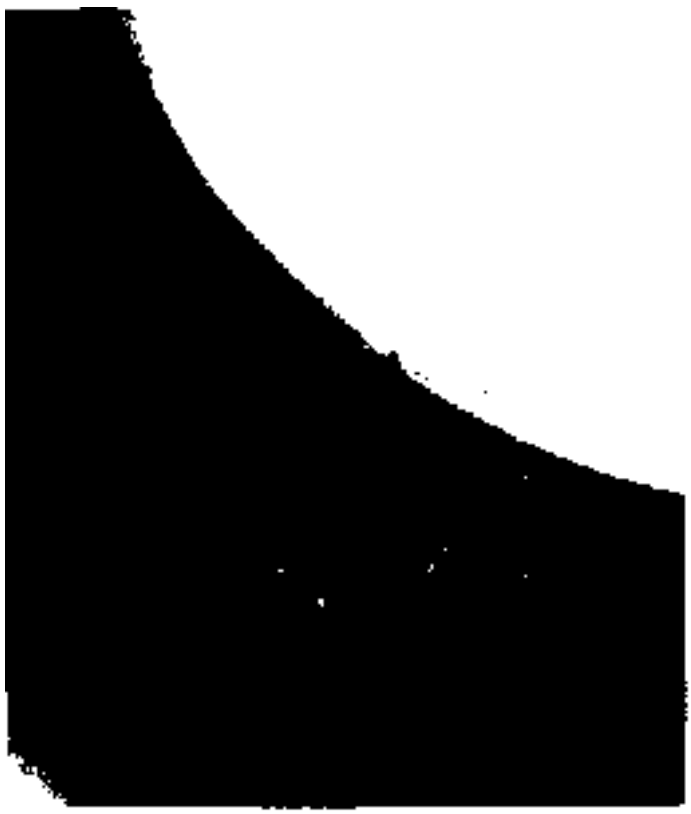


Fig. 2 Cross section of a pit as shown in Fig. 1. 6×

Chemical analysis revealed that the tube material was within specifications for type 410 stainless steel. The microstructure consisted of annealed ferrite free from any inclusions or carbides. Pitting was random and could not be correlated with any sensitization or metallurgical deficiencies in the tubes.

Conclusions. Pitting failures of both 300-and 400-series stainless steels can occur after brief exposures to chloride environments. In this case, failure was caused by chlorides in the water used to flush the tubes before service.

Recommendations. Pitting could be reduced or eliminated by avoiding the use of brackish water to flush or test stainless steel equipment.

Boxing of tubing for shipping should be done with care. All tubes should be wrapped in waterproof paper or plastic for resistance to corrodents that may be present in the air during shipping or that may be present in outdoor storage and may be washed through the tubes by rainwater. It is also possible for preservatives in wooden boxes to leak out and wash over the tubes.

A problem with an airborne corrodent occurred in one case in which packing boxes became contaminated with calcium chloride from road-deicing and dust-control operations. The corrodent washed over stainless steel condenser tubes, causing ferric chloride pitting, which perforated the tubes before they were removed from the boxes. Wrapping with waterproof paper probably would have prevented corrosion of these tubes. In another case, a zinc chloride wood preservative leached out of wooden boxes containing tubes and caused pitting of the tubes before their installation in a heat exchanger. Exposure to ammonia or other atmospheric contaminants in industrial areas could result in damage to stored components.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Corrosion

General corrosion resulting in uniform wall thinning causes considerable damage to heat-exchanger tubes. However, the principal corrosion damage in heat exchangers is usually from localized attack. Stray electrical currents may also contribute to significant corrosion of heat exchangers. Localized corrosive attack can occur as pitting, impingement, thinning, or selective leaching. More information on corrosion is provided in the article “Analysis and Prevention of Corrosion Failures” in this Volume.

General corrosion can be the result of deliberately designing a heat exchanger with a limited life. Otherwise, it can be the result of a mistaken choice of materials or of miscalculation of the corrosive effect of the medium circulating through the heat exchanger. For example, an acid-containing vapor stream may not be corrosive above its dew point, but if the stream is cooled below its dew point, severe general attack can result from acid condensation on internal surfaces of tube walls. On the outside surfaces of tubes, corrosion may be concentrated in the bottom row of tubes or in other areas where condensates can accumulate. Water vapors containing acids, hydrogen sulfide, carbon dioxide, and ammonia can be very corrosive environments.

Another cause of inadvertent general corrosion is improper chemical cleaning using uninhibited acids, excessive temperatures, or prolonged contact time. Water used to wash fire-side deposits where high-sulfur fuel oil or gas was fired will also create very acidic conditions and lead to general corrosion in areas that are not easily drained or flushed, such as under refractory linings.

Example 2: Corrosion Failure of a Tee Fitting. Wet natural gas was dried by being passed through a carbon steel vessel that contained a molecular-sieve drying agent. The gas, which contained 15% H_2S, was passed through the vessel at a pressure of 4.5 MPa (650 psi) and at a temperature of 40 to 45 °C (100 to 110 °F). After several hours, the drying agent became saturated and was taken off the line and regenerated by gas that was heated to 290 to 345 °C (550 to 650 °F) in a salt-bath heat exchanger. Figure 3(a) shows the arrangement of the 75-mm (3-in.) diam schedule 40 piping for the dehydrator system. After 12 months of service, the tee joint failed and a fire resulted.

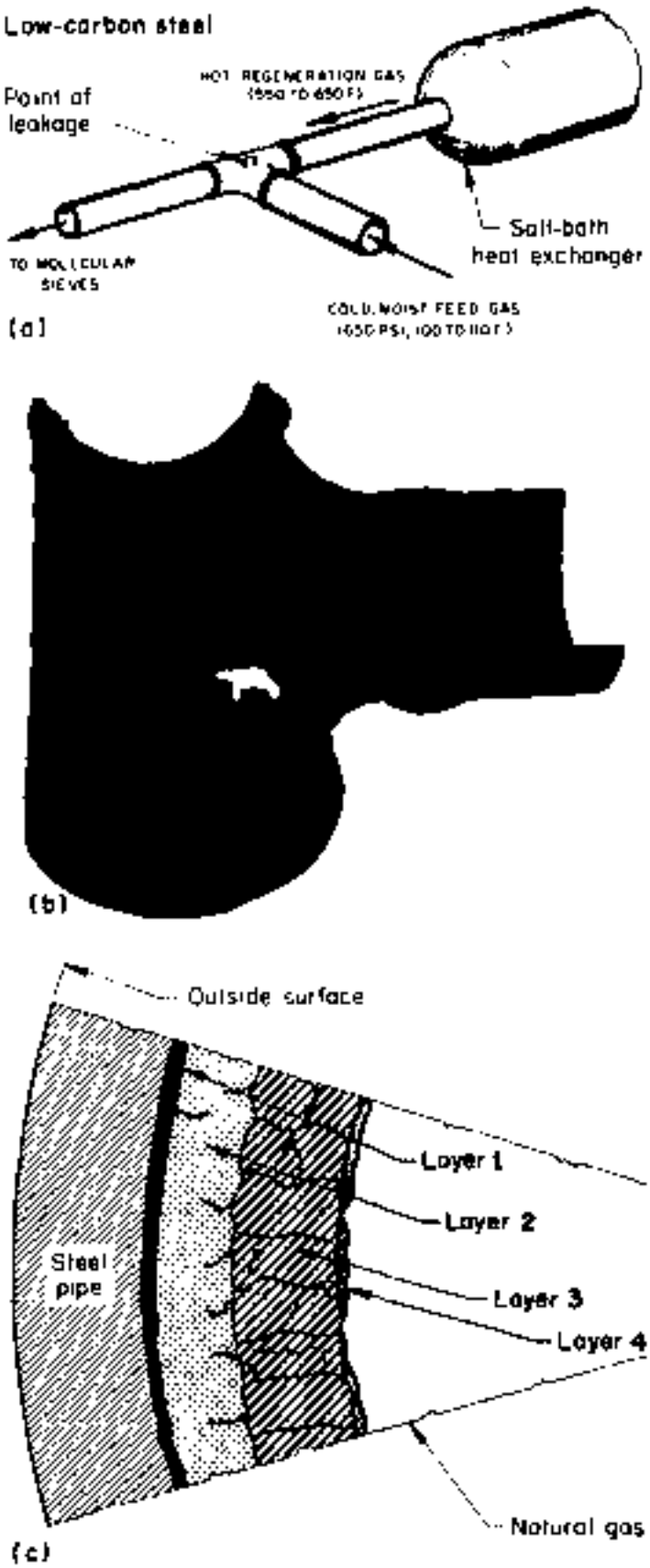


Fig. 3 Low-carbon steel tee fitting in a line leading to a natural-gas dryer that failed from hydrogen sulfide corrosion. (a) Arrangement of piping showing point of leakage in the tee fitting. (b) Inner surface of the tee fitting showing corrosion deposit and area of complete penetration through the tube wall. (c) Positions of layers of corrosion product on inner surface of piping (see Table 2 for compositions of layers)

Investigation. Inspection of the piping between the heat exchanger in the salt bath and the molecular-sieve bed revealed a hole in the tee fitting (Fig. 3a and b) and a corrosion product (scale) on the inner surface of the pitting. This scale occurred in four layers (Fig. 3c).

Chemical analysis of the scale revealed it to be iron sulfide of various compositions in four distinct layers. Table 2 lists the probable composition and thickness of each layer. The layered structure of the scale indicated hydrogen sulfide attack on the carbon steel. The sulfur content was lowest in the layer in contact with the inner surface of the carbon steel pipe.

Table 2 Compositions and thicknesses of layers of corrosion product found on inner surface of low-carbon steel gas-dryer piping

Layer (see Fig. 4c)	Carbon, % [(a)]	Sulfur, % [(b)]	Stoichiometric ratio	Probable composition	Approximate thickness	
					mm	in.
1	10	36.9	$FeS_{1.08}$	$FeS_{1.1}$	0.5	0.02
1 and 2 (composite)	…	41.7	$FeS_{1.15}$	$FeS_{1.1} + FeS_{1.2}$	…	…
2	…	…	…	$FeS_{1.2}$ [(c)]	2.0	0.08
2 and 3 (composite)	6	45.0	$FeS_{1.46}$	$FeS_{1.2} + FeS_2$	…	…
3	…	…	…	$FeS_{1.2} + FeS_2$ [(c)]	2.0	0.08
4	…	…	…	FeS_2 [(c)]	[(d)]	[(d)]

(a) Determined by combustion methods.
(b) Determined by gravimetric methods.
(c) Determined by x-ray diffraction.
(d) Layer too thin and fragile to measure its thickness

For several months before the failure, the temperature of the gas heated by the salt-bath heat exchanger was below 290 °C (550 °F) and was about 230 °C (450 °F) at the molecular sieves.

Corrosion of carbon steel by hydrogen sulfide occurs only at temperatures above 260 °C (500 °F). At 315 °C (600 °F), 15% H_2S would cause rapid attack on carbon steel. However, the carbon and sulfur found in the scale on the piping and in the molecular sieves indicated that oxygen was present in the system. At pressures of 2070 kPa (300 psi) or more, the presence of small amounts of oxygen promotes hydrogen sulfide corrosion of carbon steel, even at room temperature. Thus, oxygen combined with moisture produces conditions for attack of hydrogen sulfide on carbon steel at temperatures below 260 °C (500 °F).

Failure in the wall of the tee fitting was probably the result of turbulence, with some effect from the coarse grain size usually found in pipe fittings. Grain size usually affects corrosion rate.

Conclusions. The piping failed by corrosion in the tee fitting because of the presence of hydrogen sulfide, moisture, and oxygen in the natural gas that was dried in the system. Turbulence in the tee fitting and coarse grain size both contributed to the corrosion.

Corrective Measures. The piping material was changed from carbon steel to AISI type 316 stainless steel, which is readily weldable and resistant to corrosion by hydrogen sulfide at temperatures to 400 °C (750 °F). Chloride concentration in the plant was very low; however, postweld stress relief was used to minimize residual stresses and to avoid the possibility of SCC.

A less expensive alternative would have been to use a 5% Cr steel or a 9% Cr steel, neither of which is susceptible to SCC in the presence of chlorides. These steels are susceptible to corrosion by hydrogen sulfide, although much less so than carbon steel.

Example 3: Pitting of a Condenser Tube in a Saltwater Heat Exchanger Due to Hydrogen Sulfide Contamination. A tube sample from an aluminum brass seawater surface condenser was received for analysis. This condenser had failed due to pitting after less than 1 year of service.

Investigation. Metallographic analysis, energy-dispersive x-ray spectrometry, and x-ray diffraction were used to analyze the tube and deposit samples. A thick, nonuniform black scale was present over the entire inside surface of the tube. When this scale was removed, large pits filled with a green deposit were evident. Figure 4 shows the pitting, the thickness of the scale, and the intergranular nature of the attack. Energy-dispersive x-ray analysis of the base metal indicated that the material was aluminum brass, which is commonly used for seawater condensers. X-ray analysis of the scale, which constituted the bulk of the corrosion deposit, yielded the following results:

Element	Composition, %
Copper	69
Zinc	17
Sulfur	9
Aluminum	2.5

Chlorine	1.5
Calcium	1
Iron, silicon, magnesium	trace

The black deposit was identified as primarily copper sulfide, with zinc and aluminum sulfides also present. The green deposit was also analyzed and identified as copper chloride. Copper, chlorine, zinc, and aluminum were the major constituents.

Fig. 4 Failed aluminum brass condenser tube from a saltwater heat exchanger. The tube failed from pitting caused by hydrogen sulfide and chlorides in the feedwater. (a) Cross section of tube showing deep pits and excessive metal wastage. $2\frac{3}{4}\times$. (b) Higher magnification view of a pit showing the depth of attack and thickness of the corrosion deposit. 28×. (c) Micrograph showing the intergranular nature of the attack and undercutting of metal particles. Etched with ammonium persulfate. 110×

Discussion. Most copper alloys depend on the formation of a protective oxide scale for corrosion protection. When hydrogen sulfide is present, it interferes with the formation of such a layer, forming a sulfide layer instead. Unlike the oxide layer, the sulfide scale is not protective, and it allows corrosion, as well as diffusion and concentration of such corrosive species as chlorides, to continue beneath the deposit.

Conclusions. The combination of sulfide and chloride attack on the tubes resulted in accelerated failure. The major cause of the shortened tube life was hydrogen sulfide, which was present in seawater at this particular site at about 5 to 10 ppm.

Recommendations. Injection of ferrous sulfate just upstream of the condenser can aid the formation of protective oxide films. From the preceding discussion, the most obvious corrective measure would be to eliminate sulfide contamination. The prime source of such contamination appears to be the accumulation of sulfide-containing sludge in coral formations near the condenser inlets. Consideration might also be given to relocation of the pump suction to an area where there is continuous entry of fresh seawater, which would reduce sulfide concentrations.

Crevice Corrosion in Water. A common cause of deterioration in tubing containing water is corrosion pitting caused by differential aeration or concentration-cell corrosion (crevice corrosion). This type of corrosion occurs frequently, but is often not identified as crevice corrosion.

Corrosion by differential aeration arises from regional differences in the concentration of dissolved gases (principally oxygen) or of soluble contaminants, such as chlorine, fluorine, or sulfur, in water that contacts a part or component. The result is the establishment of an electrochemical cell, with the surface in the region of lowest concentration becoming anodic relative to the surrounding surface. This situation can develop under conditions of stagnant flow between two overlapping surfaces or beneath corrosion products or sludge on the surface. Anything that restricts free access of oxygen is likely to cause crevice corrosion.

Other factors, such as gas bubbles on the surfaces, inhomogeneity of the material, cracks in mill scale, and local breakdown of protective films, may also initiate crevice-corrosion attack. Corrosion of this type is often associated with water that contains dissolved solids in comparatively small concentrations, the amounts present being insufficient or unsuitable to cause formation of protective films or deposits but sufficient to raise the electrical conductivity of the water so that it becomes an electrolyte. Invariably, the cathodic zone is larger than the anodic zone; consequently, the rate of attack at the small anodic zone is high. Corrosion extends preferentially in depth, forming pits and cavities. The affected part sometimes becomes completely penetrated even though the total area affected by corrosion is relatively small. The corrosion products commonly form a nodule over the cavity, increasing the crevice effect. When the corrosion is associated with elevated temperature—for example, at the hot-test parts of heat-exchanger surfaces—the nodule that is formed is hard and is difficult to remove until it has attained considerable size. The temperature may be sufficient to

evaporate some of the liquid within the nodule; the corrosive action then becomes intermittent, and partial drying of the corrosion products occurs.

In the corrosion of iron-base alloys, dissolved oxygen in the water diffuses into the precipitated ferrous hydroxide and, in combination with the ferrous hydroxide and water, forms hydrated ferric hydroxide (rust). If the supply of oxygen is limited, the corrosion product may be green hydrated magnetite or black anhydrous magnetite. The nodule that forms from the corrosion product is porous; therefore, electrochemical attack proceeds as iron enters into solution in the aqueous medium beneath the nodule. The oxygen dissolved in the water is mainly consumed over the large cathodic surface in oxidizing the hydrogen evolved there; that is, it acts as a depolarizer, with the rate of oxygen consumption determining the rate at which ions leave the metal beneath the nodule.

Because of the intense corrosive attack, the corrosion nodule frequently has a stratified structure, with the composition changing from rust-colored ferric hydroxide on the outside surface of the nodule to layers of ferrous hydroxide, ferric oxide, and magnetite on the inside. The ferric oxides permit the passage of hydroxide ions and the diffusion of oxygen, with the result that electrochemical action is able to proceed, and the rate of attack is increased with an increase in temperature because of the resultant alteration in the nature of the corrosion products, the improved electrical conductivity of the water, and the enhanced rate of chemical reaction that follows an increase in temperature.

The following features are characteristic of crevice-corrosion attack:

- Nodules will generally form on the corroded surfaces. The nodules will be hard if formed on a hot surface, but may be softer if formed on a cold surface
- The composition of the nodules varies from the outside surface to the inside. Externally, they often present a rusty appearance. Internally, they are mostly black. In some cases, the appearance may be modified by the presence of salts deposited from the water, giving various colors
- The nodules are generally isolated from one another, although they may sometimes be found in confluent groups
- When a nodule is removed, the cavity in the plate beneath it is often sharply defined, and the walls may be almost perpendicular to the metal surface
- The metal surface between the cavities generally shows little or no corrosion, and in some cases the original mill scale may be present

New installations are particularly vulnerable to crevice corrosion because of insufficient time for a protective film or scale to form on the surface before the onset of corrosion.

Copper alloys are also susceptible to crevice corrosion, forming similar corrosion products and taking part in similar chemical reactions. For example, in the corrosion of a copper alloy in an aerated-water medium, copper carbonate hydroxide [$CuCO_3 \cdot Cu(OH)_2$] would be the expected corrosion product. Bicarbonate (HCO_3) and carbonate (CO_3) radicals, which are needed for the formation of copper carbonate hydroxide, are present as contaminants in most naturally occurring waters.

Example 4: Crevice Corrosion of Tubing in a Hydraulic-Oil Cooler. Leakage from the horizontal heat-exchanger tubes in one of two hydraulic-oil coolers in an electric-power plant occurred after 18 months of service. Under normal operating conditions, service life in similar units is generally about 10 years. In this plant, river water was used as the coolant in the heat-exchanger tubes. The tubes were 9.5 mm ($\frac{3}{8}$ in.) in diameter with a 0.6-mm (0.025-in.) wall thickness and were made of copper alloy C70600 (copper nickel, 10%). Five of the tubes that were leaking were removed and sectioned lengthwise. One section of each tube was sent to the manufacturer of the cooler, and the other was sent to the power company laboratory for determination of the cause of failure.

Investigation. Visual examination of the sectioned tubes revealed several nodules on the inner surface and holes through the tube wall, which appeared to have formed by pitting under the nodules. In Fig. 5(a), arrow A shows a pit that penetrated the wall of the tube, and arrow B shows a typical nodule.

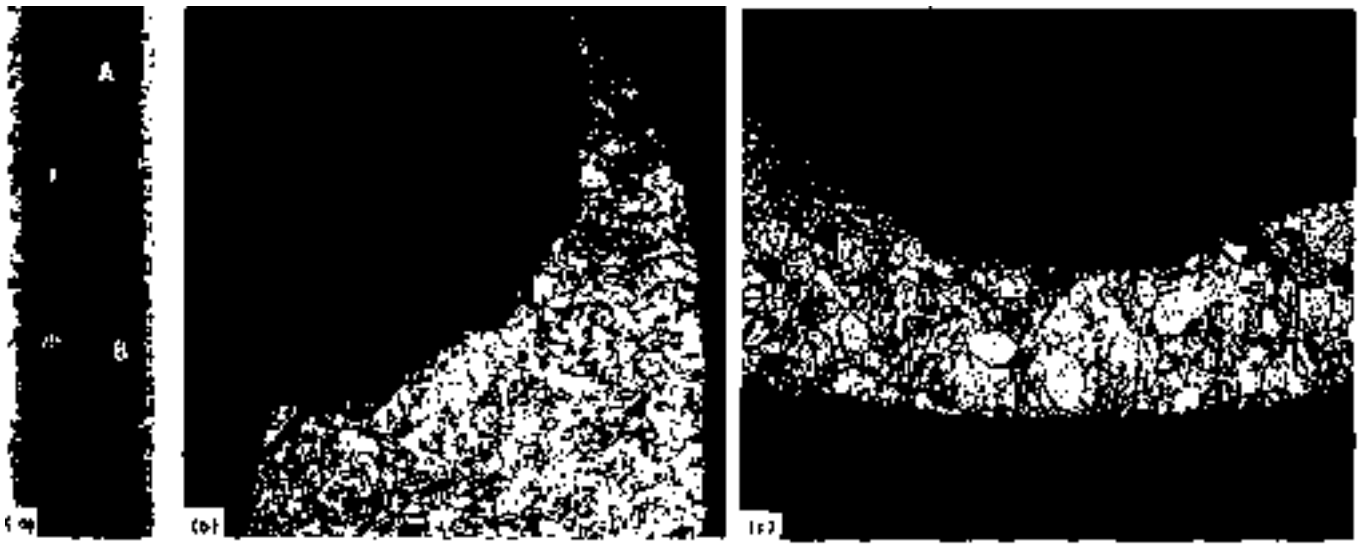


Fig. 5 Copper alloy C70600 tube from a hydraulic-oil cooler. The cooler failed from crevice corrosion caused by dirt particles in river water that was used as a coolant. (a) Inner surface of hydraulic-oil cooler tube containing a hole (arrow A) and nodules (one of which is indicated by arrow B) formed from a corrosion product. (b) Etched section through the pit at arrow A, which penetrated the tube wall. 100×. (c) Etched section through a pit beneath the nodule at arrow B, with about 35% of the wall thickness remaining. 100×

Metallographic examination was performed on three distinct areas of the tube: areas that exhibited no effects of corrosion, areas that were pitted, and areas containing nodules. The tubing contained several nodules and pits, but all were isolated from each other. A cross section through an unaffected area of the tube showed no corrosion on the inner surface and an undamaged microstructure. Most of the inner surface of the tube was in good condition.

Figure 5(b) shows a micrograph of a section through a pit that had penetrated the tube wall. The pit had steep sidewalls, which indicated a high rate of attack. A micrograph of a section through the tube beneath one of the nodules is shown in Fig. 5(c). Corrosion beneath the nodule had penetrated approximately 65% of the tube wall. The pit had a low slope angle, which indicated a low or intermittent rate of corrosion.

Spectrographic analysis of the tube metal indicated that it was copper alloy C70600. A reddish deposit was removed from the inner surfaces of the tubes and analyzed. The major constituent was iron oxide, and less than 1% was manganese dioxide. The presence of these and other constituents indicated that effluent from steel mills upstream was the source of most of the solids found in the tubes.

A greenish nodule was analyzed by x-ray diffraction. The major compound was copper carbonate hydroxide [$CuCO_3 \cdot Cu(OH)_2$]. This compound would be expected in corrosion of a copper alloy exposed to aerated natural water.

Flow of cooling water from the river through the tubes was not uniform, but ebbed and flowed in accordance with cooling demand. Under these circumstances, deposits would form on, or particles would drop on, the inner surfaces of the horizontal tubes.

Conclusions. The tubing failed by crevice corrosion. Particles of dirt in the river water were deposited inside the tubes during periods of low flow. The difference in oxygen content under the deposit and on the exposed surface of the tube established an electrolytic cell under the deposit. The end results were nodules of a corrosion product and pits that in some cases eventually became holes in the tube wall.

Corrective Measures. The tubing in the cooler was replaced, and the cooling-water supply was changed from river to city water, which contained no dirt to deposit on the tube surfaces. After the first cooler was back in service, the second cooler was shut down, and the tubes removed for investigation. Nodules, pits, and deposits similar to those found on tubes from the first cooler were found on the inner surfaces of the tubes. Although the tubes were not leaking, they would have been within a short time. For this cooler also, the tubes were replaced, and city water was substituted for river water as the coolant. At the time of this report, both coolers had been in service for about 3 years with no evidence of leakage.

An alternate solution to the problem would have been to install replacement tubes in the vertical position and to continue to use river water. With the tubes in the vertical position, solids would be less likely to be deposited on the tube walls during low flow of the river water.

Selective leaching, or dealloying, is localized corrosion of copper alloys that leaves a spongy, structurally weak mass of the more noble alloying element in place at the site of corrosion attack. Selective leaching can occur in brasses

(dezincification) and in copper-nickel alloys (denickelification). The net result of selective leaching is that sound metal is gradually changed to a brittle, porous mass of copper that has little mechanical strength. Selective leaching can sometimes be detected visually by the reddish appearance of the deposited copper; it is favored by waters having a high content of oxygen, carbon dioxide, or chloride and is accelerated by elevated temperatures and low water velocities. Excessive chlorination of cooling water may also lead to dezincification of brass.

Dezincification occurs in two forms, plug type and layer type. Plug-type dezincification is less spread out on the surface than layer type and can cause pitlike wall perforation more rapidly than layer type. An acidic environment favors layer-type attack. High-zinc brasses of the α variety, such as copper alloy C26000 (cartridge brass, 70%), are more susceptible to dezincification than copper alloys having a lower zinc content, for example, copper alloy C24000 (low brass, 80%). In copper alloys that show an α-plus-β structure, such as copper alloy C28000 (Muntz metal, 60%), the β constituent, which has the higher zinc content, may suffer dezincification.

Figure 6 shows a section through a copper alloy C26000 steam-turbine condenser tube. Uniform dezincification of the bore extended from one-half to two-thirds through the wall of the tube. The porous nature of the affected portion of the tube is shown in Fig. 6(b).

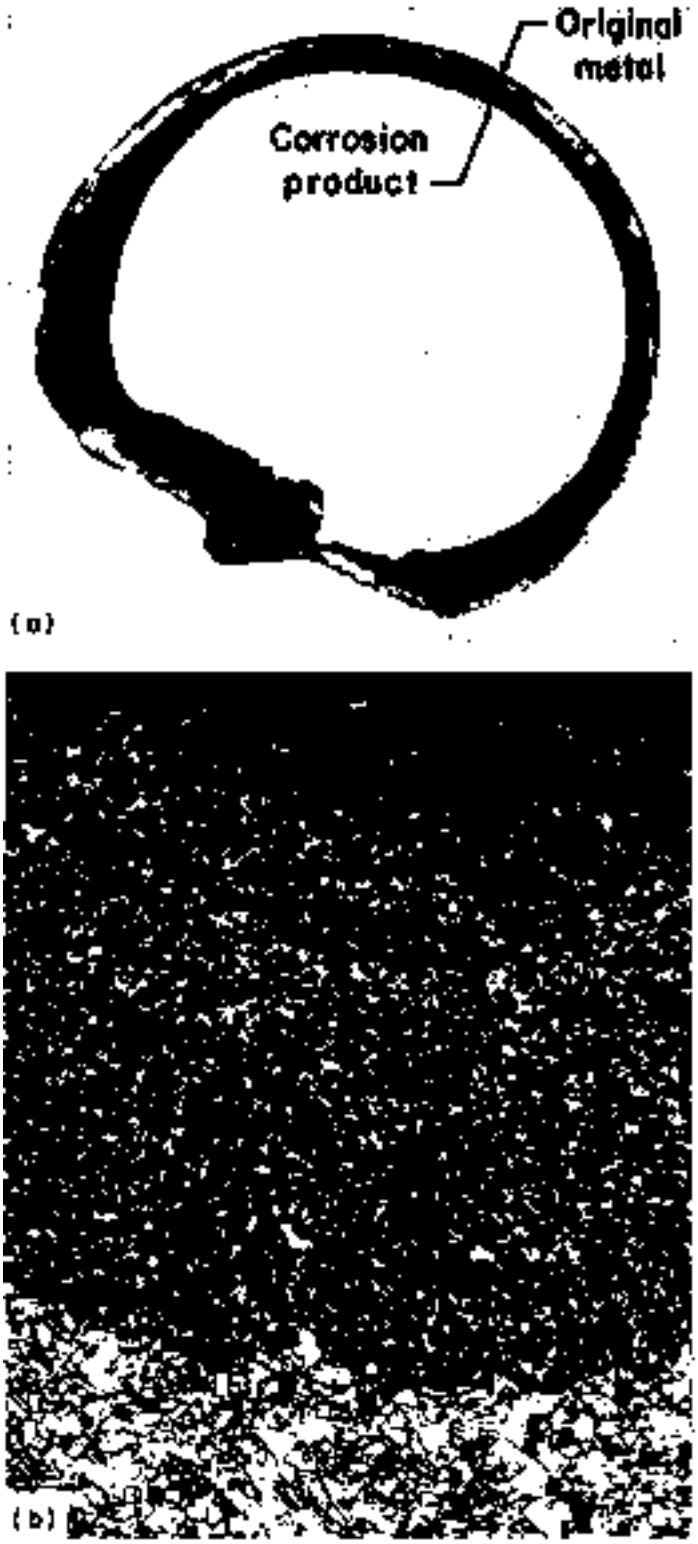


Fig. 6 Copper alloy C26000 steam-turbine condenser tube that failed by dezincification. (a) Section through condenser tube showing dezincification of inner surface. $3\frac{1}{2}\times$. (b) Etched specimen from the tube showing corroded porous region at the top and unaffected region below. 100×

Figure 7(a) and (b) show a section through a heat-exchanger tube made of copper alloy C71000 (copper nickel, 20%). Water passing through the tube and steam in contact with the outer surface caused denickelification of the alloy at both inner and outer surfaces. The dealloying occurred along the grain boundaries, which have higher energy than the grains and are more susceptible to attack. Another form of denickelification is shown in Fig. 7(c) for a copper alloy C71500 (copper nickel, 30%) heat-exchanger tube.

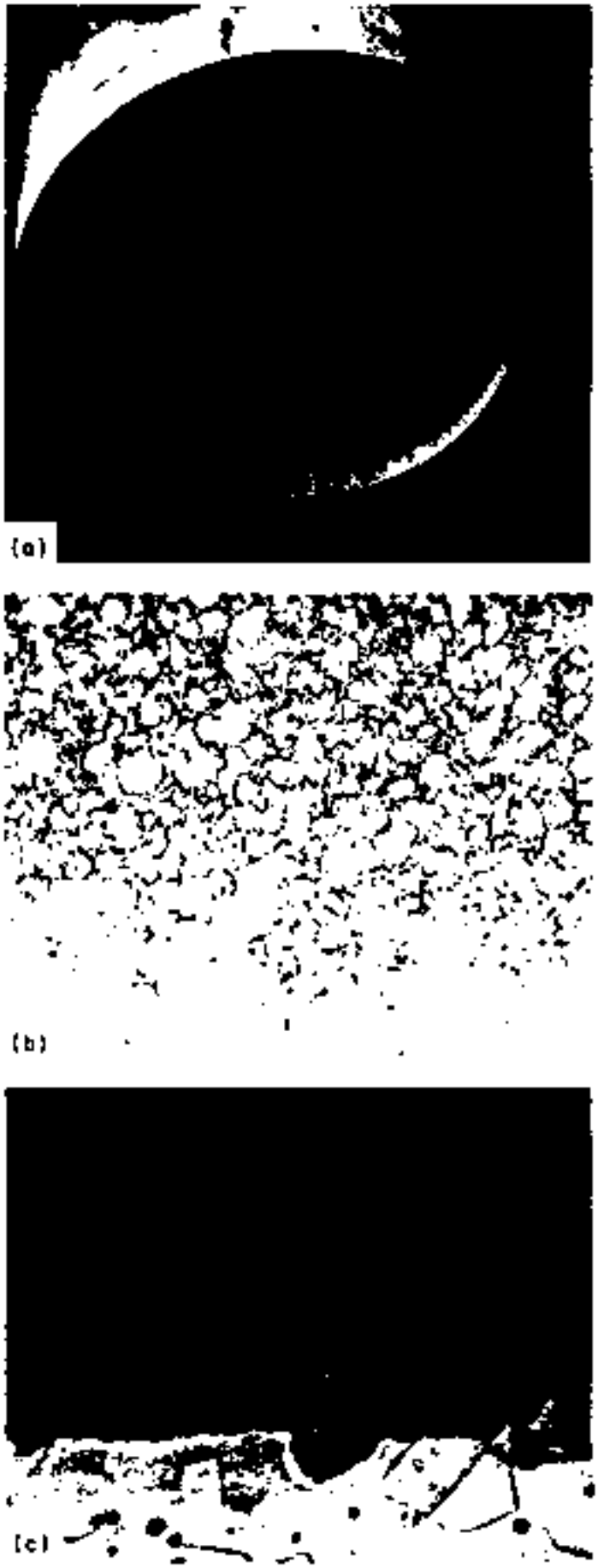


Fig. 7 Copper-nickel alloy heat-exchanger tubes that failed from denickelification due to attack by water and steam. (a) Etched section through a copper alloy C71000 tube showing dealloying (light areas) around the tube surfaces. Etched with NH_4OH plus H_2O.

3.7×. (b) Unetched section through the outer surface of the tube shown in (a); dark pattern shows dealloyed region. 85×. (c) Etched section through a copper alloy C71500 tube showing the surface that was denickelified (top) and unaffected base metal (bottom). Etched with equal parts nitric and acetic acid. 375×

Impingement attack, or erosion-corrosion, occurs where gases, vapors, or liquids impinge on metal surfaces at high velocities. The erosive action removes protective films from localized areas at the metal surface, thereby contributing to the formation of differential cells and localized pitting of anodic areas.

Impingement attack often produces a horseshoe-shaped pit. This form of attack frequently occurs in condenser systems handling seawater or brackish water that contains entrained air or solid particles and circulates through the system at relatively high velocities and with turbulent flow. The pits are elongated in the direction of flow and are undercut on the downstream side. When the condition becomes serious, a series of horseshoe-shaped grooves with the open ends on the downstream side may be formed. As the attack progresses, the pits may join one another, forming fairly large patches of undercut pits.

When impingement attack occurs in heat-exchanger or condenser tubing, it is usually confined to a short distance on the inlet end of the tube where the fluid flow is turbulent. Impingement attack can be controlled by use of corrosion-resistant alloys, such as copper alloy C71500 (copper nickel, 30%).

Impingement attack has been somewhat relieved by inserting relatively short tapered plastic or alloy sleeves inside the tube on the inlet end of the tube bundle. Other corrective measures include decreasing the velocity, streamlining the fluid flow, and decreasing the amount of entrained air or solid particles. These may be attained by a streamlined design of water boxes, injector nozzles, and piping. Abrupt angular changes in direction of fluid flow, low-pressure pockets, obstruction to smooth flow, and any other feature that can cause localized high velocities or turbulence of the circulating water should be minimized as much as possible. Sacrificial impingement baffles can minimize damage to the shell-side surfaces of tubes. More information on liquid impingement is provided in the article “Liquid Impact Erosion” in this Volume.

Example 5: Failure of Copper Alloy 443 Heat-Exchanger Tubes. Tubes in heat exchangers used for cooling air failed in an increasing number after 5 to 6 years of service. Air passed over the shell-side surface of the tubes and was cooled by water flowing through the tubes. Although sanitary well water was generally used, treated recirculating water was also available and was sometimes used in the heat exchanger. Water vapor in the air was condensed on the tube surfaces during the cooling process. The pH of this condensate was approximately 4.5. Air flow over the tubes reversed direction every 585 mm (23 in.) as a result of baffling placed in the heat exchangers.

The tubes were 19 mm ($\frac{3}{4}$ in.) in diameter with a wall thickness of 1.3 mm (0.050 in.) and were made of copper alloy C44300 (arsenical admiralty metal). Samples of the tubes were sent to the laboratory to determine the mechanism of failure.

Investigation. Visual examination of the damaged area revealed an uneven ridgelike thinning and perforation of the tube wall (Fig. 8a) on the leeward side of the tube. An enlarged view of the ridges on the surface of the tube is shown in Fig. 8(b). Metallographic examination of a cross section of the failed area (Fig. 8c) disclosed undercut pits on the outer surface of the tube. Both the ridgelike appearance of the damaged area and the undercut pitting indicate of impingement attack. Spectrochemical analysis of the tube confirmed that the material was copper alloy C44300 (arsenical admiralty metal).

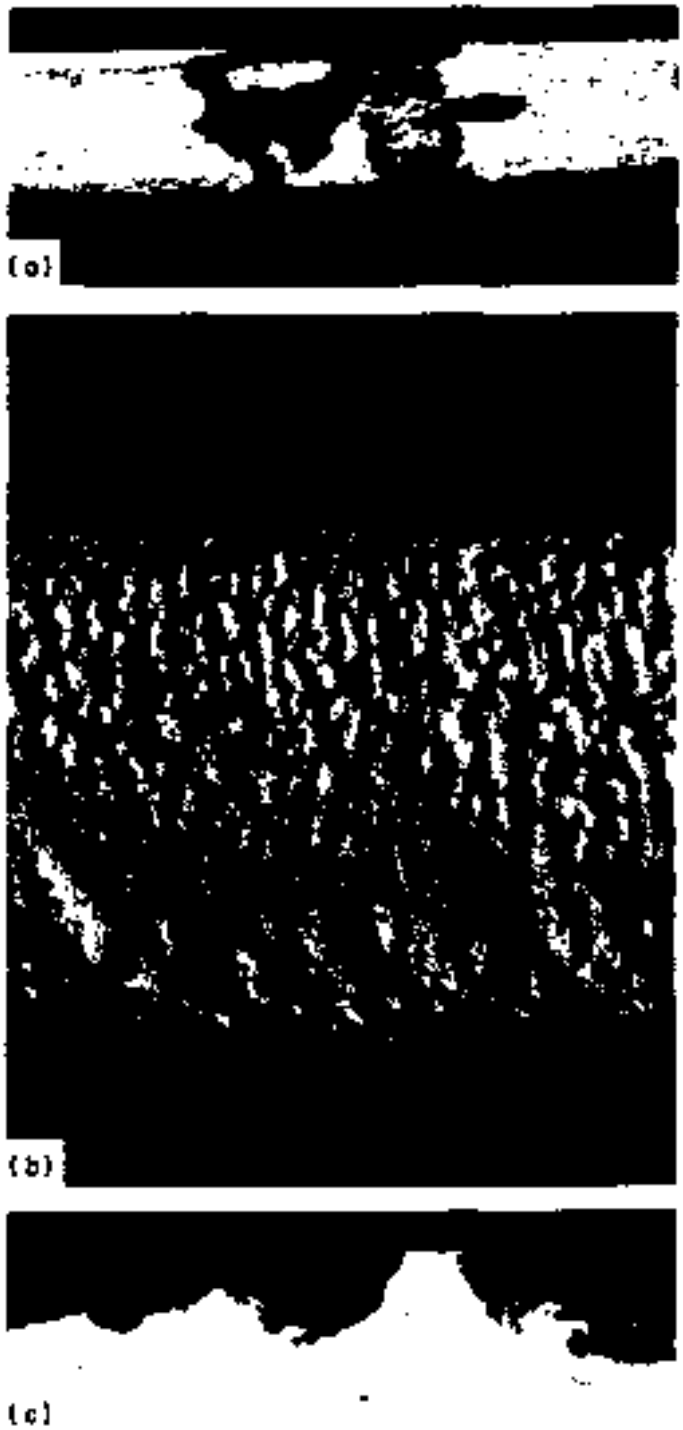


Fig. 8 Copper alloy C44300 heat-exchanger tube that failed by impingement corrosion from turbulent flow of air and condensate along the shell-side surface. (a) Shell-side surface of tube showing damaged area. (b) Damaged surface showing ridges in affected area. 4×. (c) Unetched section through damaged area showing undercut pitting typical of impingement attack. 100×

Conclusion. The tubes failed by impingement corrosion that resulted in perforation. A combination of air and water (the condensate) was rapidly agitated by changes in direction at the metal surface and accelerated the impingement corrosion. This action and the low pH (4.5) of the condensate resulted in removal of the protective film that normally formed on the tube surface during initial corrosion and ultimately produced a hole in the tube.

Corrective Measures. The heat exchanger was retubed with tubes made of aluminum bronze, for example, copper alloy C61400. At the time of this report, these tubes had been in service for more than 12 years without failure.

Revised by Robert J. Franco, Exxon Company, U.S.A.

Stress-Corrosion Cracking

Stress-corrosion cracking (SCC) occurs as a result of corrosion and stress at the tip of a growing crack. The stress level necessary for SCC is below that needed for fracture without corrosion. Additional information is provided in the article "Stress-Corrosion Cracking" in this Volume.

Evidence of corrosion, although not always easy to find, should be present on the surface of an SCC fracture up to the start of final rupture. Generally, one of several specific corrosive substances must be present for SCC to occur in a given alloy. For example, ammonia compounds have been identified as agents that produce SCC in copper-zinc alloys.

The arsenic, antimony, or phosphorus additions for inhibiting dezincification do not prevent SCC of inhibited admiralty metals (copper alloys C44300, C44400, and C44500). Cracking can be transgranular, intergranular, or mixed on the same metal part, and it may be accompanied by deposition of copper within the cracks. Arsenic does not prevent SCC in copper alloy C68700 (arsenical aluminum brass) subjected to ammonia- and copper-dosed seawater, although arsenic can have beneficial effects at stresses below the elastic limit. Low-temperature SCC has been found in brasses. The copper nickels (copper alloys C70100 to C72900), despite their proven susceptibility to ammonia, are much superior in service with respect to SCC in condenser-tube service than brasses are.

Carbon steels can experience SCC in several environments (see the article "Stress-Corrosion Cracking" in this Volume). In heat exchangers, these environments may be present at such low concentrations that they appear to be harmless. However, concentration of cracking species by evaporation can lead to equipment failure. For example, sodium hydroxide may be present in boiler water in part per million concentrations. Waste-heat boilers can allow concentration of hydroxide ions by steam blanketing, by incomplete water wetting of a vertical boiler, or by flange leaks that allow water to become steam. Failure usually occurs in welds or in highly stressed areas. Carbon steel heat-exchanger equipment can also fail by SCC in wet hydrogen sulfide environments if the hardness of the component exceeds approximately 22 HRC. Flange bolts or studs that are improperly heat treated for service in such an environment are particularly susceptible. Additional information is available in the articles "Hydrogen Damage and Embrittlement" in this Volume.

The austenitic stainless steels are subject to SCC that is induced by acidic chlorides and fluorides or by polythionic ($H_2S_xO_6$; x = 2 to 6) acids if the microstructure is sensitized. In hot systems in which there is a possibility of chloride contamination, austenitic stainless steels should not be considered. If these steels must be used, they should be in the annealed condition or should be clad over a base metal that resists cracking. In very corrosive environments, higher-nickel austenitic alloys may be required to resist both general attack and SCC. In less corrosive environments, ferritic or duplex stainless steels can be considered.

Example 6: SCC of a Stainless Steel Integral-Finned Tube. A tubular heat exchanger in a refinery reformer unit was found to be leaking after 1 month of service. The exchanger contained 167 type 304 stainless steel U-bent integral-finned tubes, 19 mm ($\frac{3}{4}$ in.) in outside diameter.

Investigation. Examination revealed cracks in the tube wall about 75 mm (3 in.) from the tube sheet and at the first full-depth fin on the tube (Fig. 9a). Hardness of the tube was 30 HRC at the inside surface and up to 40 HRC at the base of the fin midway between the roots. These hardness values indicated that the fins were cold formed and not subsequently annealed. Cold working without subsequent annealing made the tubes susceptible to SCC because of a high residual-stress level.

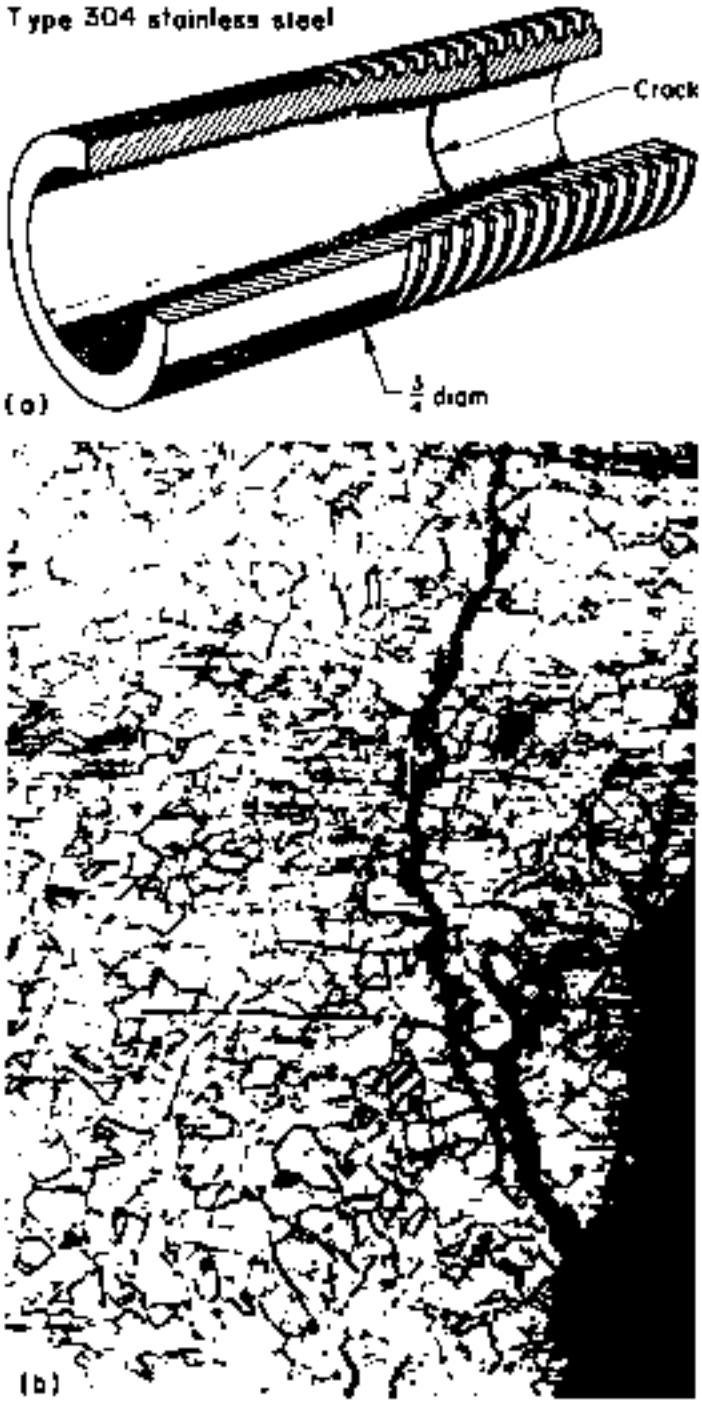


Fig. 9 Type 304 stainless steel integral-finned tube that cracked from chlorides and high residual stresses. (a) Section of integral-finned tube showing major crack (circumferential crack between fins). Dimension given in inches. (b) Branched transgranular cracking propagating from major crack. 45×

Metallographic examination of a section through the tube wall established that fracture was predominantly by transgranular branched cracking (Fig. 9b) and had originated from the inside surface. Transgranular branched cracking of austenitic stainless steels is frequently caused by chlorides in the presence of high residual stresses.

The presence of chlorides is not unusual in a reformer unit. Although steam and steam condensate injected into reformer systems are normally free of chlorides, there can be accidental carry-over of dissolved solids with the steam under periods of high demand.

Conclusion. The tubes failed in SCC caused by chlorides in the presence of high residual stresses. The fins were cold formed on the tubes and not subsequently annealed.

Corrective Measure. The finned tubes were ordered in the annealed condition. This minimized the possibility of residual stresses in the tubes.

Improper Materials Specification and Design. Heat exchangers sometimes fail by SCC because of improper specification of materials and design. In one case, a heat exchanger was part of a separator, which was designed as a single combination vessel for construction economy. The exchanger was of a vertical-tube design and was located directly beneath the separator. The separator had a 6.4-mm ($\frac{1}{4}$-in.) thick AISI type 316 stainless steel shell that extended below the partition head of the vessel, forming the shell of the heat exchanger. The designer had not provided for complete filling of the water space between the upper tube sheet and the partition head. As a result, the type 316 stainless steel shell of the exchanger was constantly wetted with water in the vapor space below the partition head. Heat from the separator caused continuous flashing of water into steam inside the exchanger, building salt deposits on the hot surfaces of the shell. The deposits contained chlorides from the city water used in the exchanger. Acting in combination with the normal fabricating stresses inherent in the vessel, chloride salts caused the type 316 stainless steel shell at the vapor space to fail because of the presence of numerous small stress-corrosion cracks. Some of the cracks penetrated completely through the wall of the shell.

This problem could have been avoided by providing for complete filling of the water space above the tube sheet or by providing sufficient distance between the separator and exchanger so that heat from the separator would not cause steam flashing inside the exchanger. Finally, it is not considered good practice to use austenitic stainless steels in hot-water service under conditions of alternate wetting and drying; the drying favors the accumulation of corrosive salt deposits.

As a corrective measure, carbon steel doubler plates were welded over the cracked region of the exchanger shell. The exchanger was eventually replaced with a vessel designed to eliminate the vapor space between the tube sheet and the partition head.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Corrosion Fatigue

Corrosion fatigue is the combined action of an aggressive environment and a cyclic stress, which results in premature failure of metals by cracking. Neither cyclic stress in air nor environmental attack applied separately produces the same damaging results as combined action. Thus, a precorroded specimen does not necessarily show appreciable reduction in fatigue life, nor does prefatiguing in air increase the corrosion rate of metals.

In general, corrosion fatigue should not be confused with stress corrosion. All metals that are susceptible to corrosion are susceptible to corrosion fatigue in any corrosive environment, whereas SCC occurs only in alloys subjected to very specific environments and normally occurs under conditions of static tensile stress rather than dynamic stress. However, the application of cyclic stresses under SCC conditions may also influence corrosion-fatigue behavior.

In general, more materials are susceptible to corrosion fatigue than to stress corrosion. Stress-corrosion cracking mainly affects the brasses and the austenitic stainless steels, but corrosion fatigue can occur in most metallic materials used in the chemical industry.

Two general sources of stress and environment are necessary for corrosion fatigue: (1) where the cyclic stress and corrosive environment result from the same sequence of events and (2) where the cyclic stress and corrosive environment arise independently. The first source is more common in high-temperature applications and relies on changes in temperature that produce thermal stresses and corrodent concentration. In the second source, cyclic stresses are usually mechanical in origin, such as rotating bending, cyclic pressurization, and start-stop centrifugal loading. Because corrosion fatigue is by definition restricted to the simultaneous action of a cyclically varying stress and a corrosion process, the damage mechanisms not covered by this term are purely mechanical stress, followed by a corrosive action in the unstressed state, and corrosion in the nonloaded state, followed by purely mechanical cyclic stressing without interaction of corrosion.

Fracture Appearance. For carbon and low-alloy steels, the corrosion-fatigue fracture surface is typically jagged, whereas the fracture surface of rust- and acid-resisting steels—for example, austenitic stainless steels—is usually flat. The number of secondary cracks influences whether the fracture surface is flat or jagged; with a great number of cracks, the fracture is jagged. The jagged surface results when several cracks penetrate, very close to one another, into the interior of the material. Because the cracks are not in the same plane, the fracture jumps from one crack plane to another, causing the jagged appearance. When the distance between each crack is large, which is usually the case with a small number of cracks, the fracture cannot jump from one plane to another. Consequently, the crack that penetrates to the greatest depth into the material continues to grow until the component fails. The flat form of fracture is, therefore, determined by the number of cracks per unit length and is not a specific property of austenitic steels.

On the other hand, the form of a fracture in an austenitic corrosion-resistant steel will be similar to that of a carbon steel if the rate of corrosion is increased. Tests carried out with an austenitic stainless steel that was exposed to cyclic loading and a spray of saturated sulfurous acid showed severe corrosion conditions, producing numerous surface cracks and a jagged fracture surface.

The number of cracks, which considerably influences the form of fracture, depends not only on the rate of corrosion but also on the state of the surface from which the crack begins to propagate. On rough surfaces, many cracks are formed, resulting in a jagged fracture surface. With alloys that are not corrosion resistant, the jagged form of fracture becomes more characteristic when the mechanical stress decreases. With a small stress, the life of a part is long; consequently, there is enough time for the corrosion to roughen the surface. Thus, the conditions for the formation of many cracks and a jagged fracture are created.

Thermal Stresses. Thermally generated cyclic stresses are as important in corrosion fatigue as mechanical stresses. Transient thermal conditions can arise when a fluid is brought into contact with a surface having a markedly different temperature or, more commonly, due to imposed system operations. Such thermal stresses result from constrained thermal expansion or contraction.

Mechanical Stresses. Mechanical cyclic stresses arise from rotating machinery and cyclic pressurization. Vessels subject to cyclic pressurization are usually penetrated by pipes welded to the vessel shell. In addition to being regions of stress concentration, these welded connections can act as sources of residual stresses, and any lack of penetration or fusion at the weld may be a site for environmentally assisted cracking. Mechanical cyclic stresses superimposed on any residual stresses and any additional thermal stresses resulting from operational transients are similarly additive. Other significant sources of cyclic stresses are fans or pumps, which generate high-frequency vibrations in coolant circuits; turbulent fluid flow through or over tubes, such as steam flow over condenser tubes; and such upsets as water hammer in steam-generating equipment (see the article “Failures of Boilers and Related Equipment” in this Volume).

Example 7: Corrosion-Fatigue Failure of U-Bend Heat-Exchanger Tubes. Two admiralty brass heat-exchanger tubes from a cooler in a refinery catalytic reforming unit were sent to the laboratory for analysis. The tubes had been in service approximately 10 years and had cracked circumferentially in the area of U-bends in the tubes.

Investigation. Both tubes showed cracks extending circumferentially about 180° on the tension (outer) side of the U-bends. Tube 1 exhibited a relatively smooth, uniform black deposit on its surfaces; tube 2 had an additional buildup of corrosion deposits on both the inside- and outside-diameter surfaces. Inspection data indicated that the samples received were typical of many cracked U-bends through the tube bundle.

Metallography of the fracture area of tube 1 showed blunt transgranular cracking with minimal branching propagating from the inside surface of the tube (Fig. 10). In addition to the large crack, several smaller cracks were found near the fracture. The appearance of these cracks was typical of a corrosion-fatigue mechanism.

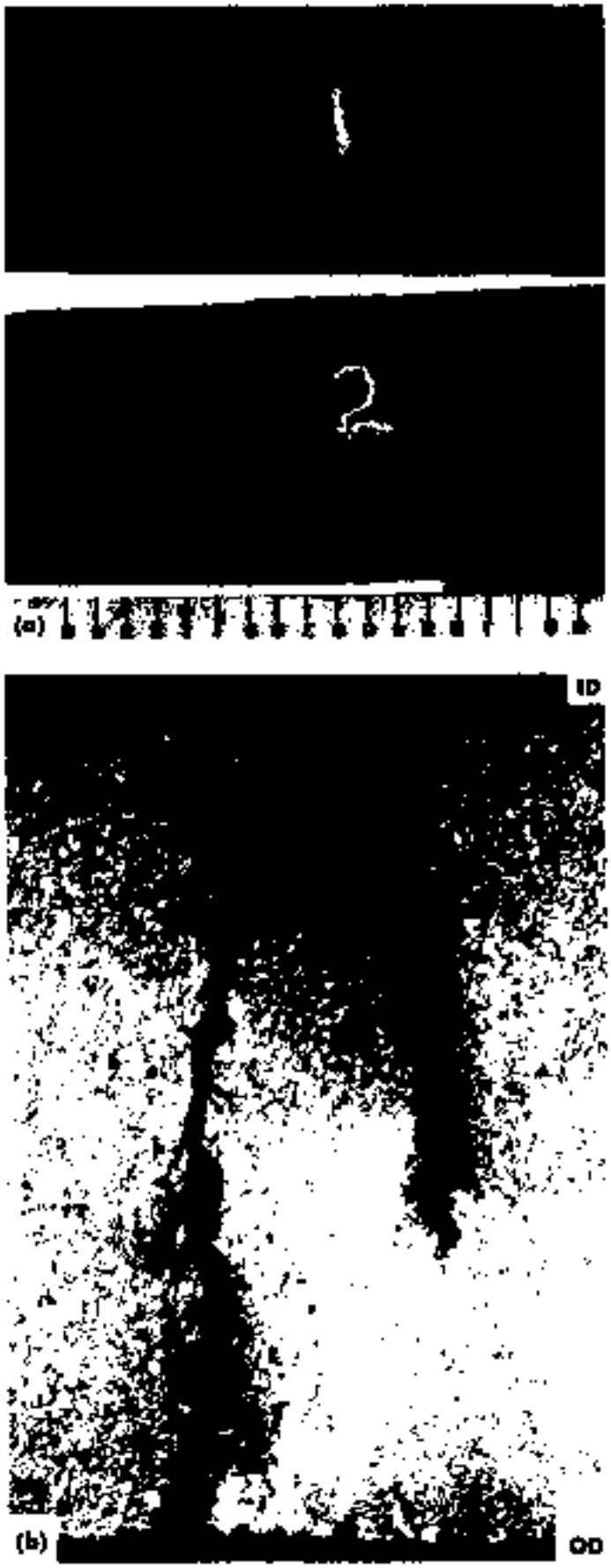

1
2
(a)
ID
(b)
OD

Fig. 10 Failed admiralty brass heat-exchanger tubes from a refinery reformer unit. The tubes failed by corrosion fatigue. (a) Circumferential cracks on the tension (outer) surface of the U-bends. Approximately $1\frac{1}{4}\times$. (b) Blunt transgranular cracking from the water side of tube 1. 40×

Energy-dispersive x-ray analysis of the corrosion deposits from tube 2 indicated the presence of copper, zinc, iron, as well as small amounts of chloride, sulfur, silicon, tin, and manganese. The dark brown and black deposits covering the surfaces of the tubes were identified as copper (II) oxide. Also present were reddish deposits identified as copper (I) oxide, as well as iron oxides. The green color of some deposits indicated that the chlorine and sulfur present were most likely in the form of copper chloride and copper sulfate.

Hardness measurements on the tube specimens ranged from 80 HRF near the crack to 99 HRF on the tension side of the tube. Typical hardness for annealed copper alloy tubing is 75 HRF. This indicated that the tubes may not have been annealed after the U-bends were formed. The appearance of cracks on the tension side of the U-bends showed that residual stresses played a role in crack formation.

Conclusions. The tubes failed by corrosion fatigue initiated by pitting at the inside-diameter surface. The presence of chlorides and sulfates in the cooling water contributed to corrosion by the formation of basic copper chlorides and sulfates. The corrosion pits acted as stress raisers that, in combination with residual stresses from tube bending and mild cyclic stresses in service, promoted crack growth.

Recommendations. Admiralty brass was recommended for continued use in the tubes due to its generally good corrosion resistance in the catalytic reforming environment. The tubes should be annealed after bending to reduce residual stresses from the bending operation to an acceptable level.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Effects of Design

The shape of the components in heat exchangers is important in producing regions of stress concentrations. The ASME Boiler and Pressure Vessel Codes are generally used for the design of pressure vessels. These codes contain relatively simple formulas that can be used for design. If the formulas are properly used, the vessel will withstand the operating pressures and temperatures anticipated. If used improperly or if used for the wrong application, these formulas can result in designs that can cause failures.

One member of the brine heater described in the following example had a high discontinuity stress that was not taken into account in the design. This discontinuity stress exceeded the tensile strength of the carbon steel shell material. Although initial cracking probably occurred during hydrostatic testing, the vessel operated at such low pressure that it was in service for about 2 years, during which time a crack propagated nearly 2.1 m (7 ft) before failure. Investigation revealed that a single welded joint, although properly made, provided a notch that acted as a site of crack initiation.

Example 8: Fracture of a Brine-Heater Shell at Welds. The brine-heater shell (Fig. 11) in a seawater-conversion plant failed by bursting along a welded joint connecting the hot well to the heater shell. Failure occurred approximately 2 years after the plant had been brought on stream.

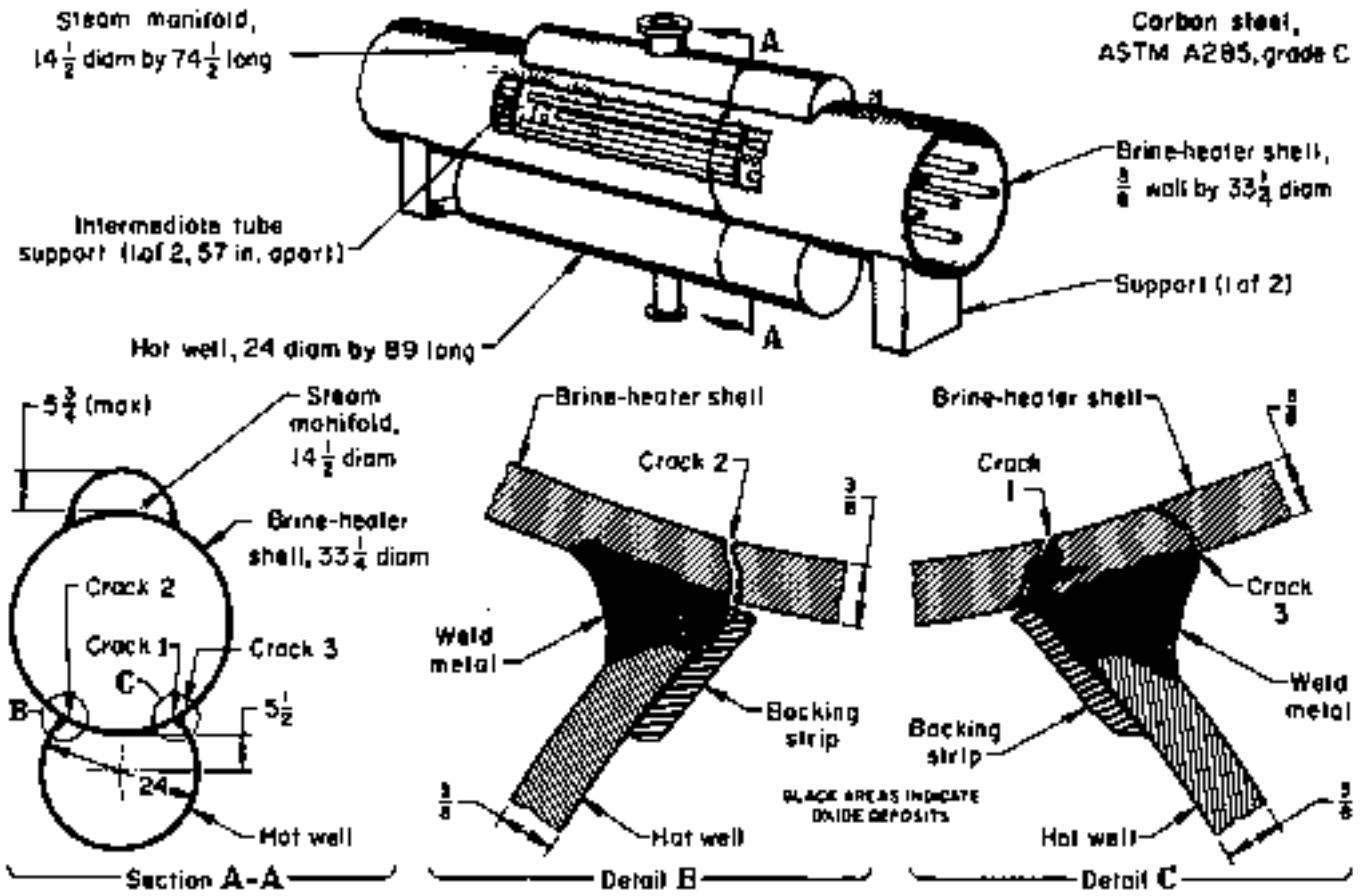


Fig. 11 Failed brine-heater shell of ASTM A285, grade C, carbon steel. The shell fractured at welded joints because of overstress during normal operation. Dimensions given in inches

Specifications required the heater shell to be made of ASTM A285, grade C, steel, which has a tensile strength of 379 to 448 MPa (55 to 65 ksi), a yield strength of 207 MPa (30 ksi), and elongation in 50 mm (2 in.) of 27% minimum. The tube side of the heater was designed for a pressure of 1205 kPa (175 psi) and a temperature of 150 °C (300 °F); the shell side, 1035 kPa (150 psi) and 185 °C (365 °F). Both sides were hydrostatically tested at 150% of design pressure. The heater tubes were made of copper alloy C70600 (copper nickel, 10%) per ASTM B 466.

On-site and laboratory investigations were conducted to determine the cause of failure and to develop corrective measures to minimize future failures.

Service History. The steam supplied to the plant was normally controlled at 517-kPa (75-psi) line pressure. For the first 18 months of service, the brine heater operated at a maximum temperature of 120 °C (250 °F) and a steam temperature of 130 °C (265 °F) with a pressure of 170 kPa (25 psi). One exception was during the scaling of the brine-heater tubes about 1 month after start-up, when the brine temperature exceeded 145 °C (290 °F) and the steam pressure was greater than 386 kPa (56 psi).

During the 6 months before failure, the maximum brine temperature was 140 °C (280 °F) and the steam pressure was 331 MPa (48 psi). About 5 months before failure, a second scaling of the brine-heater tubes occurred, and the steam pressure approached that of the steam line (517 kPa, or 75 psi).

There were 44 recorded pressure surges in the steam line. Most of these were minor (plus 69 kPa, or 10 psi, maximum). However, pressures of 621 kPa (90 psi) were recorded three times; more than 690 kPa (100 psi), four times. The last 690-kPa (100-psi) pressure surge occurred about 6 weeks before the failure.

When the brine heater failed, the operators immediately went to the control room and found that the steam-temperature valve that admits steam to the brine heater was secured. However, the chart indicated that the brine temperature was 60 °C (140 °F) and that the pressure in the steam supply had increased. The steam valve was immediately closed, and within 3 min, the valve in the steam line at the power plant was closed.

The relief valve on the steam-supply line at the power plant, which was set at 862 to 896 kPa (125 to 130 psi), was inspected; there were no indications that the valve had relieved. Also, in the power-plant control room was an alarm that sounded if the pressure upstream of the final reducing station increased to approximately 2070 kPa (300 psi). According to the power-plant operators, this alarm did not sound. During the day the failure occurred, the steam supply pressure at the conversion plant did not exceed 655 kPa (95 psi).

Visual inspection of the heater shell disclosed three cracks or failure areas in the welded joints between the heater shell and the hot well. The locations of these cracks, numbered in the probable order of occurrence, are shown in Section A-A in Fig. 11.

Crack 1 may have initiated at the built-in notch (Detail C, Fig. 11), which is characteristic of welds made with a backing strip. The fracture surfaces were covered with high-temperature oxidation, indicating that the surfaces may have been separated for some time. The crack, which extended the length of the hot well, was completely internal and could not be detected from outside the vessel. The intermediate tube supports were tack welded to the heater shell. The welds nearest the crack were broken and were covered with high-temperature oxidation, indicating that they had been broken for some time.

The heater shell burst along crack 3 (Detail C, Fig. 11), which was in the HAZ of the longitudinal weld. All the fracture surfaces of this crack exhibited bright metal, and there did not appear to be any one point or defect at which this failure started.

Crack 2 (Detail B, Fig. 11) was on the side of the hot well that did not rip open. This crack was approximately 760 mm (30 in.) long and was propagating toward the ends of the hot well. These fracture surfaces were also covered with high-temperature oxidation. The temperature-control valve was removed and inspected, but no damage was noted.

Metallurgical Tests. A section was taken through the longitudinal weld on the same side of the hot well as crack 2. A crack was observed that extended from the notch formed at the root of the weld and the backing strip to about the center of the heater-shell plate. This crack terminated at what appeared to be an inclusion in the metal. The hardness of the metal in the heater-shell plate at this section ranged from 81.5 HRB in the base-metal zone to 85 HRB in the HAZ adjacent to the weld. Hardness near the inclusion, which was in the HAZ, was 72 HRB. The hardness value of 81.5 HRB indicated that the steel should have a tensile strength of about 517 MPa (75 ksi).

Tensile Tests. The tensile strength of three specimens from the brine-heater shell was 531 to 545 MPa (77 to 79 ksi), yield strength was 372 to 414 MPa (54 to 60 ksi), and elongation in 50 mm (2 in.) was 22 to 27%. Compared to the specified requirements for ASTM A285, grade C, steel, these strengths exceeded the maximum, and ductility was less than or equal to the minimum. Chemical analysis showed a silicon content of 0.48%, which indicated a killed steel. Carbon content was 0.20%.

Stress analysis was conducted on the original design of the brine heater. A very high discontinuity stress existed at the longitudinal welds between the hot well and the heater shell. A lesser discontinuity stress existed at the longitudinal weld between the brine-heater shell and the steam manifold. The stress values were calculated for the heater-shell wall between the welds and the hot well outside the welds. These values exceeded the actual yield strength of the shell material at operating pressures of 331 kPa (48 psi) and greater. The material was no longer in the elastic region, and it deformed plastically to relieve or transfer the stress.

It was assumed that the section of the heater shell between the welds attempted to straighten. When this occurred, the root of the weld would have been under tension, which would create an ideal site for a crack to initiate.

Discussion. Cracks 1 and 2 probably originated during hydrostatic tests of the heater or shortly after the heater was put into operation. These cracks were internal and undetectable from the outside. During the first 18 months, when the brine heater was operated at 172 kPa (25 psi), these cracks would propagate at a very slow rate, if at all. During the last 6 months, the brine heater was operated at 331 kPa (48 psi), where the discontinuity stresses were greater than the yield strength of the material, and the cracks propagated at a more rapid rate. The brine heater was subjected to 44 recorded steam-pressure surges, which increased the operating pressure above the nominal 172 and 331 kPa (25 and 48 psi) needed to maintain the required brine temperature. During plant shutdown, there were probably other pressure surges similar to the one that caused the final failure that were not recorded. In addition to the high discontinuity stresses, the fact that the cracks did exist increased the stress concentrations in the heater. The loss of the support by the intermediate tube sheets at the failed portion of the shell increased the stress in the remaining shell.

Conclusions. The brine heater cracked and fractured because it was overstressed in normal operation. The two internal cracks propagated by continuing along the root of the longitudinal weld. The crack on the side that fractured had extended nearly 2 m (7 ft) before failure.

Corrective Measures. The heater design was modified to make the heater shell and the hot well two separate units. Also, the heater-shell expansion section was redesigned to minimize stress on the tubes and to minimize the resulting distortion. A tight shutoff valve was added to the brine-heater steam line to prevent steam from leaking into the heater. Holes in the support plates were enlarged to allow the heater to move in any horizontal direction.

Recommendations. A relief valve should be installed in the heater or in the steam line near the heater. At the time of failure, the relief valve was some distance upstream of the brine heater. Additional hydrostatic tests of the new brine heater should be conducted at ambient temperature and 1035 kPa (150 psi) (design pressure) at approximately 6-month intervals.

Mechanical Joints. The coefficient of expansion and yield strength at high temperature of metal components used in an assembly are important design considerations. Bolted assemblies require fasteners with a yield strength that will allow the fasteners to be tightened to such a prestress value that the assembly would remain leakproof even at high temperatures. In one case, leaks developed in an assembly that consisted of a flanged bonnet bolted to the shell of a heat exchanger and

that was operated at a temperature of 510 °C (950 °F). The bonnet was made of a ferritic steel; the tube sheet, heat-exchanger shell, and studs were made of an austenitic stainless steel. When leakage was not stopped by repeated tightening of the nuts on the studs, the joint was disassembled.

Initial inspection was done to check the surface finish on the flanges and the condition of the gasket. Both were satisfactory. The joint dimensions and metals used were then checked for compliance to specifications; these were also satisfactory. During the examinations and review of the design calculations, it was noted that the materials in the assembly had greatly differing coefficients of expansion and that the studs had a coefficient of expansion that matched only the components made of austenitic stainless steel. Studs made of conventional stainless steel have a low yield strength and high ductility, which leads to stud elongation because of the operating temperature and load. These differing coefficients of expansion caused the tension on the studs on the assembly to relax, permitting leakage. The joint was made tight by replacing the conventional austenitic stainless steel studs with those conforming to ASME specification SA-453, which calls for a high-temperature bolting material with a coefficient of expansion comparable to austenitic stainless steel but with a much higher yield strength. Use of material conforming to SA-453 permitted tightening the nuts on the studs to a torque load that would maintain tightness at the operating temperature.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Effects of Welding Practices

Welding, which is commonly used to join components of heat exchangers, produces metal in the joints that is not homogeneous. The weld consists of a dendritic structure (cast metal) and intermediate zones between fused metal and unaffected base metal. In some cases, the entire heat exchanger can be normalized or annealed, and the metallurgical structure of the weld metal and base metal may become nearly identical. However, because heat treating of the heat exchanger is not always feasible, there generally is a gradation in metallurgical condition from unaffected base metal through an HAZ to weld metal.

In addition to this intrinsic change in homogeneity of material, other factors may provide metallurgical or mechanical stress raisers. Weld defects may include voids, porosity, slag or oxide inclusions, poor penetration, and shrinkage cracks. Many welds have rough surfaces and may join the base metal with an undercut or reinforcement. If metallurgical discontinuities are minimal, if defects are eliminated, and if the external contour does not contain mechanical notches, the properties of a welded joint can be expected to approach those of the base metal.

Preparation of joint surfaces before welding can be a source of weld defects in heat exchangers. For example, a pinhole leak developed in a heat-exchanger shell adjacent to a weld deposit that joined a nozzle, made of 19-mm ($\frac{3}{4}$-in.) diam schedule 80 pipe, to the shell. When the nozzle was removed, it was found to have been installed by torch cutting a 38-mm ($1\frac{1}{2}$-in.) diam hole in the shell, inserting the nozzle and welding it in place. There was no evidence of cleanup after torch cutting the hole or of beveling of the wall edges around the hole to permit good weld penetration. Examination of the weld deposit revealed that the weld metal was generally oxidized and contained several flux inclusions. There was little or no bonding of the weld bead to the surfaces that had been cut with the torch. Examination revealed evidence that the weld was made under conditions of poor fit-up, lack of cleanness, and poor workmanship. A good joint was made by removing old weld metal and properly beveling the edges of the hole before rewelding.

Joint design can be responsible for mechanical notches at welds in heat-exchanger components. Designs that incorporate sharp corners and heavy longitudinal welds are sources of stress raisers.

Example 9: Fracture of a Carbon Steel Pipe in a Cooling Tower. A 455-mm (18-in.) diam, 7.9-mm ($\frac{5}{16}$-in.) wall carbon steel discharge line for a circulating-water system at a cooling tower fractured in service; a manifold section cracked where a Y-shaped connection had been welded. The steel pipe was made to ASTM A 53 specifications.

Investigation. Examination of the pipe revealed that cracking occurred in the HAZ parallel to the weld and across the weld as shown in Fig. 12(a). The end of the branch pipe was prepared for welding by making a transverse cut; a second cut was then made at about 15° to the longitudinal axis of the pipe, resulting in a sharp comer (Fig. 12a). The main pipe was then notched to fit the end of the branch pipe. Fitting of the saddle connection was poor, and no backing strips were used even though the pipe wall was thin; as a result, the condition of the root-weld bead was below standard, and

numerous deep crevices and pits were in the weld area exposed to the inner surfaces of the manifold (Sectional view, Fig. 12a).

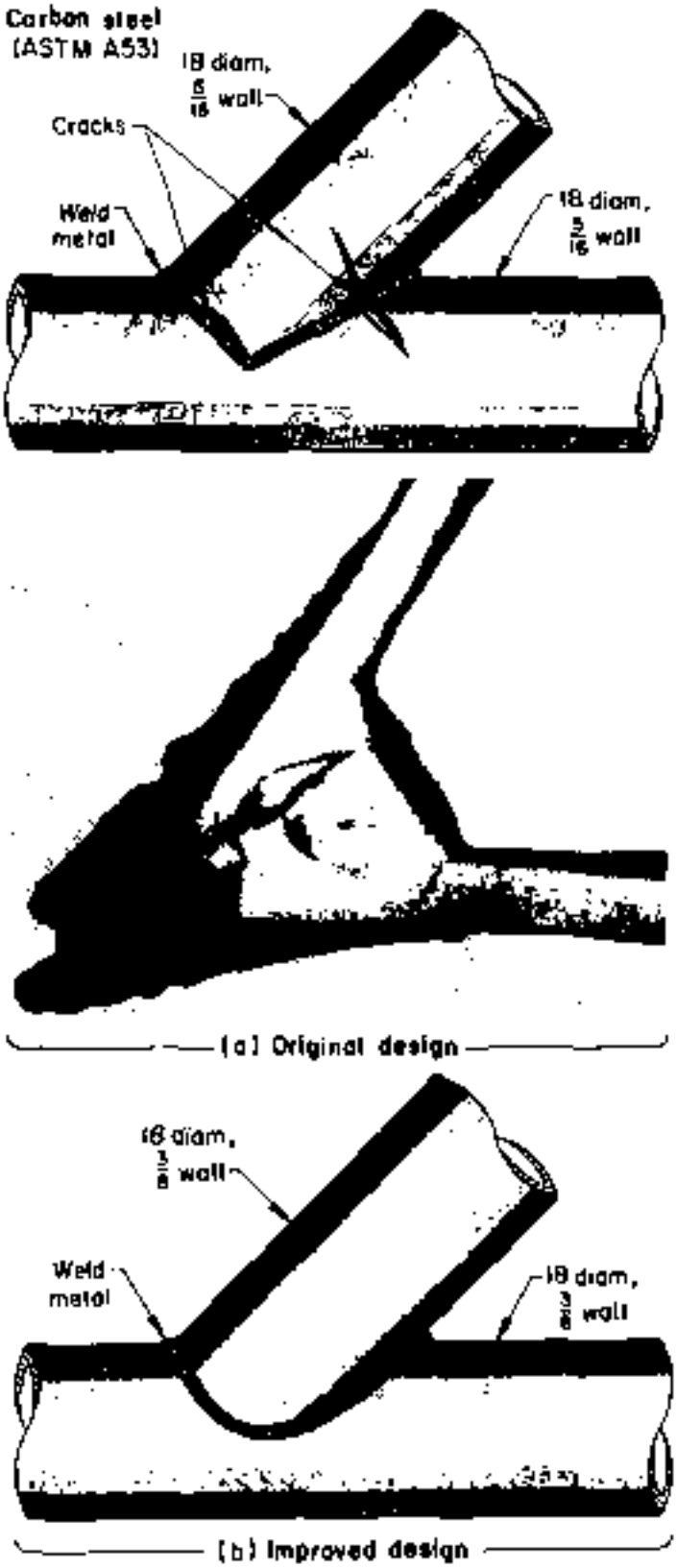


Fig. 12 Carbon steel discharge line at a cooling tower that failed because of poor fit-up at Y-joint and poor-quality welds. (a) Original joint design of pipe connection and location of cracks. Photograph is an oblique view of a section through the weldment showing the abrupt intersection of pipe walls and the voids and crevices in the weld metal. (b) Improved joint design. Dimensions given in inches

Conclusions. The pipe failed by fatigue. Cracks originated at crevices and pits in the weld area that acted as stress raisers, producing high localized stresses because of the sharp-radius corner design. Abnormally high structural stresses and alternating stresses resulting from the pump vibrations contributed to the failure.

Corrective Measures. The joint design was changed to incorporate a large-radius corner (Fig. 12b), and fitting of the components was improved to permit full weld penetration. Backing strips were used to increase weld quality. Also, the pipe wall thickness was increased from 7.9 to 9.5 mm ($\frac{5}{16}$ to $\frac{3}{8}$ in.).

Example 10: Intergranular Cracking in Heat-Exchanger Welds Due to Hot Shortness. Samples from an AISI type 316 stainless steel heat-exchanger shell were submitted for examination after radiographic testing revealed the presence of subsurface cracks in a longitudinal weld seam of the shell. The heat exchanger had been in service for 12 years.

Investigation. The as-received sample was dye penetrant tested before sectioning, and no cracks were indicated. Longitudinal sections were then taken along the centerline of the weld (Fig. 13).

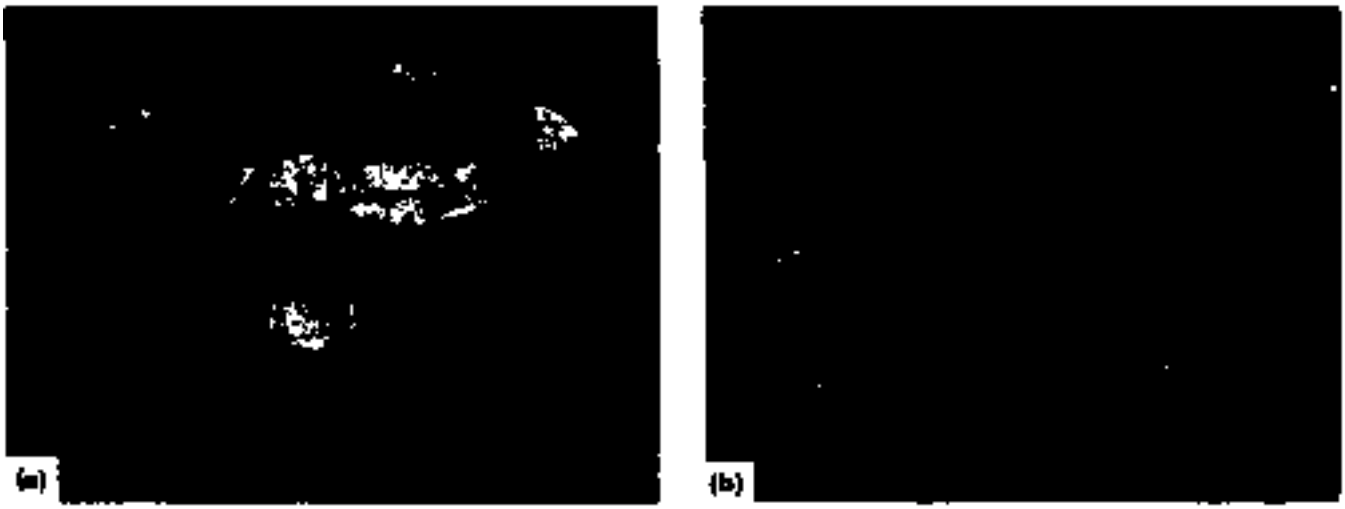


Fig. 13 Weld in AISI type 316 heat-exchanger shell that failed due to hot shortness. (a) Longitudinal section of weld; the dotted line indicated how the sample was sectioned for microexamination. Approximately $2\frac{1}{2}\times$. (b) Micrograph of section from weld. Hot shortness resulted in intergranular cracking along the large columnar grain boundaries of the root-pass fusion zone. Etched with 10% oxalic acid. 55×

Metallographic examination revealed numerous intergranular cracks associated with the root pass of the weld. The cracks had propagated both parallel to and normal to the weld seam (Fig. 13). This type of cracking is characteristic of a phenomenon termed hot shortness (see the article “Failures Related to Welding” in this Volume).

Energy-dispersive x-ray spectrometry of the areas surrounding the cracks showed that a type 316 electrode was not used for the root pass. Analysis indicated that the root pass was instead made using a nickel-copper alloy electrode. Dilution effects and subsequent welding with type 316 electrodes accounted for the observed variations in elemental concentration.

Conclusions. Cracking was due to the use of an incorrect electrode for the root pass of the weld. Fully austenitic electrode materials, such as the nickel-copper alloy used for the root pass, are quite crack sensitive, especially if overheated. The subsurface defects in the fusion zone of the weld were probably too small to be detected by nondestructive inspection at the time of fabrication, but long-term service and thermal stresses propagated the fissures to the extent that they were detected by radiographic inspection 12 years later.

Recommendations. This weld seam and others were completely ground out and were replaced with the correct electrode material.

Stiffening, which results when the weld metal is thicker than the adjacent base metal, can also lead to failure, as discussed in the following case. The design of expansion bellows in a heat exchanger incorporated a longitudinal weld across each corrugation and a circumferential weld joining each corrugation (Fig. 14a). The bellows had an outside diameter of 660 mm (26 in.) and were made of 3.9-mm ($\frac{5}{32}$-in.) thick type 321 stainless steel. The heat exchanger was shut down when leakage occurred in the bellows. Circumferential cracks were found along the HAZ of the weld and through the corrugation wall. These are areas of highest stress.

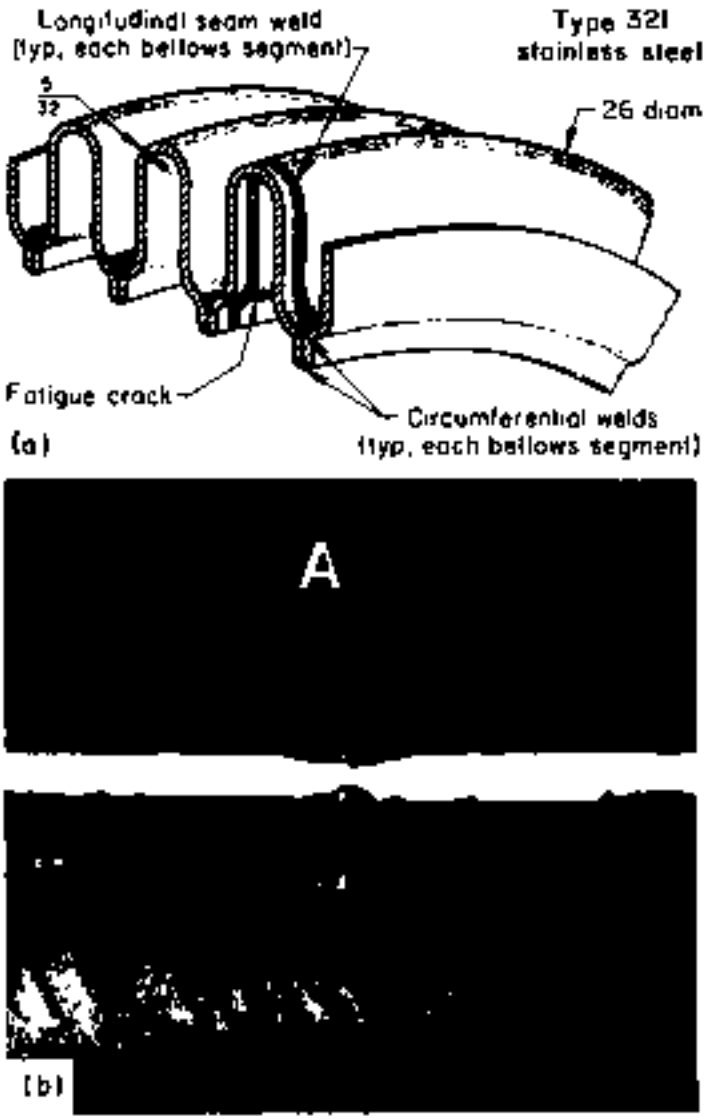


Fig. 14 Type 321 stainless steel heat-exchanger bellows that failed by fatigue originating at heavy weld reinforcement of a longitudinal seam weld. (a) A section of the bellows showing locations of the longitudinal seam weld, the circumferential welds, and the fatigue crack. Dimensions given in inches. (b) Mating fracture surfaces of the bellows. Crack initiated at the longitudinal seam weld (region A) and propagated circumferentially along the HAZ of the circumferential weld (region B). Fatigue beach marks are present along edges (at center). The fracture surfaces are oriented to show the inner surface of the bellows (top) and the outer surface (bottom). 2×

Examination of the fracture surface revealed fatigue beach marks (Fig. 14b) that initiated at the heavy weld reinforcement of the longitudinal seam weld. The unusually heavy reinforcement had a stiffening effect, which caused overstressing and transverse cracking of the longitudinal seam. Grinding the reinforcement of the longitudinal weld flush reduced the stiffening. As a corrective measure, the bellows were formed from a single sheet large enough to provide the proper number of convolutions without a circumferential weld in the working section. The only weld required was one longitudinal seam the full length of the bellows. This change in design eliminated the points of high stress that had led to failure.

Revised by Robert J. Franco, Exxon Company, U.S.A.

Effects of Elevated Temperature

Failures associated with elevated temperature generally involve a material that is too weak for the service temperature or a material that develops unexpected embrittlement at or above the service temperature. Thermal cycling can cause failures where adjacent materials differ significantly in composition and coefficient of thermal expansion. Embrittlement may also be the result of metallurgical changes that leave the material relatively ductile and tough at elevated temperature but brittle at or slightly above room temperature.

Another temperature-related problem is scaling and oxidation due to excessive temperatures. Creep and stress rupture also occur because of excessive temperatures, as discussed in the article "Creep and Stress-Rupture Failures" in this Volume. These temperature-related failure mechanisms must be accounted for in selection of material, in fabrication, and in operating procedures.

Example 11: Embrittlement of a Titanium Heater Tube. A flow stoppage in a leach heater caused brief overheating of the 89-mm ($3\frac{1}{2}$-in.) outside-diameter, 6.4-mm (0.25-in.) wall thickness titanium heater tubes (ASTM B337, grade 2). Many of the tubes showed colorful surface oxides, and some sagged slightly.

A short section of an oxidized and sagged tube was sent to the metallurgical laboratory for evaluation. The purpose of the examination was to determine whether the mechanical properties of the tube had been altered by overheating, particularly whether the tube had been embrittled, making it susceptible to fracture. A section from a similar tube removed previously from a leach heater that had not been heated above normal operating temperatures was sent to the laboratory for comparison.

Visual examination of the section of the overheated tube disclosed blue-tinted areas and patches of flaky white, yellow, and brown oxide scale. There was a dark-gray scale on the inside surface with no evidence of corrosion. The presence of the colored scale on the outer and inner surfaces was an indication of heating to approximately 815 °C (1500 °F).

Mechanical Tests. A 75-mm (3-in.) long section of the overheated tube was subjected to a flattening test. Specification ASTM B 337 requires that a welded or seamless grade 2 titanium tube be capable of withstanding, without cracking, flattening under a load applied gradually at room temperature until the distance between the platens is seven times the nominal wall thickness—in this case, 45 mm (1.75 in.). In the test, when the platens were 66 mm (2.6 in.) apart, cracking initiated at both inner and outer surfaces of the tube; when the platens were 61 mm (2.4 in.) apart, the tube fractured, with fracture initiating at the inner-surface cracks. Thus, the tube no longer met ASTM B 337 specifications. Figure 15(a) shows an end view of the flattened tube. The cracks on the outer surface are shown in Fig. 15(b).

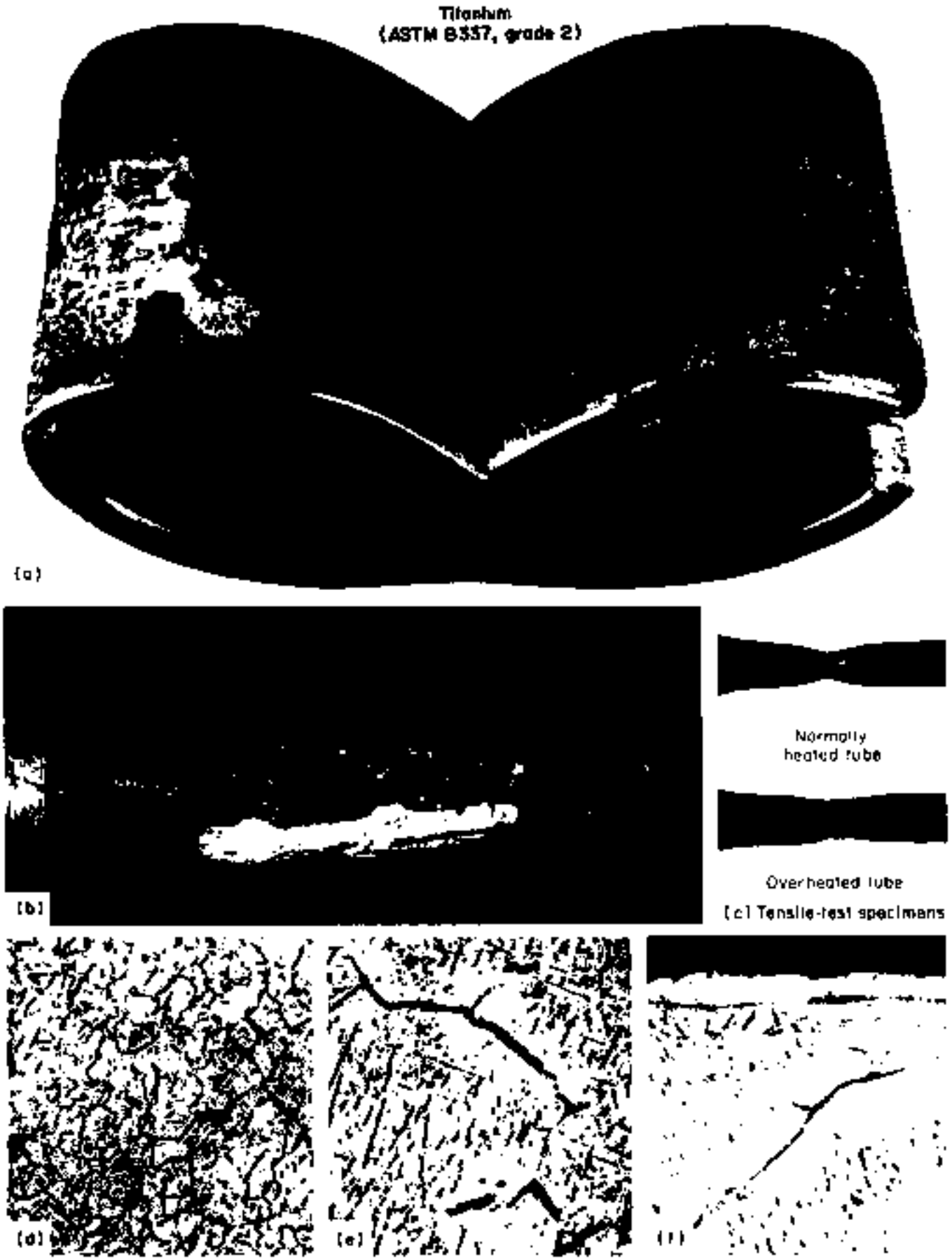


Fig. 15 Titanium heat-exchanger tube (ASTM B337, grade 2) that became embrittled and failed because of absorption of hydrogen and oxygen at elevated temperatures. (a) Section of the titanium tube that flattened as a result of test per ASTM B 337; the first crack was

longitudinal along the top inside surface. $\frac{3}{4}$×. (b) Side of flattened tube section showing cracks. $1\frac{1}{4}$×. (c) High-temperature (370 °C, or 700 °F) tension-test specimens: top, specimen from normally heated tube; bottom, specimen from overheated tube. (d) Normally heated tube specimen etched in Kroll's reagent showing fine grain size and hydride particles (faint black needles). 200×. (e) Overheated tube specimen etched in Kroll's reagent showing large grain size and hydride particles (small needles). 200×. (f) Transverse section of overheated tube specimen at outside surface showing 0.05-mm (0.002-in.) thick oxygen-embrittled layer (light band near top)

Longitudinal tensile tests were conducted at 370 °C (700 °F) on specimens taken from the overheated and normally heated tubes. The results are listed in Table 3. As shown in Fig. 15(c), the specimen from the normally heated tube necked down more than the overheated tube specimen, although elongation was approximately the same.

Table 3 Results of longitudinal tensile test

Specimen	Yield strength, 0.2% offset		Tensile strength		Elongation in 50 mm (2 in.), %	Reduction of area, %
	MPa	ksi	MPa	ksi		
Overheated tube	97	14	146	21.2	36	69
Normally heated tube	100	14.5	197	28.5	38	82

Metallographic examinations were made of specimens from both the normally heated and the overheated tubes. The normally heated tube had a fine grain size and only a few faint hydride particles in the microstructure (Fig. 15d). By comparison, the microstructure of the overheated tube (Fig. 15e) exhibited a much larger grain size and more numerous needlelike hydride particles. Also, the overheated tube had a very hard brittle surface layer approximately 0.05 mm (0.002 in.) thick (Fig. 15f), which was present on both the inner and outer surfaces. The layer was extremely hard (51 HRC) at the surface and decreased to almost normal hardness (33 HRC) at 0.076 mm (0.003 in.) from the surface. At the approximate center of the tube wall, the hardness was 83 HRB.

The flaky white, yellow, and brown oxide scale and the increased grain size indicated that the tube had been heated to approximately 815 °C (1500 °F). At this temperature, the metal readily absorbs hydrogen and oxygen. The hydride formation and the shallow surface embrittlement were indications of hydrogen and oxygen absorption. The oxygen absorbed at the surfaces created very hard and therefore brittle layers that made the tubes susceptible to cracking at ambient temperatures.

Although the high-temperature (370 °C, or 700 °F) tensile strength of the overheated tube was significantly lower than that of the normally heated tube, the yield strengths were about the same, and the ductility of the overheated tube was slightly lower. At 370 °C (700 °F), there was no evidence of the severe embrittlement shown in the flattening test at room temperature. However, at temperatures below 205 °C (400 °F), embrittlement from hydrogen and oxygen absorption occurred, as evidenced by the poor performance of the flattening test.

Conclusions. The titanium tubes had been embrittled by being heated to a temperature of about 815 °C (1500 °F), which is excessive for ASTM B337, grade 2, titanium. The most serious effect of overheating the tubes was the generally embrittled condition at low temperature, and particularly the severe surface embrittlement resulting from oxygen absorption. The brittle surface layers made the overheated tubes susceptible to cracking under start-up and shutdown. At operating temperatures of 370 °C (700 °F), the overheated tubes would not have been brittle, but the overall mechanical properties of the tubes would have been somewhat degraded.

Recommendations. If the overheated tubes remain in service, thermal stresses should be avoided during start-up and shutdown, particularly at temperatures below 205 °C (400 °F). Although at the operating temperature of 370 °C (700 °F) the overheated tubes are not in an embrittled condition, the tubes should be replaced because there is a possibility that cracks formed during start-up and shutdown would propagate through the tube wall under operating conditions. Replacement tubes should be made of a heat-resistant alloy (e.g., Hastelloy C-276), rather than of titanium.

Acknowledgment

R.J. Franco wishes to acknowledge G.M. Buchheim, Exxon Research & Engineering Company, for his assistance in preparing several of the examples in this article.

Failures of Heat Exchangers

Revised by Robert J. Franco, Exxon Company, U.S.A.

Selected References

- S.D. Reynolds and F.W. Pement, Corrosion Failures of Tubing in Power Plant Auxiliary Heat Exchangers, *Mater. Perform.*, Vol 13 (No. 9), Sept 1974

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Introduction

PRESSURE VESSELS, piping, and associated pressure boundary items of the types used in nuclear and conventional power plants, refineries, and chemical-processing plants are discussed in this article. These types of equipment operate over a wide range of pressures and temperatures and are exposed to a wide variety of operating environments.

The National Board of Pressure Vessel Inspectors compiles domestic statistics on reported accidents involving pressure vessel material. These data indicate the prevalent causes and types of failures and are summarized in Table 1 for a 4-year period beginning in 1981. Faulty design, fabrication, and inspection contribute to significant numbers of failures, particularly for shell components. Damage during shipment and storage must be considered because air and its contaminants may be corrosive. Field fabrication and erection practices should also be considered. Simple matters such as specifying and verifying the proper material and material condition are critical. One company has used portable spectroscopy equipment to examine a number of bulk items supposedly made of alloy steels and found 1 to 3% of the total items to be carbon steel. A survey of equipment that required 320,000 individual instrument readings found that 1.8% of the material was supplied incorrectly, and the most frequent error was substitution of carbon steel for alloy or stainless steel, or vice versa (Ref 1).

Table 1 Summary of statistics on pressure vessel failures compiled from 1981 through 1984 by the National Board of Boiler and Pressure Vessel Inspectors

Initial part failure	Causes							Type of failures							Numbers		
	Low water cut-off	Faulty design fabrication or installation	Corrosion or erosion	Operator error or poor maintenance	Burner failure	Pressure control failure	Other	Burned or overheated	Collapsed inward	Combination explosion	Cracked	Torn assunders (rupture)	Leakage	Other	Accidents	Injuries	Deaths
1981																	
Shell	5	18	25	36	…	12	21	…	6	21	57	26	23	3	94	19	3
Head	1	7	5	9	…	3	1	…	6	2	12	7	1	…	23	…	…
Attachments	7	5	…	4	…	7	2	…	…	2	7	5	1	…	31	1	1
Piping	10	2	1	3	1	5	18	1	…	2	10	7	15	4	41	1	…
Safety valves	…	…	…	1	…	…	2	…	…	…	…	3	…	…	3	1	1
Miscellaneous	…	2	4	7	…	4	166	1	2	5	4	9	12	…	232	5	4
1982																	
Shell	4	17	29	13	1	3	34	4	10	3	62	45	35	0	129	7	4
Head	1	4	3	11	0	4	5	0	2	2	13	8	7	1	19	3	2
Attachments	2	4	3	4	0	2	2	1	0	0	10	9	11	10	21	1	0
Piping	0	3	8	18	0	5	5	0	0	1	14	30	24	0	33	17	0
Safety valves	0	18	25	12	0	3	4	0	0	1	1	2	26	0	5	5	1
Miscellaneous	1	6	14	24	0	86	43	0	1	0	11	2	23	3	89	3	2
Total															296	36	9
1983																	
Shell	1	26	64	44	1	12	1143	15	13	1	61	31	41	22	167	75	13
Head	1	3	9	9	0	3	12	2	2	3	14	6	9	3	22	18	4
Attachments	0	2	3	2	0	2	2	0	0	0	15	3	16	4	19	2	0
Piping	1	7	14	54	2	6	24	1	0	1	16	18	72	4	62	16	0
Safety valves	0	0	28	1	0	2	23	0	0	0	1	0	39	14	3	11	0
Miscellaneous	2	25	12	47	0	6	56	1	1	0	17	20	33	15	149	18	0
1984																	
Shell	3	77	75	53	16	10	151	4	9	17	265	98	92	96	625	209	43
Head	…	17	15	22	7	2	17	…	3	…	57	10	5	11	95	124	13
Attachments	1	4	5	6	1	2	12	1	…	1	27	2	9	12	57	20	12
Piping	…	15	22	21	…	4	47	9	…	1	10	17	34	25	110	46	4
Safety valves	9	6	…	11	7	12	4	…	1	2	…	9	1	7	51	11	1
Miscellaneous	1	25	15	36	2	11	189	14	2	1	44	22	22	147	372	27	…
Total															1310	437	73

Normal operating conditions contribute to changes in materials and structures that should be accounted for in design. Operating temperatures are important because they determine whether the material is stressed above or below any ductile-to-brittle transition temperature below which relatively low energy fractures can occur. Other degradation mechanisms, such as corrosion and scaling; stress corrosion; erosion from flowing gases, liquids, and solids; hydrogen damage; creep and stress rupture; fatigue and creep fatigue; and aging, are generally accelerated with increasing temperature. Operating pressure is another important design consideration because pressure combined with structural constraints determines the stresses in pressure boundaries both at steady-state operating conditions and during transients. Verifying whether failures occurred as a result of, or during, off-normal conditions is an important part of the analysis.

Cyclic operation greatly affects the life of pressure boundaries. Intermittent operation can result from normal demand schedules or from operation during peak demand periods only. The cyclic thermally induced stresses associated with frequent start-ups and shutdowns are frequently more severe than stresses of steady-state operations, and designers cannot always predict the use cycles of equipment that owner-operators impose. Economics frequently dictate that use of capital equipment be extended beyond the original design life in many industries, with the result that repairs and renovations are attempted. Accounting for all the relevant practices of several decades that can affect life-extension programs is difficult, and failures can arise from obscure combinations of circumstances.

Intermittent operation can also affect the life of pressure vessels. Long periods of inactivity greatly enhance the probability of internal or external corrosion. During long shutdown periods, the recommended practice is to fill the system with inert or nonreactive gases, such as helium or nitrogen, in order to avoid unacceptable corrosion. Wet lay-up using deoxygenated water or the continuous circulation of demineralized deoxygenated water can also be used to minimize corrosion. Recordkeeping during such periods is as important as record-keeping during operating periods.

Unanticipated effects resulting from the design of other components of the system, such as piping and valves, can be important. When all components are incorporated into a system, the service conditions may be unacceptable for one of the components. Water pulses and vibration can damage piping systems, including hangers and supports. Heavy valves, instruments, and associated components should be properly supported so that the connecting piping system is not overstressed under fatigue loading.

Failures in other parts of a system can cause pressure boundaries to fail. Burner failures, pressure control valve failures, low water cut-offs that lead to improper coolant levels, and simple operator errors or improper maintenance of equipment account for a sizable fraction of recorded pressure boundary failures.

A large number of failures result from causes designated as “other” in the failure statistics. Unique experiences and failures due to multiple causes are not uncommon, and the analyst must be aware of multiple, interconnected mechanisms leading to failures.

Reference cited in this section

1. G. Sorell and M.J. Humphries, High Temperature Hydrogen Damage in Petroleum Refinery Equipment, in *Corrosion/78*, National Association of Corrosion Engineers, Houston, 1978

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Procedures for Failure Analysis

The basic procedures for analyzing failures of pressure vessels and pressure piping are not much different from those for investigating failures of other types of equipment (see the article “Conducting a Failure Examination” in this Volume). At the outset, a preliminary examination should be made to provide a clear overall picture of what happened. If the failure involved an explosion, a careful search should be made for all material components and fragments. The location and mass of each component or fragment should be recorded, and all burn discolorations or impact marks on surrounding objects or surfaces should be located. This ballistic information can be used to estimate launch angles and trajectories and to calculate the approximate pressure in the vessel at the instant of failure.

Because an explosive failure of a pressure vessel may result in extensive damage, system downtime, and expense, it is natural that the preliminary examination will generate considerable expression of opinions as to

the possible cause of failure. Casual opinions should be avoided, and certainly not solicited. Subsequent recollections of witnesses could be colored by the offhand comments or speculation of investigators.

As in any failure analysis, it is important that the investigator gain a clear understanding of the entire system involved, including specifications, intended functions, normal methods of operation, thermomechanical history, maintenance history, manufacturing history, and operator experience. This information, plus information from subsequent examinations and from laboratory analyses of specimens, should permit accurate determination of how and why the failure occurred.

Techniques used during a preliminary examination should not affect any subsequent examinations. For instance, liquid-penetrant inspection can make it difficult to analyze for chlorides in deposits on the part surfaces or on fracture surfaces.

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Effects of Using Unsuitable Alloys

Specification of an alloy that is not suited to the application is frequently the cause of failure of a pressure vessel. The composition of the alloy may be incorrect, or the metallurgical properties may not meet requirements. This often results from a misunderstanding of the operating conditions by the designer.

A common cause of failure of pressure vessels is use of an alloy other than the one specified. Sometimes, bars or plates are not properly marked or are accidentally stored with another alloy. Mistakes can be made when care is not taken to ascertain that the alloy used complies with the alloy specified.

In one case, several low-carbon steel nipples were inadvertently used in place of 5Cr-0.5Mo steel nipples in a pipeline containing hydrogen at 400 °C (750 °F). Failure occurred after approximately 4 years of service. The micrograph shown in Fig. 1, which illustrates a specimen from one of the low-carbon steel nipples, reveals fissuring at the grain boundaries. This phenomenon is characteristic of hydrogen attack when low-carbon steel is exposed to hydrogen at high pressures and temperatures. The piping system was completely surveyed to ensure that only the specified metal was used. Replacements were made when the wrong metal was found.

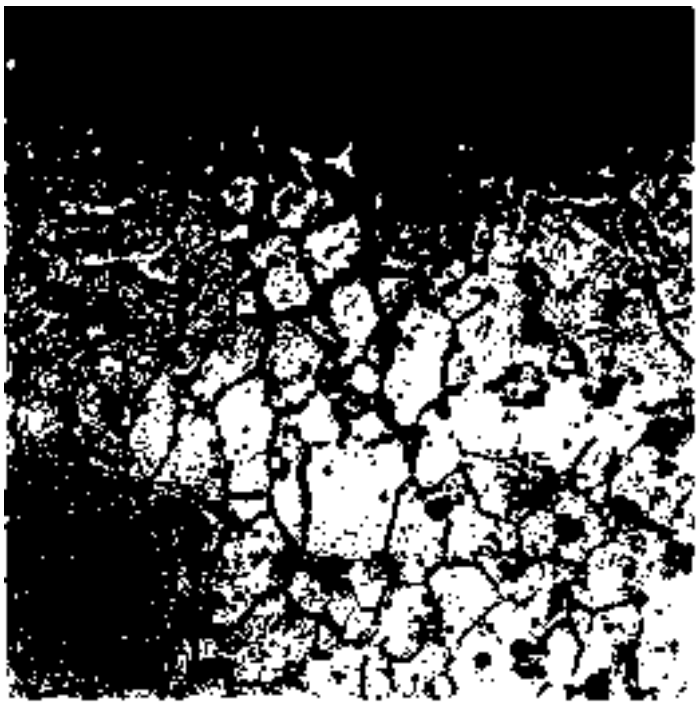


Fig. 1 A specimen from a low-carbon steel nipple showing fissuring at grain boundaries (top) caused by hydrogen attack. 80×

A common experience for major oil refiners is failures or near-failures that can be attributed to insufficient alloy content in steels for hydrogen service. In one example, a carbon steel pipe section failed that was inadvertently substituted for the specified 1.25Cr-0.5Mo steel. This 305-mm (12-in.) diam line ruptured after 16 years of service at conditions slightly above the Nelson curve limit for carbon steel, resulting in an expensive fire. The failed end of the pipe was jagged but undistorted, indicating a brittle failure (Fig. 2(a)). Metallographic examination of a fragment of the pipe revealed extensive cavities through about three-quarters of the wall thickness, with the outer surface of the pipe wall relatively unaffected (Fig. 2(b)). At higher magnification, decarburization and fissuring are visible in the damaged microstructure (Fig. 2(c)). Such incidents of hydrogen damage occur in other branches of the petrochemical- and chemical-processing industries. A severe explosion resulted from the erroneous installation of a 305-mm (12-in.) carbon steel pipe at a German ammonia plant in 1974. The pipe section, which had been installed in a 17-MPa (2500-psig) alloy steel loop, fragmented and released a jet that toppled a 254-Mg (280-ton) ammonia converter.

Fig. 2(a) Ruptured 305-mm (12-in.) carbon steel pipe, inadvertently installed in a 1.25Cr-0.5Mo circuit, that was severely damaged by hydrogen embrittlement. On-stream failure caused extensive fire damage.

Fig. 2(b) Micrograph of outside-diameter surface of failed pipe in Fig. 2(a). Hydrogen attack had progressed about three-fourths through the wall thickness. 18×

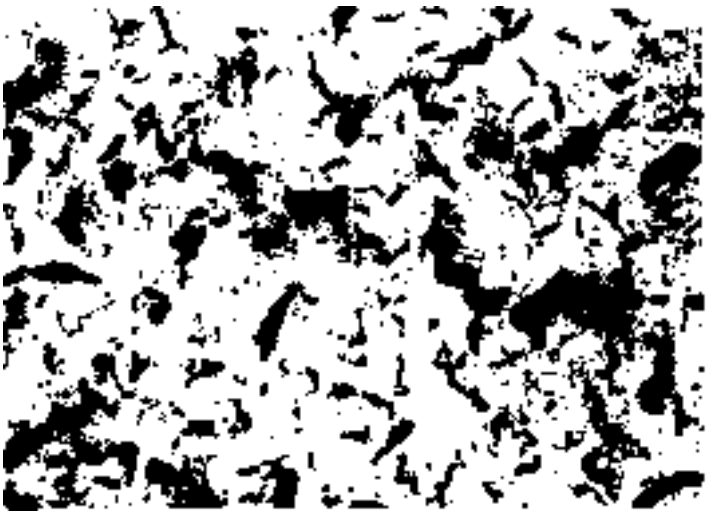


Fig. 2(c) Micrograph of inside-diameter surface of failed pipe in Fig. 2(b) showing severe decarburization and fissuring. 180×

Welding Problems. In the fabrication of pressure vessels, welding problems, such as brittle cracking in the heat-affected zone (HAZ), often result from the use of steels containing excessive amounts of residual elements that increase hardenability and susceptibility to cracking. Such steels are sometimes produced inadvertently when low-carbon steel is made in a furnace normally used to make high-alloy steels. The refractory lining of the furnace may impart sufficient residual chromium and other alloying elements to a heat of low-carbon steel that problems occur when components made from that heat are joined using standard welding procedures. Therefore, complete control of composition is of the utmost importance when welding is involved.

As an example of improper material specification or use, galvanized steel clips were welded to stainless steel piping. The galvanized (zinc) coating became molten during the welding process, and the liquid zinc attacked the grain boundaries in the stainless steel, causing intergranular cracks. The piping was dimpled and leaked at some of the welds. The locations and depths of the cracks were related to the welding procedure and the amount of welding heat used. Excessive welding heat had caused dimpling (distortion) of the stainless steel piping and further crack penetration into the piping wall. The cracks continued to propagate during service from stresses induced in the piping.

In another case, a transfer line carrying 73% sodium hydroxide developed a leak at a weld joining a cast Monel coupling to wrought Monel pipe. Metallographic examination disclosed many small microfissures in the weld metal, but no microfissures were found in the HAZs of the wrought Monel. This section of line contained high thermal stresses that resulted from injection of steam a short distance upstream from the coupling. Chemical analysis revealed that the cast Monel coupling had a silicon content of 1.97%, whereas the wrought Monel piping had a silicon content of 0.10%. Castings of high-silicon Monel generally have poor weldability and are often hot short, or sensitive to cracking in the weld HAZ. Fissures in the cast coupling were formed during welding and subsequently penetrated the section when the coupling was subjected to the thermal shock of steam injection. Where welding is required, a weldable-grade casting should be specified to ensure freedom from hot fissuring.

A welded-on blank cap began leaking during a hydrostatic test because it had cracked in the weld. Analysis showed that the cap had a carbon content of 0.56%, whereas a maximum carbon content of 0.30% was specified. The excessive carbon content had resulted in a brittle HAZ. It was found that construction personnel had substituted ASTM A53 type F butt-welded piping, which does not have a specification for maximum carbon content. Because the higher-carbon steel had been used on a random basis throughout the piping system, all piping was spark tested in the field to expedite repair. When a spark-tested section generated a questionable spark pattern, a specimen was removed for chemical analysis. All sections having a high carbon content were replaced by piping with a maximum carbon content of 0.30%.

Failure of several high-pressure nipples was attributed to the occasional inadvertent use of 4140 steel instead of the specified 1035 steel. The presence of approximately 1% Cr and 0.15 to 0.25% Mo in 4140 steel enhanced the hardenability and contributed to hardness and brittleness in HAZs adjacent to welds. All failures in the nipples occurred in these hardened zones. A spot test for chromium was made, and those nipples showing chromium contents exceeding 0.3% (estimated) were removed and replaced with nipples fabricated from 1035 steel. Although 4140 steel is a weldable alloy steel, welding procedures suitable for 1035 steel were not appropriate for the higher-alloy material.

Specification of the filler metal used for welding components of pressure-piping systems is important, as demonstrated in the following example.

Example 1: Brittle Fracture of a Clapper Weldment for a Disk Valve Due to Improper Filler Metal. After a service life of only a few months, the clapper in a 250-mm (10-in.) diam valve fractured. During operation, the valve contained a stream of gas consisting of 55% H_2S, 39% CO_2, 5% H_2, and 1% hydrocarbons at 40 °C (100 °F) and 55 kPa (8 psi).

Specifications required the clapper to be made of 13-mm ($\frac{1}{2}$-in.) thick ASTM A36 steel, stress relieved and cadmium plated.

Investigation. Fracture occurred at the welded joint between the clapper and a 20-mm ($\frac{3}{4}$-in.) diam support rod, which was also made of ASTM A36 structural steel. Examination revealed voids on the fracture surface and evidence of incomplete weld penetration. The overall appearance of the fracture surface indicated brittle fracture, but there was some evidence of fatigue beach marks. Fracture originated at a slag inclusion in the weld metal. In some areas, particularly on the weld metal, the plating material had flaked off the clapper.

Vickers microhardness surveys across the weld area revealed very narrow bands of high hardness (400 to 650 HV) at the edges of the weld metal. The normal hardness of the base metal was 150 to 160 HV, hardness in the HAZ was 150 to 180 HV, and hardness of the weld metal was 230 to 250 HV.

Chemical analysis of the material in the high-hardness bands revealed a composition of 70Fe-17Cr-9Ni-1.2Mn. This composition indicated that a stainless steel filler metal had been used, which produced a mixed composition at the weld boundaries. Although the weldment may have been stress relieved as specified, stress-relief annealing was apparently incapable of reducing the high hardness of the localized bands caused by alloy mixing. Chemical analysis of the plating material showed it to be nickel, probably electroless nickel because it had a high hardness.

Conclusions. The clapper failed by fatigue and brittle fracture because it was welded with an incorrect filler metal. Fatigue cracking was initiated at voids and high-hardness bands in the weld deposit. Also, the clapper was plated with nickel instead of cadmium.

Corrective Measure. A clapper assembly was welded with a low-carbon steel filler metal, then cadmium plated. The new clapper was installed in the valve and subsequently provided satisfactory service.

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Effect of Metallurgical Discontinuities

Nonmetallic inclusions, seams, laps, bursts, and pipes are common discontinuities found in wrought products that may cause premature failure. Shrinkage, gas porosity, and cold shuts are likely to occur in castings and can lead to failure, usually leakage in cast components.

Discontinuities can and sometimes do cause failures and should never be overlooked during failure analyses. However, except in fatigue, the number of failures caused by microstructural defects is relatively small compared with the number resulting from macroscopic defects.

Example 2: Rupture of a Ferritic Steel Return Bend Because of Inclusions. After 2 years of service, a return bend from a triolefin-unit heater ruptured (Fig. 3a). The bend was made from 115-mm ($4\frac{1}{2}$-in.) schedule 40 (6.0 mm, or 0.237-in., wall thickness) pipe of ASTM A213, grade T11, ferritic steel. The unit operated at 2410 kPa (350 psi), with a hydrocarbon feed stream (85% propylene) entering at 260 to 290 °C (500 to 550 °F) and leaving at 425 to 480 °C (800 to 900 °F). The temperature of the combustion gas was 900 °C (1650 °F).

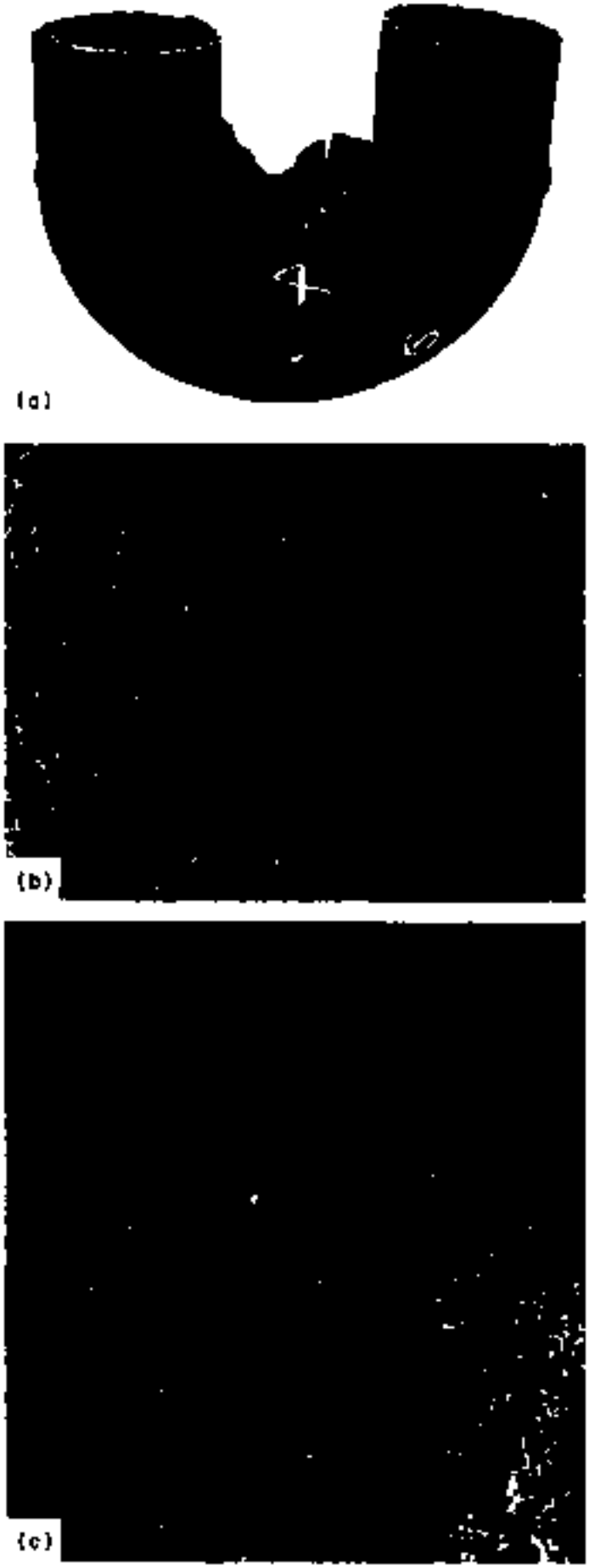


Fig. 3 Return bend made of ASTM A213, grade T11, ferritic steel that ruptured because it contained a large number of inclusions. (a) Overall view of the return bend showing

rupture. (b) Micrograph of an unetched specimen showing high concentration of inclusions. 400×. (c) Micrograph of an unetched specimen showing inclusions and cracks containing scale. 435×

This rupture occurred on the inner surface of the bend, rather than on the outer surface where erosion-induced failures usually occur. The fracture was limited to the return bend and terminated at the welds that joined the bend to the pipeline.

Investigation. Examination of the fracture surfaces with a low-power stereomicroscope revealed that very little thinning of the metal had taken place, indicating low ductility and little corrosive attack.

Metallographic specimens were prepared from regions adjacent to the fracture and at the outer surface of the bend. The steel near the fracture exhibited a high concentration of both small and large inclusions (Fig. 3b). The steel in the outer surface contained relatively few inclusions, most of which were very small.

A micrograph of a section through the fireside edge of the fracture surface is shown in Fig. 3(c). Branched cracks similar to those produced by stress corrosion of steel were observed. These cracks had apparently propagated between the inclusions in the steel.

Scale was observed over most of the crack path, even down to the fine tip of the crack. The scale provided stress raisers in two ways. First, the volume of the oxide exceeded that of the steel it replaced, placing the metal at the crack tip in tension and promoting a stress-corrosion reaction. Second, the effect of the oxide was magnified during thermal cycles because of differential thermal expansion, with the steel having a greater expansion coefficient than the scale. The high density of inclusions was assumed to be the result of improper cropping of the ingot from which the tube for the return bend was made.

Conclusions. The return bend failed as a result of stress-corrosion cracking (SCC), which propagated because of the presence of numerous inclusions in the metal. Scale formed in the cracks, abetted by thermal cycling, contributed to stresses that caused cracking.

Corrective Measures. There is no general remedy for this relatively rare type of failure, which is difficult to control because of its localized nature. It is not cost-effective to nondestructively examine all fittings for such defects, but material that is intended for critical applications where failure cannot be tolerated should be examined.

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Effects of Fabrication Practices

Mechanical or metallurgical imperfections in a workpiece can be introduced, and detected, at any stage of construction, including mill processing, fabrication, shipping, and erection. Imperfections do not always result in failure and are in fact a relatively infrequent cause of failures of pressure vessels during operation. The ASME Boiler and Pressure Vessel Codes recognize that perfection is seldom obtained in commerical materials and define acceptability standards for the producer.

Many imperfections are detected before actual operation. For example, a crack indication was found in a forged nozzle by magnetic-particle inspection before welding of the nozzle into a pressure vessel. The nozzle manufacturer had found the crack and, instead of rejecting the part, had attempted to repair it by welding, which was not permitted in the specification.

In another case, a piece of 50-mm (2-in.) diam schedule 80 (5.5 mm, or 0.218-in., wall thickness) ASTM A106 seamless carbon steel pipe that had been hydrostatically tested was about to be welded into a pipeline when the operator found a lap in the outer surface. The lap was about 75 mm (3 in.) long and extended through almost the entire wall. A slight ridge on the inner surface had prevented leakage through the lap during hydrostatic testing. A section across the lap showed light mill scale adhering to the sides with some decarburization. There was no evidence of branched cracking nor any indication of SCC. The condition of the lap surfaces indicated that the lap had been formed at a temperature above 870 °C (1600 °F), which would have occurred early in the tube-forming process. It was concluded that a foreign object had fallen on the tube as it was being formed and that the object had passed through the rollers, producing a dent in the tube wall. Subsequent forming operations produced a lap.

Example 3: Failure of a Utility Boiler Drum During Hydrotesting. A 150-cm (60-in.) inside-diameter boiler drum fabricated of 18-cm (7-in.) thick ASTM A515, grade 70, steel failed during final hydrotesting at a pressure of approximately 26 MPa (3.8 ksi). The shock wave from the vessel rupture also damaged several connected headers and

piping runs. A schematic of the crack path, including the location of three SA-106C nozzles, is shown in Fig. 4. The nozzles are numbered 1 to 3, moving away from the head.

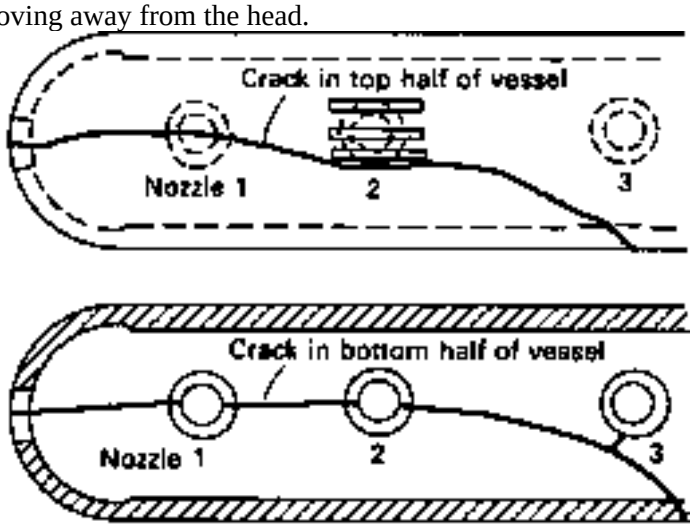


Fig. 4 Crack path in a failed utility boiler drum.

Investigation. Field examination occurred 2 days after the fracture. The fracture surfaces between nozzles 1 and 2 and through the manway (access port) in the head were classic brittle fractures. The remainder of the fracture involved tearing, and the crack led off from parallel to the axis of the vessel.

Samples were taken to allow examination of the brittle-fracture area and the fracture-initiation sites; they were flame cut and then sawed into manageable sizes for laboratory work. Chemical analyses of the materials of construction demonstrated that the original materials met the specified chemistries. Drop-weight tests, dynamic tear tests, and Charpy V-notch impact tests of the shell plate material, the head material, and the nozzle material indicated that these materials had ductile-to-brittle transition temperatures above room temperature.

Macroscopic observations of the fracture surfaces at nozzles 1 and 2 showed chevron patterns that allowed tracing the fracture to its initiation points. The fracture apparently initiated at the nozzle 1 to shell junction on the head side and on the shell side, propagated around the head to the back side of the vessel, and continued down the back side of the shell. On the nozzle side, the fracture reinitiated on the far side of each of nozzles 1 and 2 and continued down the shell.

Drain grooves had been arc gouged at the nozzle sites, and examination of the fracture surface at the grooves revealed short, flat segments of fracture area indicative of pre-existing cracks. These features were consistent with crack indications found during magnetic-particle and scanning electron microscopy (SEM) examination of the groove lips. The remainder of the edge along the groove surface exhibited small segmented shear lips.

Metallographic sections were taken through the drain groove and fracture surface at the forward sides of nozzles 1 and 2. In these sections, a thin layer of material with a dendritic structure was observed at the groove surface, and the HAZ resulting from the arc gouging was obvious (Fig. 5). Cracks were visible in the groove surface; most were confined to the outer dendritic layer, but at least one crack was observed to extend about 6 mm (0.25 in.) into the base metal. The microhardness of the dendritic layer was 489 to 591 HK and 47 to 52 HRC, while the HAZ was 179 to 258 HK and 85 HRB to 21 HRC. A qualitative microprobe analysis of the dendritic layer revealed that it contained over 1% C, higher than the carbon content of the base metal. This high hardness is consistent with the carbon content and indicates a high susceptibility to cracking upon cooling or deformation.

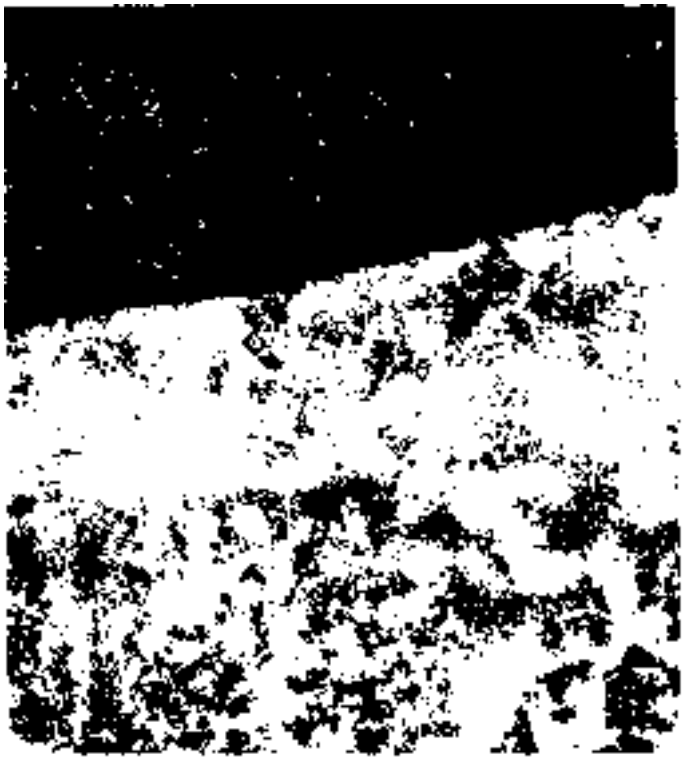


Fig. 5 Section through arc-gouged drain groove showing thin layer of dendritic structure containing cracks. Etched with nital. 55×

The microstructure of the main fracture surface in unaffected base metal tended to be straight and transgranular, characteristic of cleavage fracture. In regions in which the fracture intersected a weld HAZ, the fracture path became irregular and exhibited features suggesting ductile fracture within the zone. Ductile dimples were also observed in the HAZ at the arc-gouged drain grooves.

Conclusions. The cracks in the drain groove surface could have occurred (1) immediately after arc gouging as a result of differential thermal contraction between the resolidified layer and the base metal, (2) during the subsequent stress-relieving operation, or (3) during the hydrostatic test due to localized stresses exceeding the yield strength of the base metal. The ragged surface of the drain groove probably contributed to even higher local stresses than would a smooth groove.

Recommendations. Flame cutting of surfaces in these steels should be avoided because it can lead to local embrittlement and stress raisers that initiate major failures. In addition, the ductile-to-brittle transition temperature of materials of construction should be below the hydrostatic-test temperature.

Effect of Stress Raisers. Designers generally avoid configurations that can provide local stress raisers, but macroscopic stress raisers can occur as a result of fabrication errors. The most common are those that result from lack of complete weld penetration. Inspection practices are aimed at preventing pressure vessels containing inadvertently incorporated stress raisers from reaching service, but they occasionally do.

Example 4: Failure of a Steam Accumulator Due to Lack of Complete Weld Penetration. A 1.2-m (4-ft) OD × 4.3-m (14-ft) long steam accumulator, constructed with 10.3-mm ($\frac{13}{32}$-in.) thick SA515-70 steel heads and an 8-mm ($\frac{5}{16}$-in.) thick SA455A steel shell, ruptured after about 3 years of service, launching the body of the vessel into a neighboring house. The accumulator was used in a plastic molding operation and operated at a pressure of 827 kPa (120 psi), dropping to 780 kPa (113 psi) during the injection cycle. It was designed for 1035-kPa (150-psi) operation and had presumably been pressure tested at 1550 kPa (225 psi). Because the seals on both safety relief valves were intact, the valves were tested and found to operate properly at the required 1035-kPa (150-psi) set point.

Investigation. A field examination showed that the manhole end of the vessel had ruptured around approximately 270° of the circumference of the head-to-shell girth weld before continuing to complete separation in the head material. Visual examination indicated an extensive lack of weld penetration in this girth weld. Both sides of the weld were removed for laboratory examination (Fig. 6).

Fig. 6 Both sides of the fracture surface from a failed steam accumulator. Section of the vessel is the upper piece, and the mating head is the lower piece.

The initial laboratory examination after cleaning of the samples disclosed that the weld penetration ranged from less than 50 to 75% of the head thickness at the joint, rather than the 100% penetration required by the ASME Boiler and Pressure Vessel Code. At no location was there 100% weld penetration. Typical longitudinal cross sections through the weld are shown in Fig. 7. The joint had apparently been made in two passes, but the root pass was not always centered over the joint. There was also some misalignment of the head to the shell because of radial displacement of the shell and head centerlines. This resulted in excessive clearances between the two parts and a slight difference in the thicknesses of the parts.

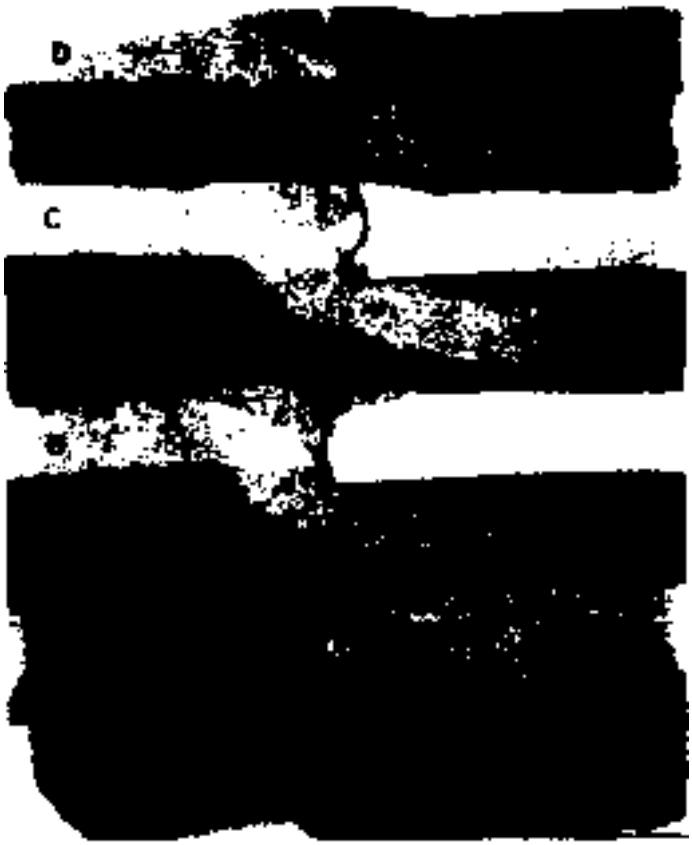


Fig. 7 Several sections through a girth weld in a field steam accumulator showing lack of weld penetration.

A hardness traverse on the shell yielded values from 84.1 to 86.5 HRB, corresponding to an approximate tensile strength of 550 MPa (80 ksi), which is within the 515- to 620-MPa (75- to 90-ksi) range required for SA455 material. The hardness of the head material ranged from 94.2 to 95.7 HRB, which indicated a tensile strength of approximately 695 MPa (101 ksi), which exceeds the 480- to 585-MPa (70- to 85-ksi) range for SA515-70 material. This higher hardness

was probably due to cold forming of the head flange. The weld-metal hardness ranged from 84.8 HRB at the root pass to 99.3 HRB at the finish-pass crown—acceptable values for this material.

There were no anomalies in the microstructure of either the weld metal or the base metal. The fracture path was clearly transgranular, with occasional secondary branches. Some small, secondary cracks were also observed. There were relatively few beach marks on the fracture surface, indicating the occurrence of only a few loading cycles. In all cases, the fracture path extended from the root of the weld through the weld metal or HAZ (up to the point at which the final-fracture path was in the head).

A stress analysis of the joint was performed, taking into account the various degrees of weld penetration and misalignment, as well as the differences in wall thickness of the shell and head. For proper alignment and complete weld penetration, the stresses in the failure region are quite low—about 25 MPa (3.57 ksi). Lack of penetration can cause the longitudinal stress to approach the yield strength of the material, and a small amount of misalignment added to lack of penetration increases the stresses to near the tensile strength of the material.

Conclusions. It was concluded that the failure could be attributed to two fabrication errors: lack of weld penetration and misalignment of mated parts. The failure was judged to be a short-cycle high-stress notch-fatigue failure.

Heat treatment may be required subsequent to a fabricating operation. Deviations from specifications are potential causes of failure. Figure 8 shows a section of tubing made of rimmed steel that had been cold swaged from 65 to 50 mm in diameter. Stress relief at 620 °C (1150 °F) after swaging was specified but was inadvertently omitted, and when the tube was expanded in a hole in a tube sheet, it broke in a brittle manner. The fracture was the result of strain aging of the steel, which can occur at room temperature and markedly decreases formability.

Fig. 8 Rimmed steel tube that failed by brittle fracture after being strain aged by cold swaging.

Repeated heat treatments, although performed at the proper temperature, may produce undesirable microstructures that can result in premature failure. For example, a stainless steel tube was solution heat treated, straightened, then reheat treated. The light surface deformation between heat treatments caused severe localized grain growth, which resulted in a coarse and less ductile grain structure. In service, grain-boundary cracking developed with relatively little swelling. Generally, this condition is harmful only when the tubing material has marginal strength or is stressed to an extent that results in measurable deformation.

An instance of improper heat treatment involved a pressure vessel made from forged HY-100 high-strength alloy steel plates that were subsequently heat treated. Heat-treating specifications called for water quenching, with the vessel being lowered into the bath in the vertical position to develop uniform properties. After heat treating, six slots were machined through the vessel shell to serve as seats for wedges designed to retain a removable closure. The plugs cut from the slots in the vessel wall were subjected to mechanical tests, which showed large variations in tensile strength and impact properties depending on circumferential location in the shell. In some locations, the values met specification; in other locations, they were considerably below specification. Investigation revealed that the vessel had been quenched in the horizontal position instead of the vertical position because there was not enough headroom over the available quench tank to permit lowering the vessel into the tank in the vertical position. The vessel was reheated and quenched in the vertical position to develop the specified properties.

An instance in which heat treatment was required after forming involved a gas-plant debutanizer tower made of 30-mm thick high-strength steel plate with properties as specified by ASTM A516, which permits, as an option, a refined normalized microstructure. Radiographic inspection of the entire vessel disclosed no imperfections. The vessel was designed to operate at 1550 kPa (225 psi) and 345 °C (650 °F). The vessel was hydrostatically tested using water at a temperature of not less than 15 °C (60 °F). When the test pressure reached 2760 kPa (400 psi), the vessel ruptured, and about one-third of the head was torn out. Chemical analysis and tensile testing showed that the composition and tensile properties of the steel complied with ASTM A516. Charpy impact tests made from 22 to 80 °C (72 to 175 °F) indicated that the ductile-to-brittle transition temperature was above 40 °C (100 °F). Metallographic examination showed a microstructure with a grain size coarser than ASTM 1, which indicated that the head had not been normalized after forming. The coarse grain size contributed to a reduction in fracture toughness of the steel. The vessel was completely

checked ultrasonically for other plates with large grains, and the head was replaced. The vessel was hydrostatically tested again, then returned to service. No further difficulties were encountered.

Overheating during fabrication can result in cracking of susceptible materials. Leakage was detected in the U-bend of a 50-mm (2-in.) diam stainless steel tube during preoperational hydrostatic testing. The tube had been heated for bending, with a specified maximum temperature of 980 °C (1800 °F). Metallographic examination of a section through the leakage area showed evidence of partial melting, which indicated that the specified forming temperature had been considerably exceeded.

Example 5: Cracking of a Fire-Extinguisher Case Because of Overheating. The top of a fire-extinguisher case (Fig. 9a) was closed by spinning. The case was made from 1541 steel tubing 170 mm (6.75 in.) in diameter and 4.2 mm (0.165 in.) in wall thickness. Leakage from the top of the case was observed during testing, and a 75-mm (3-in.) long specimen was sent to the laboratory for analysis to determine the cause.

Fig. 9 Fire-extinguisher case that failed because of ferrite streaks resulting from overheating during spinning. (a) Top of the case. Dimensions given in inches. (b) Micrograph showing ferrite streaks. 150×

Investigation. Visual examination of the specimen did not reveal any surface cracks. However, three small folds were observed on the surface, and one was sectioned for microscopic examination. A very fine transverse fissure through the section was visible. Examination of a micrograph (Fig. 9b) showed streaks of ferrite, which can be formed when steel is heated approximately to the melting point. Examination of specimens from other areas of the case exhibited a similar structure. The normal microstructure of 1541 steel for the specified heat treatment was annealed ferrite and pearlite.

The chemical composition was normal for 1541 steel. No steelmaking or tubemaking defects were found in the specimen.

Conclusions. Cracking of the top of the fire-extinguisher case was the result of ferrite streaks that were formed when the metal was overheated. The initial forming temperature of the case plus frictional heat generated during spinning resulted in extreme overheating and some localized melting.

Recommendations. The temperature of the metal must be more closely controlled so that the spinning operation is done at a lower temperature, thus avoiding incipient melting and formation of ferrite streaks.

Example 6: Failure of Piping System Cross by Intergranular Cracking Traceable to Improper Heat Treatment (Ref 2). During a routine maintenance procedure at a power station, dye-penetrant examination revealed linear indications on the outer surface of a cross in a piping system. The cross was specified as SA403 type WP 304 stainless steel that had originally been extruded as a 65-cm (26-in.) diam, 75-mm (3-in.) thick pipe section. The material had been bright dipped, sand blasted, dye penetrant inspected, and passivated by the manufacturer. A spare cross from a sister power station was examined and found to have similar indications. Both crosses had been subjected to an induction-heating stress improvement-induction heating the outer wall of the pipe to 550 to 600 °C (1020 to 1110 °F) maximum for a maximum of 12 min while the inside is cooled with water. This treatment leaves a residual tensile stress on the outer surface and a residual compressive stress on the inner surface of the pipe.

Investigation. The linear indications on the cross in service were located in 150- to 180-mm (6- to 7-in.) wide bands running circumferentially below the cross-to-cap weld and above the cap-to-discharge-pipe weld. The indications were random in orientation, measured up to 20 mm (0.75 in.) in length (excluding one very large indication that measured 90 mm, or 3.5 in., in length), and averaged about 6 mm (0.25 in.) in depth. The indications were predominantly located in the thicker wall sections of the cross.

The available specimen of the unused cross was visually inspected. It consisted of about a 180-mm (7-in.) square that had a maximum thickness of 60 mm (2.37 in.). It included a weld-preparation area that was about 16 mm (0.625 in.) thick.

One side of the specimen had been flame cut, while the other had been sawed. The outside surface appeared to be coarse ground and had cracks visible on the surface. Record photographs were taken.

A dye-penetrant examination was performed using a two-step penetrant method. There were numerous indications on the flame-cut side; none were on the saw-cut side of the specimen. The longest single indication was approximately 19 mm (0.75 in.), while the average length was about one-half that. The cracks appeared to be intergranular.

The pipe was chemically analyzed and found to conform to SA182 type F 304 stainless steel specifications. It contained 0.2% C, 1.56% Mn, <0.005% O, 0.016% S, 0.66% Si, 18.0% Cr, 9.6% Ni, and 0.12% Mo.

Knoop hardness measurements were made with a 300-g load on a cross section of the material. The minimum hardness for 28 readings was 215, the maximum hardness was 298, and the average was 267, corresponding to 23.1 HRC. This hardness corresponds to a tensile strength of 807 MPa (117 ksi), and the material should therefore meet the 517-MPa (75-ksi) tensile-strength requirement.

A cross section was ground, polished, lightly etched in 10% oxalic acid, and examined optically. Intergranular cracks filled with oxide were observed to have depths up to about 8 mm (0.3 in.); one crack appeared to be transgranular when viewed with an optical (light) microscope. The grain size of the material exceeded the 00 ASTM grain size defined by ASTM Standard E 112 (Ref 3).

To determine whether the material had been sensitized by the induction-heating stress-improvement treatment, the oxalic acid etch test, Practice A of ASTM A 262 (Ref 4), was implemented. The polished specimen was etched at a current density of 1 A/cm^2 (6.5 A/in.^{2}) for 1.5 min in 10% oxalic acid-demineralized water, using a piece of type 304 stainless steel as the anode. There was no evidence of continuous grain-boundary precipitates typical of sensitized material. The cross section was repolished and etched electrolytically with NaOH to search for other second phases, such as σ and χ; no evidence of these phases was found.

Three mechanical test specimens with cracks were cut from the specimen and notched. The specimens were soaked in liquid nitrogen and broken. The large grain size of the material was apparent on the intergranular fracture surface (Fig. 10). A rust-colored oxide film was visible on the part of the fracture surface that corresponded to the cracked volume of material. At higher magnification, evidence of some plastic deformation before fracture was found. Energy-dispersive spectrometry (EDS) scans of an oxidized area of the fracture revealed traces of sulfur and chlorine, probably from the dye penetrant, and the iron, chromium, and nickel normally present. At another location, traces of potassium and titanium were found. Analyses of other areas did not disclose definitive evidence of contaminants to support SCC as the failure mechanism.

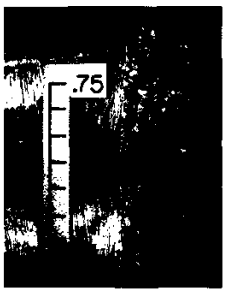


Fig. 10 Fracture surface of mechanical test specimen from piping cross. Fracture is intergranular. The coarse grain size of the material is evident. Note 0.75-in. scale.

The possibility of intergranular and hot cracking during welding is high for materials containing high δ-ferrite. Ferritescope measurements on this material yielded results in the range from 0.3 to 1.6% δ-ferrite, too low for this mechanism to have been operative.

Conclusions. The results of the analysis eliminated intergranular SCC; the material was not in a sensitized condition, and no definitive evidence of contaminants (other than the dye-check fluid) was found. No detrimental second phases were found, nor was there significant ferrite to cause hot cracking during welding. The remaining possibility was over-heating or burning of the forging, which classically results in large grain size, intergranular fractures, and fine oxide particles dispersed throughout the grains. The induction-heating stress-improvement process was suspected to have induced tensile stresses into the outside of the cross that opened pre-existing cracks so that they were found by dye penetrant.

Multiple Causes. Extensive examination of a failed vessel sometimes reveals that failure was caused not by one but by several interrelated inadequacies in fabrication, as in the following example.

Example 7: Failure of a Thick-Wall Alloy Steel Pressure Vessel Caused by Cracks in Weld HAZ (Ref 5). A large pressure vessel (Fig. 11a) designed for use in an ammonia plant at a pressure of 35 MPa (5.1 ksi) at 120 °C (250 °F) failed at 34 MPa (5.0 ksi) during hydrostatic testing. The intended maximum test pressure was 48 MPa (6.95 ksi). The vessel weighed 166 Mg, measured 18.2 m in overall length, and had an outside diameter of 2.0 m. It was fabricated from ten manganese-chromium-nickel-molybdenum-vanadium steel plates 150 mm thick, which were rolled and welded to form

ten cylindrical shell sections and three forgings of similar composition. The forgings formed two end closures and a flange for attaching one of the forged end closures to the vessel.

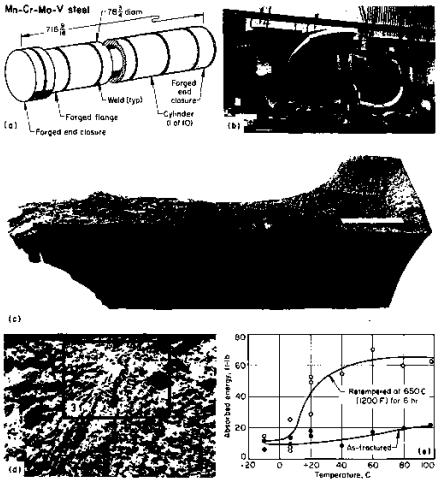


Fig. 11 Large thick-wall pressure vessel that failed because of cracking in weld HAZ. (a) Configuration and dimensions (given in inches). (b) Shattered vessel. (c) General appearance of one fracture surface; arrow points to facet at fracture origin. (d) Enlarged view of region at arrow in (c). 5.5×. (e) Charpy V-notch impact-strength values of specimens from weld area, both as-fractured and after retempering at 650 °C (1200 °F) for 6 h. Source: Ref 5

The failure did extensive damage to one end forging and three adjacent shell sections. Four large pieces were blown from the vessel; the largest, weighing about 2 Mg , penetrated the shop wall and traveled a total distance of 46 m (152 ft). Pieces of the failed vessel are shown in Fig. 11(b).

Fabrication History. The cylindrical shell sections were hot formed so that the rolling direction of the plates was perpendicular to the axis of the vessel. All plates were supplied in the normalized-and-tempered condition. The forgings were annealed, normalized, and tempered at 645 °C (1195 °F) to obtain the desired mechanical properties.

The longitudinal seams in the cylindrical shells were electroslag welded, and the welds were ground to match the curvature of the shell. The cylindrical shell sections were heated at 900 to 950 °C (1650 to 1740 °F) for 4 h, rounded in the rolls within that temperature range, and then cooled in still air for examination of the seams.

Circumferential welding was done by the submerged arc process, using preheating at 200 °C (390 °F). Each subassembly was stress relieved by being heated to 620 to 660 °C (1150 to 1220 °F) for 6 h.

Final joining of the three subassemblies followed the same welding procedures except that localized heating for stress relief was employed. During various stages of manufacture, all seams were examined by gamma radiography, automatic and manual ultrasonic testing, and magnetic-particle inspection.

Hydraulic Testing. The vessel was closed at the top, and water was admitted to the vessel through a fitting in the cover plate until the vent that was within 3 mm of the inside surface at the top showed a full-bore discharge of water, at which time it was blanked off. The specified hydrostatic proof test called for a pressure of 48 MPa (6.95 ksi) at an ambient temperature of not less than 7 °C (45 °F).

At 34 MPa (5.0 ksi), the customary halt was made, and about 30 s later, the flange end of the vessel exploded without warning. The flange forging was found to be cracked completely through in two locations, and the first two cylindrical sections were totally shattered.

Metallurgical Examination. The fracture surfaces were typical for brittle steel fracture. There were two points of origin, both associated with the circumferential weld that joined the flange forging to the first shell section.

The general appearance of one fracture surface of the flange forging is shown in Fig. 11(c). The arrow near the center indicates a flat facet (with major dimensions of roughly 9 mm, approximately 14 mm below the outer surface of the vessel, which was partly in the HAZ on the forging side of the circumferential seam weld. Figure 11(d) shows an enlarged view of the region containing this facet. A slightly larger facet was found on another fracture surface of the flange, near a fracture origin 11 mm below the outer surface. This facet was also situated partly on the forging side of the HAZ. These flat, featureless facets in the HAZ were the fracture-initiation sites.

Metallographic examination revealed that the structure immediately below each facet in the HAZ was a mixture of bainite and austenite having a hardness (1-kg load) of 426 to 460 HV. Elsewhere, the structure of the HAZ was coarse and jagged bainite with a hardness of 313 to 363 HV.

Examination of a section transverse to the weld disclosed that the structure of the flange forging contained pronounced banding, whereas the plate did not. The structure between the bands (and away from the weld) consisted of ferrite and pearlite with a hardness of 180 to 200 HV (10-kg load). In the bands, the structure was upper bainite with a hardness of 251 to 265 HV. Specimens cut from this area were austenitized at 950 °C (1740 °F) and quenched in a 10% solution of NaOH in water to produce pure martensite. A Vickers hardness traverse (1-kg load) across a band showed average values of 507 on one side of the band, 549 within the band, and 488 on the other side. A microprobe scan across the banded area showed the following differences in chemical composition: 1.56% Mn outside the band, 1.94% within it; 0.70% Cr outside, 0.81% within; and 0.23% Mo outside, 0.35% within. These differences indicated higher hardenability within the bands, which suggested a greater susceptibility to cracking, especially where the bands met the HAZ of the circumferential weld.

The hard spots within the forging suggested that during stress relief the vessel had not actually attained the specified temperature. To check this, specimens from the HAZ were accordingly heated to a variety of tempering temperatures to observe at what point softening would occur. Vickers hardness tests (1-kg load) on hard-spot specimens showed no deviation from the as-failed hardness scatter band for temperatures up to 550 °C (1020 °F), but gave definitely lower values after retempering at or above 600 °C (1110 °F). A similar behavior occurred in specimens taken outside the hard spots. It was therefore concluded that stress relief had not been performed at the specified temperature level.

Other studies of microstructure established that the mode of crack propagation was transgranular cleavage, which occurred not only in prior-austenite grains but also in ferrite grains and pearlite colonies. Branching and subsidiary cracks were also found, all very close to the main fracture surface. The distance from the fracture to the crack farthest away from it was 0.5 mm (0.020 in.).

To further investigate the effect of the stress relief on the properties of the vessel, standard Charpy V-notch specimens were prepared from the flange forging, the plate, and the weld. All specimens were oriented with their length parallel to the circumference of the vessel so that the induced fractures would be parallel to the main fracture. Tests were performed on as-received vessel material over the temperature range of -10 to 100 °C (15 to 212 °F), yielding impact values that met the specifications for the forging and plate, but the weld metal exhibited inferior impact strength (lower curve, Fig. 11e). None of the materials displayed a sufficient capacity for energy absorption (41 J, or 30 ft · lb, for the forging; 69 J, or 51 ft · lb, for the plate; and 16 J, or 12 ft · lb, for the weld) to have arrested the crack propagation that led to failure of the pressure vessel. Other specimens of the weld metal, which were retempered at 650 °C (1200 °F) for 6 h and tested over the same temperature range as those discussed above, showed a marked improvement in impact strength at or above 20 °C (70 °F) (upper curve, Fig. 11e). This confirmed the previous deduction that the stress relief of the vessel was at a lower temperature than the specified temperature.

Conclusions. Failure of the pressure vessel stemmed from the formation of transverse fabrication cracks in the HAZ of the circumferential weld joining the flange forging to the first shell section. The occurrence of the cracks was fostered by the presence of bands of alloying-element segregation in the forging, which had created hard spots, particularly where they met the HAZ. Stress relief of the vessel had been inadequate, leaving residual stresses and hard spots and providing low notch ductility, especially in the weld.

Recommendations. The final normalizing temperature should be Ac_3 + 50 °C (Ac_3 + 90 °F), and the tempering temperature should be 650 °C (1200 °F), where Ac_3 represents the temperature at which transformation of ferrite to austenite is complete during heating. After completion of each seam weld, the weld preheat should be continued for a short distance along the seam. The critical flaw size for this alloy and section thickness should be determined by fracture mechanics, and the capacity of available nondestructive testing methods for detecting flaws of this size should be assessed. The possible benefits of a more effective solution treatment or a revised forging practice should be explored to eliminate the banding in the flange forging.

Example 8: Failure of a Reheat Steam Piping Line at a Power-Generating Station. A 75-cm (30-in.) OD by 33-mm (1.31-in.) thick pipe in a horizontal section of a hot steam reheat line ruptured explosively with serious damage and loss of life after approximately 15 years of service. The line nominally operated at 4135 kPa (600 psi) and 540 °C (1000 °F). The section that failed was manufactured from rolled plate of material specification SA387, grade C. The longitudinal seam weld was a double butt weld that was V welded from both sides. The failure propagated along the longitudinal seam and its HAZ for a distance of 2.7 m (9 ft.) in both directions from the point of origin, creating a clam-shell opening 2.1 m (7 ft) wide. On one end, the failure traveled through a circumferential seam that deflected its path and stopped about 150 mm (6 in.) past the weld. On the other end, the fracture stopped several feet short of a circumferential seam while traveling in the longitudinal seam HAZ. The fracture in the pipe was pointed directly toward occupied regions of the plant.

Investigation. A systematic review of the circumstances of the failure revealed a number of factors that may have played a role in the failure. In the design of the system, the original material specified was ASTM A155 fusion-welded pipe that had been substituted with SA387, grade C, plate, rolled and welded with filler metal added. At the time of the design of the system, the allowable design stress for SA387, grade C, material was 48 MPa (6.9 ksi) at 540 °C (1000 °F). Approximately 3 years after the plant was put into service, this was reduced to 46 MPa (6.6 ksi) at 540 °C (1000 °F).

The hydrostatic test pressure specified for the system was sufficiently high that, if it was actually applied, it is possible that the pipe cylinder and its weldment were overstressed, perhaps to the point of being damaged. The actual test pressure

could not be determined in the original investigation. Two other unconfirmed possibilities are operation of the pipe at higher-than-anticipated service temperatures for short times during its service life and earthquake loadings acting on the pipeline.

Several metallurgical factors were found during a laboratory analysis of samples from the failed pipe. However, two key pieces, the fracture area containing the fracture-initiation point and the end of the crack in the longitudinal seam weld, were not available for this analysis. Three samples containing parts of the fracture surface were received: a 165- × 285-mm (6.5- × 11.25-in.) section near the end of the fracture in the longitudinal seam weld as well as 205- × 405-mm (8- × 16-in.) and 180- × 330-mm (7- × 13-in.) sections from either side of the point of origin of the fracture. One 150- × 305-mm (6- × 12-in.) section, including the undamaged longitudinal weld and a circumferential weld to a fitting, was also received. These samples were removed by flame cutting, and care was taken to avoid the heat-affected cut zone in all testing.

The fracture surface near the inner wall of the pipe had a bluish gray appearance, while the fracture surface near the outer wall was rust colored. Two of the samples had a distinct gray color halfway through the thickness of the reheat pipe. Before sectioning, all fracture surfaces were coated with clear acrylic, which was later removed with acetone for optical microscope examination. The rust colored oxide could not be cleaned off, while the bluish gray coating appeared thick at magnifications of 10 to 70×. It was very brittle with a mud-crack appearance, and appeared to consist of a mixture of oxides.

The base metal, which was analyzed in three locations, contained:

Element	Composition, %
Carbon	0.14
Manganese	0.51-0.54
Phosphorus	0.016-0.018
Sulfur	0.011-0.015
Silicon	0.54-0.57
Copper	0.023-0.028
Tin	0.003-0.006
Nickel	0.022-0.024
Chromium	1.39-1.43
Molybdenum	0.50-0.54
Aluminum	0.000
Vanadium	0.005-0.006
Niobium	0.006-0.011
Zirconium	0.001-0.002
Titanium	0.000
Boron	0.0001-0.0003
Cobalt	0.004-0.006

The weld filler metal, including both inside-diameter and outside-diameter samples, contained:

Element	Composition, %
Carbon	0.096-0.10
Manganese	0.81-0.91
Phosphorus	0.015-0.018
Sulfur	0.013-0.015
Silicon	0.47-0.54
Copper	0.13-0.40
Tin	0.005-0.013
Nickel	0.10-0.12
Chromium	1.28-1.43
Molybdenum	0.50-0.54
Aluminum	0.003-0.007
Vanadium	0.007-0.008
Niobium	0.009-0.013
Zirconium	0.001-0.002
Titanium	0.000
Boron	0.0003-0.0005
Cobalt	0.007-0.009

These contents conform to the requirements of SA387-67, grade B and C, as well as SA387-82, grade 11. The weld-metal analyses were from the centers of the welds and may differ from other areas of the weld; the weld metal and base metals were thought to be compatible.

Each of the samples was polished and etched for metallographic examination in the planes normal to the longitudinal and/or circumferential welds. Cracks were found in and around the root-pass areas of the three samples containing the longitudinal seam weld. There were several secondary cracks on the fracture side and a short internal crack in one specimen (Fig. 12(a)).

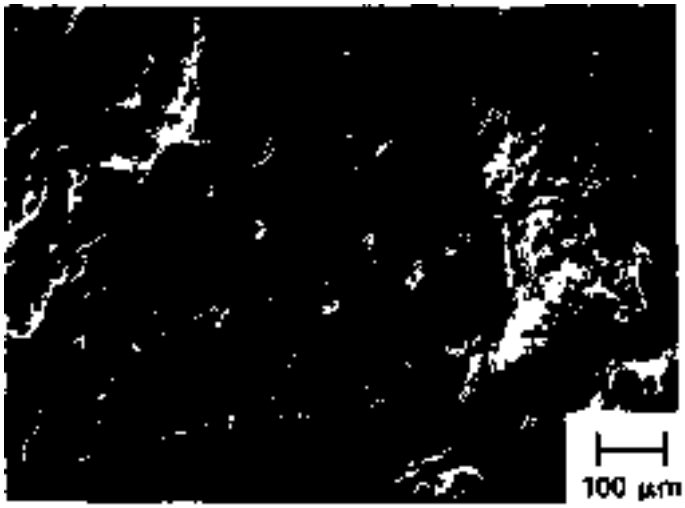


Fig. 12(a) Fracture surface of reheat steam pipe showing corrosion products covering early-fracture region and freshly exposed fracture surface of weld metal.

The sample containing the circumferential weld was examined using radiography and was found to be acceptable. Radiography of the samples containing the longitudinal seam weld detected only one sample with a crack, which ran parallel to the weld seam.

Longitudinal and transverse tensile specimens were cut from the pipe, and transverse-to-the-weld specimens were cut from the longitudinal and circumferential welds from the undamaged specimen. All room-temperature tensile properties were within specification: tensile strengths ranged from 476 to 517 MPa (69 to 75 ksi), yield strengths ranged from 265 to 305 MPa (38.5 to 44.3 ksi), and elongations ranged from 24 to 27%, except for the transverse-to-the-weld specimen from the longitudinal seam weld, which had a value of 19% (compared with 22% from the specification). That sample has 34% reduction of area and a shear-type failure rather than the cup-cone failures in the 63 to 70% reduction-of-area range of the other samples. It was concluded that the welded longitudinal seam exhibited embrittlement.

A Rockwell hardness traverse across the longitudinal weld from the undamaged region did not disclose a systematic variation in hardness across the weldment; hardnesses from 72 to 75 HRB were reported. A similar traverse across the circumferential weld revealed higher hardness in the weld metal and HAZ, reaching a value of 87 HRB. Hardnesses along the fracture edge of a longitudinal weld sample ranged from 70 to 80 HRB.

Several sections were cut from the fracture-surface samples and prepared for both optical microscopy and SEM. The rust spots were removed with phosphoric acid, which also removed or attacked some of the other corrosion products, changing their color from bluish gray to black. The microstructure of the material is predominantly an equiaxed ferritic matrix, with corrosion pits and grain-boundary precipitates. The fracture surface in the vicinity of the origin of fracture, near the outside diameter, is mainly ductile dimples, but corrosion pits and cracks are also present; by contrast, the tensile sample from the undamaged region had only ductile dimples. The pictures taken of the fracture using a scanning electron microscope (Fig. 12(a), 12(b), 12(c), and 12(d)) were consistent with a low-ductility intergranular fracture that progressed through the weld metal and had been in existence long enough to allow the formation of corrosion products on the fracture surface. Energy-dispersive x-ray analyses (EDAX) of the corrosion products showed them to be high in iron, with silicon, phosphorus, calcium, and manganese in small amounts, together with light elements, such as oxygen or carbon. Examination of the structure of the longitudinal weld near the fracture-initiation point revealed the presence of a grain-boundary phase (Fig. 12(d)), whose EDAX spectrum was high in silicon, sulfur, chromium, manganese, and iron.

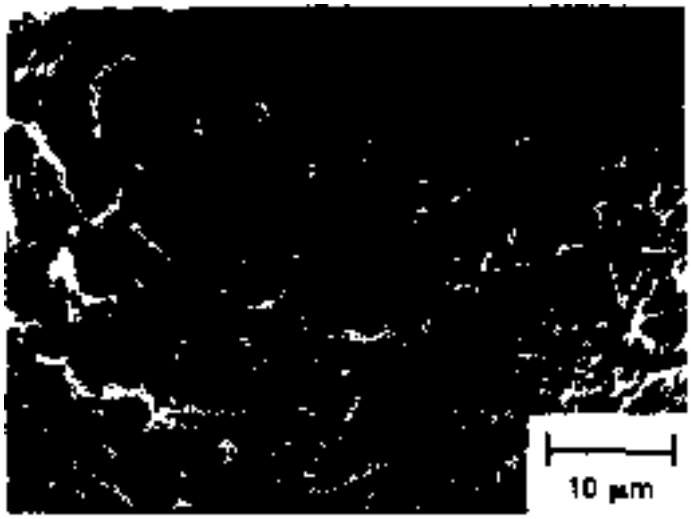


Fig. 12(b) Close-up of weld metal showing intergranular cracks and appearance of brittle failure.

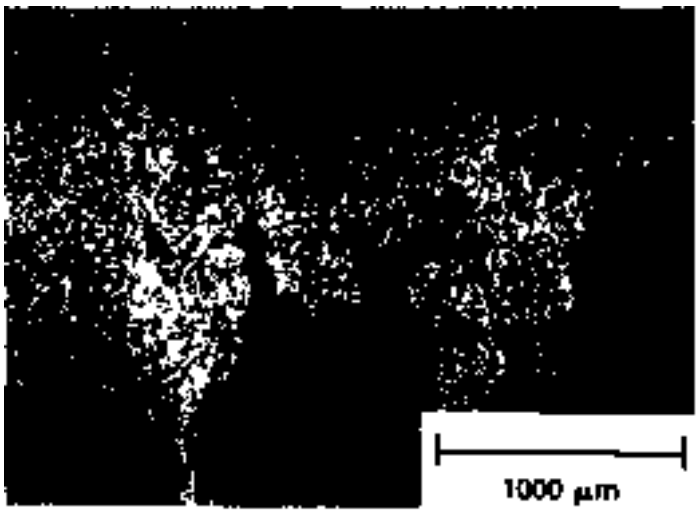


Fig. 12(c) Fracture surface near root pass showing corrosion products on part of fracture surface.

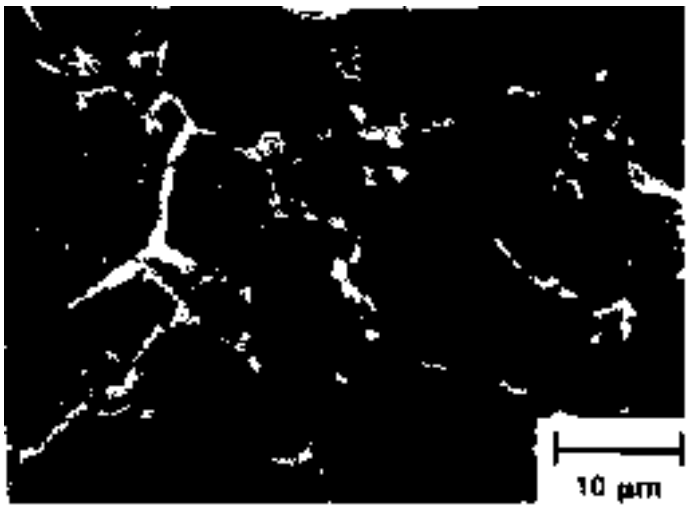


Fig. 12(d) Microstructure of longitudinal weld metal near fracture-initiation point. Note the white phase along grain boundaries.

Conclusions. The metallurgical analyst concluded that low ductility was responsible for the original initiation of cracks in the pipe. The low ductility may have been due to the presence of carbides and sulfides in grain boundaries, and corrosion may have played a major role in the deterioration of the steel in the weld-joint area. The extent and amount of corrosion products indicated that the cracks were in existence for some time before the final failure. Further work was deemed desirable to confirm these findings. It was not clear which of the design and construction, operating, and metallurgical factors played the dominant role in this failure, and further investigation of these factors may be required.

Example 9: Shell and Head Cracking in Gray Cast Iron Paper Machine Dryer Rolls. Unlike the individual examples of failures discussed throughout this article, several failures of dryer rolls will be addressed in this example. Most of these gray iron structures were cast in the 1890–1920 time frame when they were not covered by pressure vessel codes. The rolls are typically 0.3 to 0.9 m (1 to 3 ft) in diameter and range to about 3.7 m (12 ft) in length, with ground outer cylindrical surfaces on which the paper web is dried. They are unique pressure vessels in that they rotate about their longitudinal axes at speeds from 50 to 250 rpm while containing saturated steam from 35 to 380 kPa (5 to 55 psi). During its service life, a roll may be subjected to a variety of unrecorded operating conditions. Repair, regrinding, and wear in service all change the dimensions and stresses applied under given loads. Several constructions are found. The heads may be cast integrally with the cylindrical bodies or cast separately and bolted to the shells. The wall thicknesses of the shells, and especially those of the heads, may vary; artistically cast company names, logos, and raised reinforcements are frequently found.

Service failures typically occur in the shell body, in a head near a hand hole or a manhole opening, or in a head near the journal-to-head interface. Examples of these failures are shown in Fig. 13 and 14. Detailed fractographic analyses using SEM techniques invariably show a cleavage fracture, regardless of the driving stress for failure. Fracture surfaces may exhibit the chevron marks typical of fatigue or may have raised points or tears pointing in the direction of the probable origin of failure. Heavy rust, patches of oil, or other contaminants can indicate the existence of cracks older than the final failure.

Fig. 13 Failures of gray cast iron paper-roll driers. (a) Axial-shell failure. (b) Circumferential-shell failure

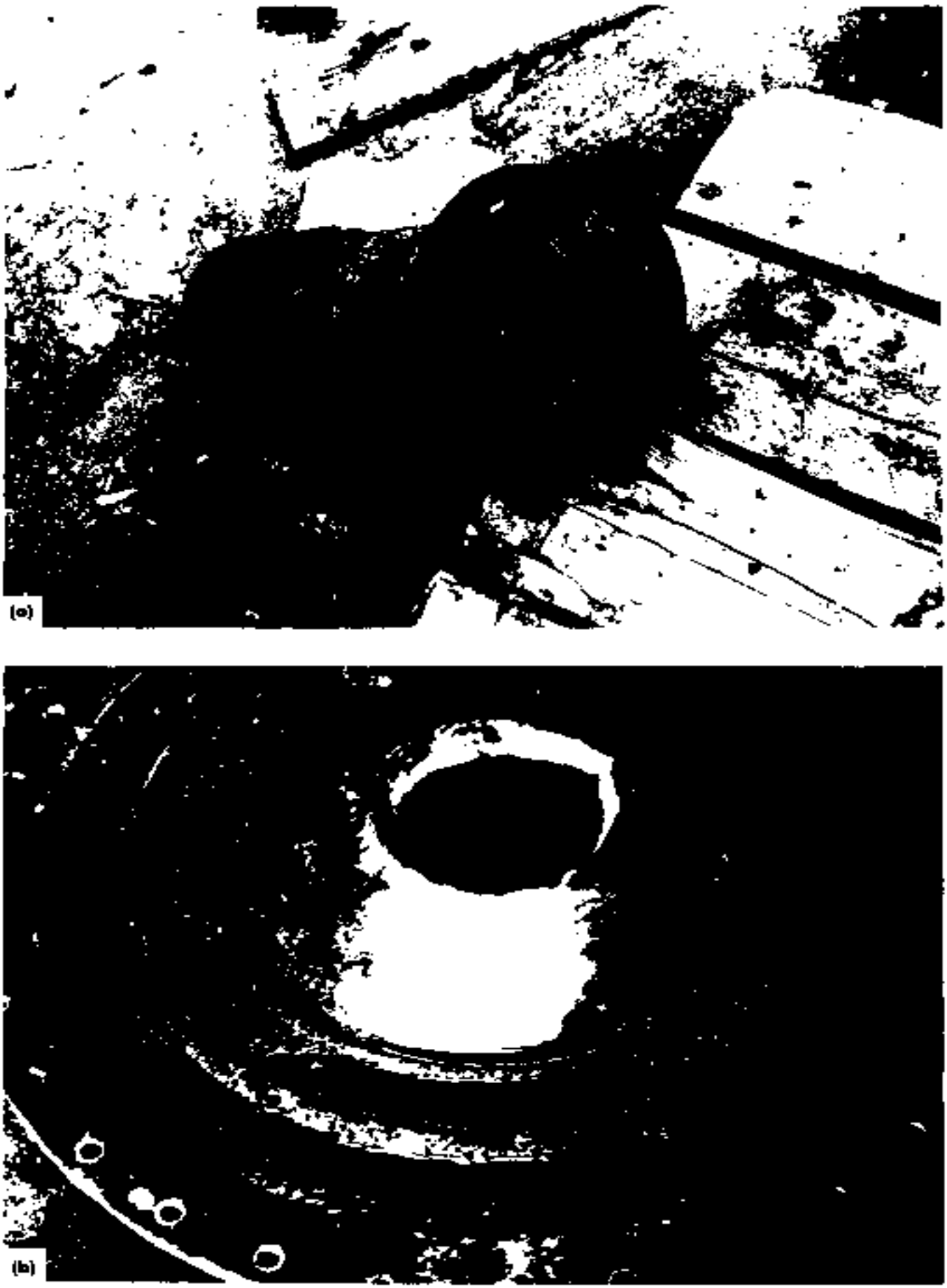


Fig. 14 Both sides of a journal-to-head failure in a cast iron paper-roll dryer. (a) Bottom of failed dryer. (b) The failed bolted-on head

When analyzing one of these failures, the analyst should be aware of the following characteristics of thin-wall cast iron structures:

- Castings are frequently poured from multiple ladles; therefore, the chemical composition and composition-dependent properties can vary widely within a given roll
- The cooling rates and solidification directions can vary dramatically with locations, producing anisotropic structures with variations in tensile strength up to 20% in different principle directions. The shapes and orientations of inclusions and other defects also vary widely. A single orientation of samples for metallographic analysis is not generally sufficient to characterize a microstructure
- The combinations of these factors can lead to tensile-strength variations up to 20% within a shell (excluding anisotropy effects)

For any specific failure, an immediate cause may be determinable, either in the material or in the recent service history. However, the analyst should be wary of the long histories of these vessels. Owners and operators of these vessels are concerned with developing end-of-life criteria and inspection methods for these and similar older pressure vessels.

Inspection Practices. The lower limits of detectability of various nondestructive inspection tools are of major importance in failure prevention. The effectiveness of radiographic, ultrasonic, and eddy current techniques depends largely on size and orientation of the discontinuity in the metal.

Touching the surface of a pressure vessel with an energized coil or probe used to magnetize a component for magnetic-particle inspection may produce an arc burn, which is a potential stress raiser.

The water used in hydrostatic proof testing and the liquid vehicles used in liquid-penetrant inspection can leave chloride residues that can later cause cracking unless the parts are carefully cleaned and dried after testing. For example, a type 304L stainless steel pipeline used for transferring caustic materials was hydrostatically tested after fabrication and was found acceptable. However, after about 2 months of service, leaks were detected at several circumferential weld joints. The pipe had been installed in a horizontal position, and the leaks occurred predominantly at the bottoms of the joints (6 o'clock position), with a higher incidence of leakage at low points in the pipeline. Leakage occurred in both shop and field circumferential welds; none was observed in the longitudinal seam in the welded pipe.

Investigation revealed that the leaks had resulted from chloride pitting. The source of the chlorides was determined to be the water used in hydrostatic testing of the pipeline. After the hydrostatic test had been completed, the water, which originally contained less than 100 ppm chlorides, was drained from the system. It was determined that some water had collected at low points in the pipeline and at discontinuities resulting from the weld reinforcement on the pipe inner surface. The piping was allowed to sit in this condition for about 4 to 5 months before being used to transfer caustic. Evaporation and chloride concentration, followed by corrosion and leakage, could have been eliminated if the pipeline had been completely drained after hydrostatic testing.

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Pressure Vessels Made of Composite Materials

Composite materials are used in pressure vessels to reduce cost and to conserve strategic material. Also, they provide vessels strong enough to resist deformation under pressure with surfaces that resist attack by the contained substance. Rigidity is usually provided by a carbon steel shell. A stainless steel or corrosion-resistant alloy lining is common.

Linings used for pressure vessels made of composite materials are applied by surface welding, roll cladding, plug welding, or explosive bonding. Surface-welded linings are produced by facing the interior carbon steel surface with an overlay of a corrosion-resistant alloy, using one of several processes for surface welding. These processes include manual shielded metal arc welding, automatic and semiautomatic submerged arc welding, and plasma arc welding.

Roll cladding is usually accomplished by one of two methods. In one method, stainless steel is placed on a carbon steel ingot and bonded to it by hot rolling. The other method is essentially the same except that a flux is placed between the two materials before hot rolling. Plugwelded linings are produced by securing roll-formed sections of a corrosion-resistant alloy to the interior surface of a carbon steel shell by plug welding.

Explosive bonding produces an excellent bond and an excellent product. The layer of cladding material is bonded to the base metal by a shock wave generated in a suitable medium by a controlled explosion.

Welding. When composite materials are joined by welding, special techniques are required, and metallurgical problems can arise. With roll-clad material, the overlay is machined away from the weld site, and the carbon steel shell is welded by the usual methods from the outside. The inside welds are made with stainless steel filler metals, such as ER308, ER309, ER310, or ER312. In plug weldings, similar techniques are implemented. Either 300 or 400 series stainless steel sheets may be used, and these sheets are welded together and to the carbon steel pressure vessel with stainless steel filler metal.

Metallurgical problems arise from the effects of dilution of the alloying constituents in the corrosion-resistant alloy coating by the carbon steel base metal. Fully austenitic stainless steel filler metals are susceptible to microfissuring during welding. A duplex structure of austenite and a small amount of ferrite in the weld deposit prevents microfissuring. Therefore, the ferrite content is usually kept between 5 and 9%. This duplex structure may be obtained by suitably balancing the chromium and nickel contents in the weld metal. Alloy dilution can be controlled by varying the welding technique and the electrode composition. Small dilutions can produce fully austenitic structures, but too much produces structures containing martensite.

Sigma phase is sometimes found in austenitic stainless steel weld metal. Although σ phase can occur in the as-welded condition, it is more commonly encountered after long exposures at temperatures of 595 to 870 °C (1100 to 1600 °F). Sigma phase may embrittle austenitic stainless steel at room temperature; at elevated temperatures, it can cause low ductility failures during creep deformation. Sigma phase is usually associated with ferrite; alloying elements, such as chromium, niobium, and molybdenum, which stabilize ferrite, also promote σ-phase formation.

Care must be used when welding Monel to carbon steel to avoid excessive dilution of the Monel with iron, such as in weld overlays. Excessive iron dilution embrittles the Monel and causes cracking. Iron dilution can be minimized by using a barrier layer of pure nickel weld deposit or by carefully controlling the welding procedure with the use of small-diameter electrodes and a stringer-bead technique.

Example 10: Cracking in Plug Welds That Joined a Stainless Steel Liner to a Carbon Steel Shell (Ref 6). Repeated cracking occurred in the welds joining the liner and shell of a fluid catalytic cracking unit that operated at a pressure of 140 kPa (20 psi) and a temperature of 480 °C (895 °F). The shell was made of ASTM A515 carbon steel welded with E7018 filler metal. The liner was made of type 405 stainless steel and was plug welded to the shell using ER309 and ER310 stainless steel filler metal.

Investigation. Examination of the microstructure in the plug-weld zone revealed a good duplex structure with what appeared to be fine cracks starting inside the weld zone and spreading outward through the weld and toward the surface (Fig. 15a). At higher magnifications, some of the fine cracks appeared to be precipitates (Fig. 15b). When cracking had completely penetrated the plug weld, corrosion of the carbon steel occurred, spreading radially. Also, there was decarburization and graphitization of the carbon steel at the interface.

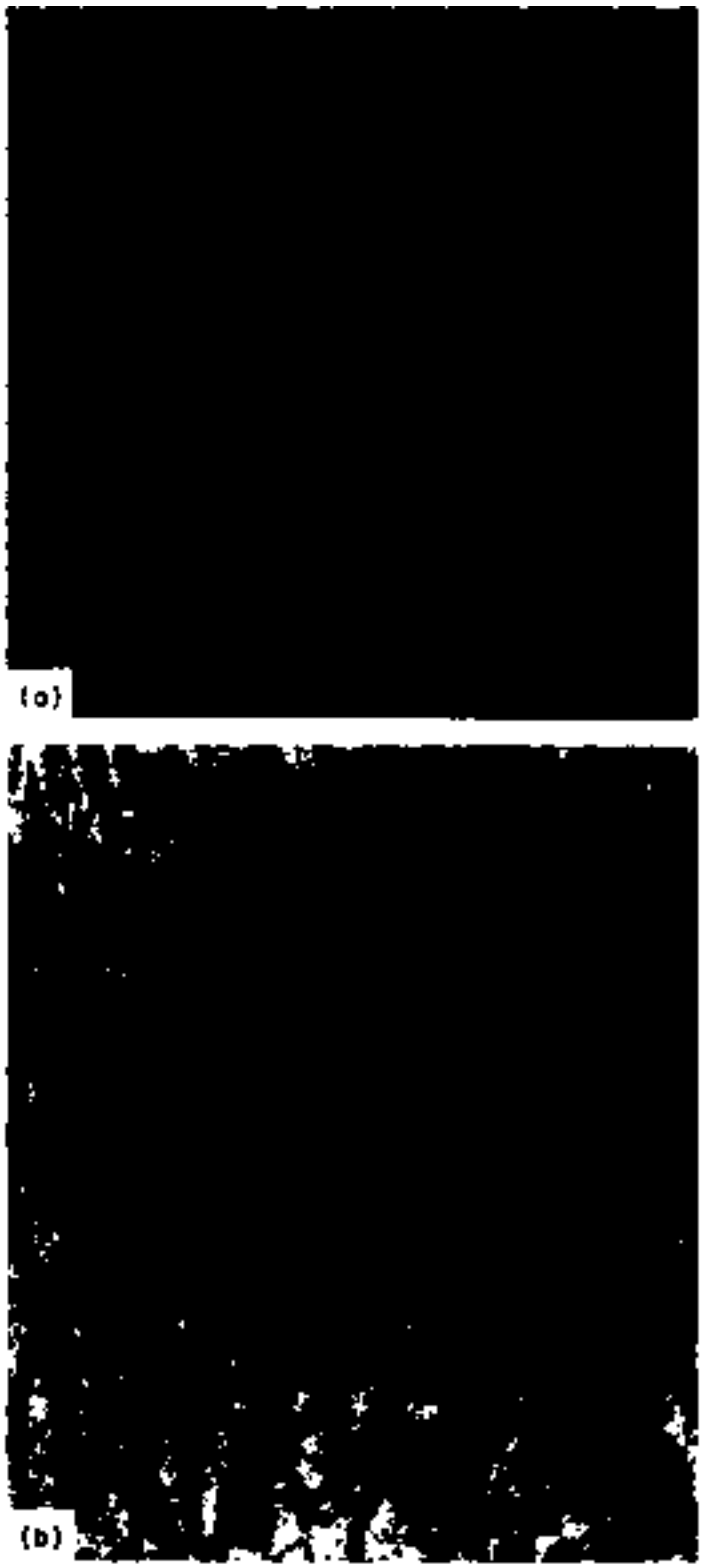


Fig. 15 Cracks in pressure vessel made of ASTM A515 carbon steel lined with type 405 stainless steel. Failure occurred at plug welds because of dilution of weld metal. (a) Micrograph of specimen through weld area etched in acid cupric chloride showing ASTM A515 carbon steel (top), interface region (center), and duplex weld structure containing fine cracks (bottom). 100×. (b) Micrograph of an area of the specimen in (a) showing cracks and precipitates. 250×. Source: Ref 6

The body of the weld had a hardness of 45 HRC. This, combined with the decarburization of the carbon steel, was assumed to be indicative of carbon migration out of the HAZ and into the weld metal, thus preventing the formation of martensite. Further into the weld metal, the high carbon level allowed martensite to form. In the area where the grain-

boundary precipitates appeared heaviest, between the martensite band and the duplex structure, the structure was probably austenitic. These phase changes are predicted from the Schaeffler diagram (in the article "Hardfacing, Weld Cladding, and Dissimilar Metal Joining" in *Welding, Brazing, and Soldering,* Volume 6 of the *ASM Handbook*).

The precipitates followed the austenitic regions into the duplex structure and eventually disappeared. They undoubtedly contributed to the cracking that occurred.

The composition of the precipitates was analyzed using an electron microprobe. Silicon was not present in significant amounts, nor was oxygen, except near a filled branch crack that also showed small amounts of carbon. Phosphorus was absent, and iron, chromium, and nickel were present in amounts commensurate with the weld material. Approximately 4% S was present in the filled branch crack, and the precipitates contained about 12% S.

Thus, several mechanisms were operative in cracking and damaging of the vessel, many of which were the result of microstructural changes in the weld alloy at the interface caused by dilution of the alloy. Austenite was formed at a low dilution level, and the presence of sulfur produced grain-boundary precipitation of sulfides, thus causing hot shortness. Further dilution produced a brittle martensitic phase, which is more susceptible to failure and is a likely source of cracking initiation. This phase was not present at the interface, probably because of carbon migration into the weld, as evidenced by the decarburization and graphitization of the base material. Finally, the differential thermal expansion of the carbon and stainless steels produced the necessary internal stress to produce extensive cracking. Once the carbon steel was exposed to the contents of the pressure vessel, corrosion occurred.

Conclusions. Cracking in the plug welds was the result of excessive dilution of the weld metal. Dilution of the weld metal permitted formation of a martensitic phase in which cracks initiated.

Corrective Measures. Periodic careful gouging of the affected areas followed by repair welding proved sufficient to keep the vessel in service. However, in vessels operated at higher pressures or containing more aggressive corrosive materials, such measures would not suffice, and careful selection of the filler metal and welding procedure would be necessary to prevent formation of the martensitic phase. A lower sulfur content would prevent the grain-boundary precipitate. Carbon migration would still occur, but in itself might not cause a major problem.

Reference cited in this section

6. R.D. Wylie, J. McDonald, Jr., and A.L. Lowenberg, "Weld Deposited Cladding of Pressure Vessels," paper presented at the 19th Petroleum Mechanical Engineering Conference, Los Angeles, Sept 1963

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Stress-Corrosion Cracking (SCC)

Stress-corrosion cracking occurs when stress and a specific corrodent for the material are present in amounts large enough to cause failure by the combined action of both but too small to cause failure by either one acting alone. Highly branched transgranular cracks are characteristic of stress corrosion; under certain conditions, intergranular cracking occurs. Equipment that fails by SCC usually must be replaced because the extent of cracking is difficult to determine and because the cracks are difficult to repair by welding.

The source of the stress may be internal or external. Internal stress can occur on a microscopic scale—for example, the stresses induced by a martensitic transformation. Also, residual stresses produced during thermal cycling, welding, and straightening or bending may be very important. Externally applied stresses arise from applied mechanical loads, vibrations, or pressure.

Effect of Dissolved Solids. Corrosion may arise from many sources, some of which appear to be innocuous or nonaggressive. Even relatively pure water can be corrosive or can accelerate corrosion. If any concentration mechanism is at work, very small amounts of dissolved solids can be disastrous.

Example 11: SCC of a Nuclear Steam-Generator Vessel at Low Concentrations of Chloride Ion (Ref 7, 8). A small leak was found in a nuclear steam-generator vessel constructed of 100-mm (4-in.) thick SA302, grade B, steel [0.25% C (max), 1.15–1.5% Mn, 0.035% O (max), 0.04% S (max), 0.15–0.4% Si, 0.45–0.6% Mo] during a scheduled shutdown. The leak originated in the circumferential closure weld joining the transition cone to the upper shell. The unit had provided approximately 3 years of effective full-power operation in the 6 years following its installation.

Investigation. Examination of the leak area disclosed a hole approximately 16 mm (0.6 in.) long and 5 mm (0.2 in.) wide. Further visual inspection, ultrasonic examination, and magnetic-particle examination of that closure weld, as well as others on similar steam generators at the plant, disclosed about 100 cracks associated with each closure weld, although the

only through-crack was the leak. Figure 16 shows a macrophotograph of a polished-and-etched plug sample of the failure. The details are obscured by erosion damage from the escaping coolant, but the failure clearly passes through both weld and base metal.

Fig. 16 Cross section through leak in steam-generator wall. Crack extends across weld metal, base metal, and HAZ. Erosion obliterated much of the original crack detail.

The welds had been fabricated from the outside by the submerged arc process with a backing strip. The backing was back gouged off, and the weld was completed from the inside with E8018-C3 electrodes by the shielded metal arc process. The weld was continuously stress relieved for 12 h at 540 °C (1000 °F). Hardness traverses on the weldments suggested that the actual heat-treatment temperature may have been less than 540 °C (1000 °F), leaving relatively high residual stresses in the weldments. Metallographic analyses demonstrated that the base metal had a tempered martensitic structure consistent with A302, grade B, steel and that the weld metal had dendritic structures normal for both the shielded metal arc and submerged arc portions of the weld. The HAZ had a more acicular structure normally associated with untempered martensite.

Scanning electron microscopy of the failure surface of the plug sample revealed striations of the type normally associated with progressive or fatigue-type failures (Fig. 17), including beach marks that allowed tracing the origin of the fracture to the pits on the inner surface of the vessel. Examination of discolorations with EDS disclosed that copper deposits with zinc were also present.

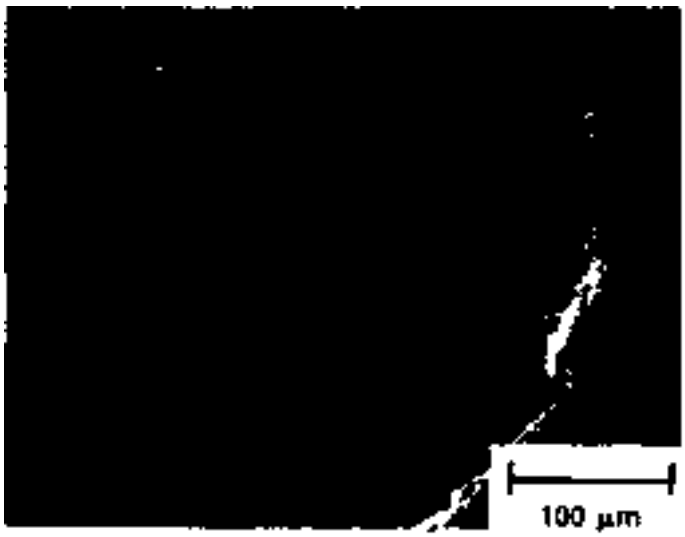


Fig. 17 Fracture morphology of steam-generator wall in a region where regularly spaced striations typical of fatigue are evident.

Three samples were also taken from cracked regions of another steam generator. They exhibited similar features, including varying degrees of pitting at the origin of the cracks (Fig. 18). Examination of the leading edges of the cracks by EDS revealed traces of copper and silicon, as well as a few scans with zinc or nickel present. The area of initiation of one crack had iron, chromium, nickel, silicon, sulfur, copper, and zinc.

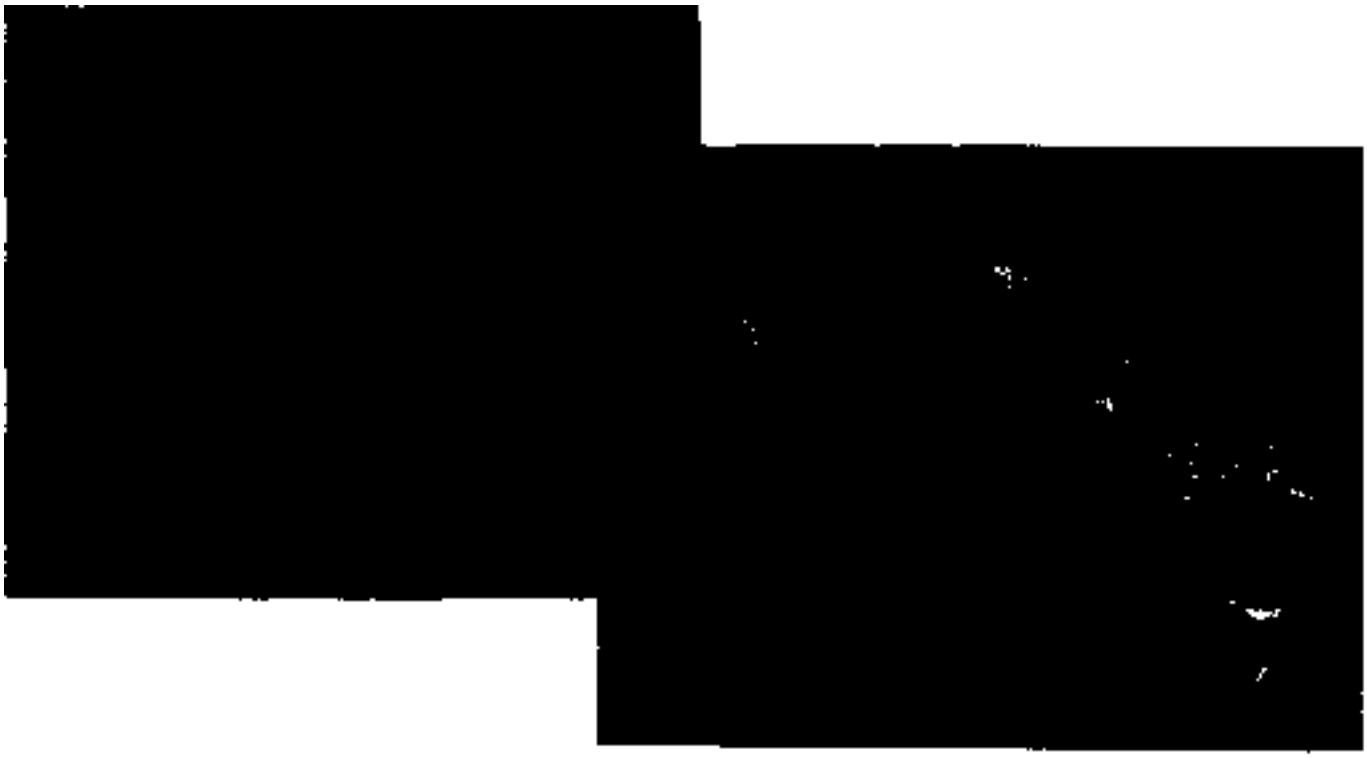


Fig. 18 Fracture surface of a failed steam-generator sample. The fracture morphology is characteristic of fatigue cracking, but 1 ppm of Ce^{-} under constant extension rate testing produced this same fracture morphology in laboratory tests. The initiation site is on the inside surface of the steam generator and apparently emanates from surface pits.

Several historical factors were found to influence the purity level of the coolant water in the steam generators. Brackish river water had corroded copper alloys in the condenser tubes, and some leakage into the system was indicated. Both cuprous oxide ($Cu^{2}O$) and α-hematite were found in the sludge, indicating that oxygen control in the steam generators had been poor for some time and that hydrazine levels had been kept low due to environmental concerns. These factors were

thought to contribute to pitting. In addition, a turbine-blade failure occurred and damaged about 50 condenser tubes, releasing sufficient chloride into the steam generators that values up to 350 ppm were measured. This may have also played a role in pitting.

At this point, the analyst was uncertain whether crack propagation was due to corrosion fatigue or stress corrosion: therefore, a series of tests was performed on welded plate of the same specification grade that failed. A series of controlled crack-propagation-rate stress-corrosion tests was performed at temperatures ranging from 30 to 268 °C (85 to 514 °F) for flat tensile specimens in various heat-treatment conditions. These tests led to the conclusion that A302, grade B, steel is susceptible to transgranular stress-corrosion attack in constant extension rate testing with as low as 1 ppm chloride present (as cupric chloride) at 268 °C (514 °F). Welded SA302, grade B, in the stress-relieved condition is susceptible to stress-corrosion failure at stress levels of about 70% of the yield stress in 268-°C (514-°F) water containing 325 ppm of chloride (as cupric chloride). At lower concentrations of cupric chloride (1 and 5 ppm of chloride), there is a beneficial effect of stress relief on weldments in minimizing this kind of attack; however, the base metal is susceptible to SCC in this environment. A base-metal specimen that failed in water containing 1 ppm of chloride had a fracture morphology nearly identical to that of the failed steam generator.

Recommendations. To alleviate the possibility of additional cracking in the field, the coolant environment should be maintained low in oxygen and chloride, and the copper ions in solution should be eliminated or minimized.

Deaerating Feedwater Heater Failures. Recently, there have been several catastrophic failures and some through-wall leaks in deaerating feedwater heaters used to degas water for tube boiler installations operating at pressures exceeding 1035 kPa (150 psi). It has been estimated that up to 60% of the vessels that have been properly inspected in service have shell cracks. The deaerators operate at pressures ranging from subatmospheric to about 1380 kPa (200 psi). No general correlation has yet been established with such factors as vessel age, size, capacity, manufacturing techniques, service pressure, materials, and mode of operation.

The typical deaerating feedwater heater has both the deaerating zone vessel and the storage vessel made of low-carbon steel. For older vessels, SA212, grade A, B, and C, was commonly used, and for recent vessels, SA285, grade C, as well as all grades of SA515 and SA516 are frequently specified. The vessels, which are usually designed and fabricated in accordance with Section VIII, Division 1, of the ASME Boiler and Pressure Vessel Code, are usually of welded construction and are frequently thin enough that postweld heat treatment or stress relieving is not required.

Cracks in deaerating feedwater heaters have been found on internal surfaces at welds and in the HAZs of the welds; the cracks run both parallel and transverse to the welds. Head-to-shell circumferential welds, circumferential welds joining plate courses, longitudinal seam welds, and nozzle welds have all been found cracked. Failures have been examined *in situ* and in laboratories using both metallographically prepared samples (boat samples) and whole sections of plates. Figure 19 shows the types of cracks that are typically disclosed by field inspection of the insides of vessels. Wet fluorescent magnetic-particle examination is effective on a surface that has been brushed, ground, or blasted free of oxide and dirt without closing the surface indications or cracks. The typical cross section of a crack is shown in Fig. 20. The crack morphologies include both transgranular and intergranular modes of separation; both types of cracks are usually filled with corrosion products.

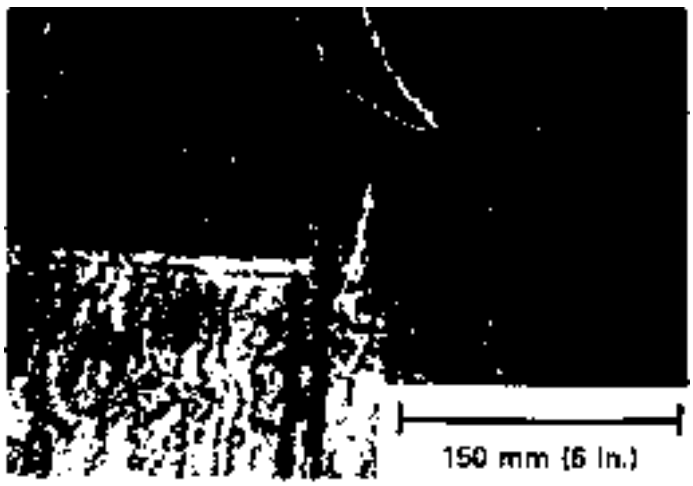


Fig. 19 Typical cracking found by fluorescent magnetic-particle inspection of the internal surface of a feedwater heater.

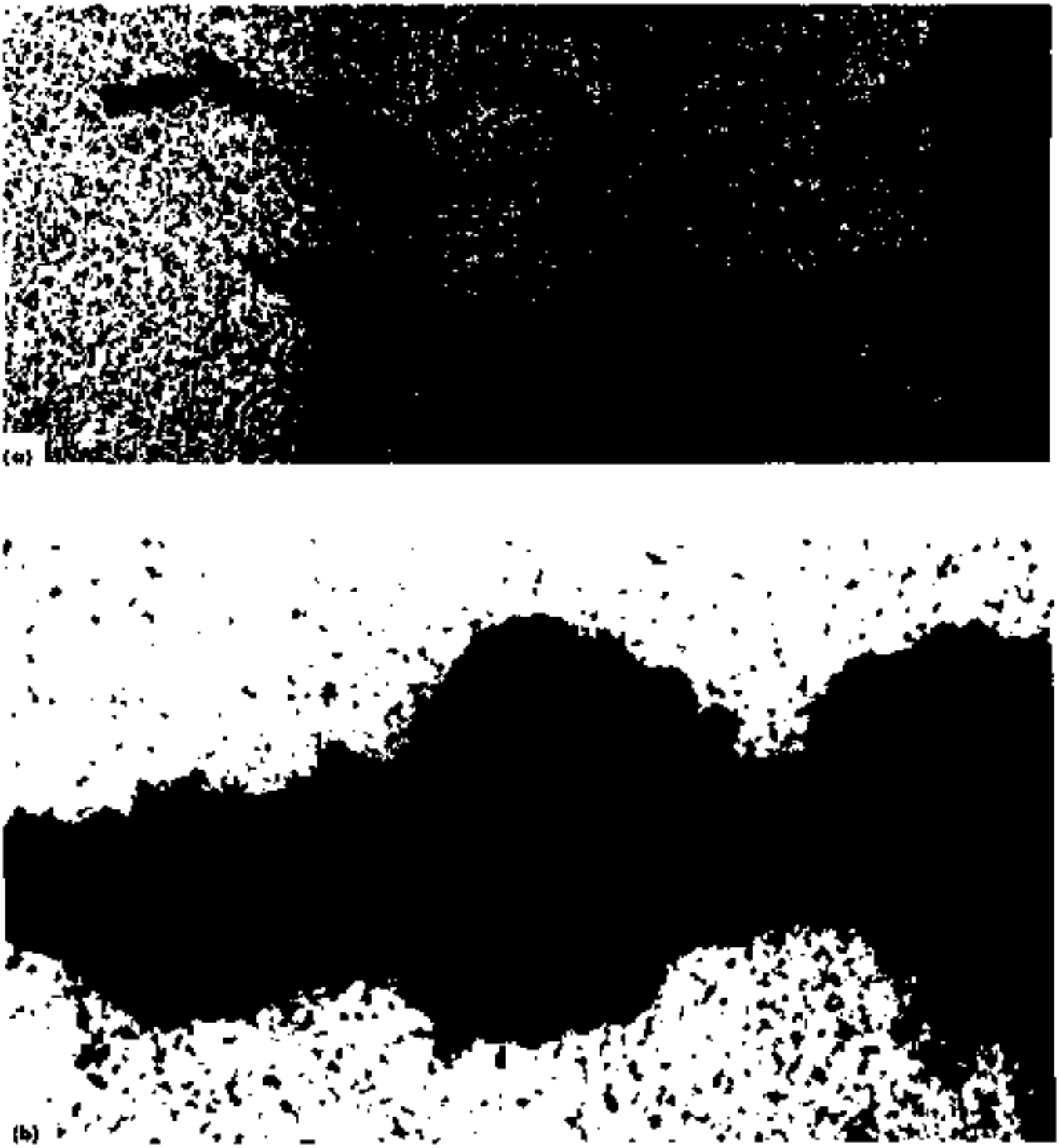


Fig. 20 Typical micrographs of cracks in feedwater heater steels. (a) Cracks identified as corrosion fatigue mixed with SCC. 50×. (b) Corrosion-fatigue crack morphology alternating with corrosion pits and transgranular cracking. 100×

The mechanism of crack initiation and propagation has been classified by various authors as stress corrosion and corrosion fatigue. The exact mechanism of cracking has not been defined, nor have reliable methods been developed for preventing these failures. Current practice is to inspect the tanks periodically and repair or replace them as needed.

Caustic Embrittlement. The embrittling effects of mercury on brass and aluminum alloys are well known. Ammonia produces season cracking in cold-worked brass. Alkaline compounds attack high-strength aluminum alloys. Stainless steels are readily attacked by hot chlorides in aqueous solution, which even in very small amounts can cause severe damage to stressed austenitic stainless steels (see the section “Stress-Corrosion Cracking From Hot Chlorides” in this article). Such steels can also be attacked by concentrated sodium hydroxide (NaOH), as can low-carbon steels. Stress-corrosion cracking of Hastelloy C-276 has occurred in environments containing organic compounds and aluminum chloride ($AlCl_3$) at 100 °C (212 °F). No cracking occurred after the $AlCl_3$, present as an impurity, was removed.

Example 12: Failure of a Low-Carbon Steel Pressure Vessel From Caustic Embrittlement by Potassium Hydroxide. A large pressure vessel (Fig. 21a) that had been in service for about 10 years as a hydrogen sulfide (H_2S) absorber developed cracks and began leaking at a nozzle. The vessel contained a 20% aqueous solution of potassium hydroxide (KOH), potassium carbonate (K_2CO_3), and arsenic at 6550 kPa (950 psi) maximum and 33 °C (91 °F). The cylindrical portion of the vessel was 168 cm (66 in.) in inside diameter, 1020 cm (402 in.) in length, and 64 mm ($2\frac{1}{2}$ in.) in wall thickness. A few months before failure occurred, portions of the exterior surface of the vessel had been exposed to a fire.

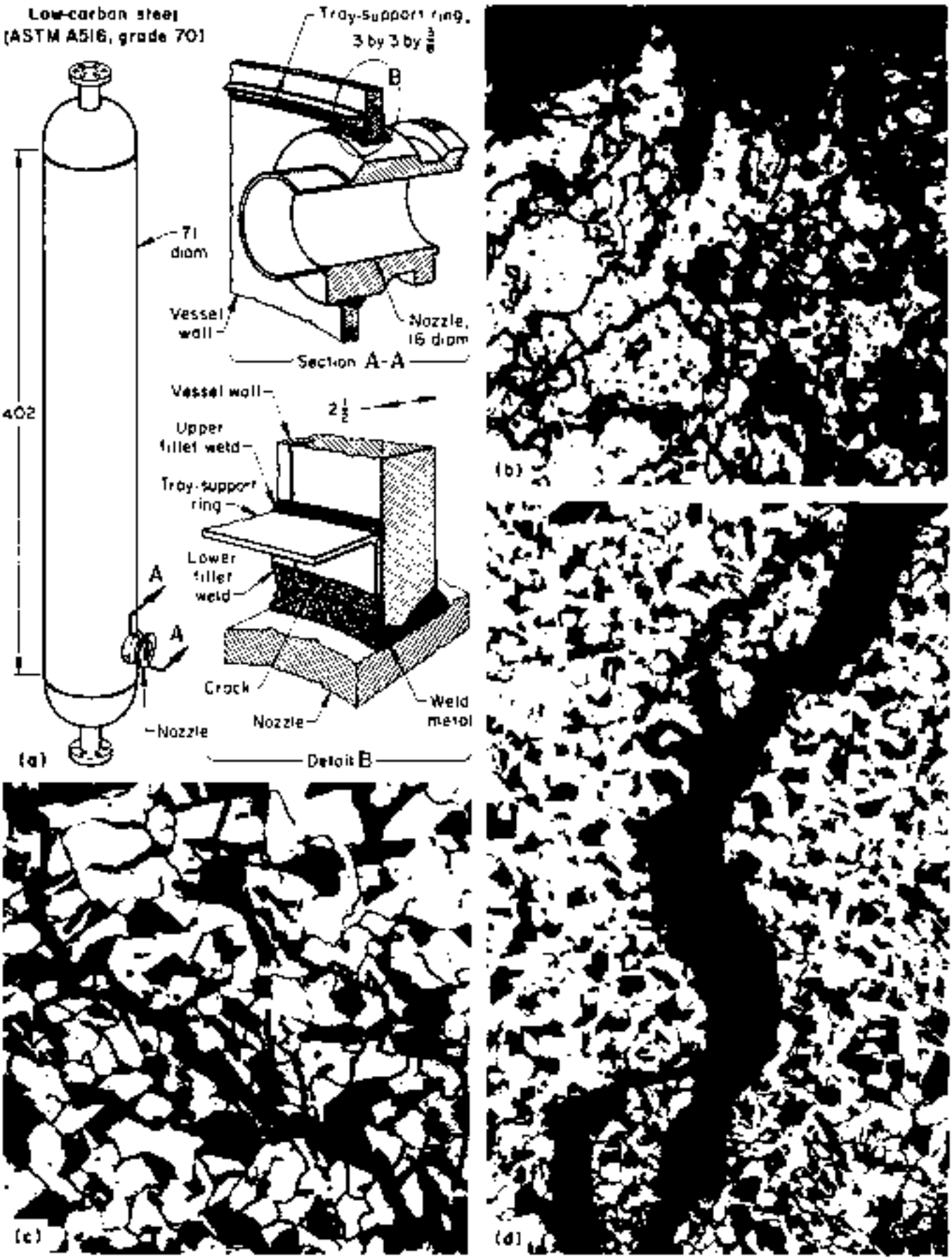


Fig. 21 Large enclosed cylindrical pressure vessel that failed by SCC because of caustic embrittlement by potassium hydroxide. (a) View of vessel before failure and details of nozzle and tray support. Dimensions given in inches. (b) Micrograph showing corrosion pits at edge of fracture surface and inclusions. 600×. (c) Micrograph showing auxiliary

cracks in tensile-overload region of main crack. (d) Micrograph showing auxiliary cracks in stress-corrosion zone of main crack. 100×

The vessel was examined at the job site and was found to be empty, but the inside surface was covered with a layer of sulfur. A sample was removed from near the end of the crack at the nozzle and was sent to the laboratory. The nozzle itself was not cracked. The vessel was returned to the fabricator and was stress relieved before repair work was begun.

The vessel wall, in which the fracture occurred, consisted of 64-mm ($2\frac{1}{2}$-in.) thick ASTM A516, grade 70, low-carbon steel plate. To provide a tray support, a steel angle, 75 × 75 × 10 mm (3 × 3 × $\frac{3}{8}$ in.), had been formed into a ring and continuously welded to the inside wall of the vessel at the heel and the toe of one leg (Section A-A, Fig. 21a).

Investigation. In the groove formed by the junction of the lower tray-support weld and the top part of the weld around the nozzle was a crack about 55 mm ($2\frac{1}{4}$ in.) long. Each end of the crack branched into a Y (Detail B, Fig. 21a). The crack portions in the weld metal around the nozzle were arrested, and the nozzle was not damaged.

The field sample submitted for laboratory analysis contained the remote end of one of several cracks that propagated into the upper tray-support weld. The surfaces of this crack were covered by a tightly adhering dark-brown and black scale. The surfaces contained two triangular regions that were smooth and that showed no evidence of ductility, one at the top of the weld and the other at the outer surface of the wall. The remainder of the fracture surface showed cleavage planes with some evidence of ductility, indicating that failure had resulted from tensile overload.

A dirty yellow scale was found under a piece of steel angle that was part of the field sample, but no scale was noticeable elsewhere after the angle had been removed by air carbon-arc gouging. No additional cracks were found in the welds for the tray support.

Tensile and impact tests performed on specimens from the sample indicated that the physical properties, after stress relief, were adequate for the service conditions.

Because of the proximity of the crack to the nozzle, which was saved for muse, only fragments from the tray-support side of the crack were available. How representative these were of the original microstructure was a moot question because they had been annealed during removal. Microscopic examination revealed pits and scale near the crack origin; the pits contained corrosion products. Although difficult to characterize because of the small grain size, cracking was principally transgranular.

The surface of a section through a crack that started in a groove formed by the junction of the lower tray-support weld and the weld around the nozzle was examined under a microscope. The first 6 mm (0.25 in.) of the crack surface sloped at an angle of about 45° to the cut surface, was relatively smooth, had a thin orange-brown scale, and, at higher magnification, exhibited plastic flow at some points and corrosion pits at others. Figure 21(b) shows corrosion pits in the edge of the fracture surface as well as inclusions consisting of large dark orange-brown conglomerates and smaller gray globules that appear to have nucleated in the grain boundaries. These features suggested that periods of corrosion alternated with sudden instances of cleavage, under a tensile load, along preferred slip planes. Cracking in such a manner is typical of the stress-corrosion process, according to one theory.

After the first 6 mm (0.25 in.), the crack progressed abruptly straight downward into the weld and was filled with a thick gray scale. The crack surface was wavy and contained branching cracks filled with corrosion products. In this crack, corrosion had obviously predominated, with cleavage playing a lesser role or none at all. This condition is in accordance with an alternative, strictly electrochemical, theory on the mechanism of stress corrosion.

Micrographs of sections across the remote end of the crack in the field sample that propagated into the material showed both a zone that failed by tension overload (Fig. 21c) and a stress-corrosion zone (Fig. 21d). These two distinguishable fracture zones were confirmed by a gray corrosion product similar to that found near the crack origin, which was present in auxiliary cracks in the stress-corrosion zone but not in the tension-overload zone. Cracking in both zones was clearly transgranular.

Discussion. Because the surface of the steel angle adjacent to the vessel wall had been sealed and because only one point of seepage under it had been found, the crack at the grooved area between the lower weld joining the angle to the wall and the weld around the nozzle must have formed first. Although the vessel had been postweld heat treated before being put into service, residual stress was apparently sufficient in this grooved area for it to act as a notch or a stress raiser. With this configuration and enough time, the combination of the residual plus operating stresses and the amount of KOH present would have caused stress corrosion—in this case, as a result of caustic embrittlement.

The fire to which the vessel had been exposed before failure, the only evidence of which was observable bulging of the vessel skirt, could have been instrumental in initiating cracks by any or all of the following conditions affecting the critical grooved area: (1) an increase in temperature to above 120 °C (250 °F), (2) an increase in concentration of the caustic solution to greater than 50%, and (3) addition of thermal stresses to existing residual stresses. On the other hand,

cracking may have been initiated simply by the combination of existing residual stresses and normal operating conditions over a long period of time.

Caustic embrittlement appeared most plausible as the cause of failure because of the presence of KOH and the known role of KOH in stress-corrosion failures in plain carbon steel vessels. However, the presence of arsenic can induce hydrogen embrittlement under alkaline conditions. The probability of sulfur, a lesser by-product of the removal of H_2S, inducing hydrogen embrittlement under alkaline conditions is slight. Hydrogen embrittlement from arsenic may therefore have been the added influence that helped produce transgranular corrosion rather than the simpler mechanism of intergranular corrosion.

Conclusion. Failure was caused by stress corrosion, which probably resulted from caustic embrittlement by KOH.

Recommendations. The tray support should be installed higher on the vessel wall to prevent coincidence of the lower tray-support weld with the nozzle weld, thus avoiding the formation of a notch.

Stress-Corrosion Cracking From Hot Chlorides. Equipment made of austenitic stainless steel may sustain SCC when contacted by hot chlorides in aqueous solution—even in concentrations of only a few parts per million. Therefore, chloride attack is generally insidious, and operating personnel may not be aware of such a problem until failure occurs.

Chlorides are commonly introduced into pressure vessels as dissolved salts in water or wet (saturated) steam. At temperatures below about 50 °C (120 °F), small quantities of chloride salts create few problems. At higher temperatures in systems containing austenitic stainless steels, chlorides can be disastrous. Water is most safely introduced to pressure vessels made of austenitic stainless steel in the form of either clean steam condensate or dry superheated steam. Generally, there is no problem as long as ample quantities of condensate or dry steam are available. Unfortunately, during refinery turnarounds, superheated steam can be in short supply because the heaters are shut down. As the various units are put back on stream, the boiler plant may lack sufficient capacity to meet all immediate needs. A related problem can occur in steam methane reformers during start-up. Unless special start-up procedures are carefully followed, steam introduced to the reformer may contain water droplets with entrained chloride salts (carryover). When the wet steam subsequently flashes against a hot surface farther down the system, a residue of chloride salts remains. If the surface is hot, SCC will occur quickly in some grades of austenitic stainless steel.

In one case, SCC caused failure of a type 347 stainless steel shaft in a hydrogen-bypass valve in a steam methane reformer plant. Chloride-salt residues were found on the fracture surfaces, together with highly branched cracks typical of stress-corrosion failure (Fig. 22). Wet steam introduced to the system during start-up operations was the source of the chlorides. The shaft had a hardness of 245 HB, indicating that it had been made from cold-finished bar stock. Undoubtedly, the higher surface residual stresses inherent in cold finishing made the shaft more sensitive to SCC. The fractured shaft was replaced with an annealed type 316 stainless steel shaft having a hardness of 169 HB. Also, changes were made in the reformer start-up procedure, minimizing the probability of introducing wet steam.

Fig. 22 Chloride SCC in a type 347 stainless steel shaft in a hydrogen-bypass valve.

Example 13: SCC of an Inconel 600 Safe-End on a Reactor Nozzle. Cracking occurred in an ASME SB166 Inconel 600 safe-end forging on a nuclear-reactor coolant-water recirculation nozzle while it was in service. The safe-end was welded

to a stainless steel-clad carbon steel nozzle and a type 316 stainless steel transition metal pipe segment (Fig. 23). An Inconel 600 thermal sleeve was welded to the safe-end, and a repair weld had obviously been made on the outside surface of the safe-end to correct a machining error. Initial visual examination of the safe-end disclosed that the cracking extended over about 85° of the circular circumference of the piece.

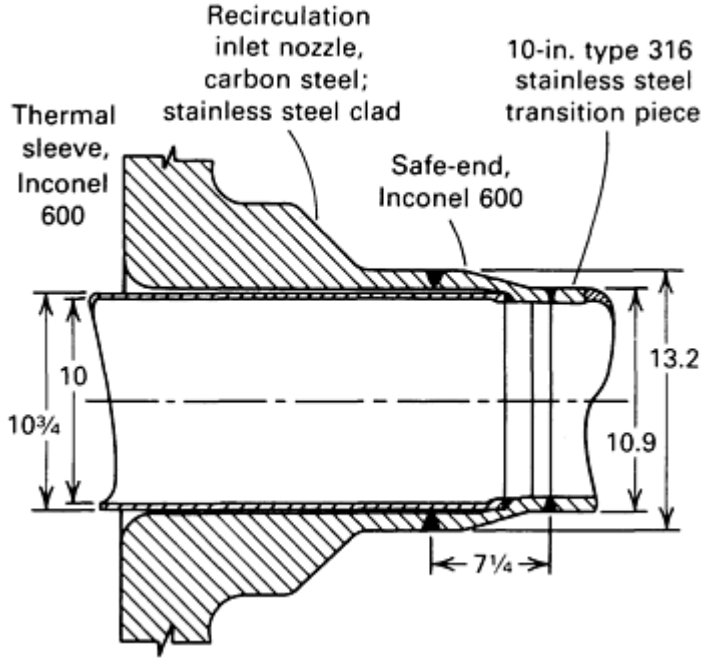


Fig. 23 Cross section through recirculation inlet nozzle of reactor vessel. Shown are the nozzle, the safe-end that failed, and the thermal sleeve that created susceptibility to crevice corrosion. Dimensions given in inches

Investigation. After the safe-end had been cut from the reactor, it was radiographically inspected on-site to confirm the extent of cracking. This inspection indicated major cracking over about 280° of the circumference. Limited ultrasonic inspection performed in the laboratory with a 1.5-MHz transducer, directing the incident sound beam from the large end of the safe-end toward the small end, demonstrated that the cracking could be detected nondestructively over about the same region that was indicated radiographically.

Eleven longitudinal cross sections were prepared for macroscopic examination, including locations near each end of the through-wall portion of the crack and sections selected on the basis of radiographic crack indications. Major cracking was found in all sections examined. At all locations, the crack originated near the tip of the crevice between the thermal sleeve and the safe-end, adjacent to the attachment weld for the safe-end, and extended outward through the safe-end wall. Over about 35° of the circumference, cracking propagated through the repair weld, while over a different 50° of the circumference, the propagation path was entirely within the safe-end base metal, penetrating the outside surface adjacent to the fusion line of the repair weld.

Metallographic examination of selected cross sections showed the effects of repeated thermomechanical cycling during welding on the precipitate structure in the HAZ of the Inconel 600. There were very few matrix precipitates immediately adjacent to the fusion line of the heat shield attachment weld; this condition was true of the entire HAZ near the root pass. The cracking generally initiated in the HAZ some distance away from the tip of the tight crevice and, in its early stages, generally followed a path parallel to the fusion line (Fig. 24). Depending on the location, the crack could be found in regions of heavy or light carbide precipitation. Metallographic cross sections and SEM studies of the fracture surface were in agreement: the cracking was all intergranular in the base metal and intercolumnar where it penetrated the repair weld metal. No secondary cracking was found, nor was any discontinuity found that might have initiated the cracking.

Fig. 24 Polished-and-etched cross section through the failed safe-end of the recirculation inlet nozzle. Note the intergranular nature of the failure. Etched in 8:1 phosphoric acid. 53×

Microanalyses disclosed the presence of sulfur at locations corresponding to the later stages of cracking, but not on the crevice surface between the safe-end and the heat shield. Chlorine was present at one point on the crevice surface. Other deposits on the crevice surface and the crack surface near the crevice were rich in iron.

Conclusions. These observations led to the conclusion that the cracking mechanism was intergranular SCC. The literature indicates that Inconel 600 is susceptible to intergranular stress corrosion in high-purity water, but it is only likely at stresses above the yield stress of the material. The analysis by the reactor vendor showed the primary bending plus membrane stresses due to service loading to be 78% of the yield strength. Further, butt welds in austenitic piping materials can leave residual stresses of the order of the yield strength; therefore, the mechanism is reasonable. Whether or not the material is in a sensitized condition does not greatly influence susceptibility to cracking, and cracks propagated in the lightly sensitized region adjacent to the fusion line and in the heavily sensitized regions some distance from the fusion line. Therefore, the furnace-sensitized condition of the material was not considered to be a significant contributing factor in the cracking problem.

The role, if any, of the observed contaminants could not be clearly identified from this study, although there is a suspicion that they contributed to the failure. Further surface analysis would be required to clarify this point.

The role of the tight crevice between the safe-end and the thermal sleeve thus comes into question. It was considered unlikely that a classical differential aeration cell could develop in the crevice in the uniform high-purity water environment, but such regions could entrap air during outages, causing locally high oxygen levels upon start-up. The presence of anionic contaminants could cause acidification of the crevice region, which might also enhance stress-corrosion mechanisms. The crevice is considered to be the principal contributing factor in this case.

Protection During Off-Stream Periods. When subjected to stress in aqueous chloride solutions, austenitic stainless steels fail by highly branched transgranular cracking. When the same steels are exposed in refinery processing units to weak sulfur acids, more commonly known as polythionic acids, intergranular cracking can occur. The austenitic stainless steel must be in a sensitized condition—that is, with heavy intergranular carbides—for polythionic acids to cause cracking. Polythionic acids form directly from sulfide corrosion products inside desulfurizing, hydrotreating, and hydrocracking units at ambient temperatures during refinery shutdowns, but are not a problem at operating temperatures. These types of

units are frequently clad with austenitic stainless steel for on-stream corrosion protection, but the steel must be protected during offstream periods. Procedures used for this protection are discussed below.

First, when a stainless-clad desulfurizer, hydrotreater, or similar reactor is shut down but not opened, it should be kept under positive nitroger pressure. If the reactor is opened but not dumped, it should be turned over with ammonia three or four times before opening. Also, the work should be completed before the walls of the reactor reach the ambient dew point. If this is not practical, the reactor should be kept under positive nitrogen pressure during the entire downtime. Second, during dumping of the reactor catalyst, 5000 ppm of ammonia should be added to the nitrogen purge. Once the catalyst is removed, a 2% solution of soda ash with 0.5% sodium nitrate should be used to fill the reactor and soak all stainless steel. Third, any austenitic stainless steel in such components as recycle-furnace tubes, feed-heater furnace tubes, feed or effluent piping, and tubes or cladding in feed-effluent exchangers is also susceptible to intergranular cracking. Provisions should be made to keep this equipment under positive nitrogen pressure if unopened or to fill and soak with soda ash solution before opening.

Because polythionic acids are most likely to cause cracking in austenitic stainless steels that are in a sensitized condition, cracking can be minimized in pressure vessels of welded construction by using stabilized stainless steels, such as type 321 or type 347, or by using low-carbon grades, such as type 304L or type 316L.

References cited in this section

7. C.J. Czajkowski, “Investigation of Shell Cracking on the Steam Generators at Indian Point Unit No. 3,” NUREG/CR-3281, also BNL-NUREG-51670, U.S. Nuclear Regulatory Commission, Washington, June 1983

8. C.J. Czajkowski, “Constant Extension Rate Testing of SA302 Grade B Material in Neutral and Chloride Solutions,” NUREG/CR-3614, also BNL-NUREG-51736, U.S. Nuclear Regulatory Commission, Washington, Feb 1984

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Hydrogen Embrittlement

When corrosive reactions occur at a metal surface, hydrogen may be produced, particularly if acids or strong alkali solutions are involved. The hydrogen evolved is preferentially absorbed by steel and possibly by other metals. This absorption can cause hydrogen embrittlement. Consequently, SCC and hydrogen embrittlement are not only associated but also sometimes indistinguishable, particularly in sour gas service in which H_2S is the active corrodent.

Example 14: Hydrogen-Embrittlement Cracking in a Large Alloy Steel Vessel. Figure 25(a) shows a vessel that had been in service about 3 years as a hydrogen reformer when cracking occurred in the weld between the shell and the lower head. As shown in Detail A in Fig. 25, one crack, which was 150 mm (6 in.) long, developed at the bottom of the weld at the inner surface of the head (crack A), and another, which was 0.9 m (3 ft) long, developed at the top of the weld at the outer surface of the shell (crack B). Crack B later extended to a length of 2.7 m (9 ft). Also, some small blisters were observed on the inner surface near the weld after attempts had been made to repair the crack.

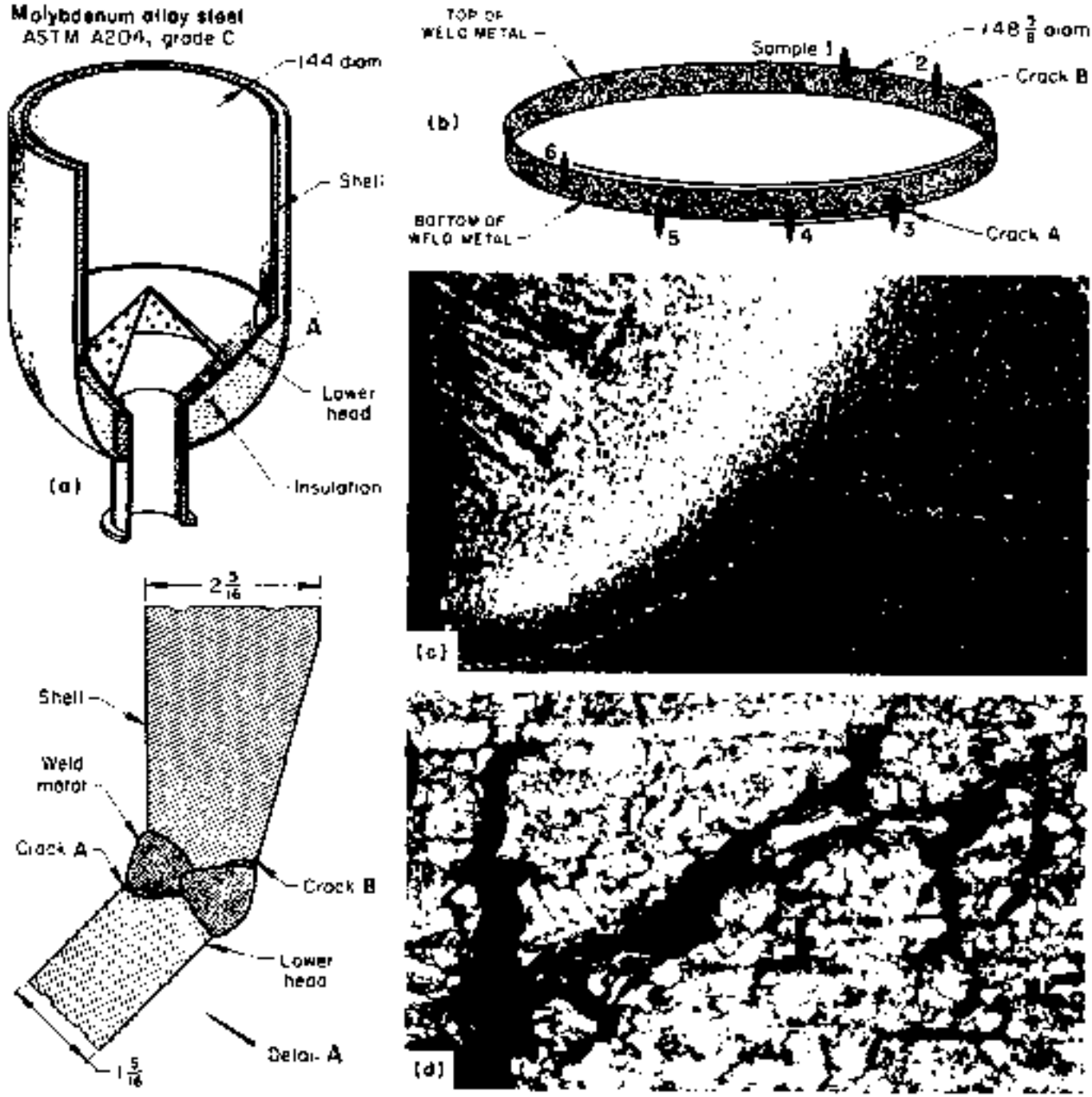


Fig. 25 Vessel made of ASTM A204, grade C, steel that failed as the result of hydrogen embrittlement. (a) Portion of tank; Detail A shows locations of cracks, at welds joining shell to lower head. Dimensions given in inches. (b) Schematic illustration of the weld area showing locations of cracks and of samples removed for examination. (c) Macrograph of a nital-etched specimen from the inner surface of the head showing a crack in the HAZ. 6.5×. (d) Micrograph of a nital-etched specimen showing normal microstructure and hydrogen embrittlement. 100×

The vessel contained ceramic balls in the lower head and a catalyst in the shell section. Temperature of the incoming gas was 690 °C (1275 °F); the outgoing gas, 620 °C (1150 °F). Design temperature of the vessel, which was insulated inside, was 120 °C (250 °F), and design pressure was 4480 kPa (650 psi). No liquid phase was present. A temperature of 340 °C (640 °F) had occurred at two hot spots 100 mm (4 in.) in diameter on the outside surface just above the area of the failure.

The vessel was made of ASTM A204, grade C, molybdenum alloy steel. The head was 33 mm ($1\frac{5}{16}$ in.) thick, and the shell was 59 mm ($2\frac{5}{16}$ in.) thick.

Metallurgical Investigation. Six samples were removed from the outside surface of the vessel at right angles to the weld, using an abrasive cutoff wheel; as shown in Fig. 25(b), three were cut across the weld line on the shell side and three on the head side. The samples were triangular in cross section, about 13 mm ($\frac{1}{2}$ in.) wide on each surface, and 75 mm (3 in.) long. A sample at the end of the crack on the inner surface of head was removed parallel to the weld. This sample was also triangular in shape, 22 mm ($\frac{7}{8}$ in.) wide at the head surface, 16 mm ($\frac{5}{8}$ in.) deep, and 240 mm ($9\frac{1}{2}$ in.) long. Also, three 21-mm ($\frac{13}{16}$ in.) diam plugs were trepanned through the complete wall thickness, one each from the head, weld, and shell.

Of the six samples taken from the outer surface, only sample 3 was visibly cracked. This crack, about 5 mm ($\frac{3}{16}$ in.) deep, initiated at the edge of the weld in the coarsegrain portion of the HAZ. Microscopic examination revealed a small crack in sample 1, which was taken from the shell side. This crack was also located in the coarse-grain portion of the HAZ. No cracks were discovered in the other samples taken from the outer surface. No microfissures, decarburization, or characteristics of hydrogen embrittlement were observed in any of the six samples from that surface.

Examination of a section through the sample from the inner surface of the head showed that the crack was at the edge of the weld, completely in the HAZ of the head (Fig. 25c). Fusion of the weld metal on both the head and shell sides was good on both the inner and outer surfaces. A micrograph of the start of the crack, which was at the inner surface, revealed elongated grains, indicating that the metal had undergone plastic flow before failure. Figure 25(d) shows a micrograph of a portion of the same section farther along the crack. The microstructure was normal in appearance, but was severely embrittled, as shown by the region near the fracture surfaces. Additional micrographs near the crack showed that the metal had become severely gassed in an area extending up to 3 mm ($\frac{1}{8}$ in.) on each side of the crack.

Microscopic examination of the plug trepanned from the head showed decarburization, which was present in nearly equal amounts on both inner and outer surfaces and therefore was not the result of service conditions. The inner surface was coated with a dark-brown scale. The microstructure of the metal in the head was banded, except near the surfaces, and contained spheroidal carbides, indicating that the metal had been held at temperatures above 540 °C (1000 °F) for a long period of time.

The plug from the shell also had a dark-brown scale on the inner surface and showed a microstructure that was uniform throughout and typical for the metal. The carbides in the pearlite areas had lost some of their lamellar form, which indicated that the material had been held at a high temperature for some time, but the breakdown was not as severe as that observed in the head.

Hardness tests were performed on one of the samples taken from the outer surface, the sample from the inner surface, and the three trepanned plugs. Rockwell B hardness values and equivalent tensile-strength values are given in Table 2.

Table 2 Hardness and tensile properties of samples taken from a failed pressure vessel

Location of hardness test	**Hardness, HRB**	Equivalent tensile strength	
		MPa	ksi
Outer-Surface Sample			
Shell plate	86	586	**85**
Head plate	86	586	**85**
HAZ[(a)]	92–98	724	**105**
Weld metal	92–96	690	**100**
Inner-Surface Sample			
Head plate	76	465	**67.5**
HAZ[(b)]	89	565	**85**
Weld metal	90	638	**92.5**
Trepanned Plugs			
Shell:			

Inner wall	84	569	**82.5**
Center	81	517	**75**
Outer wall	85	579	**84**
Head:			
Inner wall	76	465	**67.5**
Center	78	490	**71**
Outer wall	79	500	**72.5**
Weld metal:			
Inner wall	96	769	**111.5**
Center	94	748	**108.5**
Outer wall	**95**	**758**	**110**

(a) Both shell and head sides.
(b) Head side

Discussion. Hardness data and microscopic examination indicated that the lower head had been overheated for a sufficiently long time to reduce the tensile strength below the 517-MPa (75-ksi) minimum required for ASTM A204, grade C, steel. The wide difference in tensile strength between head and weld metal (including HAZ) formed a metallurgical notch that enhanced the diffusion of hydrogen into the metal in the cracked region. The resultant embrittlement and associated fissuring led to further weakening and eventually to failure.

Because the fracture was brittle, internal cracking must have taken place before plastic deformation on the inner surface. Also, the major crack must have propagated internally because it did not intersect the inside surface near its end. Cracking on the outside surface was simply the result of overloading the embrittled material alter internal failure had already occurred.

The total absence of decarburization on the trepanned plugs at the inside surface of the shell and the minor decarburization of the head suggested that hydrogen diffusion was limited to the head regions adjacent to the HAZ or that hydrogen diffused into the metal at temperatures below 205 °C (400 °F). The tensile tests of the head material would have detected any embrittlement. Had hydrogen diffusion occurred at higher temperatures, more fissuring (gassing) or considerably more decarburization would have been in evidence. Because no serious damage was detectable in the trepanned plugs, heat treatment would relieve any embrittlement that may have occurred in the shell and head. Grains need not be separated for embrittlement to occur, and provided no fissures have formed, the pressure can be relieved by allowing the gases to diffuse out of the metal during careful heat treatment; otherwise, blisters will result. The damaged region adjacent to the weld on the head side should be removed, because these fissures cannot be repaired by heat treatment.

Conclusion. Failure of the vessel was chiefly the result of hydrogen embrittlement with over-heating as a contributory factor.

Corrective Measures. The vessel was wrapped in asbestos and heated to 455 °C (850 °F) at a rate of 55 °C (100 °F) per hour, held at 455 to 480 °C (850 to 900 °F) for $2\frac{1}{4}$ h, heated to 595 °C (1100 °F) at a rate of 55 °C (100 °F) per hour, and held for $2\frac{1}{4}$ h. The vessel was then cooled at a rate prescribed by the ASME code. This treatment caused hydrogen to be diffused out of the metal, thus removing the source of embrittlement.

The weld metal on both sides of the crack and 5 mm ($\frac{3}{16}$ in.) of base metal in the shell and the head were removed. The area thus prepared was repair welded using E7018 filler metal, then stress relieved.

Example 15: Cracking in Carbon-Molybdenum Desulfurizer Welds. Welds in two carbon-molybdenum (0.5% Mo) steel catalytic gas-oil desulfurizer reactors cracked under hydrogen pressure-temperature conditions for which hydrogen damage would not have been predicted by the June 1977 revision of the Nelson Curve for that material. As a result of this experience, a major refiner instituted regular ultrasonic inspections of all welds in its Fe-C-0.5Mo steel desulfurizers.

Investigation. During a routine examination of a naphtha desulfurizer by ultrasonic shear wave techniques, evidence of severe cracking was found in five weld-joint areas (Ref 9). The inspection involved stripping all the insulation from the vessel and testing all seams with a beam propagated perpendicular to the welds. Defect indications were found in longitudinal and circumferential seam welds of the 113-mm ($4\frac{7}{16}$-in.) thick ASTM A204, grade A, Fe-C-0.5Mo steel sheel. Indications were also found in the 57-mm ($2\frac{1}{4}$-in.) thick top and bottom heads that were formed by welding pie sections of plate. No records were available on the welding method used in the mid 1950s. The vessel had a 2.8-mm (0.109-in.) thick type 405 stainless steel liner for corrosion protection that was spot welded to the base metal at about 38-mm ($1\frac{1}{2}$-in.) centers, and all vessel welds had been hand overlaid with type 309 stainless steel.

The vessel was emptied and visually inspected, but no surface damage was evident. To inspect the weld metal, the type 309 stainless steel weld overlay was ground off along a 965-mm (38-in.) strip of a longitudinal seam weld in an area that had ultrasonic indications. This exposed a 330-mm (13-in.) long longitudinal crack in the weld metal, as well as transverse cracks. When about 3 mm ($\frac{1}{8}$ in.) of the base-metal weld was ground off and the surface was dye checked, hundreds of transverse cracks were observed. An external shear wave test was then performed on the outside of the vessel, but perpendicular to the normal scan direction, including weld areas that had not shown indications in the original testing; indications of transverse cracks were found. The testing was extended to include all welds, and all were found to have indications of transverse cracks.

Conclusions. Samples were taken from a girth weld for examination. It was found that weld metal from the cracked region contained 0.09% C, whereas the base plate contained 0.18% C as compared with a maximum 0.25% C for ASTM A204 material. This was taken as an indication of severe decarburization in the crack area. Subsequent metallurgical examination of a cross section of a longitudinal crack after polishing and etching with 2% nital disclosed a decarburized region on either side of the crack. The cracking was clearly intergranular in nature (Fig. 26). It was concluded that the damage was caused by a form of hydrogen attack.

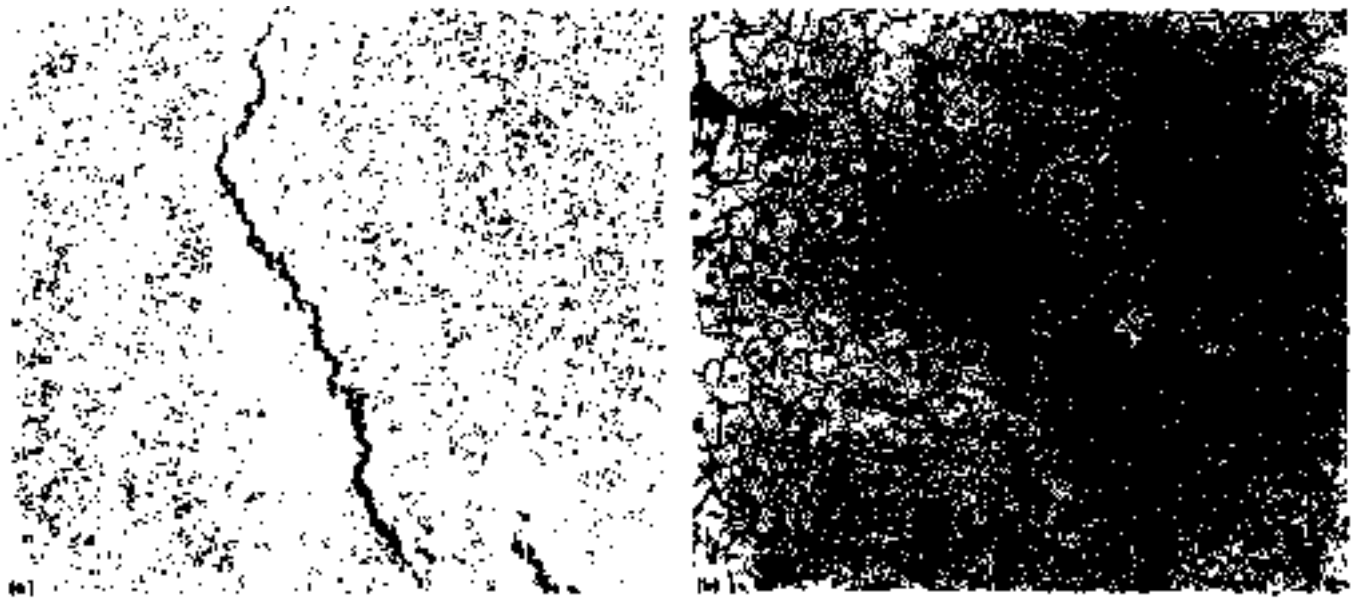


Fig. 26 Cracks in naphtha desulfurization reactor attributed to hydrogen damage. (a) Note decarburized region adjacent to crack. 25×. (b) Higher-magnification view. 200×

Corrective Measures. The nominal service conditions for the desulfurizer were approximately 345 °C (650 °F) and 2690 kPa (390 psig) of hydrogen partial pressure, conditions that would not be expected to produce hydrogen damage based on the limits of the 1977 revision of the Nelson Curve for Fe-C-0.5Mo steel. The solution to this material problem was installation of a used 2.25 Cr-1Mo steel vessel with a type 347 stainless steel weld overlay. The use of an available vessel satisfied certain time constraints on the allowable down time for the plant.

Reference cited in this section

9. D.B. Bird, “Cracking in Carbon $\frac{1}{2}$ Moly Desulfurizer Reactor Welds,” paper presented at the 1985 Fall Meeting of the American Petroleum Institute Operating Practices Committee, Subcommittee on Process Operations

---

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Brittle Fractures

Materials that are ordinarily ductile can undergo brittle fracture suddenly and without warning, depending on certain conditions and factors, which are discussed in the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals” in this Volume. Brittle fractures are most likely to occur in body-centered cubic (bcc) structures, such as ferritic steels, and in relatively thick parts that contain adequately large flaws. Strains rates required for brittle fracture depend somewhat on the temperature of the part; at low temperatures, strain rates are generally low, whereas at high temperatures strain rates are higher.

The origin of a brittle fracture will almost always be a pre-existing crack or similar type of notch or stress raiser because the concentrated triaxial stresses developed at such points are the single most important factor in determining whether a crack can propagate under the applied load. Brittle fracture generally occurs under an applied stress that is less than the yield strength, and the energy expended in propagating the crack is low.

Most sudden brittle fractures occur by cleavage, which produces a fracture surface with a bright granular or crystalline appearance. A characteristic chevron pattern, which points to the origin of fracture, is often present. Brittle fractures can be differentiated from ductile fractures by the short time required for the crack to achieve critical size and by the high rate of crack propagation. Although most brittle fractures are transgranular, intergranular brittle fracture can occur if the grain boundaries are embrittled by chemical segregation or the presence of a brittle second phase.

Brittle fractures in pressure vessels can be correlated with the notch toughness of carbon or low-alloy steels, and they occur at temperatures below the transition temperature measured by such testing. The transition temperature can increase in service due to microstructural or metallurgical changes. In carbon and low-alloy steels, the ductile-to-brittle transition temperature can be near or above the ambient temperature during fabrication and during hydrostatic testing. Factors controlling the transition temperature include chemical composition and the effects of deoxidation, forming, and heat treatment. Poor-quality steel or improper welding practice are frequently the material factors that cause brittle fracture, but externally applied stresses are required to cause failure.

Corrective action that can be used to guard against the possibility of brittle fracture in pressure vessels includes attention to all the factors noted above plus stress relief after fabrication. Metallurgical changes in the material during service should be avoided. Such changes might result if, for example, carbon steel were used instead of alloy steel in an application in which temperatures occasionally exceed the limit of operating temperature for carbon steel.

Embrittlement can result from graphitization in carbon steels and 0.5% Mo steels, welding or fabrication steps, exposure to atmospheres favoring hydrogen embrittlement, or formation of σ phase in stainless steels. Thermal fatigue caused by different amounts of expansion in adjacent dissimilar metals can result in crack propagation and eventual brittle fracture. Severe and repeated temperature changes in such components as chemical-system quench tanks can result in thermal shock, the severity of which might be sufficient to cause a brittle failure.

Brittle fractures are not restricted to steels, but have also occurred in components made of other materials, such as welded aluminum rocket-fuel tanks, because of residual shrinkage stresses. These stresses were caused by weld design that was improper for the relative stiffness of the elements being joined. Microcracks that formed during cooling of the welds provided the brittle fracture initiation sites that occurred during hydrostatic proof testing.

Example 16: Brittle Failure of a Pressure Vessel After Welded Modifications Without Stress-Relief Heat Treatment (Ref 10). A spherical carbon steel fixed-catalyst bed reactor failed after 20 years of service while in a standby condition (Fig. 27). At the time of failure, the unit contained primarily hydrogen at about 13 °C (55 °F) and 2400 kPa (350 psig) and was operating at an ambient temperature of 5 °C (41 °F) with no gas flow. The reactor fractured into six pieces in a single event. There was no evidence of an upset in operating conditions or introduction of oxygen to form an explosive mixture.

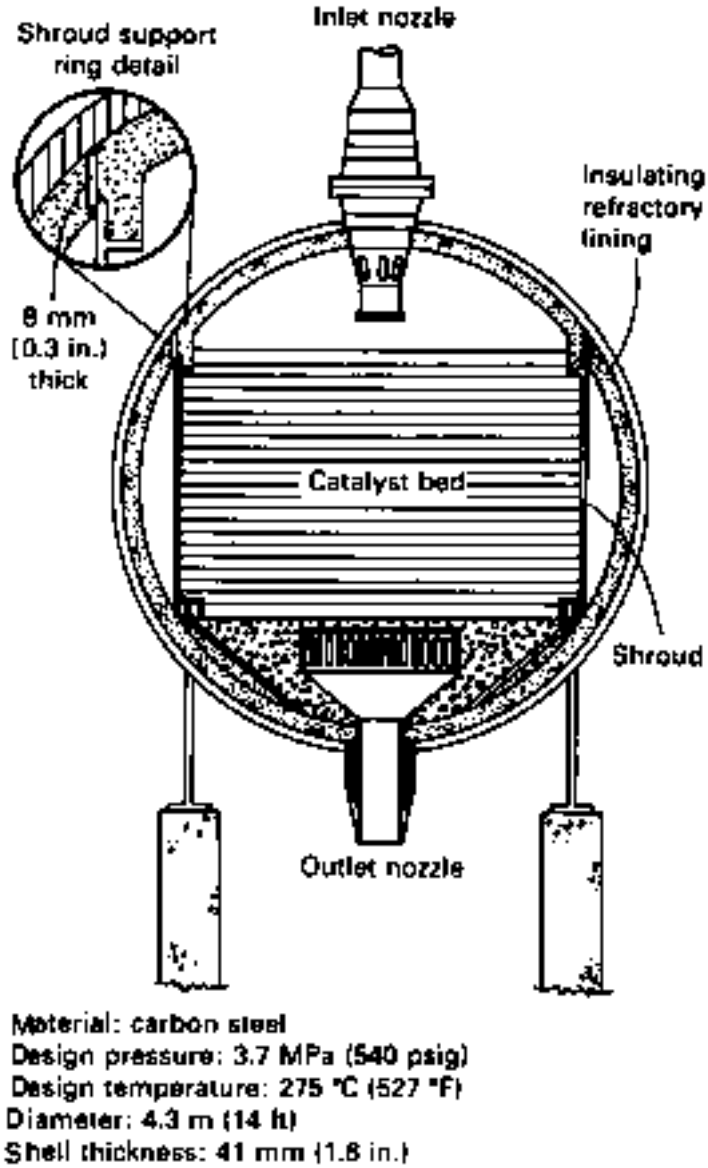


Fig. 27 Configuration and dimension of spherical fixed-catalyst bed reactor that failed after 20 years of service.

The reactor had been fabricated in 1958 from French steel A42C-3S, approximately equivalent to ASTM A201, grade B, with a 434-MPa (63-ksi) tensile strength and a 241-MPa (35-ksi) yield strength. The steel mill tests showed a Charpy V-notch energy absorption of 30 J (22 ft · lb) at 20 °C (68 °F). The plates had been hot formed into segments during construction of the reactor. The inside of the vessel was refractory lined to keep its temperature below 275 °C (527 °F) during operation. The vessel had a type 304 stainless steel shroud around the catalyst bed as protection against the overheating that was possible if the gas bypassed the bed through the refractory material.

Investigation and Conclusions. A visual examination of the fracture showed that the failure propagated through all of the plates of the shell (Fig. 28). The fracture followed the line of the shroud-support ring in four plates (plates A to D, Fig. 28), then branched into the rest of the vessel. All fracture surfaces were classical flat fractures, with no shear lip and chevron markings on all fractures except the fracture immediately at the shroud-support ring. At the ring, the chevron patterns suggested that the crack had propagated from the inside of the reactor to the outside wall. Weld seams did not provide a path for fracture. The failure clearly began at the toe of the shroud-support ring weld, but no evidence of a large pre-existing crack that could have triggered the failure could be found.

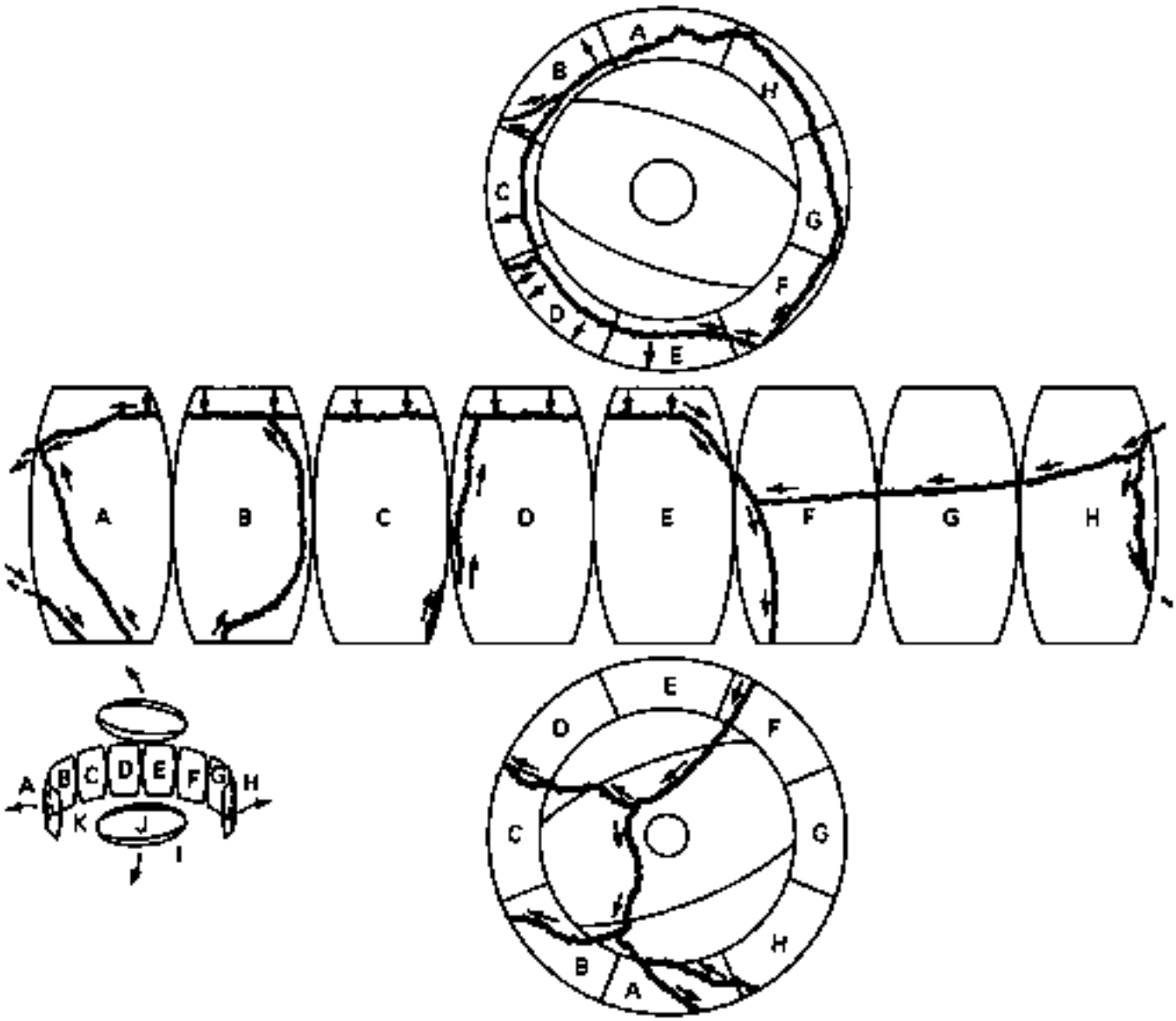


Fig. 28 Fracture path of failed pressure vessel. The arrows indicate the direction of crack propagation as determined from the chevron markings on the fracture faces. The letters identify the individual plates.

The shroud-support ring weld was ultrasonically inspected where it had not failed and was found to contain a number of 3-mm (0.12-in.) or smaller cracks at the root of the weld. Subsequent metallographic investigation revealed these indications to be lamellar tears that originated at nonmetallic inclusions on planes parallel to the plate surface, caused by weld shrinkage stresses. Figure 29 shows a typical lamellar tear.

Fig. 29 Lamellar tearing at root of weld for shroud-support ring.

Scanning electron microscopy confirmed the cleavage mode of fracture and the origin of the failure. Cleavage failure clearly initiated at the root of the weld, and it continued through the wall thickness. Along the root of the weld were indications of multiple sites of crack initiation at the pre-existing shallow lamellar tears and minor weld discontinuities resulting from arc breaks and excessive or incomplete penetration.

Metallographic cross sections of the fracture surfaces revealed the presence of extensive secondary cracking and twinning where the fracture followed the line of the shroud-support ring. The twinning was identified as Neumann bands, which in ferritic steels are caused by the highly localized deformation associated with fast fracture or dynamic stresses. This led to the conclusion that the fracture initiated at multiple sites within one plate, starting at the root of the weld, then initiated in neighboring plates in rapid succession.

The possibility of hydrogen attack was eliminated by the fact that the subsurface cracks were transgranular and were only present immediately below the fracture surface. This conclusion was substantiated by bend tests and toughness tests performed on specimens of the vessel material.

Chemical analyses were performed on the plates, which were found to be within specifications for this variety of carbon-manganese steel. Analyses for arsenic, molybdenum, nickel, copper, and chromium indicated no abnormality. The steel was not required to be fine grained; in addition, the soluble aluminum content responsible for controlling grain size and reducing susceptibility to strain aging varied among the plates, but was very low for three of the plates (<0.005%). The hydrogen contents of the steels were normal.

Toughness tests were performed on numerous plate samples from locations near the bottom of the reactor, at the middle, and at the top adjacent to where the fracture followed the shroud-support weld. The lowest values measured at 5 °C (41 °F) were 6.8 and 9.5 J (5 and 7 ft · lb), respectively. The plate in which the fracture initiated had a 27-J (20-ft · lb) transition temperature of 30 °C (86 °F), while the plates in which the fracture developed shear lips and ran off into the vessel had 27-J (20-ft · lb) transition temperatures below -5 °C (23 °F). The coarsest grained steels exhibited the lowest toughness values. In addition, the Pellini drop-weight test procedure (Ref 11) was used to study the nil-ductility temperature (the highest temperature at which small flaws can propagate as brittle fractures) of three plates. Two plates were at or below their nil-ductility temperature.

Refinery maintenance and inspection records showed that the carbon steel shroud-support ring had been removed for a hydrotest, the existing weld metal ground off, and the shell surface inspected using dye penetrant. The new ring was welded from one side. The weld was made with six passes by using low-hydrogen electrodes equivalent to E60-18 without preheat. The root pass was made with 2.4-mm ($\frac{3}{32}$-in.) electrodes, subsequent passes with 3.6-mm ($\frac{9}{64}$-in.) electrodes. The completed weld was visually inspected and not postweld heat treated, because postweld heat treatment would not have been required by applicable pressure vessel codes for new construction.

A numerical stress analysis of the weld was made for the design geometry and geometries that included sharp notches at the shroud-support ring-to-shell junction. These analyses included a condition meant to simulate a hot spot from the missing refractory lining and a combination of hot spot plus notch. The operating conditions investigated included thermal loadings from start-up and shutdown, internal pressure, axial friction from the catalyst bed, and a distributed load on the shroud due to the catalyst weight. The finite element analyses indicated a peak stress of 110 MPa (16 ksi) under an applied pressure of 2400 kPa (348 psig). The maximum stress occurred at the weld-to-ring junction, rather than at the weld-to-vessel junction, where failure occurred; the lower stresses on the reactor shell eliminated fatigue as a failure mechanism, in agreement with the metallographic observations.

The residual-stress pattern in the shroud-support ring attachment weld was measured at about 180° around the vessel from the crack-initiation site, where no cracking of the weld had occurred. Surface stresses were measured using a center-hole rosette strain gage technique, and through-thickness stresses were measured by a block-removal and layering technique. The maximum transverse tensile stress that would be applied across a radial defect in the shell near the root of the attachment weld was on the inner surface of the vessel wall near the toe of the weld. It was at least 276 MPa (40 ksi), which is equivalent to the yield strength of the base metal, and decreased linearly over the first 40% of the wall thickness.

Fracture mechanics tests were used to predict the mode of failure of the various plates. Measurements of crack-tip opening displacement indicated that ductile failure would occur in at least one plate that sustained brittle failure. No pre-existing defect could be found in the size range of ≥11 mm (0.43 in.), which would have been required to cause a brittle failure.

All of the evidence pointed to a change in the susceptibility of the plates to brittle failure near the point of failure initiation. Charpy V-notch specimens that were aged under either tensile or compressive stress at 180 °C (360 °F) for 168 h were found to lose about half their impact energy absorption from about 11 J (8 ft · lb) to about 5.5 J (4 ft · lb), compared with unaged controls. A series of subsize and full-size crack-tip opening displacement specimens employing a 0.5-mm (0.020-in.) deep by 0.15-mm (0.006-in.) wide welded-over notch was prepared. Large decreases in ductility could be ascribed to the welded-over notch. Complete cleavage failure could be induced in these samples by testing at high strain rates. A small crack that initiated in the embrittled zone could thus propagate to critical size and precipitate an unstable fracture.

The strain-aging tendencies of these steels were demonstrated by making microhardness traverses adjacent to the tips of welded-over notches. The microhardness near a welded-over notch was much higher than that in the adjacent base metal. Thin-foil transmission electron microscopy (TEM) was also used to demonstrate that high dislocation densities occurred in the vicinity of welded-over notches. The dislocations were bowed and pinned, an observation that is not exclusively associated with strain-aging embrittlement but that supports the proposed failure mechanism.

The final cause of failure was felt to be cracking that developed during the installation of the new shroud ring. The cracks caused by the root weld pass were not by themselves large enough to initiate fracture, but subsequent thermomechanical cycling during the remainder of the welding process caused dynamic strain aging and localized embrittlement at the crack tips. Three plates had relatively low toughness. Under the combination of high residual stresses, pressurization stresses, and a low metal-shell temperature, the cracks extended and caused total destruction of the vessel.

The recommendation from this analysis was to perform stress-relief heat treatments to reduce residual-stress levels after welding non-pressure retaining components to certain pressure shells. This treatment should reduce residual stresses and lessen the embrittlement effects of strain aging.

References cited in this section

10. A.R. Cuiffreda and R.D. Merrick, Brittle Fracture of a Pressure Vessel: Study Results and Recommendations, in *Proceedings of the American Petroleum Institute,* Vol 62, 1983

11. "Standard Method for Conducting Drop-Weight Test to Determine Nil-Ductility Transition Temperature of Ferritic Steels," E 208, *Annual Book of ASTM Standards,* Vol 03.01, ASTM, Philadelphia, 1984, p 346–365

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Ductile Fractures

Ductile fracture occurs with a significant amount of macroscopic plastic deformation. It is considered to take place by a shear mechanism in which the metal is torn with considerable expenditure of energy. Even though

final rupture may occur suddenly, most ductile failures of pressure vessels or piping develop gradually and give warning during propagation by means of a bulge or a leak.

The surface of a ductile fracture generally exhibits a dull, fibrous appearance, and there is evidence of plastic deformation, such as local bulging or distortion. A good indicator of deformation is a significant reduction of area at the fracture. When such an indicator is present, the failure may have resulted from some long-term change or deterioration of the material or from a gradual increase in the load. However, the mere classification of a fracture as ductile does not explain why it occurred, but simply suggests the factors that may have caused it. For example, gradual reduction in wall or section thickness by corrosion, erosion, or creep can render the material unable to sustain the operating load. To this extent, the classification of a failure as ductile depends on the time span over which the failure takes place.

Roy T. King, Cabot Corporation

Creep and Stress Rupture

Creep, or time-dependent deformation, is normally anticipated by designers of equipment that operates at elevated temperatures. The useful life of a pressure vessel or pressure boundary material can occur as a result of reaching or exceeding some design limit on strain. This is known as a creep failure, and it may involve only excessive deformation, not a fracture. Stress ruptures, complete failures under creep loadings, can breach a pressure boundary. The main variables determining the rate of deformation under creep conditions are temperature, time, and stress. It is not unusual to find other contributing factors, such as corrosion, fatigue, or material defects involved in creep and stress-rupture failures.

Creep usually occurs in time-temperature regimes for which diffusion-controlled processes are important. Within grains, the interactions of dislocations with solute atoms, second phases, and other defects can determine deformation rates; microstructures can change dramatically during the creep process. Deformation at grain boundaries, together with diffusion processes, can lead to the nucleation and coalescence of voids, causing grain boundaries to separate, sometimes at very low strains. The interpretation of the remaining useful life of a material that has undergone creep is a difficult, but important, concept.

Equipment that operates in the creep regime is frequently too expensive to discard without consideration of repair or partial replacement of material that has suffered creep damage. Life-extension programs are common in many industries and have recently become important in the electric power industry. While many other failure mechanisms must be considered in such programs, examination for creep damage is critical.

Example 17: Preventive Analyses of Croloy 1 $\frac{1}{4}$ Pressure Parts. It is generally not possible to predict the maximum life of base-loaded headers and piping until they have already developed microcracking. When cracking does occur, it can be monitored over a period of several years by a periodic inspection program (Ref 12). Typical elements of such a program include the following:

- An engineering stress analysis of the header, tube legs, and the associated piping system is conducted. Numerical analysis, a modern tool, can be used to determine which areas of the component should be tested and inspected during downtimes and to determine static and cyclic stresses accurately
- Dimensional examinations can be performed to determine the extent of time-dependent swelling, bowing, or other deformation that may have occurred
- Visual examination is performed to disclose the extent of other defects. Borescopes and fiberscopes can be used to examine otherwise inaccessible areas of a component. These examinations may be supplemented with dye-penetrant inspections in areas suspected of cracking
- Field replication is recommended to examine metallurgical microstructure *in situ.* Surfaces to be examined must be polished and etched by metallographic methods, usually under difficult field conditions and without regard to specimen flatness. The surface is then recorded by coating the surface with a resin that hardens and is stripped off, bearing the image of the specimen topography. The image is then coated to enhance detail and studied by optical or electron microscopy. Field replication is useful for detecting the early stages of creep, fatigue, grain-boundary separations, and other problems. It can also be used to detect spheroidization or graphitization of steels
- Boat samples of headers and piping are generally removed only after cracking has been detected by other means. Tube stubs may be removed to examine the more highly stressed regions of headers

In one example, this type of routine disclosed cracks caused by creep swelling in the stub-to-header welds in the secondary superheater outlet headers of a major boiler. The header was constructed of SA335-P11 material (1.25Cr-0.25Mo silicon steel). A review of the design of this unit showed that it had been designed according to allowable stresses that had been established in 1951 from a relatively small data base. As more information became available, the ASME Boiler Code allowable stresses for the material were reduced in mid 1965 for temperatures above 525 °C (975 °F); therefore, the unit was underdesigned. After about 30 years of service, repairs had been necessary on 61 of the 222 header welds, and a permanent solution to the problem was sought.

To document the amount of creep swelling that had occurred, the outside diameter of the header was measured and found to have increased 1.6% since its installation. It was re-measured after another 6 months and found to have swelled another 0.14%. Surface inspection revealed ligament cracks extending from tube seat to tube seat (Fig. 30). Cracks originated from inside the header, extending axially in the tube penetrations and radially from those holes into the ligaments. The ligament cracking of the header was a separate event from the stub-to-header weld cracking, although both were caused by high stresses in the header. Dye-penetrant inspection of the latter welds revealed cracks in 94 locations, ranging from small radial cracks to full 360° cracks. An example of this cracking is shown in Fig. 31. The unit was operated under reduced-temperature conditions and with less load cycling than previously until a redesigned SA335-P22 (2.25Cr-1Mo-0.25-Si) header was installed.

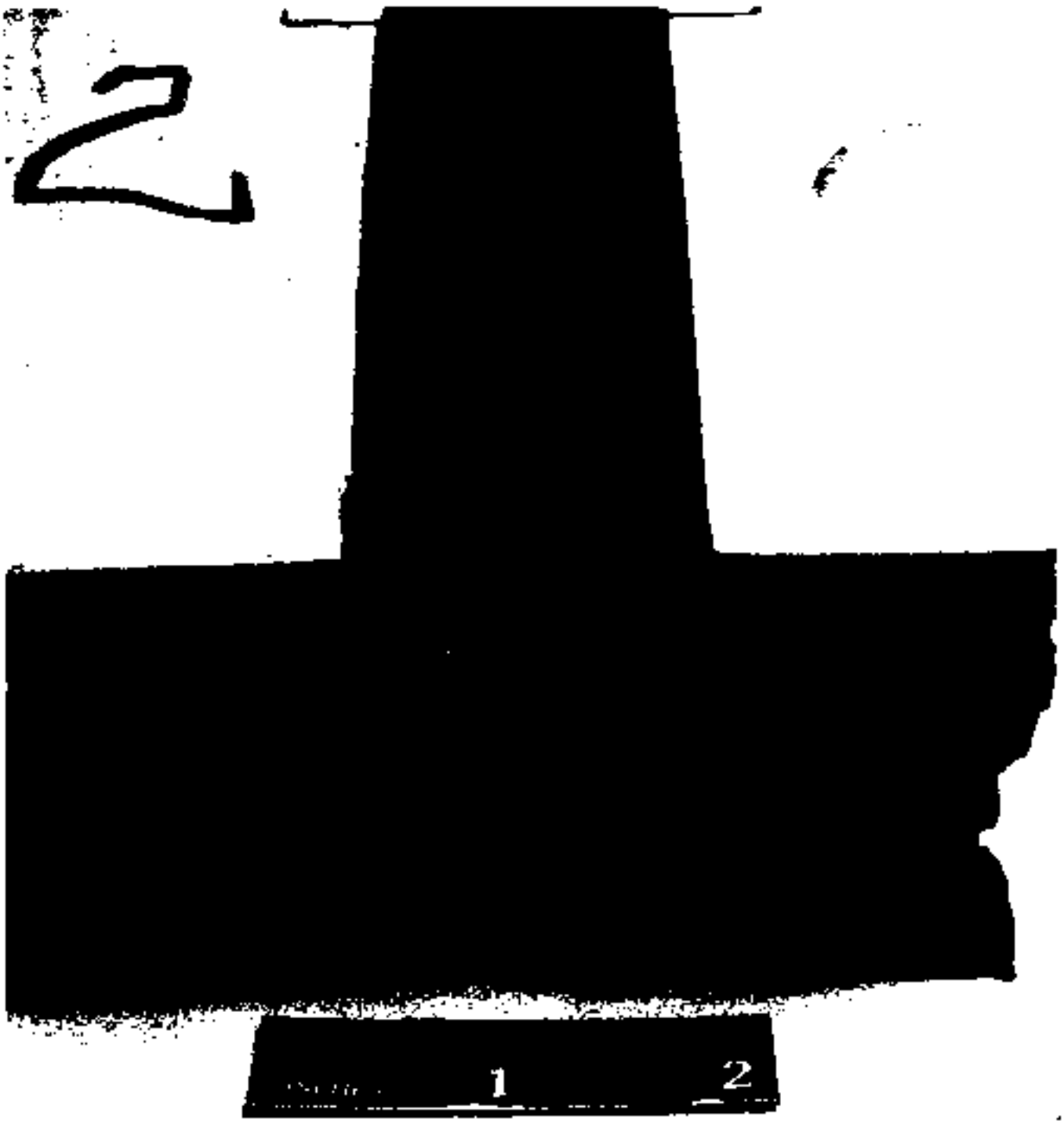


Fig. 30 Interligament cracking in a failed secondary superheater outlet header from a boiler.

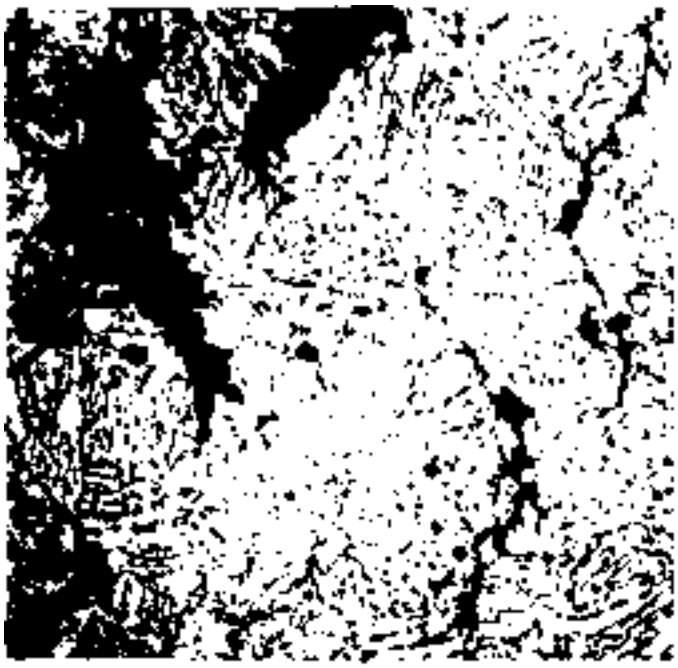


Fig. 31 Microstructure, linked voids, and split grain boundaries in the failed outlet header shown in Fig. 30. 400×

Example 18: Failure of a Main Steam Line of a Power-Generating Station. A main steam pipe was found to be leaking during a start-up after a 1-day shutdown. The leak was a large circumferential crack in a pipe-to-fitting weld in one of two steam leads between the superheater outlet nozzles and the turbine stop valves, a line made of SA335-P22 material (2.25Cr-1Mo steel) with an outside diameter of 475 mm (18.75 in.) and a wall thickness of 95 mm (3.75 in.). The design operating conditions were 25 MPa (3.6 ksi) at 540 °C (1000 °F).

Investigation. A boat sample was cut from one side of the main crack and subjected to an initial examination. Visual and magnetic-particle examinations revealed secondary cracks roughly parallel to the main failure in the weld metal and adjacent base metal. The main crack surface was rough, oriented about normal to the outside surface, and had a dark oxidized appearance. The topography of the crack surface reflected the weld-bead patterns. A metallographic section of this sample shows the main crack surface and the secondary cracks (Fig. 32). The cracking is predominantly intergranular, with some grain boundaries exhibiting unlinked intergranular voids that should eventually link to form complete grain-boundary separations. This microstructure was thought to be typical of elevated-temperature stress rupture in the material.

Fig. 32 Metallographic cross section through failure in 2.25Cr-1Mo weld main steam line of power plant. Secondary cracking in base metal indicates that failure is not uniquely the result of weld-metal properties.

Microscopic examination of the cross section revealed distinct shiny bands that etched slower than the remainder of the sample at the top of each individual weld bead. These bands were located within the interbead HAZ and consisted of localized areas with a large-grain ferrite matrix and relatively large carbide particles. The bands frequently contained small cracks and microvoids that did not extend into the surrounding fine-grain material. These features suggest that the ferrite bands served as initiation sites for the failure and that they lowered the rupture strength of the weld metal. The Knoop hardnesses of the bands tended to be in the range 138 to 144 HK, while the adjacent areas of the interbead HAZ were 161 to 168 HK. Qualitative analyses of the ferrite bands by EDS revealed no significant differences in composition between the bands and surrounding material.

While these observations identified the mechanism of cracking at the main failure, they did not provide insight into the cause of the premature service failure. The work was therefore expanded to include other weld metal and base metal of the steam line. The samples included a full ring section containing the failed weld and a quadrant of a weld that had not cracked. Smaller boat samples and pyramid samples were removed from other welds. Each was characterized, at a minimum, by chemical analyses and microstructural examinations. Stress-rupture tests and tensile tests were performed on some samples to determine the strength properties of the materials of construction. The results of these examinations are summarized as follows:

- The chemical analyses and mechanical properties were within specifications or reasonable for the materials of construction after service
- The failed weld and three others had an unusually high manganese content that identified them as having been fabricated by the submerged arc process. Three other welds were low in manganese, identifying them as having been fabricated by the manual shielded metal arc process. The submerged arc welds were cracked in the weld metal, but the submerged arc welds were cracked in the adjacent base metal or HAZ. The difference in the failure loci was apparently a result of the relative strengths and ductilities of the specific base metal/weld metal combinations that were present
- All observations are consistent with a mechanism of intergranular creep rupture at elevated temperature

- The crack shape indicated that cracking initiated on the pipe exterior, then propagated inward and in the circumferential direction in response to a bending moment load. There was no indication of a pre-existing flaw or defect at the initiation sites other than the microscopic hardness heterogeneity noted above

Conclusions. The primary cause of failure was the occurrence of bending stresses that exceeded the stress levels predicted by design calculations and that were higher than the maximum allowable primary membrane stress specified in the ASME Boiler and Pressure Vessel Code, Section I, and the ANSI Power Piping Guide.

Reference cited in this section

12. G. Harth, Header and Piping Inspection Guides Boiler Life Extension, *Power*, Vol 109, March 1965, p 99

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

Fatigue

Fatigue fractures result from cyclic stressing, which progressively propagates a crack or cracks until the remaining section can no longer support the applied load. Pressure vessels and pressure piping are subject to high static stresses arising from the pressure of contained liquids or gases, to stresses resulting from misalignment of components, and to residual stresses induced during welding. The cyclical component may be added mechanically—by vibration of associated equipment, pulsation from a compressor, or thermal cycling. Fatigue fractures generally originate at stress raisers, such as a discontinuity or a notch, which produces triaxial stressing in the material. Thermal fatigue occurs when cyclic stresses are produced by uneven heating or cooling, or when part of a vessel is restrained and cannot expand or contract uniformly.

Over a period of years, several low-carbon steel thick-wall vessels equipped with cooling jackets cracked when operated as batch reaction vessels at 400 °C (750 °F). These vessels operated with sharp thermal gradients across the walls. A small amount of sodium hydroxide could have been present in the vessels, and it was thought that cracking might have been the result of caustic embrittlement.

To increase the life of the vessels, the material was changed from low-carbon steel to a carbon-molybdenum steel. Occasional cracking still occurred. The material was then changed to 1Cr-0.5Mo steel. However, cracking occurred in a shorter time than with vessels made of carbon-molybdenum steel.

A stress analysis of these jacketed vessels indicated that cyclical stresses produced during operation might have been high enough to have caused cracking in the vessel walls. Furthermore, such stresses would be high enough to produce failure in any available grade of weldable structural steel. Thus, life of the vessels could be increased by reducing the stress but not by upgrading the material. Reducing the wall thickness would lower thermal stresses by decreasing the temperature differential in the wall.

Example 19: Failure of a Main Steam Line by Thermal Fatigue (Ref 13). Visual inspection of a type 316 stainless steel main steam line of a major utility boiler system revealed cracks on its outer surface near a hanger lug. At that time, the system had accumulated 130,520 h of high-temperature operation and had been subjected to 326 start-up cycles, many of which were from forced shutdowns. The steam line was a 230-mm (9.05-in.) outside-diameter pipe with a 64-mm (2.525-in.) minimum wall thickness.

Investigation. One third of the pipe was immediately subjected to demineralized water penetrant visual inspection, including all sections of pipe from the heat that had cracked. No further cracking was found.

A 3.4-m (11-ft) section including the cracked area was removed from the line. From this, boat samples and several ring sections were removed for examination, including the cracked region. Conventional metallographic techniques were used to assess the extent of cracking and the nature of the cracks. The damage in all specimens was essentially the same: cracking initiated at the outside of the pipe wall or immediately beneath the surface. The cracks were all intergranular, oriented normally or nearly so to the main hoop stress direction. The cracks were usually filled with oxide (Fig. 33).

Fig. 33 Oxide-filled intergranular cracks oriented normally to the hoop stress direction in the main steam line.

Examination of the metallographic samples at higher magnification revealed that the microstructure of the failed pipe consisted of a matrix precipitate array, probably $M_{23}C_6$, and large σ-phase particles in the grain boundaries. Although void formation and propagation along the σ/austenite interface is cited as a potential failure mode for this σ-phase morphology, the cavitation here occurs primarily in the ligament region between σ-phase particles. Scanning electron microscopy of the polished-and-etched cross sections of areas away from the cracks clearly showed this. The cavities were found to be decorated with small particles having a high sulfur content and were found to be faceted.

Charpy impact specimens and compact-tension specimens were cut from a section of the steam line that contained several cracks. The Charpy impact tests showed the inside pipe wall to be tougher than the outside, where the cracking and grain-boundary separations had occurred. There was no indication of a fracture-mode transition. Only slow, stable crack growth occurred in the compact-tension tests. One compact-tension specimen contained a natural crack from the steam line, but it behaved the same as specimens that had fatigue-sharpened notches on their crack fronts.

The insulation was examined to determine whether the impurities contained might have contributed to the failure. Its pH, leachable sodium, potassium, and lithium were analyzed. Colorimetric and the x-ray fluorescence spectroscopy were used to determine sulfur and chloride contents. The material was basic and had sufficiently low leachable impurities that intergranular stress-assisted corrosion was eliminated as a cause of failure.

Field metallographic techniques were developed, using a repeated polish-and-etch technique, to examine the remainder of the piping. A portable grinding tool was used to prepare the surface, followed by swab etching. A total of 72 locations in the piping system were examined, with the result that all material upstream of the boiler stop valve had oriented cracking of the type previously mentioned, while all locations downstream of the valve were free of damage. The surface metallography was validated by removal of boat samples that demonstrated the replication technique to provide microstructures representative of bulk material. A chemical analysis of the piping showed compliance with the material specification.

Significantly, the pipe showed no evidence of swelling due to creep. This fact, coupled with the lack of damage downstream from the stop valve and the high ductility of the inner-wall material, cast doubt on classical creep due to overstressing or overtemperature as the cause of failure.

Residual-stress measurements were made using a hole-drilling technique and strain gage rosettes. The holes were drilled with 50-μm alumina abrasive to minimize stresses introduced during drilling. Relatively large tensile axial residual stresses were measured at nearly every location investigated, but a large residual hoop stress was found for locations before the stop valve.

Conclusions. It was theorized that one or more severe thermal downshocks might cause the damage pattern that was found. A transient heat transfer and thermal-stress analysis was performed with numerical methods and software identified as CREPLACYL and with material data taken from the *Liquid Metal Fast Breeder Reactor Materials Handbook* (U.S. Department of Energy). The resultant elastic/plastic/creep deformation analysis indicated that a severe thermal downshock could explain the pattern and that one or more downshocks of less than 260 °C (500 °F) but more than 65 °C (150 °F) could have been involved. Thus, the root cause of the failure was thermal fatigue, with associated creep relaxation.

Reference cited in this section

13. J.E. Bynum, F.V. Ellis, M.H. Rafiee, W.F. Siddall, T. Daikoku, H. Haneda, and J.F. DeLong, "Failure Investigation of Eddystone Main Steam Piping," paper presented at the 65th AWS Convention, Dallas, 1984, and publication TIS-7540, CE Power Systems, Windsor, CT

Failures of Pressure Vessels

Roy T. King, Cabot Corporation

References

1. G. Sorell and M.J. Humphries, High Temperature Hydrogen Damage in Petroleum Refinery Equipment, in *Corrosion/78*, National Association of Corrosion Engineers, Houston, 1978

2. C.J. Czajkowski, "Evaluation of Intergranular Cracks on the Ring Header Cross at Grand Gulf Unit #1," Technical Report A-3500, Prepared for the U.S. Nuclear Regulatory Commission, Region II, under contract, Brookhaven National Laboratory, Upton, NY, March 1986

3. "Standard Methods for Determining Average Grain Size," E 112, *Annual Book of ASTM Standards*, Vol 03.03, ASTM, Philadelphia, 1984, p 120–152

4. "Standard Practices for Detecting Susceptibility to Intergranular Attack in Austenitic Stainless Steels," A 262, *Annual Book of ASTM Standards*, Vol 03.02, ASTM, Philadelphia, 1984, p 1–27

5. Brittle Fracture of a Thick Walled Pressure Vessel, *BWRA Bull.*, Vol 7 (No. 6), June 1966

6. R.D. Wylie, J. McDonald, Jr., and A.L. Lowenberg, "Weld Deposited Cladding of Pressure Vessels," paper presented at the 19th Petroleum Mechanical Engineering Conference, Los Angeles, Sept 1963

7. C.J. Czajkowski, "Investigation of Shell Cracking on the Steam Generators at Indian Point Unit No. 3," NUREG/CR-3281, also BNL-NUREG-51670, U.S. Nuclear Regulatory Commission, Washington, June 1983

8. C.J. Czajkowski, "Constant Extension Rate Testing of SA302 Grade B Material in Neutral and Chloride Solutions," NUREG/CR-3614, also BNL-NUREG-51736, U.S. Nuclear Regulatory Commission, Washington, Feb 1984

9. D.B. Bird, “Cracking in Carbon $\frac{1}{2}$Moly Desulfurizer Reactor Welds,” paper presented at the 1985 Fall Meeting of the American Petroleum Institute Operating Practices Committee, Subcommittee on Process Operations

10. A.R. Cuiffreda and R.D. Merrick, Brittle Fracture of a Pressure Vessel: Study Results and Recommendations, in *Proceedings of the American Petroleum Institute,* Vol 62, 1983

11. “Standard Method for Conducting Drop-Weight Test to Determine Nil-Ductility Transition Temperature of Ferritic Steels,” E 208, *Annual Book of ASTM Standards,* Vol 03.01, ASTM, Philadelphia, 1984, p 346–365

12. G. Harth, Header and Piping Inspection Guides Boiler Life Extension, *Power,* Vol 109, March 1965, p 99

13. J.E. Bynum, F.V. Ellis, M.H. Rafiee, W.F. Siddall, T. Daikoku, H. Haneda, and J.F. DeLong, “Failure Investigation of Eddystone Main Steam Piping,” paper presented at the 65th AWS Convention, Dallas, 1984, and publication TIS-7540, CE Power Systems, Windsor, CT

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Introduction

ORTHOPEDIC IMPLANTS are unique in that they are exposed to the biochemical and dynamic environment of the body and their design is dictated by anatomy and restricted by physiological conditions. Orthopedic implants are artificial devices (allentheses) that are mounted to the skeletal system of the human body or an animal for various purposes, such as supporting bone, replacing bone or joints, and reattaching tendons or ligaments.

Depending on the duration of their function, the two major categories of such devices are prostheses and internal fixation implants. Prosthetic devices (Fig. 1) are implants designed to remain in the body for a lifetime. They serve as joint replacements (replacing the total joint or only one component) or as bone replacements (replacing larger sections of bone, for example, after tumor resection). Internal fixation devices (Fig. 2) are implants designed to provide temporary stabilization. They are used for maintaining the shape of the reconstructed bone for treatment of fractures or for corrective orthopedic operative procedures. After healing, internal fixation devices, having served their purpose, are commonly removed.

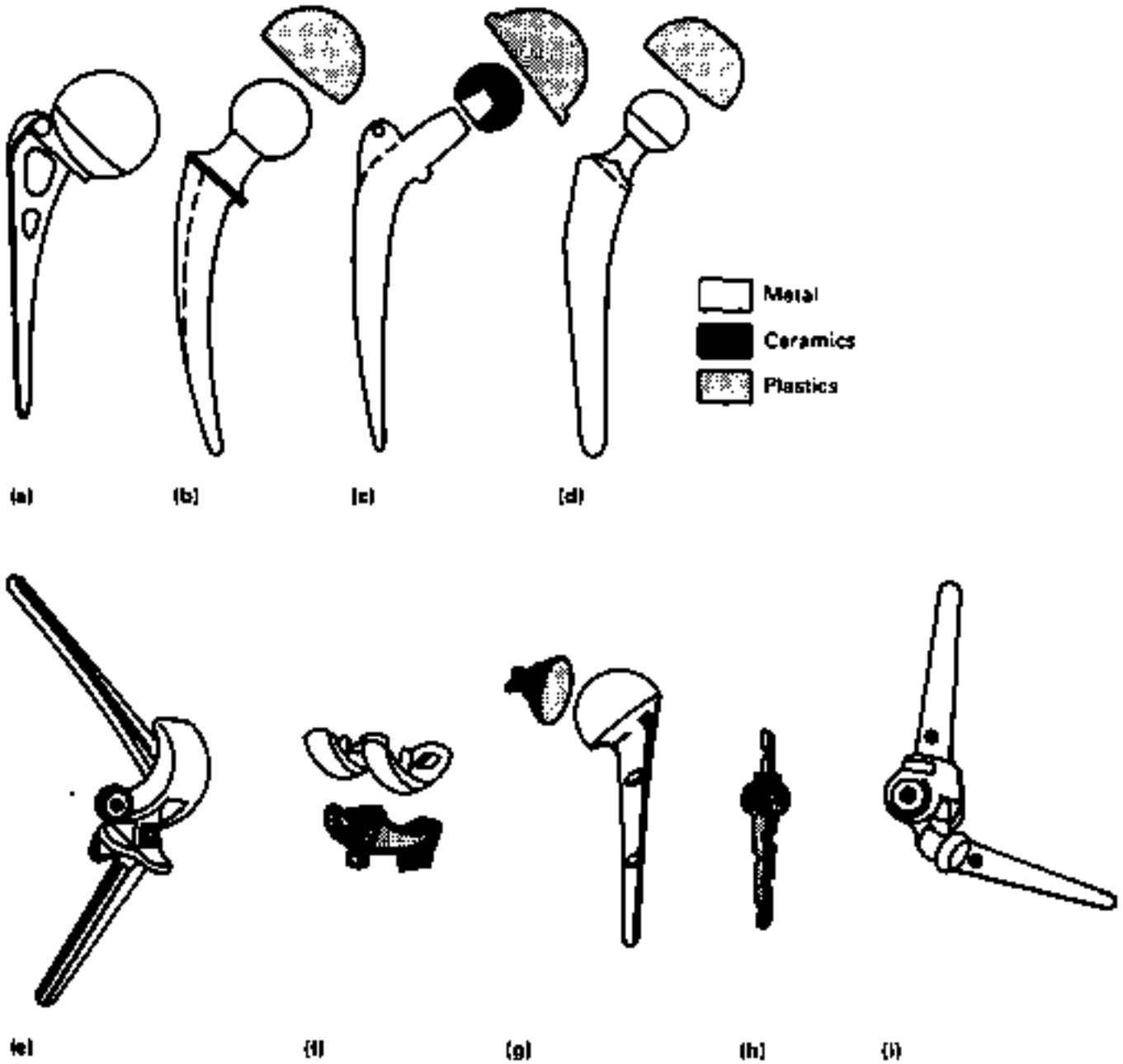


Fig. 1 Typical examples for joint prostheses (schematic). (a) Classic Moore hip endoprosthesis. (b) Müller total hip prosthesis (metal against polyethylene acetabular cap). (c) Weber total hip prosthesis with movable head and metal, ceramic, and polyethylene components. (d) Müller total hip prosthesis with straight stem. (e) Hinge-like knee joint prosthesis; metal against metal (G.E.U.P.A.R.). (f) Sliding knee joint prosthesis; metal against plastic (Geomedic). (g) Total shoulder joint prosthesis; metal against plastic (St. Georg). (h) Total finger joint prosthesis with metal and plastic components (St. Georg). (i) Total elbow joint prosthesis; metal-to-metal (McKee)

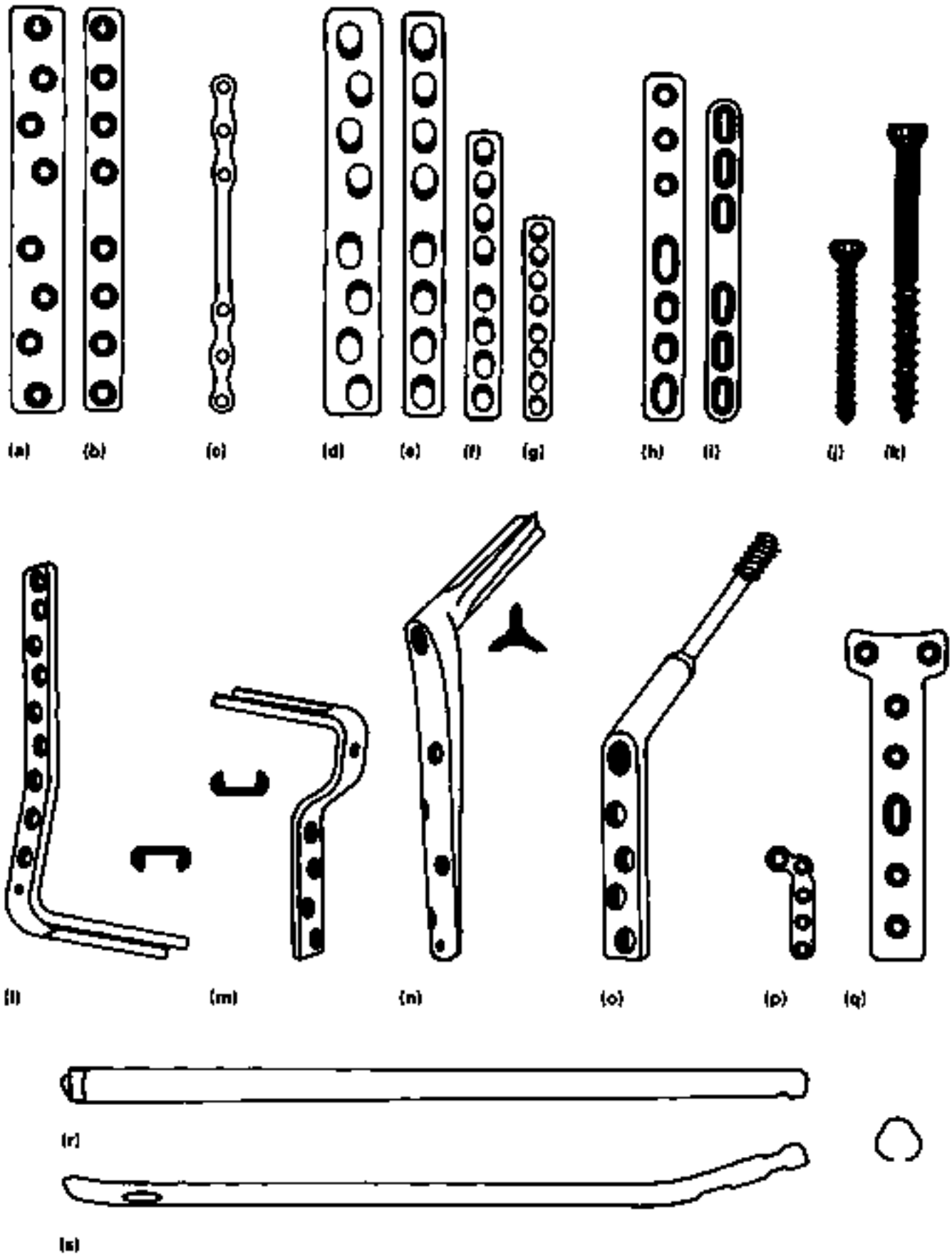


Fig. 2 Typical examples for orthopedic internal fixation devices (schematic). (a) and (b) Round hole bone plates (can be used with compression devices). (c) Classical Sherman bone plate. (d) to (g) Dynamic compression plates of various sizes. (h) Compression bone plate with glide holes. (i) Classical Bagby compression bone plate. (j) Cortical bone screw.

(k) Cancellous bone screw (with shaft to produce compression). (l) Condylar angle blade plate. (m) Hip plate for osteotomies. (n) Jewett nail plate with three-flanged nail. (o) Two-component dynamic hip screw plate. (p) Miniature L-plate—for example, for hand surgery. (q) T-plate for the humeral and tibial head. (r) Straight Küntscher intramedullary femoral nail. (s) Intramedullary tibia nail. Note cloverleaf profile on both (r) and (s).

Ortrun E.M. Pohler, Institut Straumann AG

Prosthetic Devices

Artificial joints are mainly used to replace painful joints that are degenerate and inflicted with diseases. Severe joint injuries are occasionally treated with prosthetic components.

When prostheses replace the entire joint, two artificial joint surfaces are anchored at the bone ends that remain after resection, and the prosthesis simulates the joint motion. Occasionally, only one joint surface, such as the femoral head, is replaced by an artificial component, which works against a natural articulation surface. Most joint replacements are performed on the hip. Because the normal femoral head is almost spherical, its motion is relatively easy to simulate. More complicated is the articulation of the knee and the various types of knee prostheses that compromise the original motion range. Prostheses are also available for the humeral head (shoulder), elbow and finger joints, the ankle, and the jaw (Fig. 1).

Most total joint systems have articulation components made from dissimilar materials so that, for example, metal or ceramics work against a plastic material, such as polyethylene (Ref 1). Finger joints, as well as some other prostheses, are made completely from plastics or metal-reinforced plastics (Ref 2).

Many prostheses are anchored by using a bone cement in a bed that has been mechanically prepared in the remaining bone. Recently, various designs have been developed for the cementless anchoring of prostheses by a tight fit. Certain prosthesis systems have porous surface coatings on their shafts to increase the surface area and to allow microanchorage of the bone. Developments in the field of joint prostheses are progressing rapidly. In this article, only metallic components will be considered.

In tumor resections, large defects are created by extensive bone and joint removals, which are often bridged with custom-made prostheses, modified standard joint prostheses, or fracture-treatment implants. Bone cement and/or other filling materials are sometimes added to the reconstruction for more stability. In the treatment of tumor resections, the surgeon is often forced to be inventive and to choose unorthodox techniques of reconstruction. Frequently, the remaining bone has been irradiated and lacks its normal regenerative vitality.

References cited in this section

1. D.C. Mears, *Materials and Orthopaedic Surgery,* The Williams & Wilkins Co., Baltimore, MD, 1979

2. L.R. Rubin, Ed., Chapters 23–27, in *Biomaterials in Reconstructive Surgery,* The C.V. Mosby Co., St. Louis, MO, 1983

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Implants for Internal Fixation

Implants for internal fixation are characterized by their versatility (Fig. 2). They are used to maintain the shape of the reconstructed bone in fracture treatment of the injured patient or, in corrective orthopedic surgery, to stabilize osteotomies (sectioning and realignment of bone) or to perform arthrodeses (stiffening of joints by fusion of bones). Various types and sizes of bone plates, bone screws, intramedullary rods (nails), pins, and orthopedic wire are available (intramedullary refers to use of the marrow space of the bone for support). Some implants have very specific uses; others have a wide range of applications with different functions.

Under the impact of a trauma, bones can fracture in any configuration and combination. In each patient, the fractures are unique and require individual treatment. This explains the necessity of versatile implants. Nevertheless, certain fracture classifications and methods for the treatment of typical fracture configurations have been established (Ref 3). In the selection of the appropriate implants and operation technique, not only the type of fracture and its location but also the condition of the soft tissue, the structure and consistency of the bone, the age and condition of the patient, and any present complication must be considered. Even when the best possible reconstruction has been achieved, the weight-bearing stability of an internal fixation device can vary widely, and the weight-bearing schedule of the patient must be correspondingly adjusted.

Depending on the situation, the patient may be allowed one of the following postoperative modes of motion: exercise only, partial weight bearing, or full weight bearing. The rate of bone healing depends on different factors and determines the further course of postoperative weight bearing. Because of the above-mentioned complexity, the possibility of an implant-related complication is comparably higher in fracture treatment than in corrective surgery or joint replacement, which are both usually carried out under standard conditions. Unless a pathological situation is present, the surgeon can achieve primarily stable biomechanical conditions when performing a corrective osteotomy or a joint replacement. When these conditions remain, the probability of an implant failure is minimal. The stabilization difficulties can be listed in ascending order, as follows:

- Joint replacements (endoprostheses)
- Corrective orthopedic surgery
- Fracture treatment
- Tumor resection (most difficult)

Reference cited in this section

3. M.E. Müller, M. Allgöwer, R. Schneider, and H. Willenegger, *Manual of Internal Fixation*, 2nd ed., Springer Verlag, 1979

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Complications Related to Implants

Internal fixation or joint replacement procedures can fail for such reasons as:

- Failure of the bone to heal
- Bone resorption
- Breakage of bone
- Loosening of implants
- Bending of implants
- Breakage or disintegration of implants

These events can be interrelated and usually require an intervention with reoperation to effect clinically successful results.

The term implant failure can be misleading. Therefore, in this article, the term implant failure is used to indicate that an implant has broken or disintegrated in a manner such that it cannot fulfill its intended function. Implant

failure is not identical with a faulty implant, because there can be purely biological and biomechanical reasons for implant breakage or malfunction. In most cases, only a thorough analysis of the clinical conditions under which an implant failed can explain the reason for failure.
Implants can undergo surface attack by fretting, fretting corrosion, or wear. These types of attack can be relatively mild; they often occur only on the microscopic level, do not interfere with the functioning of the implant or the healing of the bone, and do not require reoperation. On prosthetic devices, however, wear of the articulation surfaces has occasionally been found to be so intense that the components have had to be replaced. Corrosion of implants involving dissolution that requires intervention is found only with materials that are excluded by the official standards for materials for orthopedic implants. In this article, the term degradation will be associated with the surface attack of implants.

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Metallic Implant Materials

A number of metals and alloys have proven to be satisfactory as implant materials during years of surgical application. They are specified as implant materials by standards of the American Society for Testing and Materials (ASTM) and the International Organization for Standardization (ISO), as listed below, and by other national standards. These materials are corrosion resistant and well accepted by body tissues (biocompatible) and therefore satisfy two of the basic requirements for implants. These two properties are generally related because the less substance the metal surface releases, the better the material is accepted by the tissue. Experimentation has proven that the different pure metals of the periodic table exhibit cell toxicity at different concentrations in tissue and organ-culture tests (Ref 4, 5, 6). Some of the metals play important roles in the body metabolism, although they can become toxic beyond certain concentrations (Ref 7). For other metals that behave indifferently in the body, such as titanium, no metabolic function is yet known. Information on the toxicity of various metals is available in the article “Toxicity of Metals” in *Properties and Selection: Nonferrous Alloys and Special-Purpose Materials,* Volume 2 of the of *ASM Handbook.*
Fatigue resistance is another general requirement for implants, but the critical loading is different for the various types of implants and applications. The required mechanical strength of orthopedic implants also varies and depends on the shape of the implant and the application. Good ductility (discussed below) is often desired.
The following is a list of major standards for orthopedic implant materials:

- *Stainless steel:* ASTM F 55-82, ASTM F 56-82, ASTM F 138-82, ASTM F 139-82 (contains remelted Special Quality), ISO/DIS 5832/1 (1986)
- *Unalloyed titanium:* ASTM F 67-83, ISO 5832/II (1984)
- *Titanium alloy Ti-6Al-4V ELI:* ASTM F 136-79, ISO 5832/III (1978)
- *Cast cobalt-chromium-molybdenum alloy:* ASTM F 75-82, ISO 5832/IV (1978)
- *Wrought cobalt-base alloys:* ASTM F 90-82, ASTM F 562-84, ISO/DIS 5832/6, ASTM F 563-83, ISO/DIS 5832/8

Certain applications—for example, in the skull—involve use of tantalum and niobium, both of which have high corrosion resistance and are biologically well accepted. Currently, clinical trials are being performed with modified austenitic stainless steels and titanium alloys, which may become accepted implant materials in the future. Additional ASTM standards specify cast and forged conditions for stainless steels and other materials; they are found among other implant-related standards in Ref 8.

References cited in this section

4. H. Gerber *et al.,* Quantitative Determination of Tissue Tolerance of Corrosion Products by Organ Culture, *Proceedings of the European Society for Artificial Organs,* Vol 1, 1974, p 29–34

5. H. Gerber and S.M. Perren, Evaluation of Tissue Compatibility of In Vitro Cultures of Embryonic Bone, in *Evaluation of Biomaterials,* G.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

6. V. Geret, B.A. Rahn, R. Mathys, F. Straumann, and S.M. Perren, A Method for Testing Tissue Tolerance for Improved Quantitative Evaluation, in *Evaluation of Biomaterials,* G.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

7. G.K. Smith, Systemic Aspects of Metallic Implant Degradation, in *Biomaterials in Reconstructive Surgery,* L.R. Rubin, Ed., The C.V. Mosby Co., St. Louis, MO, 1983

8. Medical Devices, *Annual Book of ASTM Standards,* Vol 13.01, ASTM, Philadelphia, 1985

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

The Body Environment and Its Interactions With Implants

The Biochemical Environment. Orthopedic implants are exposed to the biomechanical and biochemical forces of the body, and certain interactions take place between the implants and the biological environment. Figure 3 indicates the ionic composition of blood plasma and interstitial and intracellular fluids. Of the components shown, the chlorine ions are the most critical for metal implants. The normal pH of the body liquids is about neutral—in the range of 7.2 to 7.4 pH (Ref 10). At injured sites, the pH shifts to acidic values of about 5.2; in hematoma, it can drop to 4.0. In cases of infection, the pH shifts to alkaline values (Ref 11). At sites of local fretting corrosion, the environment can also become acidic through corrosion products and thus enhance further corrosion.

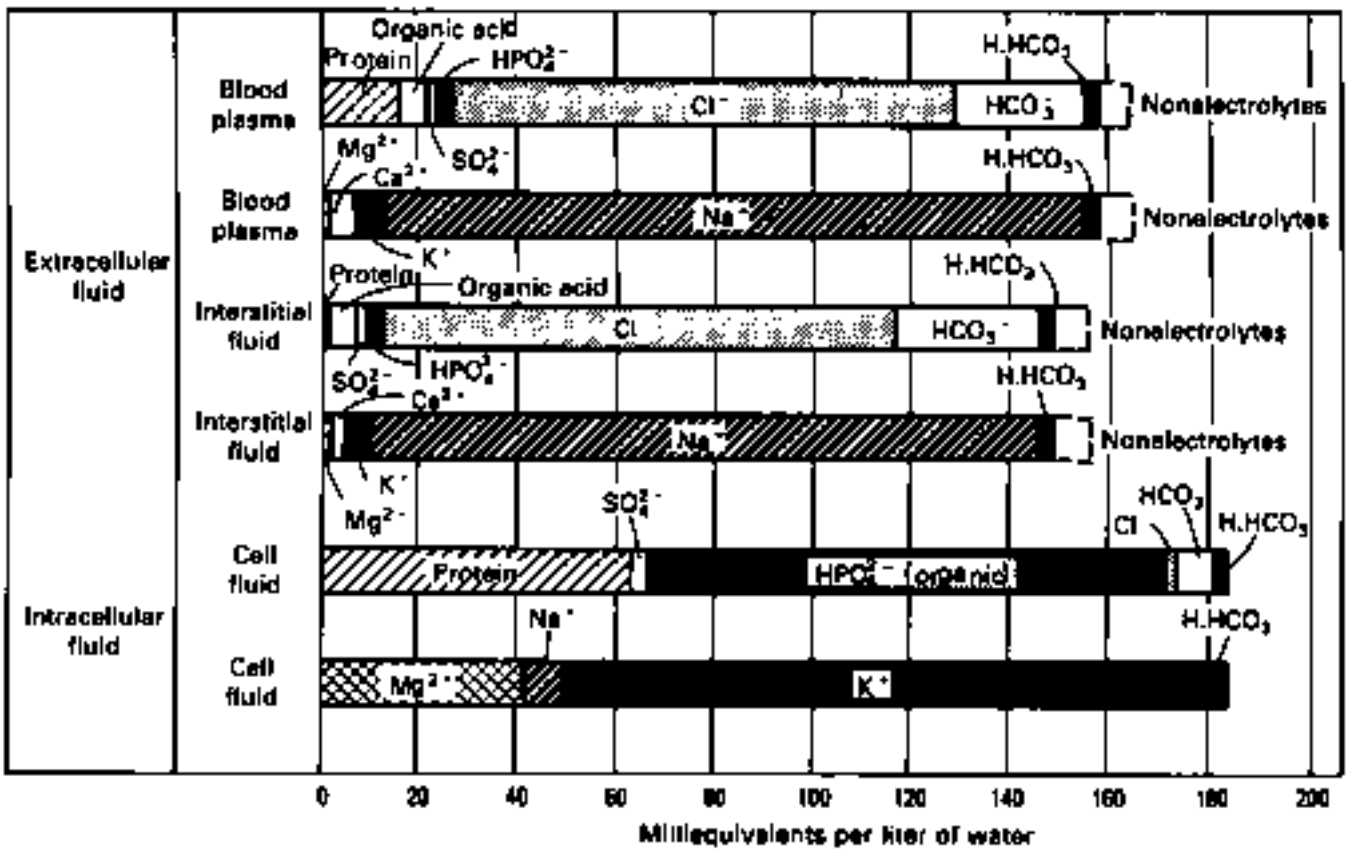


Fig. 3 Ionic compositions of blood plasma, interstitial fluid, and intracellular fluid. Source: Ref 9

Chromium steels, low-grade austenitic stainless steels, and sensitized stainless steels are not sufficiently corrosion resistant in the body environment. This has been observed on retrieved implants that were manufactured 20 to 40 years ago. High-grade austenitic stainless steels as specified in the standards (ASTM F 138/139, ISO 5832/1; 1986) exhibit overall corrosion resistance and are presently the materials most frequently used for internal fixation devices.

The Dynamic Environment. Various types of forces act on the implants and the bone. In the intact musculoskeletal system, the acting forces are balanced. The macroscopic shape and the microscopic structure of the bone are adjusted to the forces of normal motion and loading. Muscles and tendons counteract bending forces that would overload the bone. When loaded physiologically under weight bearing, the bone undergoes small elastic deformations, which act as a fine stimulus that keeps the bone in a healthy condition. For example, a trained athlete develops heavier and stronger bones than an untrained person, or if a patient is bedridden for a long period of time, his bone structure becomes porous because it is not exposed to the normal loading forces.

When a bone is fractured, the balance of forces is destroyed. This can be well studied on the dislocations of injured bones; in these cases, the muscle forces pull the bone fragments in various directions. After operative reconstruction of a fractured bone, the fragments are stabilized with implants. If the bone is perfectly reduced, the entire implant is supported by bone, and the acting forces are again in equilibrium. Relatively small and uncritical loads are then exerted on the implant, and the risk of implant-related complications is minimal.

However, if the bone is not completely reconstructed—that is, fracture gaps are present or fragments of bone are missing—the weight-bearing forces are not completely balanced and evenly distributed. As a result, bending and torsional stresses can concentrate in areas of the implant where bone support is missing. Under weight bearing, the implant undergoes cyclic loading in these zones, and a potential risk of fatigue damage may arise. For fatigue cracks to develop, it is not necessary for the implants to be loaded in the plastic deformation range. Local stresses that occur under loading in the elastic deformation range of the material are sufficient to initiate fatigue cracks in the surface of the implant.

The development of fatigue damage depends on the number of load cycles and the intensity of loading. In clinical terms, this means that implant fatigue depends on the width of the bone gaps, on the length of the lever arms, and on the intensity and duration of the weight bearing, if a critical fatigue condition develops.

Table 1 gives an estimation of the number of cycles an implant may undergo during a given time period, during which a certain amount of motion per day is anticipated. To estimate the loading, static and dynamic forces must be considered. In one study, a maximum force of 2600 N (585 lbf) was determined on the femoral head in the loading phase during a test of a 60-kg (135-lb) patient (Ref 12). This is more than four times the body weight.

Table 1 Estimation of load cycles for one leg during walking

For these calculations, it is assumed that one load cycle per leg takes 2 s, which means that every second one step is made with alternating legs (other authors estimate 1 to 2.5×10^6 load cycles per year per leg).

Motion per day, h	**Cycles per day**	**Cycles per 1 month**	**Cycles per 4 months**	Cycles per year
1	1,800	54,000	216,000	**657,000**
3	5,400	162,000	648,000	**1,971,000**
6	**10,800**	**324,000**	**1,296,000**	**3,942,000**

As long as bone healing progresses normally, an implant will not undergo fatigue fracture, because the loading decreases successively as the bone becomes more supportive. Normal bone has a remarkable inherent regeneration potential. However, certain conditions must be fulfilled so that the bone heals and regenerates. The course of bone healing is not fully predictable and depends on various factors.

Bone Healing. Normal bone regenerates continuously at a relatively slow rate. About 3% of the Haversian systems of the bone structure are usually in a stage of remodeling (Ref 13, 14). In the injured bone, the remodeling and regenerative processes are highly activated. Figure 4 shows a microradiograph of the cortical bone structure in the remodeling stage. One important precondition for bone healing is immobilization. Bone tissue tolerates only minimal amounts of elastic motion, particularly during repair. This explains the high success of operative fracture treatment by internal fixation. Even in severe joint injuries, functional healing is achieved.

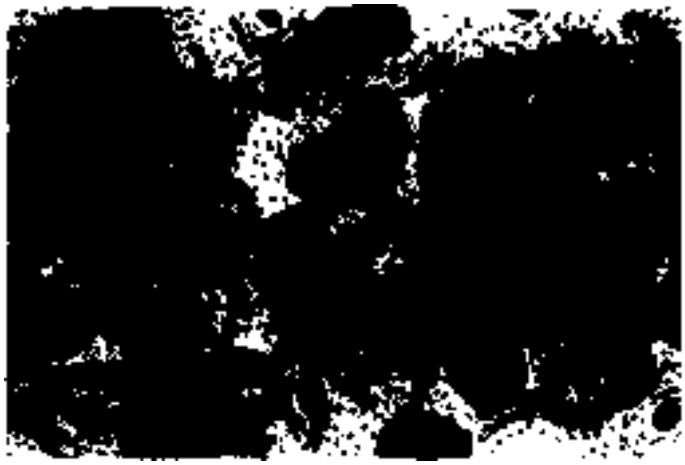


Fig. 4 Microradiograph from thin section of cortical bone. The varying x-ray density of the Haversian systems and the enlargement of some blood vessel cavities indicate that the bone is in a stage of remodeling. 77×

Depending on the local conditions, different bone repair mechanisms can occur (Ref 13, 14). Generally, one can distinguish secondary and primary bone healing. Secondary bone healing takes place if relatively wide fracture gaps exist and a certain amount of mobility is present. The bone heals through callus formation; that is, fibrous tissue and cartilage form first, followed by bone. Under too much motion, complications can arise because bone does not form, and nonunion, or pseudarthrosis, develops (pseudarthrosis is the formation of a false joint between bone fragments). If this happens and an implant is present, enough load cycles can accumulate at the fracture zone to cause the implant to fail through fatigue. In most such cases, a reoperation is inevitable to treat the nonunion—for example, by application of compression—regardless of whether a broken implant is present or not.

Example in this article discusses a case in which the implant was removed to treat a nonunion. Microscopic analysis revealed that a fatigue crack had already developed at the hole of the plate, which was situated near the unstable fracture site. Another plate was inserted under compression, and the bone healed without further complications.

Primary bone healing develops without callus formation, if the bone fragments are in contact with each other and are rigidly stabilized with implants. The bone fragments unite by means of transmigrating blood vessels and subsequent remodeling of the bone structure. These healing processes can be disturbed if the patient has a pathological bone condition, if an infection is present, or if necrotic (dead) bone fragments are resorbed.

Necrosis (the pathological death of living tissue) occurs when the blood supply of the bone fragments is damaged by the injury. Such fragments loose their nourishment and vitality. Complications can arise when necrotic fragments are resorbed: the fracture gap widens, relative motion increases, and bone healing is prevented or significantly delayed. Again, an implant can be subjected to a long period of increasing cyclic loading with a chance of fatigue failure. To treat such a situation and to enhance bone healing, implantation of cancellous (spongy) bone grafts and compression of the fracture site may be indicated. Although a second operation involves discomfort for the patient, it is sometimes necessary in order to ensure bone healing and satisfactory functional results.

Bone Resorption Through Motion. Clinical and experimental observations indicate that persistent motion beyond a certain degree causes bone resorption when mechanical instability is present (Ref 15, 16). Thus, fracture gaps can widen and implants, or portions of implants, can become loose. The widening of fracture gaps means increased cyclic loading of the implant unless weight bearing is reduced. Implant loosening can cause either local stress concentration with increased risk of fatigue or relaxation of the stresses in the implant. Under the latter conditions, the chance of implant failure is reduced. Local bone resorption followed by mechanical degradation of bone cement and partial loosening of prosthesis stems is a typical sequence of events that is responsible for fatigue failures of prostheses.

Mechanical Properties of Bone. As a construction material, bone has properties that differ from those of metals. Bone is viscoelastic, can resorb or regenerate, and thus may change its mechanical properties and volume with time. Comparative mechanical measurements show that the ultimate tensile strength, the fatigue strength, and the elastic modulus of cortical bone are about ten times less than that of cold-worked stainless steel. Cancellous bone is spongy and much softer than cortical bone and exhibits mechanical behavior that is different from that of cortical bone, but has a higher healing rate (Ref 17). Bones derive their overall strength from their geometric configuration. The typical compressive strength of bone is the result of the composite structure of organic collagen and apatite crystals.

Combined Dynamic and Biochemical Attack. Dynamic loading in the presence of body liquids can cause wear and fretting corrosion at implant junctions, such as screw heads and plate holes, or at artificial articulation surfaces. Extensive research has been devoted to the study of wear on artificial joints and its degradation products. Corrosion fatigue is always a subject of argument and will be discussed in the section “Fatigue Properties of Implant Materials” in this article.

It is generally agreed that stress-corrosion cracking is not a concern in high-quality implant materials. It can be observed only in combination with other failure modes, such as fatigue.

References cited in this section

9. A.C. Guyten, *Textbook of Medical Physiology,* W.B. Saunders Co., Philadelphia, PA

10. J.B. Park, *Biomaterials: An Introduction,* Plenum Press, 1979

11. P.G. Laing, Biocompatibility of Biomaterials, in *Orthopaedic Clinics of North America,* Vol 4, C.M. Evarts, Ed., 1973

12. F. Pauwels, *Der Schenkelhasbruch,* Enke, Stuttgart, 1935

13. R. Schenk, Fracture Repair—Overview, in *9th European Symposium on Calcified Tissues,* 1973

14. R. Schenk and H. Willenegger, Zur Histologie der primären Knochenheilung, *Unfallheilkunde,* Vol 80, 1977, p 155

15. S.M. Perren and M. Allgöwer, *Nova Acta Leopoldina,* Vol 44 (No. 223), 1976, p 61

16. S. Perren, R. Ganz, and A. Rüter, Mechanical Induction of Bone Resorption, in *4th International Osteol. Symposium,* Prague, 1972

17. H. Yamada, *Strength of Biological Materials,* The Williams & Wilkins Co., Baltimore, MD, 1970

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Design of Internal Fixation Devices

Leading internal fixation systems contain different types of devices that are available in various configurations and sizes. Some major implant types are specified by standard-bearing organizations, such as ASTM and ISO, but many established implants are not standardized. Figure 2 shows a selection of characteristic internal fixation devices.

Straight bone plates (a to i, Fig. 2) often must be contoured to fit the curvature of the bone. This requires ductility and moderate rigidity of the plates. In many fracture situations, it is necessary to compress the fracture site with the aid of the bone plate to achieve stability and to promote bone healing. Therefore, certain plates have specially formed screw holes (compare d to g, Fig. 2) that allow compression of the fracture gap while the screws are inserted. Certain types of plates can be applied with compression devices to achieve the same effect. One-piece angled plates, two-component angled plate systems, and plates with specially formed heads (l to q, Fig. 2) are designed for the fixation of short fragments at the bone ends, specifically the femur.

Bone screws have two characteristic types of threads (j and k, Fig. 2), depending on intended use in cortical or cancellous bone. Intramedullary nails (r and s, Fig. 2) are inserted in the bone cavity. They are increasingly used in an interlocking fashion for comminuted fractures, in which the bone is splintered or crushed into numerous pieces. Pins and wires of various dimensions belong to the standard equipment of the orthopedic surgeon. A variety of implants exist, and some are for specific fracture types and corrective operations.

When implants are designed, the following factors must be considered:

- General anatomy and its individual deviation
- Operative approach
- Operative technique
- Available space at the fracture site
- Variability of the fracture situation
- Local bone consistency

- Healing tendency of the particular bone regions
- Probable complications
- Physiology and biomechanics
- Dynamic loading and weight-bearing conditions
- Response of the bone to the implant
- Properties of the implant material

Due to all the physiological restrictions, the ideal implant design, from an engineering point of view, is sometimes not feasible. In such cases, only the best possible compromise can be reached.

In addition to biological compatibility, the endurance of the implant is one of the basic requirements. Because of the above listed factors, implants cannot be designed with high mechanical safety factors for all possible loading conditions, because the volume and rigidity would exceed the biological limits. For example, a bone plate displaces soft tissue; in areas where the soft tissue coverage is thin, a plate with too large a cross section could cause soft tissue damage and complications. An implant designed only for high mechanical and fatigue resistance could become very rigid and would shield the bone from the physiological loading stresses. Consequently, rarefaction of the bone structure could occur. If the elasticity of the plate and the screws are not in the right proportion to each other or to the bone, screws could be pulled out of the bone or could break. On the other hand, if a plate is too flexible, nonunion could occur. If too large an area of the bone surface is covered with a plate, the blood supply may be impaired.

Usually, implants are so designed that they maintain their shape unless a dramatic event takes place, such as a second trauma. As discussed above, the fatigue strength of implants is not unlimited. Stresses in the elastic loading range are sufficient to initiate fatigue. As long as the well-reduced bone fragments are in contact, the implants are supported. Critical cyclic loading occurs only when the internal fixation is unstable in the presence of fracture gaps and bone defects. Specific surgical techniques were developed to achieve stable fixation (Ref 3). Figure 5 shows the different degrees of stability in internal fixations.

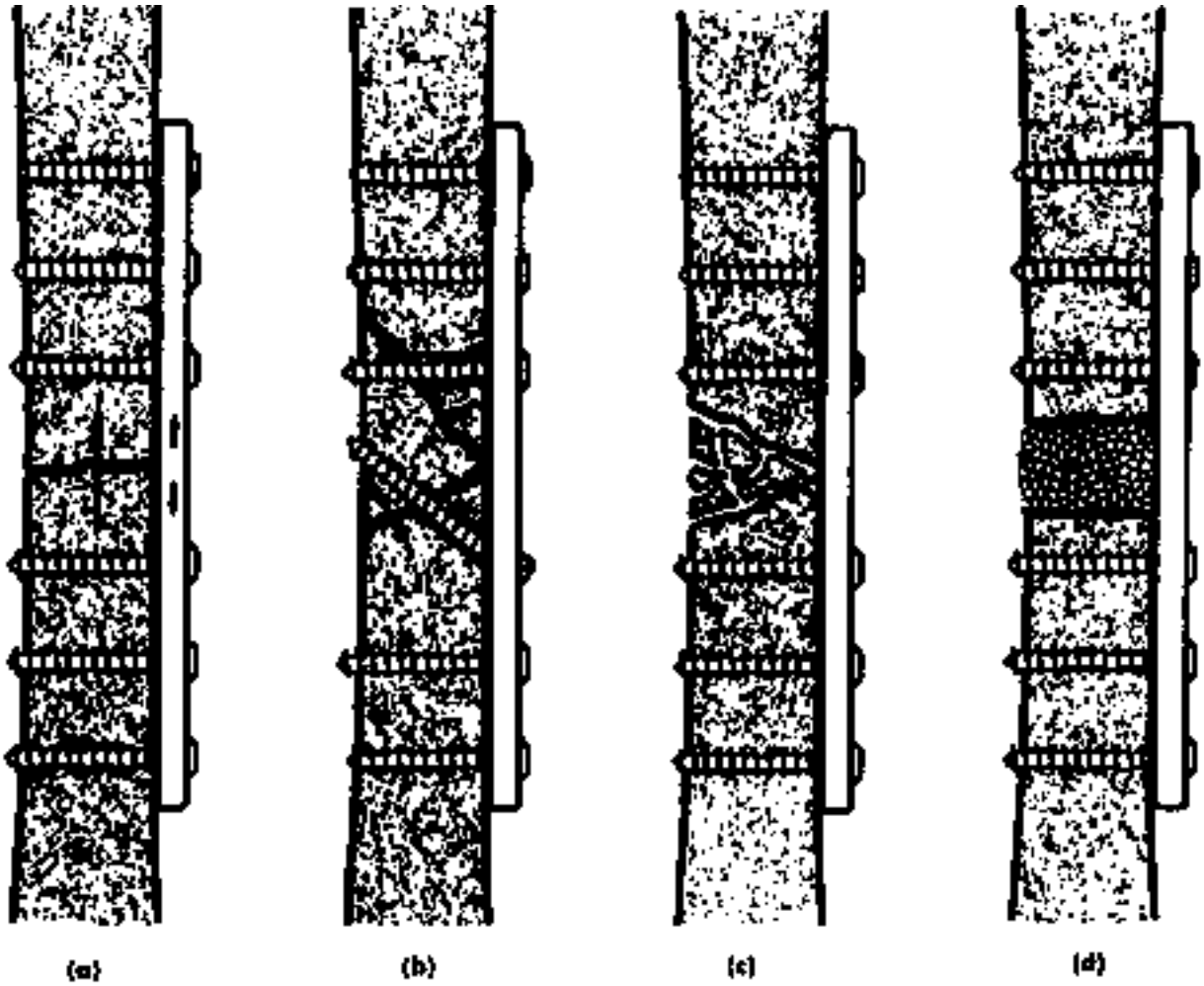


Fig. 5 Schematic of different degrees of stability in internal fixations. (a) Best possible stabilization with tension band fixation. The plate is on the tension side of the bone. The fracture site is compressed by means of the plate, which was applied under tension. If the

bone surface is straight, the plate is prebent at the center to exert compression and on the opposite cortex when the plate is put under tension. (b) Good stability. The two screws that stabilize the small third fragment are inserted in lag screw fashion and compress the fracture gaps. If no bone defects are present, the reduced bone and the plate support each other. See Ref 3 for techniques. (c) Unstable situation where plate has only adapting and bridging functions because the fragments are too small to be stabilized individually. Primary bone grafting may be indicated. This situation does not differ significantly from that shown in (d). (d) Plate has bridging function in leg-lengthening procedure. The gap between bone ends has been created intentionally for lengthening. Except for very young patients in whom a bone regenerate appears spontaneously, the gap is filled with bone grafts. Implant breakage has been rarely reported for this technique.

There are clinical situations in which comminution (reduction to particles) or loss of fragments does not allow stable fixation. Implantation of bone grafts is then indicated to promote bone healing and to develop stability.

Reference cited in this section

3. M.E. Müller, M. Allgöwer, R. Schneider, and H. Willenegger, *Manual of Internal Fixation*, 2nd ed., Springer Verlag, 1979

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Analysis of Failed Internal Fixation Devices

The metallurgical investigation of failed implants can reveal the failure mechanism. Because of the complexity of the biological conditions discussed earlier in this article, the cause of implant failure can in most cases be assessed only when the clinical history and the corresponding radiographs are analyzed by a clinical expert. This is particularly true in cases of fatigue failure. One should be cautious not to mistake small deviations from material standard specifications or small flaws as the reason for an implant failure when the real cause is heavy cyclic overloading. As experiments demonstrate, fatigue cracks can initiate on the implant surface without the presence of flaws. Judging the implant design as a possible cause of failure necessitates biological, biomechanical, and surgical knowledge.

Implants should be disinfected before examination. Although implants are usually rinsed in the hospital, biological residues can adhere to the surface and obscure the details of the fracture surface. Ultrasonic cleaning with 10% oxalic acid, a clinical disinfectant, or an enzymatic digestion is suitable for the highly corrosion-resistant implants. Probable corrosion products must be detected before such a procedure. However iron-rich blood residues should not be mistaken for rust if local microanalysis of the chemical composition is performed.

Under the microscope, electropolished implant surfaces usually exhibit small craters where inclusions were dissolved. Occasionally, these structures are erroneously interpreted as corrosion pits.

During the internal fixation and later at the retrieval, implants are handled with metallic tools, and this handling results in scratches and mechanical marks on the implant. It is often not possible to determine whether the marks were produced during retrieval or fixation.

The ASTM standard F 561 on implant retrieval and analysis may be taken as a guideline for the analysis of failed implants. Not all portions of the suggested investigation protocol are equally applicable for the different implants. An ISO standard on the same subject is in preparation.

Ortrun E.M. Pohler, Institut Straumann AG

Failures Related to Implant Deficiencies

Most of the failed implants submitted for analysis are manufactured from high-quality implant materials and are free of metallurgical defects. Implant breakage of current materials is usually due to mechanical or biomechanical reasons. However, there are occasionally implants received that are manufactured from inadequately processed materials or that were produced many years ago when metallurgical techniques were less advanced. Examples, , , and in this article document such cases.

Example 1: Severe General Corrosion on Historic Lane Plate Made From Chromium Steel. The four-hole Lane plate shown in Fig. 6(a) was inserted 46 years ago and remained in the body for 26 years. A large portion of the plate disintegrated and consisted mainly of corrosion products. Figure 6(b) shows the extensive corrosion holes on a metallographic section of the plate. The plate, manufactured from a chromium steel, exhibited transformation structures and carbides. The screws, which are formed like wood screws, were made from a soft austenitic 304 stainless steel and exhibited minimal surface corrosion. The corrosion products of the plate impregnate the surrounding tissues, as seen on the microprobe analysis shown in Fig. 6(c).

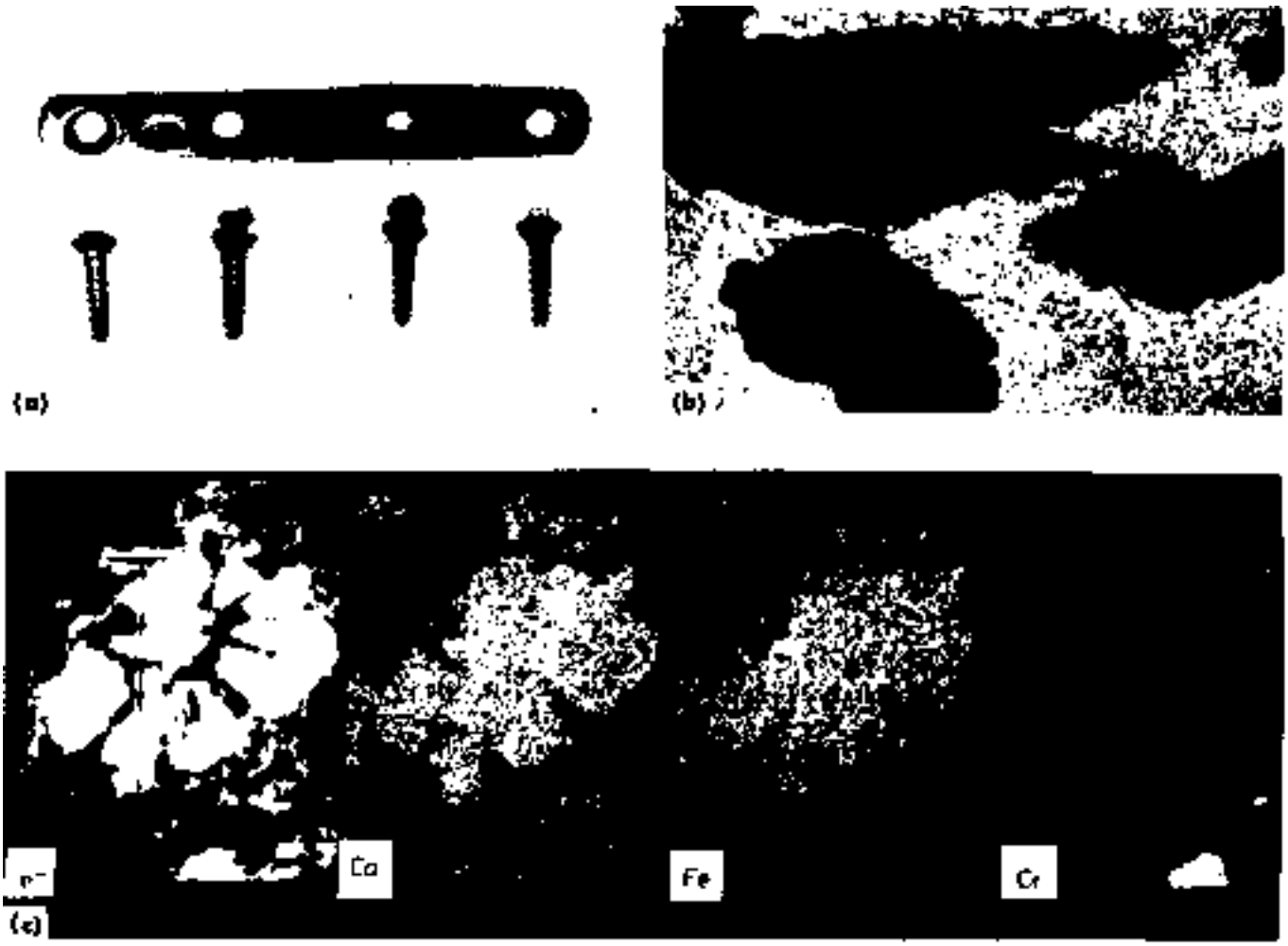


Fig. 6 Failed historic Lane plate. (a) Heavily corroded Lane plate from chromium steel. Implant was retrieved after 26 years. (b) Longitudinal section parallel to plate with large corrosion holes. 190×. (c) Microprobe analysis of tissue surrounding the plate. Chromium and iron corrosion products have impregnated the tissue. At the location shown, the question is whether the corrosion products penetrated bone structures or the impregnated tissue mineralized.

This is a case of general corrosion due to improper material selection. The corrosion may have been enhanced because of the austenitic stainless steel screws. In contrast, a recently retrieved intramedullary tibia nail of the type shown in Fig. 2(s), which had been in the body for 24 years, showed no signs of corrosion. This nail was produced from a low-carbon remelted type 316L stainless steel.

Example 2: Retrieved Bone Screw Made From Cobalt-Chromium-Molybdenum Alloy with Casting Defects. Portions of the threads of the screw shown in Fig. 7(a) had broken off and other threads had cracked. The screw was made from a cast cobalt-chromium-molybdenum alloy. A longitudinal section through the screw revealed gas porosity, segregation of primary inclusions, and oxides (Fig. 7b and c). The screw threads broke because of mechanical weakness from gas holes, brittleness, and dissolution of oxides. With today's advanced casting techniques and testing methods, such implant defects can be avoided and eliminated.

(a)

(b)

(c)

Fig. 7 Retrieved screw of cast cobalt-chromium.molybdenum alloy (type ASTM F75). (a) Defective screw threads from casting deficiencies. (b) Longitudinal section through threads showing porosity. 15×. (c) Enlarged thread of section shown in (b) with gas holes, segregation of primary phases, and dissolved oxides. 155×

Example 3: Broken Adjustable Moore Pins Made From Cobalt-Chromium Alloy. Two of four adjustable Moore pins, which had been used to stabilize a proximal femur fracture, were broken and deformed at their threads. The pins were made from a cobalt-chromium alloy and were not in the same condition. One pin contained brittle precipitates in the grains and grain boundaries. Correspondingly, the fracture occurred partially along the grain boundaries (Fig. 8a). Scanning electron microscopy (SEM) showed grain-boundary separation (Fig. 8b).

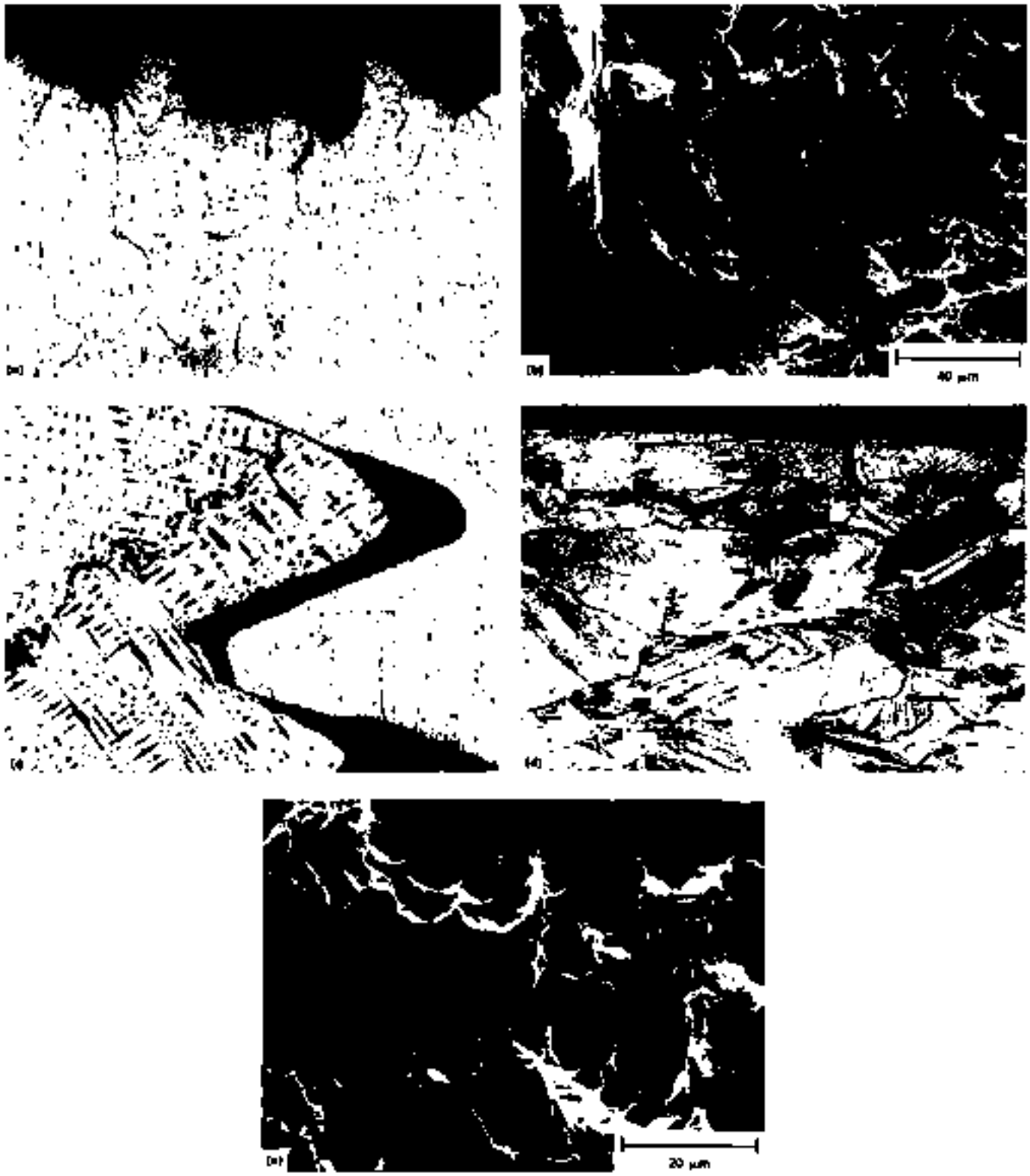


Fig. 8 Two broken Moore pins from cobalt-chromium alloy. (a) Longitudinal section through fracture surface showing grain-boundary precipitates and a partially intercrystalline fracture. 63×. (b) SEM fractograph indicating grain-boundary separation. Compare with (e). (c) Longitudinal section through screw and nut. The nut shows as-cast structures of cobalt-chromium-molybdenum alloy (type ASTM F75). 160×. (d) Longitudinal section of the other broken pin in the cold-worked condition with fewer grain-boundary precipitates, lines of primary inclusions, and a small surface crack (there were more cracks along the surface). 100×. (e) Corresponding fracture surface with dimples and fatigue striations. Compare with (b).

The nut used with the pin was made from a cast cobalt-chromium-molybdenum alloy and had segregations in the microstructure and gas holes and precipitates in the grain boundaries (Fig. 8c). The nut was not specifically loaded and did not fail.

The other broken pin was made from a cold-worked cobalt-chromium alloy. The grain boundaries contained few precipitates, but lines of primary inclusions were randomly distributed (Fig. 8d). The fracture surface exhibited a mixed fracture mode with intermingled dimples and fatigue striations (Fig. 8e). The pins did not exhibit any corrosion.

This example demonstrates how the different conditions of cobalt-chromium alloys affected failure behavior. In one of the Moore pins, brittle fracture was aided by grain-boundary separation.

Example 4: Intercrystalline Corrosion on Cerclage Wire of Sensitized 304 Type Stainless Steel. The wire was used with two screws and washers for a tension band fixation (Ref 3) in a corrective internal fixation (Fig. 9). The cerclage wire was found broken at several points and corroded when it was removed after 9 months (Fig. 10a). Metallographic analysis revealed that the wire that was made of type 304 stainless steel without molybdenum, as seen in the energy-dispersive x-ray analysis (Fig. 10b); this material does not comply with the standards. The screws and washers were intact and were made of remelted implant-quality type 316L stainless steel corresponding to the standards. The microstructure of the wire, which was in the soft condition as required, showed signs of sensitization, with chromium carbide precipitates at the grain boundaries (Fig. 10c). Typical intercrystalline corrosion with pitted grains was evident through SEM fractography (Fig. 10d).

Fig. 9 Radiograph showing a tension band fixation containing a cerclage wire, two screws, and washers beneath the screw heads. See also Fig. 10.

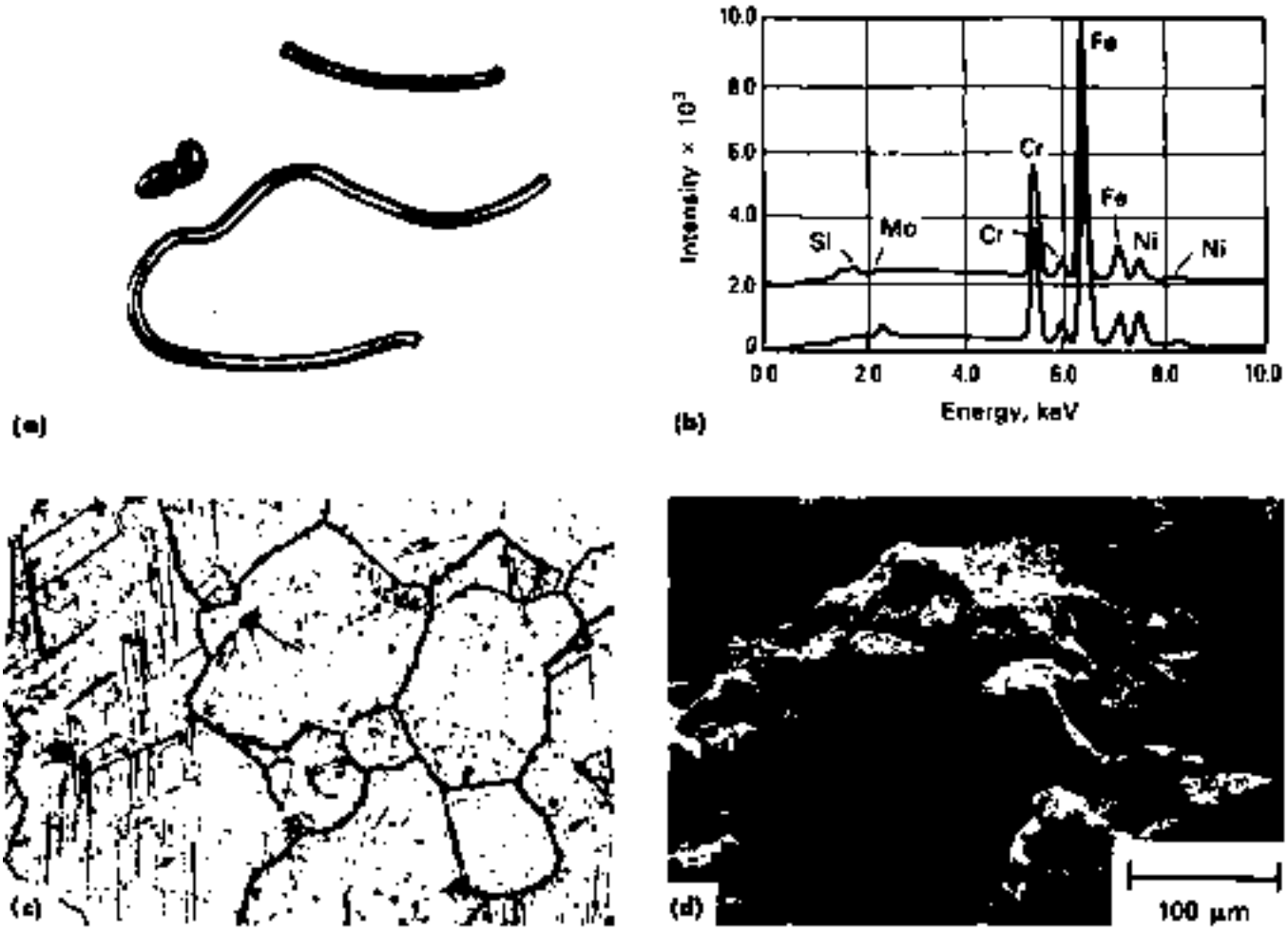


Fig. 10 Intercrystalline corrosion of a type 304 stainless steel cerclage wire. (a) Broken cerclage wire. (b) Energy-dispersive x-ray analysis under the scanning electron microscope. Lower spectrum: reference material type 316L screw material. Upper spectrum: type 304 steel with less nickel content and no molybdenum from investigated implant wire. (c) Cross section of sensitized wire, with grain boundaries and deformation lines heavily attacked by etching because of chromium carbide precipitates. 180×. (d) Fracture surface under scanning electron microscope indicating intercrystalline corrosion with pits on grain surfaces

For this internal fixation, orthopedic wire of an insufficient stainless steel type was used. Improper heat treatment of the steel lead to intercrystalline corrosion and implant separation. The wire came from a source different from that of the screws and washers and did not meet current standards for orthopedic wire.

Reference cited in this section

3. M.E. Müller, M. Allgöwer, R. Schneider, and H. Willenegger, *Manual of Internal Fixation*, 2nd ed., Springer Verlag, 1979

Ortrun E.M. Pohler, Institut Straumann AG

Implant Failures Related to Mechanical or Biomechanical Conditions

Bone screws fracture in two modes—by mechanical forces during insertion or removal or by fatigue failure. Examples for both are given below.

Mechanically forced shearing fractures are a typical failure in screws. Although stainless steel bone screws are normally manufactured from cold-worked material, their shearing ductility is usually considerable. Ductile shearing fractures are characterized by a spiral texture of the fracture surface, flow lines at the fracture edges, overload dimple structure on the fracture surface, and heavily deformed microstructure at the fracture edge in longitudinal sections. Occasionally, screw heads can be sheared off by repeated tightening, particularly when the cortex (dense outer bone wall) is thick, because the thicker the cortex, the higher the retention of the screw. Excessive oblique positioning of the screw in a plate hole can also cause fracture if the screw head is bent by the plate hole during insertion.

Example 5: Shearing Fracture of a Type 316LR Stainless Steel Screw. The cortical bone screw shown in Fig. 11 broke during the internal fixation procedure. It may have been inserted very obliquely in the screw hole and therefore sheared off. The fracture surface (Fig. 11a) exhibited extensive spiral deformation. Figure 11(b) shows the dimple structure characteristic of a ductile failure mode. The dimples are oriented uniformly in the deformation direction. Figure 11(c) shows a magnification of the flow lines at the fracture edge. The longitudinal metallographic section perpendicular through the fracture surface (Fig. 11d) exhibited a zone of heavily deformed grains at the fracture edge. The microstructure and the hardness of the screw corresponded to the standards.

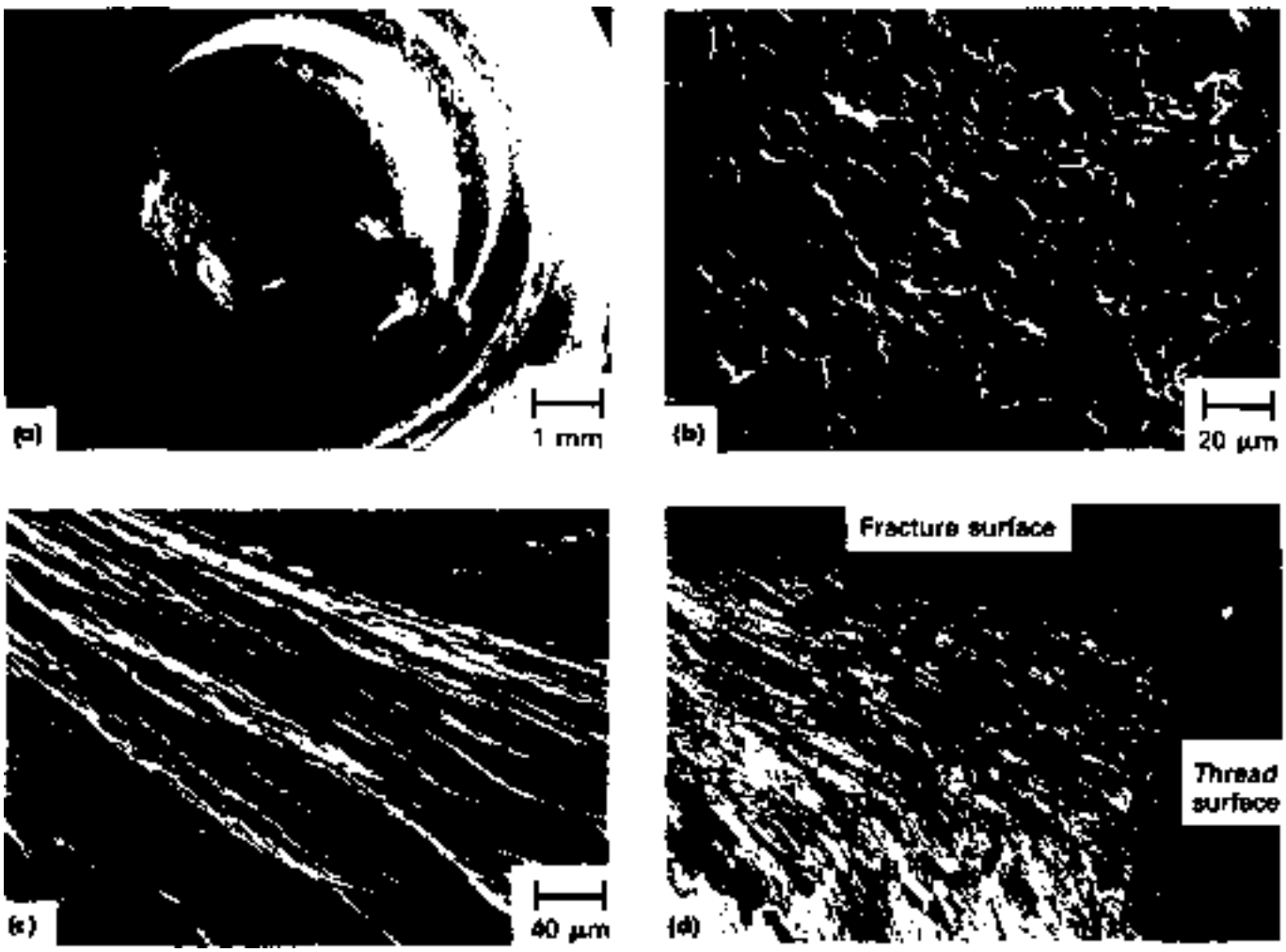


Fig. 11 Mechanically forced shearing fracture of type 316LR stainless steel screw. (a) Fracture surface with typical spiral deformation texture. SEM. (b) Close-up of fracture

surface with shear dimples oriented in twisting direction. (c) Fracture edge with flow lines. (d) Longitudinal metallographic section through fracture surface. Deformation zone from shearing is adjacent to the fracture edge. Original deformation structure of the screw is visible at the bottom of the micrograph. 62×

For comparison, the shearing fractures of a commercially pure titanium screw and a cast cobalt-chromium-molybdenum alloy (ASTM F75) screw are shown in Figures 12 and 13. The spiral-textured fracture surface of the titanium screw was smooth (Fig. 12a) and had uniform, shallow shearing tongues and dimples indicative of a ductile fracture mode. (Fig. 12b).

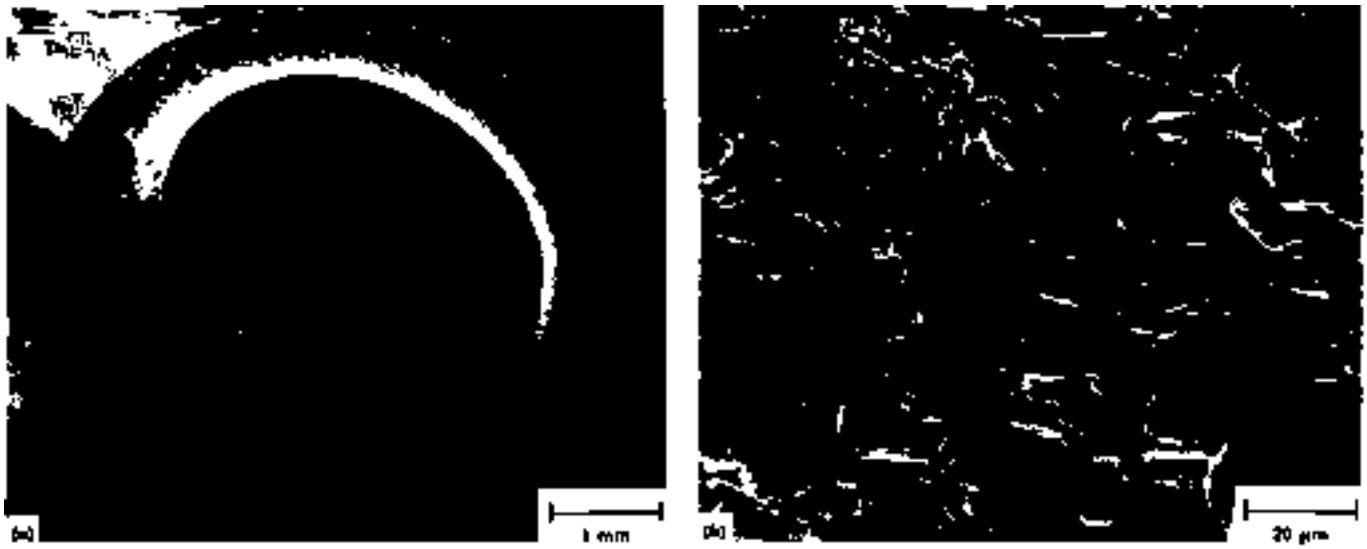


Fig. 12 Shear fracture of a commercially pure titanium screw. (a) SEM fractograph showing spiral textured fracture surface of sheared-off screw. Typical deformation lines are fanning out on the thread. (b) Uniformly distributed shearing tongues and dimples

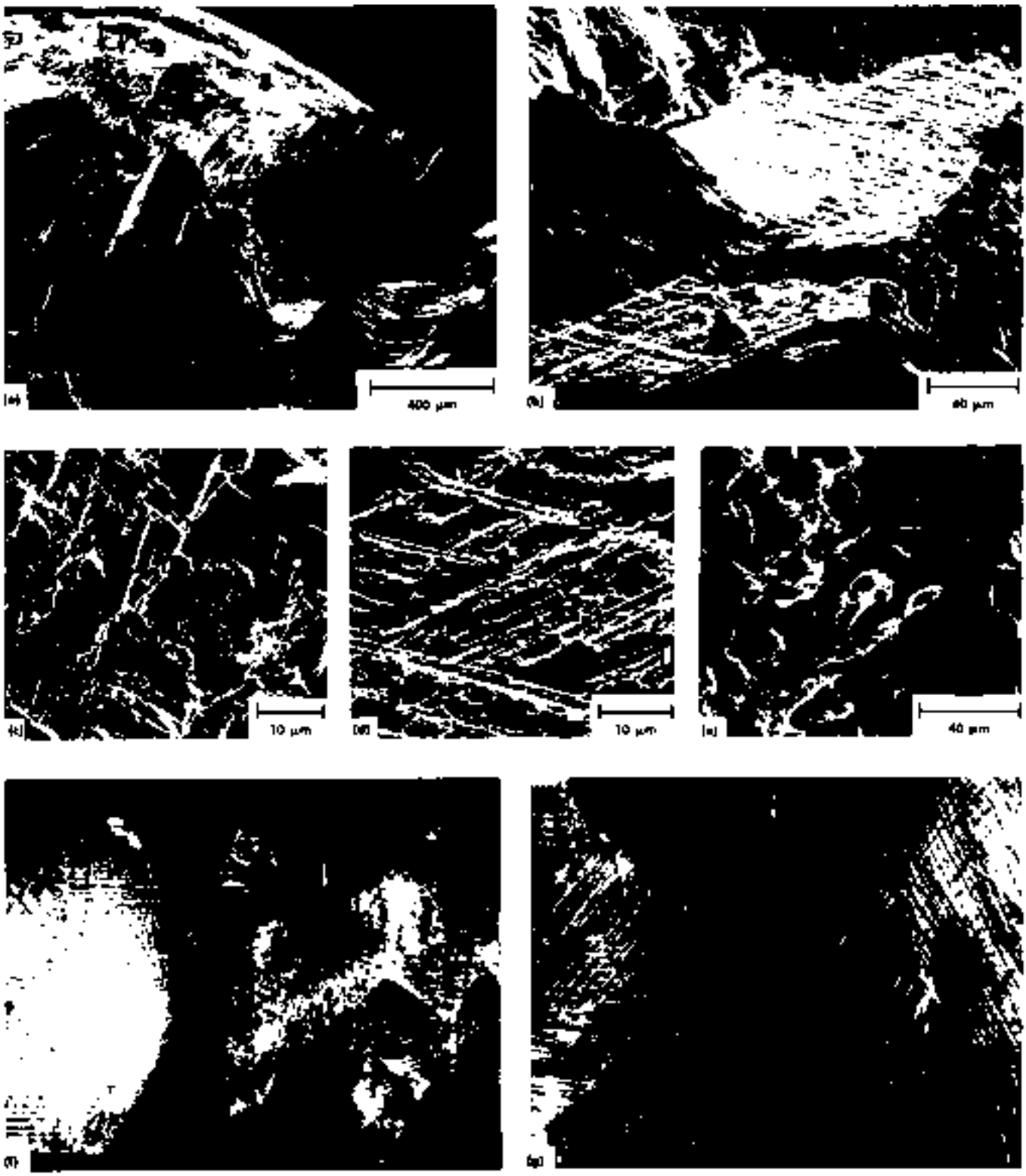


Fig. 13 Sheared-off cast cobalt-chromium-molybdenum screw. SEM fractography. (a) Overview of portion of rough fracture surface. (b) Area with fracture planes of three differently oriented grains (single arrow, a). (c) Shearing structures and dimples in grain identified by the numeral 1 in (b). (d) Slip traces in grain identified by the numeral 2 in (b). (e) Gas holes with dendritic freezing surfaces (double arrows, a). (f) Longitudinal section through identical screw with starting shearing damage. 10×. (g) Shearing crack (arrow, f). 130×

In contrast, the fracture surface of the cast cobalt-chromium-molybdenum alloy screw had an erratic structure (Fig. 13a), which can be explained by the large grain size and the tendency of the grains to fracture along crystallographic planes.

Figure 13(b) shows such fracture planes of differently oriented grains. These planes exhibited crystallographically oriented mixed shearing structures and dimples (Fig. 13c), as well as traces of slip (Fig. 13d). Gas holes with dendritic freezing surfaces are shown in Fig. 13(e). In the longitudinal section of a similar screw (Fig. 13f), large grains can be observed, along with deformation between the threads, which leads to shear rupture. Figure 13(g) shows a starting shear crack propagating along slip planes.

Fatigue Failures of Screws. As discussed in the following example, fatigue failure at varying thread levels is also a problem sometimes encountered. One of the causes of such failures is unstable fixation.

Example 6: Characteristic Observations on Type 316LR Stainless Steel Screws That Failed by Fatigue. Fatigue fracture can occur on different thread levels, depending on the loading situation (Fig. 14a). No deformation zone was created at the fracture site shown in the longitudinal metallographic section in Fig. 14(b). Occasionally, the initiation of secondary fatigue cracks is found parallel to the fracture plane or in a nearby thread groove. Figure 14(c) shows one of the fracture surfaces with fatigue striations.

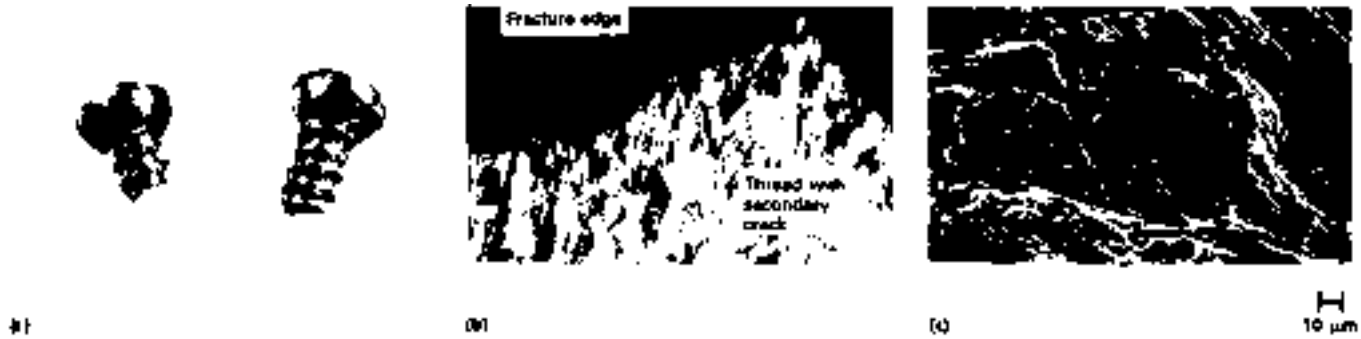


Fig. 14 Type 316LR stainless steel screws that failed by fatigue. (a) Fatigue fractures at different thread levels. (b) Longitudinal section perpendicular to fracture surface without deformation zone at fracture site. A small secondary crack is shown at thread site (arrow). 55×. (c) Fatigue striations on fracture surface

Corresponding to the energy-dispersive x-ray analyses, the screws (Fig. 14a) were manufactured from cold-worked remelted (R) type 316L stainless steel and showed no structural or manufacturing defects. The material meets ISO and ASTM F 318 Special Quality specifications. The screws were used with a relatively rigid plate to treat a fracture complication in the upper end of the femur. Radiographs indicated various signs of unstable fixation, which explains the fatigue failures. The two screws were situated at the distal end of a compression screw plate, where the proximal lever arm was the longest compared to the other screws.

Fatigue Damage of Bone Plates. Most broken plates have been loaded under unilateral bending, because bone defects in the opposite cortex are a typical cause of instability and breakage. Thus, tensile and shearing stresses are created in the top surface of the plate. Correspondingly, this is the surface where crack initiation is usually found. Depending on the fracture situation and the location, torsional stresses can also be created. This is especially true in forearm fractures because of the rotational motion that takes place. Examples, , and show three stages of fatigue damage on bone plates.

Example 7: Fatigue Initiation on Type 316LR Straight Bone Plate. The plate shown in Fig. 15(a) and (b) was used to treat a pseudarthrosis in the proximal femur. Because healing did not progress, the plate was removed and submitted for investigation. The bone was replated with an angled blade plate, and compression was exerted on the pseudarthrosis to promote bone healing.

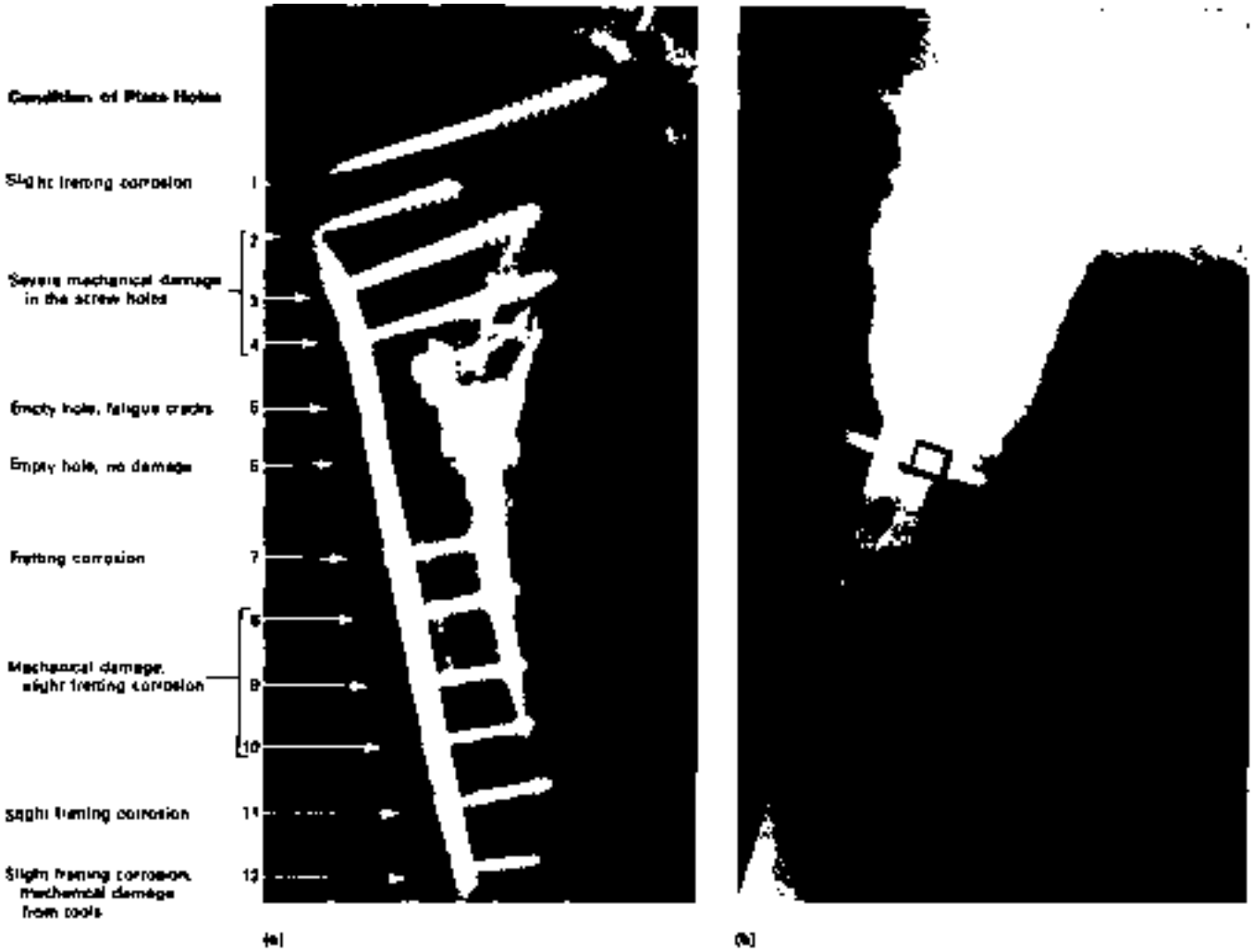


Fig. 15 Crack initiation on type 316LR stainless steel dynamic compression plate. (a) Anterior-posterior radiograph. The plate was used to treat the nonunion of a fracture between the fourth and seventh screws. The plate was bent intraoperatively to fit the contour of the bone. (b) Radiograph with lateral view showing wide bone gap, indicating instability. The boxed area indicates the location of initiating fatigue cracks.

Investigation. The plate seemed to be macroscopically intact. At higher magnification, fatigue cracks were apparent on the top surface of the small section of the plate (Fig. 16a and b) at the fifth screw hole indicated on Fig. 15(a). In the background of the surface, slip systems are visible, partially in conjunction with initial cracks.

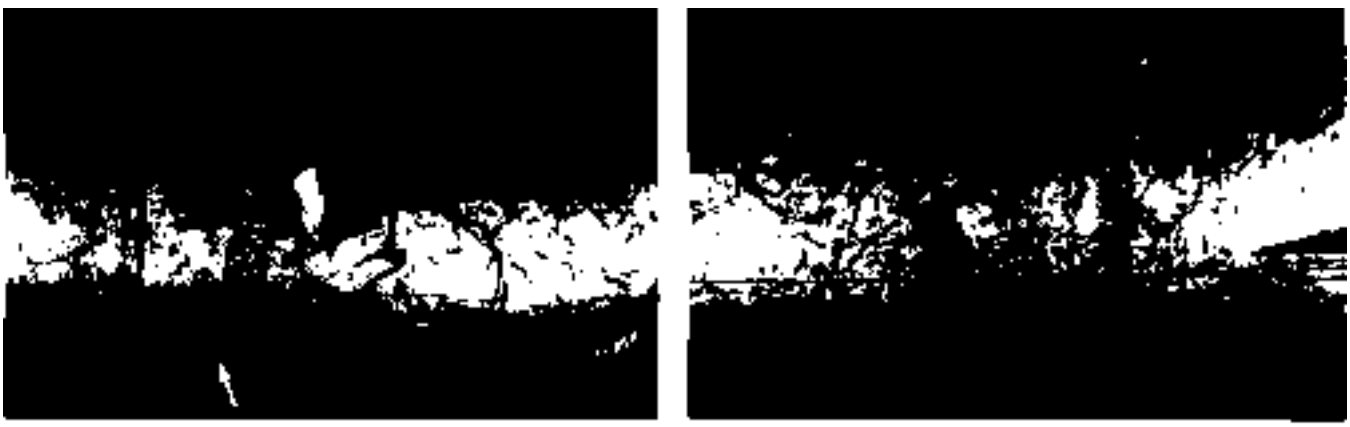


Fig. 16 Two views of the plate surface with fatigue cracks at the fifth hole. White arrow in the left-hand figure indicates tool mark from bending that does not interfere with fatigue damage structure. See text for further description. Both 110×

From crack-initiation studies, it is known that persistant slip systems appear first on the surface that is loaded under critical cyclic stresses. Precracks and initial cracks develop from these slip systems. Some of the initiation cracks propagate further and become major cracks that penetrate the cross section. The extent of this fatigue surface damage is representative for the stress intensity.

Figure 16 shows structures as were described above for the fatigue initiation process. The intensity of these structures, compared to results of systematic crack-initiation experiments, suggested that the plate was rather heavily loaded. Morphological comparison and fatigue initiation curves indicated that a local stress concentration of 500 to 600 MPa (72 to 87 ksi) was present. This load would still be below the elastic limit of the cold-worked stainless steel. No fatigue cracks were visible on the broad section of the asymmetrically placed hole. Thus, fatigue damage occurred earlier in the portion with the smaller cross section. In addition, rotational forces are to be expected in this area.

Fatigue-bending tests on such broad plates showed that the fatigue life of plates with asymmetrically arranged holes is at least as long as for plates with holes situated in the center. At a centered screw hole, fatigue initiates at about the same time on both sides of the hole. At the asymmetrically placed holes, fatigue begins at the large section only after a fatigue crack begins to propagate into the small plate section.

Discussion. Although fatigue cracks are observed more frequently at the edges of the plate holes, crack initiation often takes place on the total surface area adjacent to the middle section of the plate hole. This is demonstrated on another broken plate (Fig. 17), in which a large secondary crack had developed parallel to the main crack in the center of the surface. Another small crack was visible on the edge of the plate hole. A large portion of the plate surface at the fracture edge showed deformation from fatigue damage. Crack initiation does not take place in the plate holes and is not related to the fretting that may occur between the screw head and plate. An exception may be some special plate hole designs for which corrosion fatigue was reported (Ref 18).

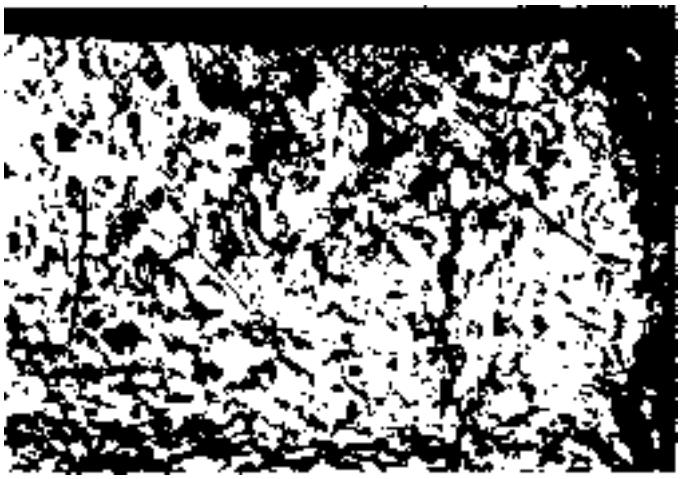


Fig. 17 Top surface of broken plate of type 316LR stainless steel. Fatigue cracks parallel to the fracture edge and a wide area exhibiting primary fatigue deformation are visible. 65×

In Fig. 16, no crack initiation is visible at the pits caused by electropolishing. In addition, mechanical marks from the plate bending did not act as stress raisers. The stainless steel has no particular notch sensitivity, and the fatigue behavior is forgiving with respect to mechanical damage.

In the radiographs shown in Fig. 15(a) and (b), extensive bone deficiency is present in the area where two plate holes were not filled with screws. In Fig. 15(b), a wide gap is visible. The empty screw hole (No. 6 from the top) situated directly over the defect does not show fatigue damage, but hole No. 5, which sits at the edge of the proximal bone fragment, has the described fatigue cracks. Because screw hole No. 5 was situated at the transition between the supporting bone and the defect, stress concentration was higher here than at the empty hole No. 6, where the elastic deformation could occur more uniformly. In the defect zone of the original fracture, between screws No. 4 and 7, the plate is not supported by the bone, and the load is transmitted only through the plate, which caused cyclic loading of the implant and fatigue-crack initiation.

It is also characteristic that fretting corrosion occurred at the screw/plate interface at screw hole No. 7. This screw was closest to the bone defect and was secured only in one cortex. Compared to all the other screws, this one was the most likely to undergo relative motion. As outlined in the legend in Fig. 15(a), at all other contact areas between screw head and plate hole either minimal or no fretting, or fretting corrosion and mechanical damage, was found. For this analysis, only the two radiographs shown in Fig. 15 were available. They are not adequate to estimate the clinical postoperative history, but they are sufficient to explain the fatigue-crack initiation due to instability.

Example 8: Fatigue Crack on a Type 316 Stainless Steel Bone Plate and Corresponding Broken Screw. A narrow bone plate was used to stabilize an open midshaft femur fracture in an 18 year old patient.

Investigation. As seen in Fig. 18(a), a radiograph taken 13 weeks after the operation, the plate had a crack at a plate hole next to the fracture site. The plate was slightly bent in the horizontal plane, and the fracture gap was considerably open. In the lateral view, a broken screw is visible at the second plate hole above the fracture gap (Fig. 18b).

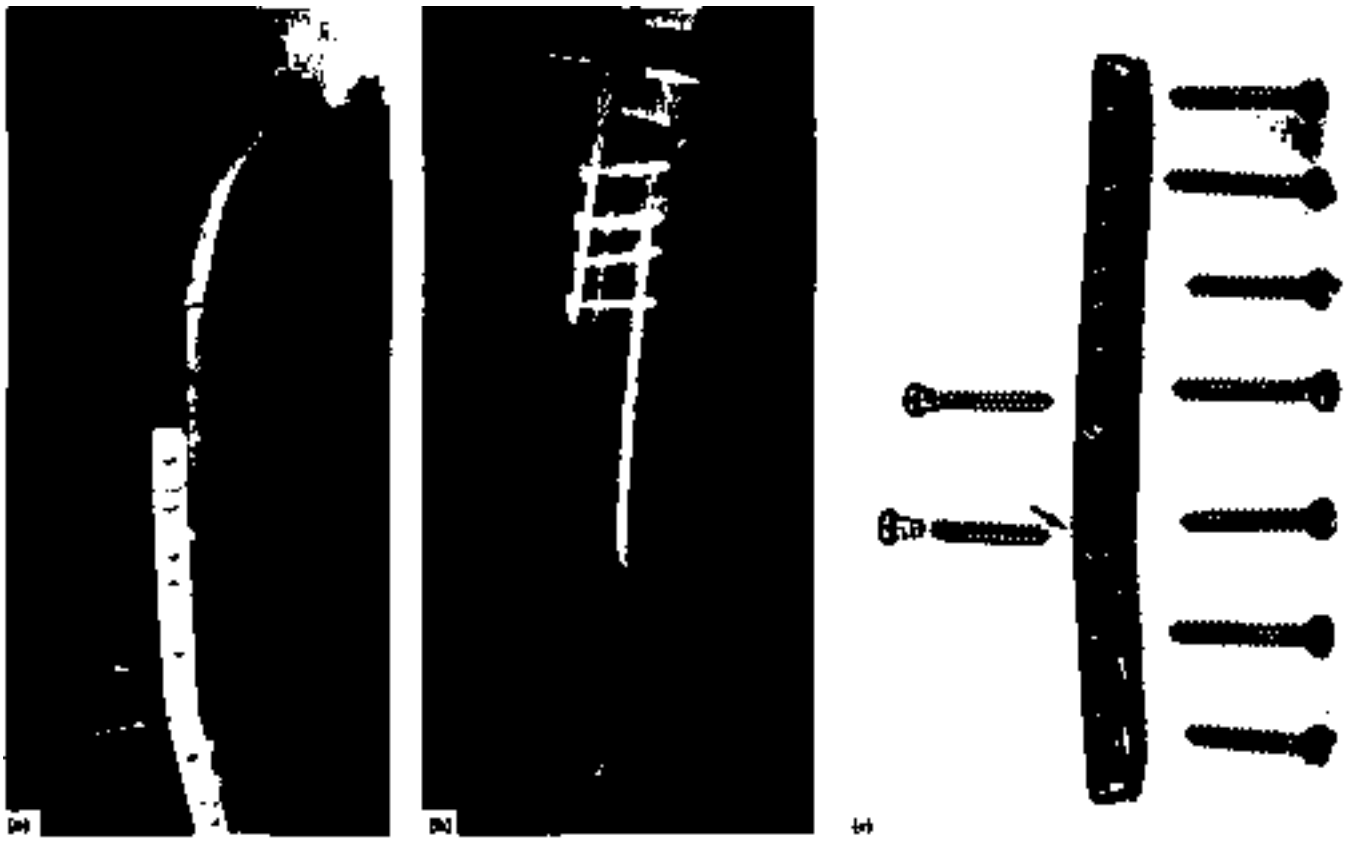


Fig. 18 Stainless steel bone plate with fatigue crack and broken screw. (a) Radiograph taken 13 weeks after operation. Anterior-posterior view. Arrows indicate crack in plate and open fracture gap. (b) Corresponding lateral view. Arrow indicates broken screw. (c) Bend in plate in the horizontal plane at the unfractured side of a screw hole (arrow). Deformation was due to dorso-lateral acting forces. Screws are not placed at original position. See also Fig. 19.

The screws and plate shown in Fig. 18(c) were supplied by three different manufacturers. The compression holes of the plate are a modification of those of other implant systems. Energy-dispersive x-ray analyses indicated that all implants were made of type 316 stainless steel. However, the microcleanliness of the materials was different. The minimal primary inclusion content found in the plate and in one of the screws was typical of high-quality remelted steel that meets the requirements in ASTM standard F 138. The remaining screws from a different source, including the broken screw, had a high primary inclusion content and did not comply with this standard.

Figure 19(a) shows a longitudinal section perpendicular to the fracture surface of the broken screw. The local crack formation appeared to have been influenced by the presence of larger inclusions. From the striations found on the fracture surface under the scanning electron microscope, it was concluded that the screw failed through a fatigue mechanism. In addition to the fatigue striations, many dimples were present (Fig. 19b), which were created primarily through inclusions. In areas with final overload structures, the dimple formation was also influenced by inclusions (Fig. 19c).

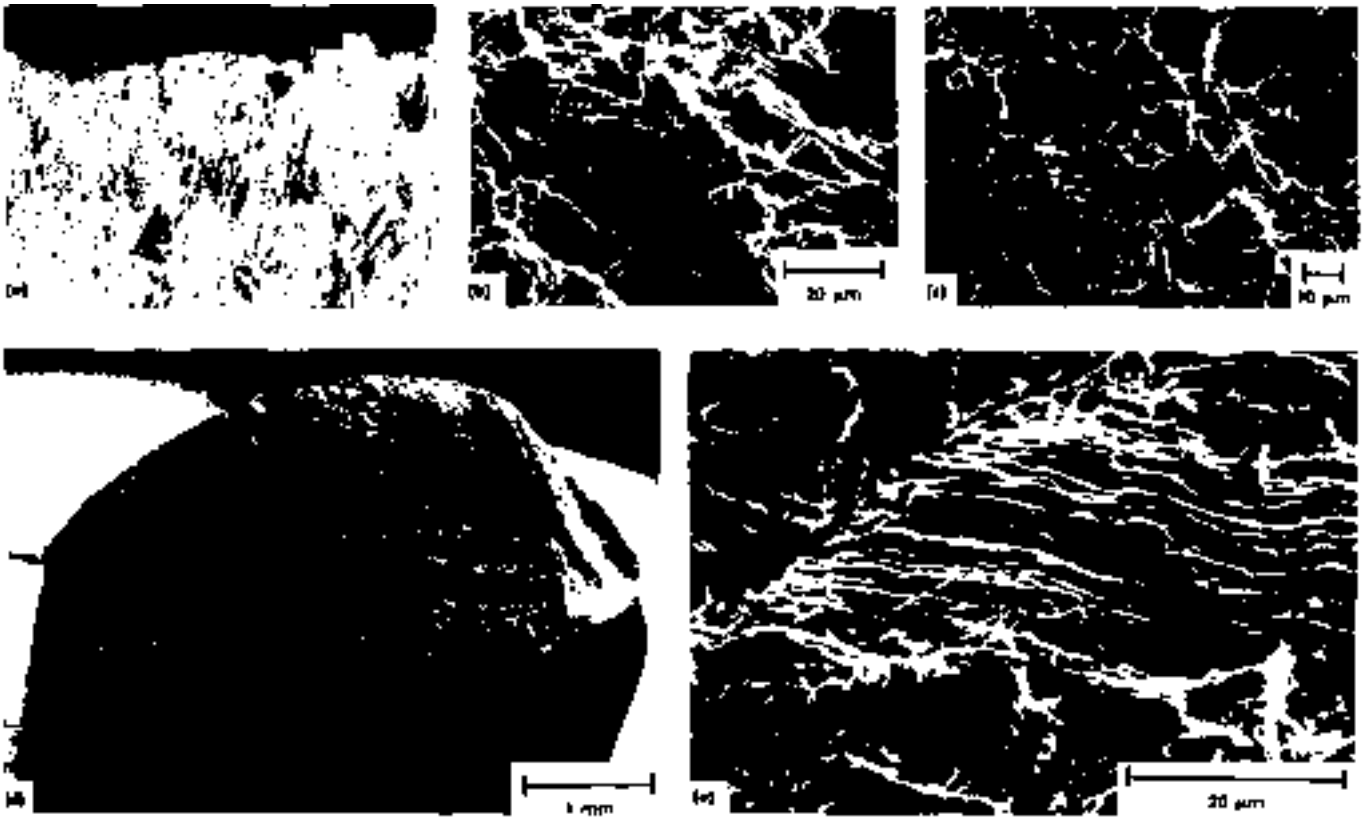


Fig. 19 Fracture surfaces of the failed screw and bone plate shown in Fig. 18. (a) Longitudinal section through fractured screw showing edge of fracture surface and high inclusion content. A large slag inclusion was present at the void under the fracture edge. 55×. (b) Fracture surface of screw showing fatigue striations (containing secondary cracks) and dimples. (c) Fracture surface of screw and zone showing overload fracture containing large dimples partly caused by inclusions. (d) Plate fracture surface. Arrow points to crack origin, which is indicated by beach marks. (e) Fatigue striations from center of plate fracture surface (running originally transverse to loading direction)

The crack in the plate originated at the upper, outer corner of the plate, as seen from the beach marks in Fig. 19(d). This indicates the action of asymmetric bending and rotational forces. The beach marks were oxidized, and the fracture surface was covered with fatigue striations (Fig. 19e).

Discussion. Not all clinical details of this case are known, but from a biomechanical viewpoint based on the radiographs (Fig. 18a and b) the fatigue failures of the plate and screw can be attributed to the unusual placement of the narrow bone plate. Because of the open fracture situation with severe soft tissue damage, the narrow plate was placed on the frontal aspect of the femur, which is not a typical configuration. A broad plate is usually applied at the lateral side of the femur if plate fixation is indicated. Physiologically, the lateral side of the femur is under tension, and if the plate is mounted laterally, exerting compression on the fracture site creates a tension band effect, which provides excellent stability (see Fig. 5a and Ref 3).

In this case, the plate was not on the tension side, and obviously, no compression (or not enough compression) had been applied. Because of the anatomic conditions and the fracture configuration, the plate was loaded under asymmetric bending and torsional forces, which created maximum stresses at the lateral edge of the plate hole near the fracture site. This is confirmed by the location of the fatigue-crack origin shown on the fracture surface in Fig. 19(d).

The instability of the internal fixation was indicated by the callus formation, by the rounded edges of the fracture ends, and by some screw loosening. The first screw over the fracture gap was definitely loose, as deduced from the resorption hole in the cortex beneath the plate. This caused additional loading on the next screw, which failed by fatigue. Because of the different sources of the implants, the head design of some of the screws was not compatible with the screw hole geometry of the plate. This did not cause the instability, but may have hastened its progress. Similarly, the inclusion content of the screw did not cause the fatigue failure, but may have reduced the fatigue strength to some extent.

This example shows that a smaller plate and an unusual positioning had to be chosen because of the critical soft tissue condition, which leads to decreasing stability and eventual fatigue failure. It is important to realize that at the time of

injury the soft tissues have priority because of the risk of serious complications. After plate removal, the fracture was stabilized with an intramedullary nail, and bone healing proceeded.

Example 9: Fatigue Fracture of a Type 316L Stainless Steel Angled Plate. Figure 20(a) shows a broken angled plate that was applied to the proximal femur. The fatigue fracture occurred at a plate hole. The fatigue damage at the fracture edge at the top surface of the plate indicated that symmetric cyclic bending forces had acted. Figure 20(b) shows slip bands produced on the surface during cyclic loading. Figure 20(c) reveals fatigue striations.

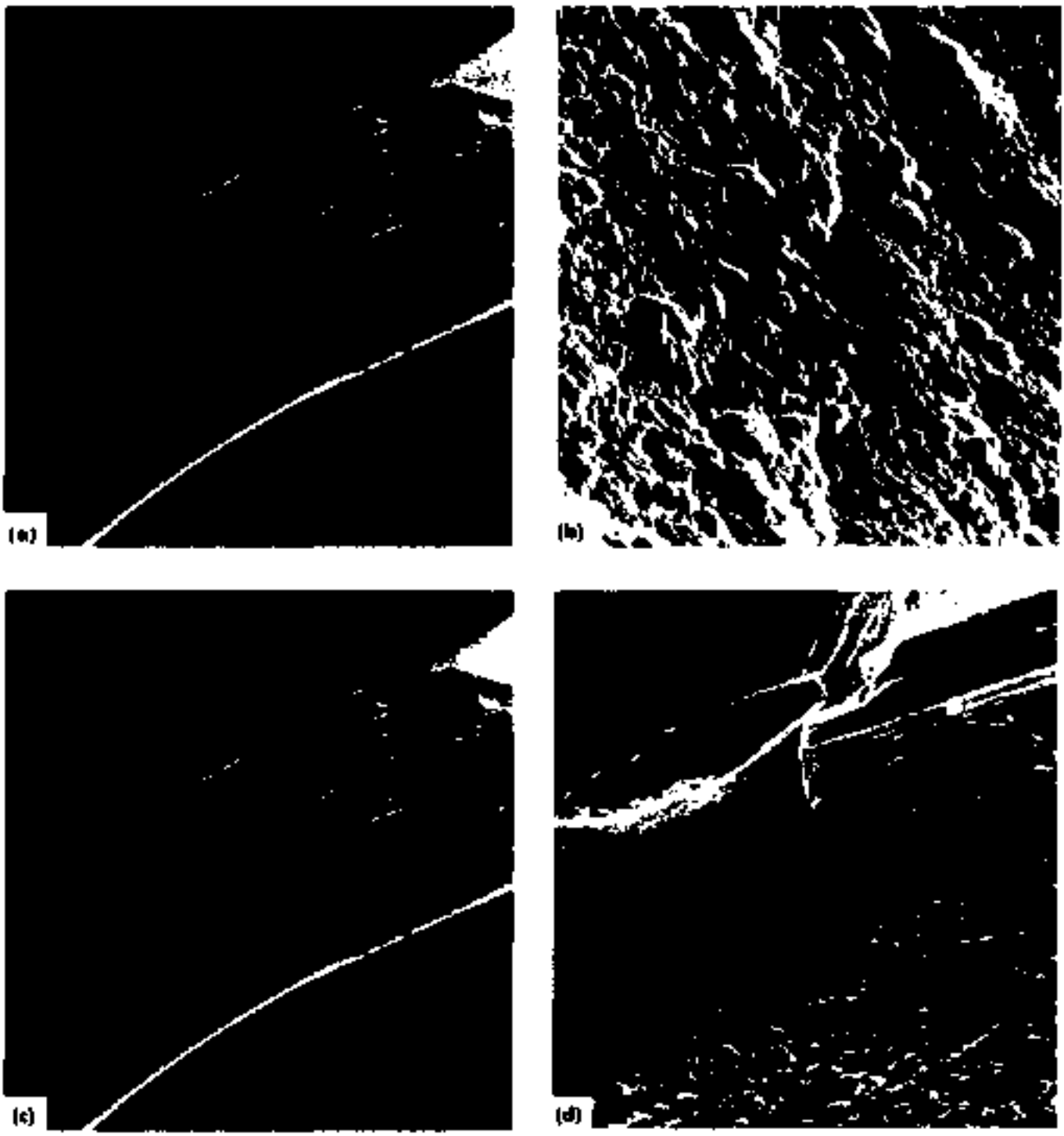


Fig. 20 Angled blade plate that failed by fatigue. (a) Fatigue fracture at plate hole. (b) Top plate surface at fracture edge showing intense fatigue-surface damage (slip bands), indicating that the plate was heavily loaded. (c) SEM of fracture surface. Fatigue striations can be seen in the grain at the lower right-hand corner. Other fracture features are crystallographically oriented. (d) Microstructure of cross section reveals that plate was made from cold-worked stainless steel of high microcleanliness. 80×

The microstructure of the plate had a high degree of cleanliness (Fig. 20d). The deformation structure showed that the material was in the cold-worked condition, which was confirmed by Vickers hardness tests. Energy-dispersive x-ray analysis indicated a chemical composition corresponding to a type 316 stainless steel (the carbon content was not determined).

The plate material corresponded to the current standards for surgical implants and was not the cause of the implant failure. Because no radiographs or other clinical details were available, it could not be evaluated if the plate size was adequate for the clinical conditions.

References cited in this section

3. M.E. Müller, M. Allgöwer, R. Schneider, and H. Willenegger, *Manual of Internal Fixation*, 2nd ed., Springer Verlag, 1979

18. L.E. Sloter and H.R. Piehler, Corrosion-Fatigue Performance of Stainless Steel Hip Nails—Jewett Type, in *STP 684*, ASTM, Philadelphia, 1979, p 173

Ortrun E.M. Pohler, Institut Straumann AG

Fatigue Properties of Implant Materials

Remelted Type 316LR Stainless Steels. Because metal fatigue accounts for most implant fractures, some fatigue properties of type 316LR stainless steel, based on experimental studies (Ref 19, 20), will be addressed.

The *S-N* fatigue curves for the steel in recrystallized and cold-worked conditions are shown in Fig. 21. As shown in Fig. 21(a), the fully reversed bending tests were performed on flat specimens in air and in Ringer's solution (a lactated 0.9% sodium chloride solution that contains about the same concentration of Cl^- ions as body fluids). The improved fatigue strength in the cold-worked condition is of particular interest for implants. Plates, screws, pins, and certain nails are produced mainly from cold-worked material. The fatigue limits determined in rotating bending beam tests on the same materials are comparably higher than those found in the bending tests (Ref 19).

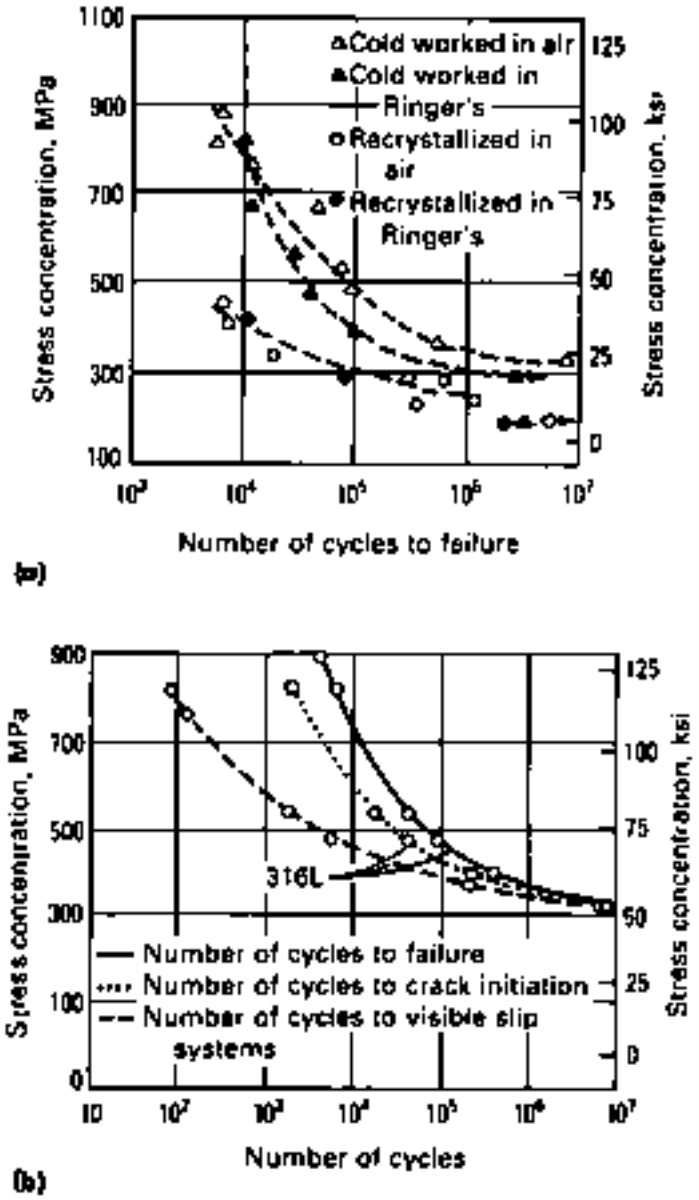


Fig. 21 Fatigue curves of type 316LR stainless steel implant material tested in bending mode. (a) *S-N* curves for stainless steel in cold-worked and soft condition that was tested in air and aerated lactated Ringer's solution. (b) Fatigue curve for number of cycles to failure as shown in Fig. 21(a) for cold-worked stainless steel that was fatigued in air. The two additional curves show the number of cycles to initiation of visible slip systems and the number of cycles to crack initiation. Source: Ref 20

In the recrystallized condition, the fatigue strength of the steel is not significantly affected by Ringer's solution. In the cold-worked condition, fatigue life in the high-cycle range is reduced in Ringer's solution. Corrosion fatigue in the usual sense, however, does not occur, because, morphologically, no signs of corrosion are detectable by optical microscopy and SEM. Ringer's solution appears to accelerate the individual fatigue stages. This complies with observations of fracture surfaces of broken plates that do not show a corrosive component combined with their fatigue structures and which do not differ from structures produced experimentally in air.

Additions of 0.25% clottable fibrinogene to the Ringer's solution have been shown to neutralize its fatigue-life reducing effect. The corresponding *S-N* fatigue curves were similar to those generated in air (Ref 21).

On retrieved implants, however, it has been found that fatigue-fracture surfaces can secondarily corrode when crevice or fretting conditions develop in the fracture gap. Such corrosion structures can be locally observed. In the transition areas, fatigue striations with superimposed corrosion pits can be seen (Fig. 22a and b).

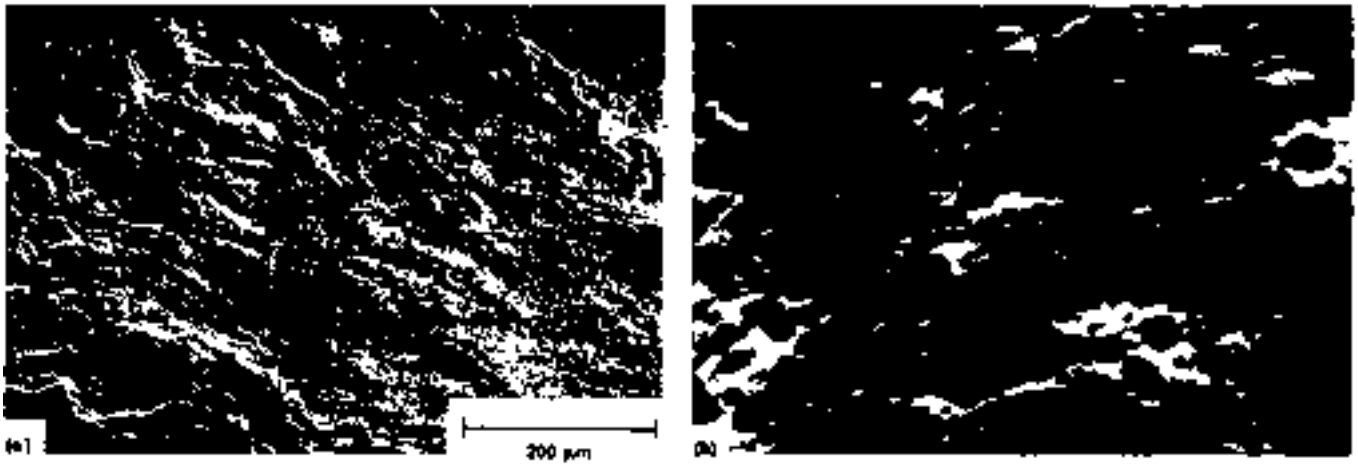


Fig. 22 Secondary corrosion attack on fatigue-fracture surface. (a) Fracture surface of 5-mm (0.2-in.) long crack in an intramedullary tibia nail made of cold-worked type 316LR stainless steel. The crack developed during the early postoperative stage when the fixation was unstable and bending and rotational forces acted on the nail. The crack did not impair fracture healing or nail removal. The fracture surface was covered with fatigue structures except for a small zone where shallow corrosion pits superimposed the fatigue structures. Locally, critical crevice conditions must have developed in the crack. Under weight-bearing, small relative motion can take place in the gap, damaging the passive film. In addition, stagnation and acidification of the body liquid is possible. Arrows indicate the transition between fatigue and corrosion structures. The microstructure of the nail showed no irregularities and was not connected with the corrosion. (b) Shallow secondary corrosion pits superimposed on fatigue striations.

Implants rarely fail under conditions where fatigue loading exceeds the yield stress and plastic deformation takes place. Table 2 lists the relationships between the ultimate tensile stress, yield stress, and fatigue limits for the recrystallized and cold-worked stainless steel used to generate the fatigue curves shown in Fig. 21(a).

Table 2 Relationship between the ultimate tensile stress, yield stress, and fatigue limit for cold-worked and recrystallized type 316LR stainless steel

Material condition	**UTS[(a)]**		**Yield stress[(b)]**		**Fatigue limit[(b)]**		**Ratio of fatigue limit to UTS**	Ratio of fatigue limit to yield stress
	MPa	**ksi**	**MPa**	**ksi**	**MPa**	**ksi**		
Cold worked	1050	152	994	144	300	43.5	0.28	**0.30**
Recrystallized	**590**	**85.5**	**266**	**38.5**	**180**	**26**	**0.30**	**0.67**

(a) UTS, ultimate tensile stress.
(b) Yield stress = $\sigma_{0.2\%}$ yield stress

The fatigue endurance limit of the cold-worked stainless steel corresponds to about 30% of the yield stress, whereas that of the recrystallized steel condition corresponds to about 67% of the yield stress. If at stresses below the yield stress deformation structures develop on the specimen surface before fatigue-cracks form, one must conclude that these deformation structures are created by local repeated elastic deformations. Examination of test specimen surfaces during the fatigue tests described above shows how fatigue cracks develop from cumulative surface damage (Ref 20). This is illustrated on replicas of a recrystallized stainless steel specimen that was fatigued in Ringer's solution (Fig. 23a to d). The following stages of fatigue surface damage were observed: formation of persistent slip steps, slip bands, increasing quantities of multiple slip systems, precracks, short secondary cracks, and one or several major cracks. A main crack will then propagate through the cross section of the specimen. The same processes are morphologically observed on specimens fatigued in air; however, they proceed more slowly in air. This mechanism explains how fatigue cracks initiate without flaws or particular stress raisers.

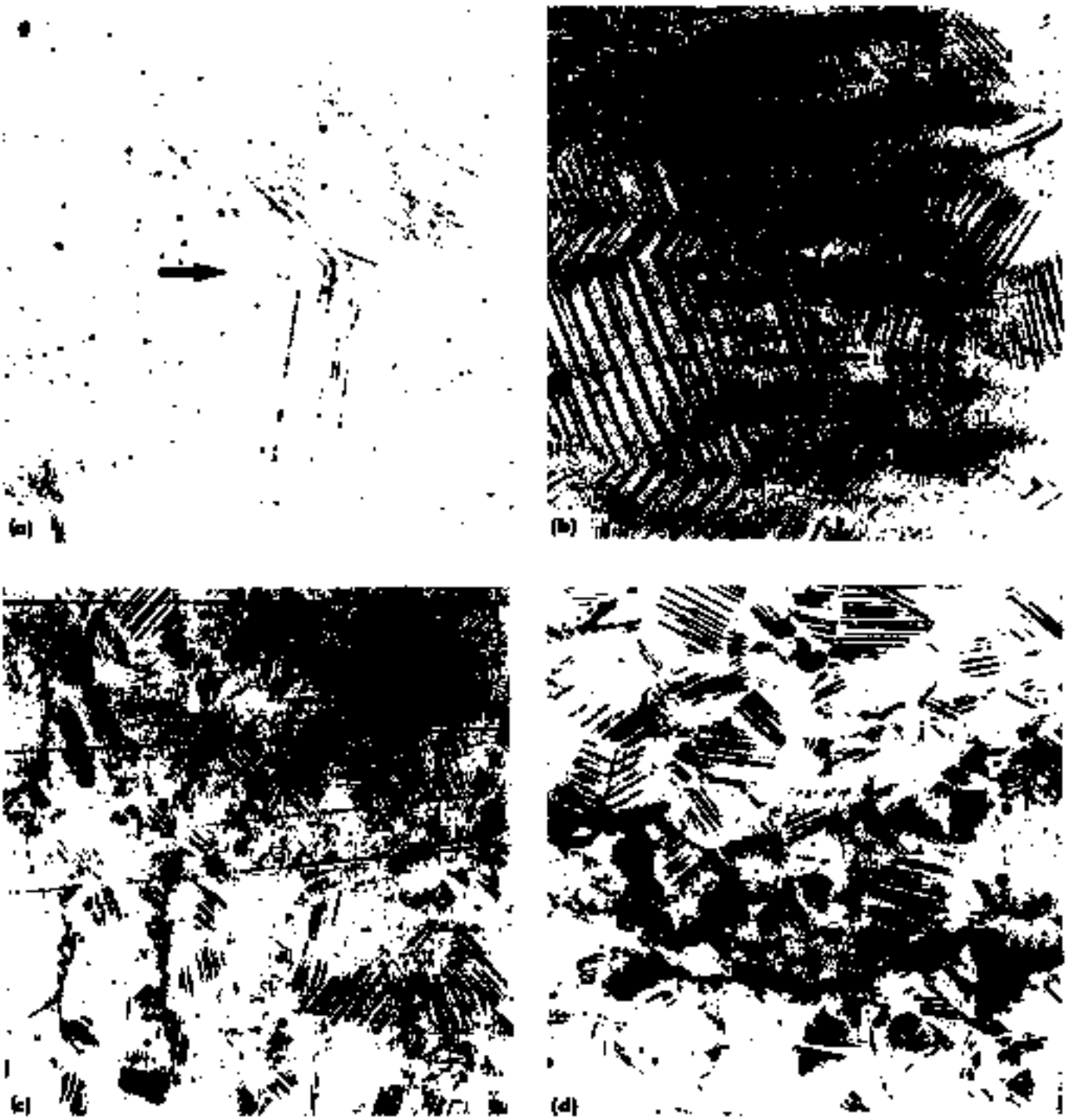


Fig. 23 Free surface replica showing the development of fatigue-surface damage on recrystallized type 316LR stainless steel in aerated Ringer's solution at 38 °C (100 °F), at applied stress of 250 MPa (35.5 ksi). (a) The first visible slip systems developed at a triple point (decorated persistent slip steps) after 500 load cycles. 225×. (b) Increased density of multiple slip systems after 29,500 cycles. 275×. (c) Short cracks that developed in a zig-zag pattern between glide systems after 40,700 cycles. Loading axis is horizontal to the micrograph. 110×. (d) Major cracks that were partially developed through coalescence of short cracks after 60,000 cycles. Loading axis is vertical to the micrograph. 90×. The specimen failed after 69,113 cycles.

The slip bands created by fatigue often have the character of extrusions that are accompanied by intrusions. These intrusions can act as micronotches and initiate cracks. On retrieved broken stainless steel implants, the fatigue damage structure described above is preserved on the surface in the vicinity of the fracture edge.

From the systematic evaluation of test specimens during the fatigue process, *S-N* curves of early fatigue damage were generated (Ref 20). As shown in Fig. 21(b) for the cold-worked stainless steel, curves for the number of cycles to visible slip system formation and crack initiation are similar to the *S-N* curve for the number of cycles to failure. The curves indicate that the lower the load, the longer the time spent in the initiation stage.

Micrographs of failed fatigue specimens of the cold-worked stainless steel show that the intensity of fatigue-surface damage adjacent to the fracture edge depends on the load (Fig. 24a to c). At low stress levels, few glide systems develop. This surface morphology can be used to estimate if an implant was loaded in the high or low cyclic fatigue range.

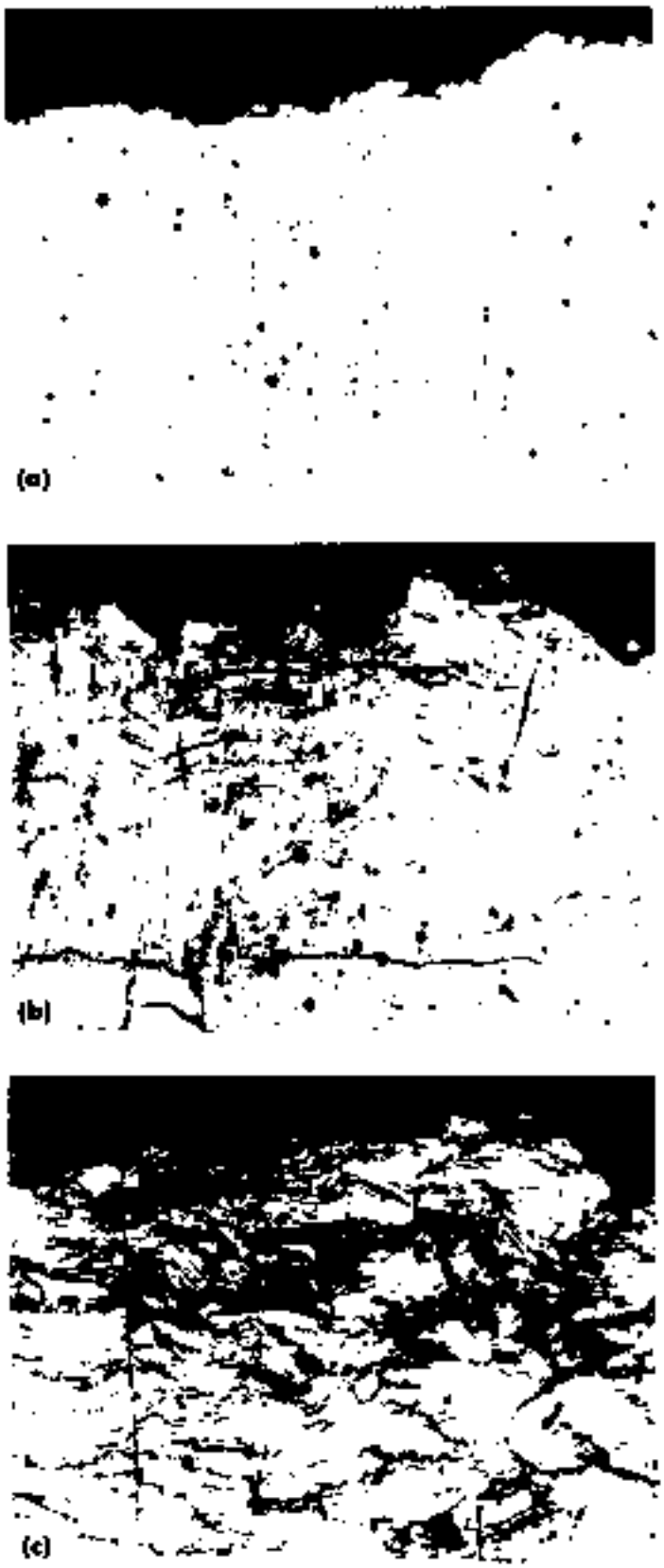


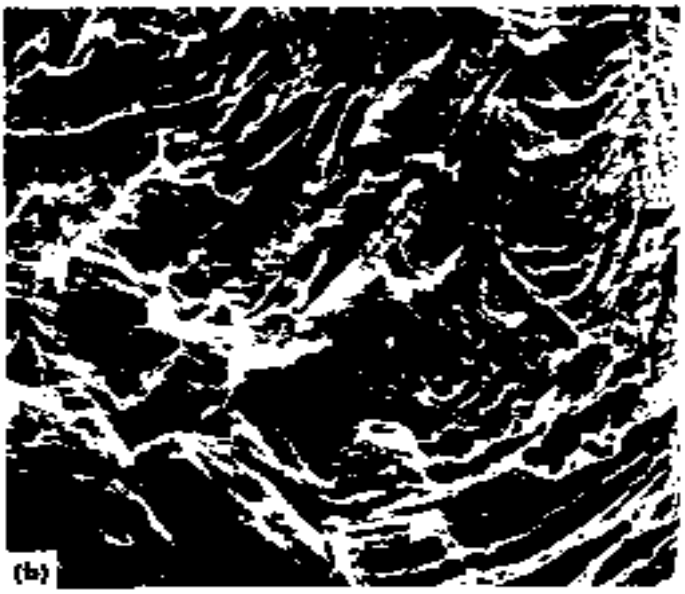
Fig. 24 Cold-worked type 316LR stainless steel that was fatigued in air at different stress levels. Surfaces of broken specimens at fracture edge are shown. (a) Failure at an applied stress of 330 MPa (47.8 ksi) after 7,682,434 load cycles. Only a few glide systems adjacent to the fracture edge are visible. 70×. (b) Failure at an applied stress of 600 MPa (87 ksi) after 105,308 cycles. Multiple glide systems and additional cracks near the fracture edge are shown. 55×. (c) Failure at an applied stress of 800 MPa (116 ksi) after 5308 cycles. Heavy deformation structure and multiple additional cracks appear. 55×

On stainless steel implants, fatigue striations are abundant on the fracture surfaces, unless the surface has been subjected to secondary mechanical damage. In addition to fatigue striations, crystallographic fracture structures are found. The latter appears before a stable crack front forms. Secondary cracks, which run parallel to fatigue striations, do not indicate stress corrosion or corrosion fatigue (Ref 19).

Wrought Type ASTM F563 Cobalt Alloy. The fatigue-crack initiation mechanism in wrought cobalt alloys was assessed with the same type of tests described for stainless steel. Morphologically, the process is the same as that for stainless steel, but all initiation steps are shifted to higher stress values because of the higher strength and fatigue resistance of alloy ASTM F563.

Compared to stainless steels, the morphology of the fatigue-fracture surfaces is characterized by very dense patterns of fatigue striations. These patterns superimpose the normally dominating crystallographic structures. It is sometimes difficult to distinguish fatigue striations from glide or tearing structures. This is particularly true when the glide systems have an orientation that accommodates the formation of striations (Fig. 25a to c). The toughness of these types of wrought cobalt alloys is assumed to account for this behavior.

(a)



(b)

(c)

Fig. 25 Fatigue-fracture structures on wrought type ASTM F563 cobalt-alloy test specimens that fatigued in air. (a) Very fine fatigue striations are superimposed on crystallographically oriented fracture structures. 2480×. (b) Crystallographically oriented fracture morphology showing twin structures. 620×. (c) Close-up view of the twin structure indicated by arrow in Fig. 25(b). The fine parallel lines within the twin structure are slip (glide) bands, which are not easily distinguished from fatigue striations.

Commercially Pure Titanium. The morphology of fatigue initiation and propagation of titanium is more complex than that of stainless steel. The limited number of glide systems of the hexagonal close-packed crystal structure of the titanium is assisted by twinning processes to allow plastic deformation (Ref 22). Correspondingly, the range of ductile behavior of titanium is more limited than that of stainless steel.

The corrosion resistance of titanium in the body and its tissue compatibility are excellent. There have been no reported cases of allergies to titanium. Unfortunately, the wear properties of titanium and its alloys are poor; therefore, they are not usually chosen as material for the articulation surfaces of prostheses.

The same fatigue experiments described for type 316LR stainless steels were performed on commercially pure titanium specimens. For comparison, Fig. 26 shows the fatigue-surface damage on a broken titanium test specimen. The polarized light reveals that twinning, wavy glide, and grain-boundary distortion contribute to the fatigue damage on the specimen surface.

Fig. 26 Specimen surface of recrystallized titanium at fracture edge. Specimen was fatigued at a stress of 600 MPa (87 ksi) in air. Twinning, wavy glide deformation, and grain-boundary distortion are visible on this relatively heavily loaded specimen. Polarized light. 330×

The fatigue-fracture surfaces of titanium can show very different features, depending on overall and local stress conditions, crystallographic orientation, and material conditions. Figures 27(a) to (c) show fractographs of the surface of a cold-worked titanium specimen fatigued at a stress level of 600 MPa (87 ksi). Depending on local stress distribution and grain orientation at the crack front, extremely finely spaced fatigue striations (Fig. 27a) can form in the vicinity of very

coarse striations (Fig. 27b), which appear to be in transition to glide formation. Figure 27(c) shows an area of overload fracture with prismatic and dimple tearing structures. These fatigue structures indicate ductile behavior.

(a)
10 μm
(b)
10 μm
(c)
10 μm

Fig. 27 Fracture surface of commercially pure titanium test specimens that failed at an applied stress level of 600 MPa (87 ksi) in air. (a) Very fine fatigue striations. (b) Coarse fatigue striations probably in transition to glide bands. (c) Overload tearing structures

Figures 28(a) to (c) illustrate an area of the fracture surface of a titanium bone plate that suffered fatigue failure in an unstable internal fixation situation. The bone plate showed mixed fracture structures, such as fatigue striations, tearing structures, and terraces. Figure 28(a) shows an overview, and Fig. 28(b) shows ledges and tearing structures of the terrace.

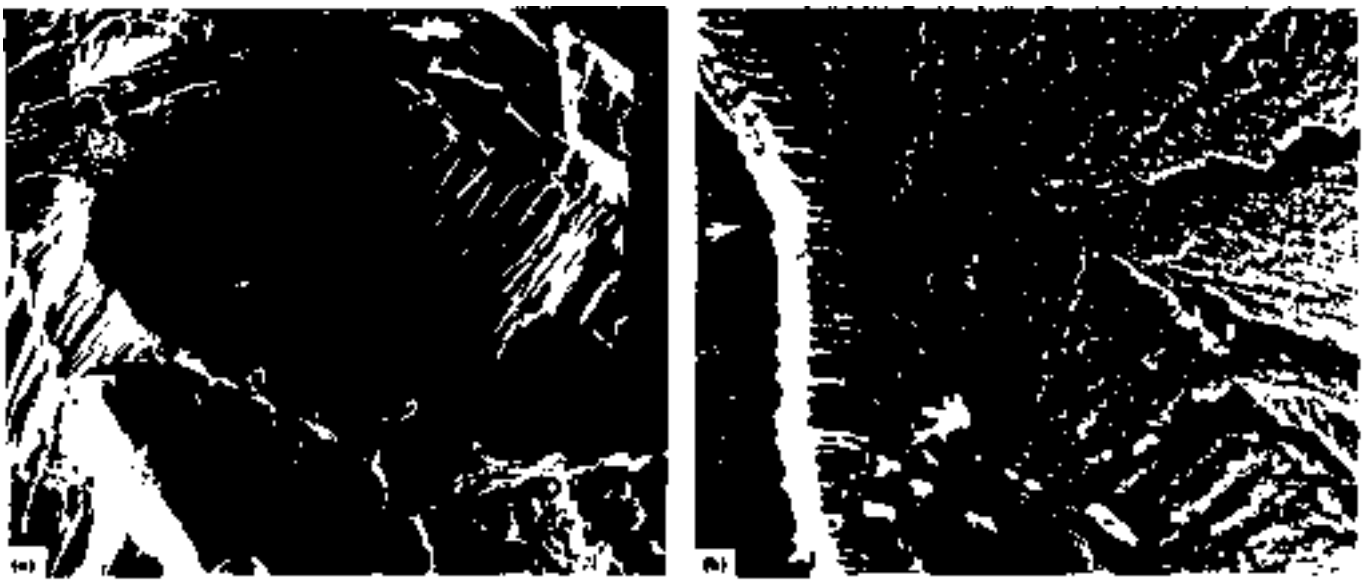


Fig. 28 Fatigue-fracture surface of broken commercially pure titanium bone plate with mixed fracture morphology. (a) Fracture surface shows fatigue striations, terraces, and tearing ridges, depending on the local crystallographic orientation. 250×. (b) Higher magnification view of the area indicated by arrow in (a) showing large and small ledges with partial tearing configurations. 1250×

References cited in this section

19. O.E.M. Pohler and F. Straumann, Fatigue and Corrosion Fatigue Studies on Stainless Steel Implant Material, in *Evaluation of Biomaterials,* C.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

20. O.E.M. Pohler, "Study of the Initiation and Propagation Stages of Fatigue and Corrosion Fatigue of Orthopaedic Implant Materials," dissertation, Ohio State University, 1983

21. E. Powell, M.S. thesis, Ohio State University, 1983

22. R.W. Hertzberg, Deformation and Fracture Mechanics of Engineering Materials, John Wiley & Sons, 1976

Ortrun E.M. Pohler, Institut Straumann AG

Degradation of Orthopedic Implants

Metallic substance can be released from implants by such mechanisms as:

- General corrosion

- Galvanic corrosion
- Intercrystalline corrosion
- Maintenance of passive film of highly corrosion-resistant implant materials
- Stress corrosion combined with fatigue
- Corrosion fatigue
- Crevice corrosion
- Fretting corrosion
- Fretting
- Wear

Reports on *in vivo* and *in vitro* studies and discussions of the degradation and corrosion of implants are found in Ref 23, 24, and 25.

Implant degradation can lead to mechanical failure of the implant when severe disintegration takes place, as illustrated in Example. However, with the use of improved implant materials and metallurgical processing, visible general corrosion, intercrystalline corrosion, and galvanic corrosion are seldom observed today.

Combined attack of fatigue with stress corrosion, fretting, or fretting corrosion can also cause implant breakdown, but these failure modes are rare. Slight branching of cracks found on broken implants may not indicate stress corrosion, but can be explained by fatigue-propagation processes. In experiments with magnesium chloride solution, stress-corrosion cracks in type 316L stainless steel are characterized by extensive branching. The fracture surfaces of these cracks exhibit facets and multiple fine cracks that in most cases are crystallographically oriented (Ref 19, 23). The combination of fatigue with stress corrosion has been observed in few cases on type 316L stainless steel. Very little metal is released into the tissue by this form of attack.

The generation of wear, fretting, or corrosion products can evoke local physiological reactions that may necessitate reoperation. This is, however, seldom experienced. Local reactions in relation to fracture-treatment implants are occasionally reported, but usually cease after resting of the limb and after routine removal of the implant upon healing of the fracture. The systemic effects of metals that are released from implants are the subject of much research and speculation. Unambiguously diagnosed general allergies to implant metals are rarely reported. However, mere contact with the metal is sufficient to create such reactions in patients susceptible to allergies. If implants remain in the passive condition, the release of metal ions is very low, corresponding to the current density related to the corrosion potential of the passive stage. These metal traces are usually below the concentrations that are detectable with modern analytical techniques.

Crevice corrosion, fretting corrosion, fretting, and wear typically occur at the contact areas of multiple-component implants. Wear is found on the articulation surfaces of prostheses. In a long-term study on total hip prostheses that were in the body for about 8.3 years, up to 2.8 mm (0.11 in.) of material loss was found, amounting to a wear rate of 0.34 mm/year (0.013 in./year). These wear rates were measured on radiographs and involve the abrasion of the metal head and the polymer acetabular cup, and the polymer creep.

The wear behavior of a prosthesis depends on the surface finish, the contact area, the loading, the loading directions, and the wear properties of the material. Many improvements have been made to minimize the wear on artificial joints, and researchers and producers have designed various wear simulators (Ref 26, 27). If prosthesis stems become loose, wear and fretting or, sometimes, shallow fretting corrosion can take place at various areas of the stem (compare Fig. 33c and d in Example).

The contact areas between screw heads and plate holes are usually susceptible to fretting and fretting corrosion when relative motion occurs between the screw and the plate. Compared to artificial joint surfaces, the contact areas between screws and plates are smaller and therefore are subject to higher surface pressures. Thus, it is difficult to prevent fretting at screw-head/plate-hole junctions by using special surface treatments. Mechanical attack is found on all implant materials at such junctions, but the degree of additional corrosion varies with the local condition and the type of material.

On commercially pure titanium, mechanical abrasion is found with oxidation of the wear particles (Example). On Ti-6Al-4V alloys, wear and wear fatigue are found; in these cases, wear fatigue is indicated by cracking of the worn material surface layers. Because fretting easily destroys the passive film, stainless steel can develop pitting corrosion in connection with fretting. In addition, the geometry between the plate hole and the screw head provides crevices. The intensity of pitting corrosion in connection with fretting depends on the quality of the steel. For example, type 304 stainless steel or steel with a high inclusion content is prone to more severe pitting than a type 316LR steel (compare Examples and). On high-quality stainless steels, mechanical attack and shallow pitting corrosion are observed. On screw-plate combinations m wrought cobalt alloys, pitting is observable only with SEM. Also, on cast cobalt-chromium-molybdenum alloys, fretting and slight corrosive attack have been observed.

On two-component hip-screw plates, fretting and fretting corrosion can also be observed. Reference 23 provides the results of analyses of fretting and fretting corrosion conditions on retrieved implants made from various materials. The following examples illustrate fretting and fretting corrosion conditions on screw-plate junctions for stainless steel and titanium.

Example 10: Heavy Pitting Corrosion on a Type 304 Stainless Steel Screw. Figure 29 shows a screw head that exhibits heavy pitting corrosion attack. Deep tunnels penetrated the screw head (Fig. 29a) and followed the inclusion lines (Fig. 29b). This screw was inserted in a plate made of type 316LR stainless steel that showed some mechanical fretting and very few corrosion pits.

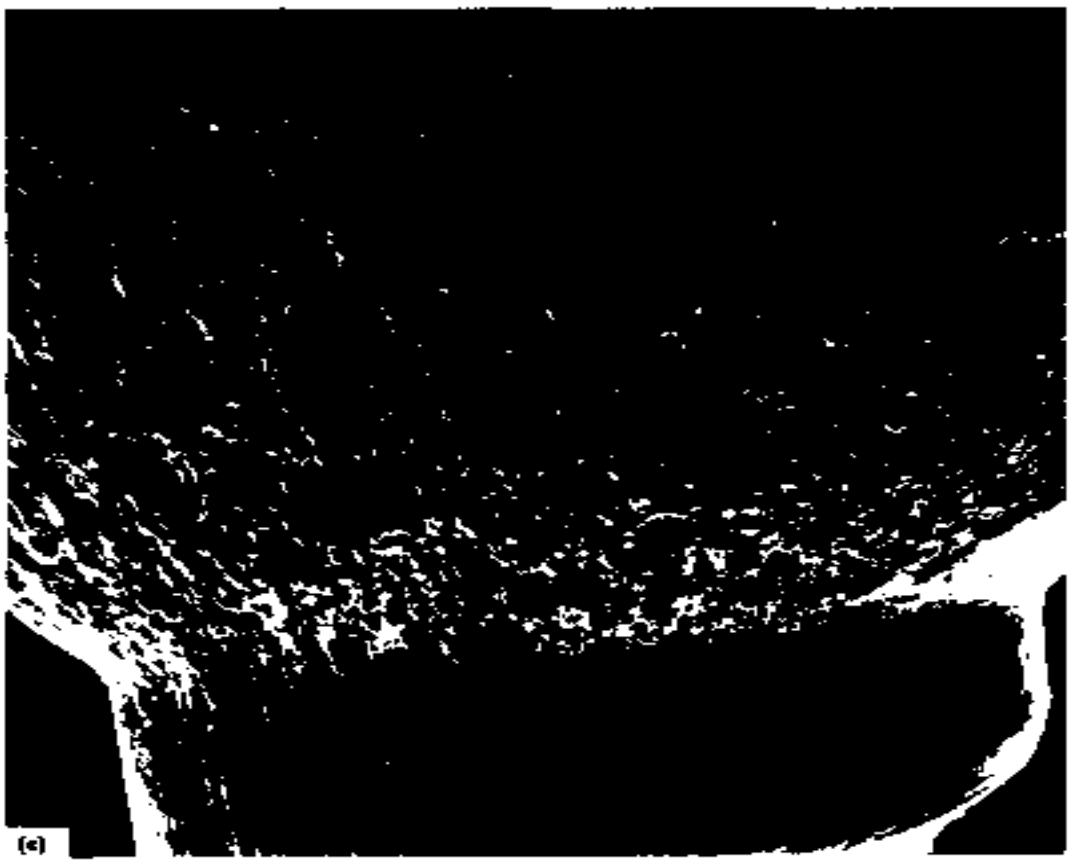


Fig. 29 Heavy pitting corrosion on type 304 stainless steel bone screw. (a) Longitudinal section through head of bone screw showing corrosion tunnels. (b) Etched longitudinal section showing the many primary inclusion lines and corrosion tunnels that follow the inclusions. (c) SEM overview of corrosion attack on screw head

This is another example where type 304 stainless steel is not satisfactory as an implant material and where inclusion lines contribute to corrosion, which is also observed on steels of higher quality but with larger amounts of primary inclusions.

Example 11: Screw Hole With Fretting and Fretting Corrosion of a Type 316LR Stainless Steel Plate. Figure 30 shows a plate hole with the area that was in contact with the screw head. In contrast to Example, the attack on this high-quality type 316LR stainless steel was only shallow. Figure 30(a) shows an overview indicating that a large portion of the contact area exhibited only mechanical grinding and polishing structures. The fine corrosion pits in the periphery are shown at a higher magnification in Fig. 30(b). The intense mechanical material transfer that can take place during fretting is visible in Fig. 30(c) and (d). In front of this structure is a corrosion pit that is surrounded by a burnished surface texture. The burnished surface may have broken open at this point. Observation of other implant specimens confirms that material layers are smeared over each other during wear and can be attacked by pitting corrosion. These pits are then covered by a new burnished material film that can break open again. Thus, the material is worn and degradated.

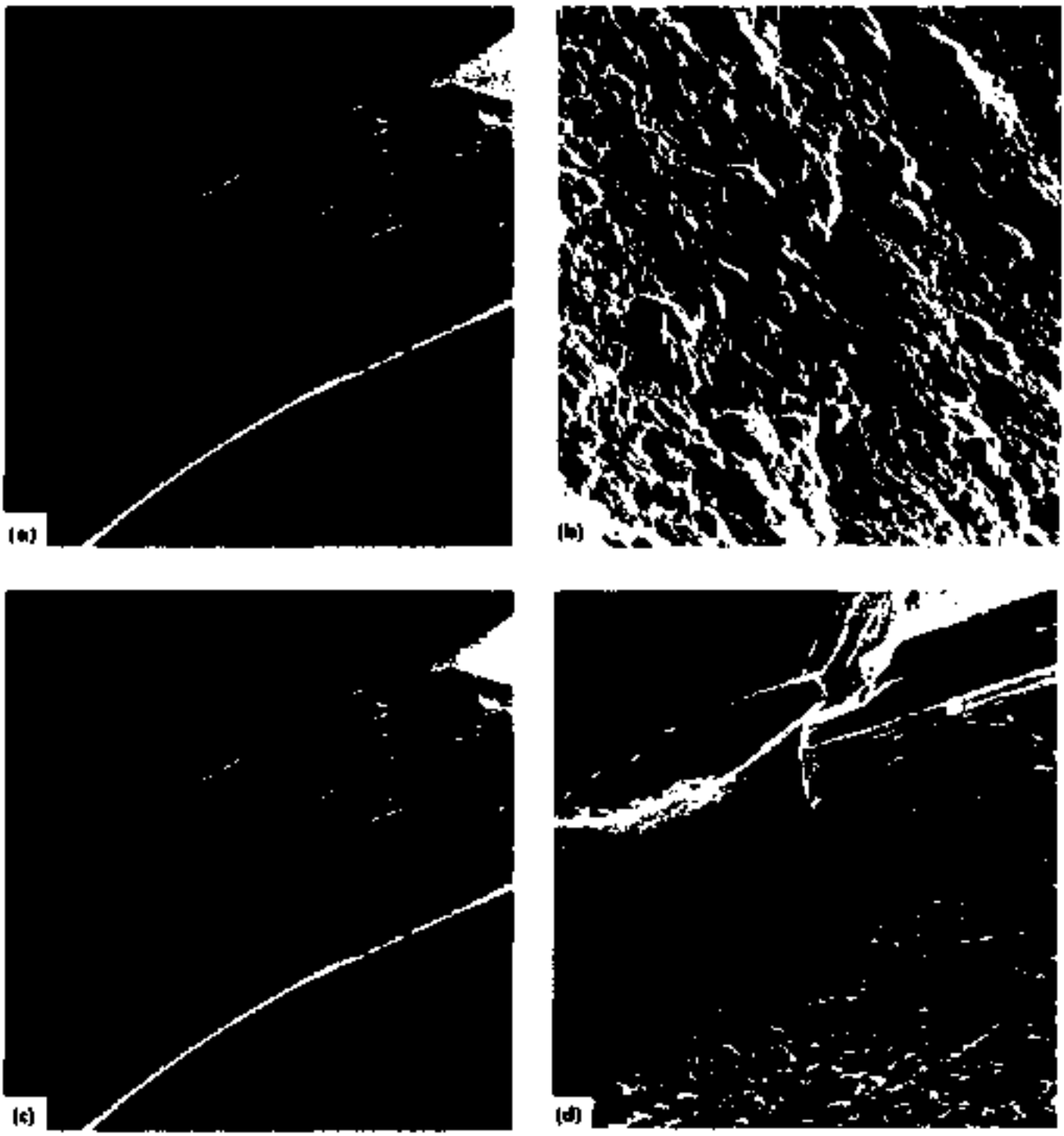


Fig. 30 Fretting and fretting corrosion at the contact area between the screw hole of a type 316LR stainless steel bone plate and the corresponding screw head. (a) Overview of wear on plate hole showing mechanical and pitting corrosion attack. 15×. (b) Higher-magnification view of shallow pitting corrosion attack from periphery of contact area. 355×. (c) Contact area showing the effects of mechanical material transport (material tongue in right upper corner), fretting and wear (burnished areas), and corrosion (shallow pitting). 355×. (d) Higher-magnification view of Fig. 30(c). 650×

Figure 31 illustrates how the corrosion and wear products are transported in the tissue. These products are found extracellularly and phagocytized in the fibrocytes of the connective tissue in this example.

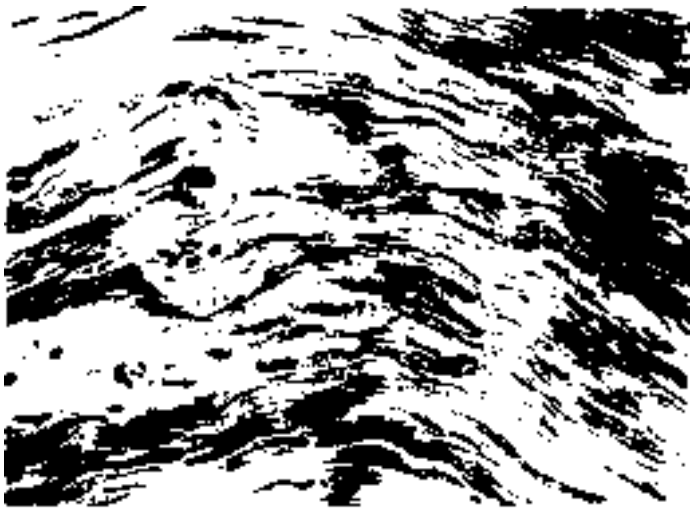


Fig. 31 Connective tissue near stainless steel bone plate with impregnation of corrosion products. These products are found extracellularly and in the connective tissue cells. 230×

Example 12: Titanium Screw Head With Fretting Structure at Contact Area With Plate Hole. Figure 32(a) shows a portion of a titanium screw head with a lip of material that was transported by fretting at a plate-hole edge. A flat fretting zone is visible on the screw surface over the material lip which is magnified in Fig. 32(b). A cellular wear structure containing wear debris is present. Figure 32(c) shows material destruction and the formation of thin flakes of wear debris. Mechanical deformation is superimposed on wear structures in Fig. 32(d). No morphological signs of corrosion have been observed in connection with fretting structures. This has been typically found on all contact areas of plate holes and screw heads of titanium which were investigated.

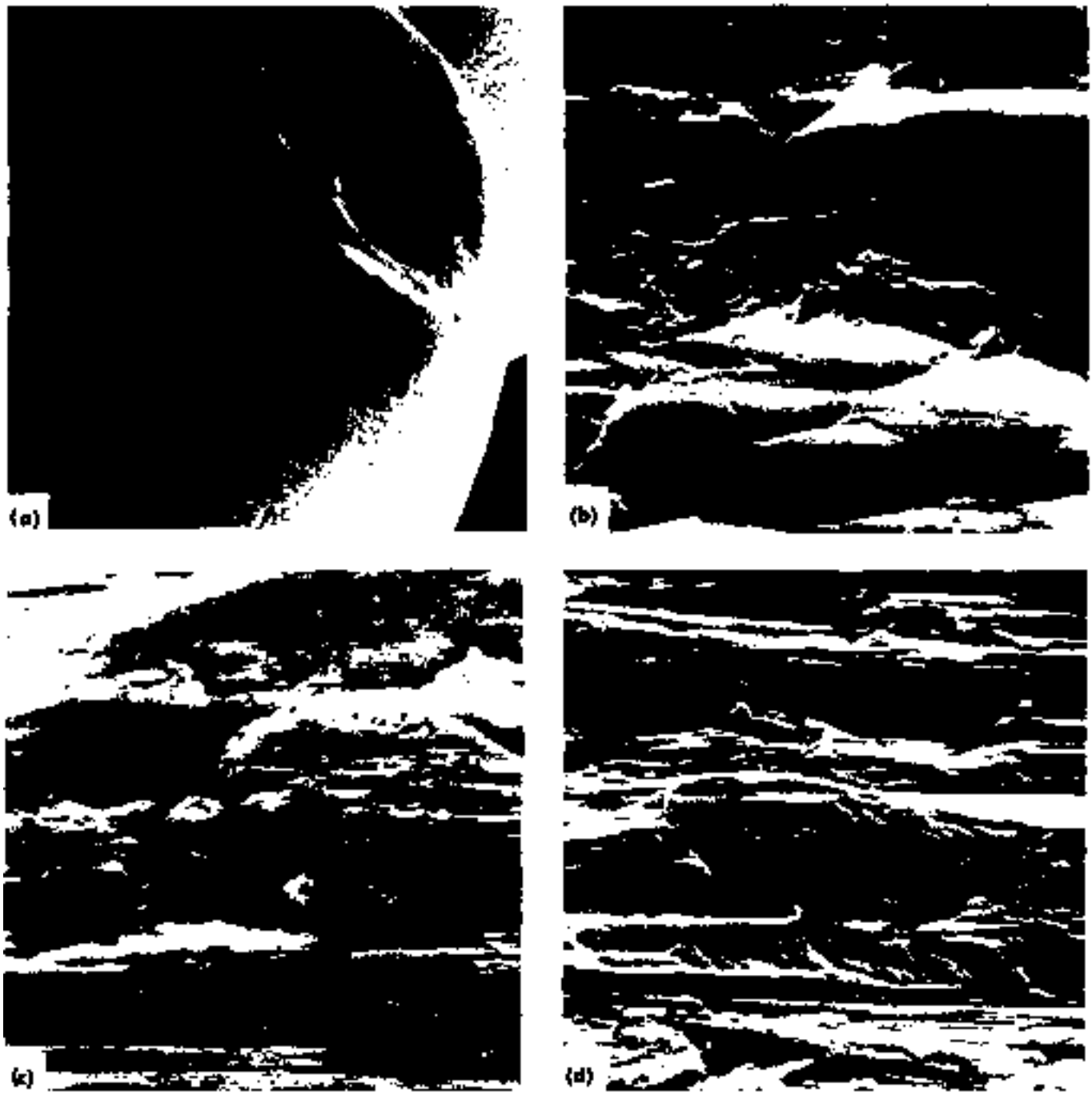


Fig. 32 Wear on head of titanium screw. (a) Material transport and fretting zone. (b) Close-up view of wear structures showing fine wear products. 120×. (c) Wear structures showing generation of small wear particles. 1200×. (d) Wear structures with additional fretting structures. 305×

References cited in this section

19. O.E.M. Pohler and F. Straumann, Fatigue and Corrosion Fatigue Studies on Stainless Steel Implant Material, in *Evaluation of Biomaterials,* C.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

23. O.E.M. Pohler, Degradation of Metallic Orthopaedic Implants, in *Biomaterials in Reconstructive Surgery,* L.R. Rubin, Ed., The C.V. Mosby Co., St. Louis, 1983

24. B.C. Syrett and A. Acharya, Ed., *Corrosion and Degradation of Implant Materials,* STP 684, ASTM, Philadelphia, 1979

25. A. Fraker and C.D. Griffin, Ed., *Corrosion and Degradation of Implant Materials,* STP 859, ASTM, Philadelphia, 1985

26. J.O. Galante and W. Rostoker, Wear in *Total Hip Prostheses, Acta Orthop. Scand. Suppl.,* Vol 145, 1973

27. M. Ungethüm, Technologische und biomechanische Aspekte der Hüft- und Kniealloarthroplastik, in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 9, Verlag Hans Huber, Bern Stuttgart Wien, 1978

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

Fractures of Total Hip Joint Prostheses

The hip joint prosthesis is the most common joint replacement. The function of the normal hip joint, the design of the hip joint prosthesis, and the related complications have been extensively studied compared to other joints.
A retrospective multicenter study on the performance of 39,000 total hip replacements over a period of 10 years was published in 1982 (Ref 28). A biomechanical concept of the total hip prosthesis based on clinical experience that gives a background for the understanding of complications can be found in Ref 29.
One of the complications encountered in total hip replacements is the fracture of prosthesis stems. Lateral bending and torsional forces act on the prosthesis stem when it is not supported. All fractures of the hip prosthesis stems are reportedly caused by a fatigue-failure mechanism (Ref 27). The fracture of prosthesis stems is a secondary event that follows loosening of the stem.
Different factors, which can also act in combination, promote stem loosening. These factors include bone resorption, degradation of bone cement, and unfavorable positioning of the prosthesis. Blood clot inclusions, holes in the cement, and overheating of the bone can also contribute to stem loosening.
Information on the performance and testing of hip stem material can be found in Ref 27, 30, 31, and 32. Standard techniques for the cyclic bending test of hip prosthesis stems have also been developed (Ref 30) and an ISO standard is in preparation. Figures 33, 34, 35, and 36 show an example of a failed prosthesis stem that involves fatigue and stem loosening due to bone resorption and bone cement degradation.

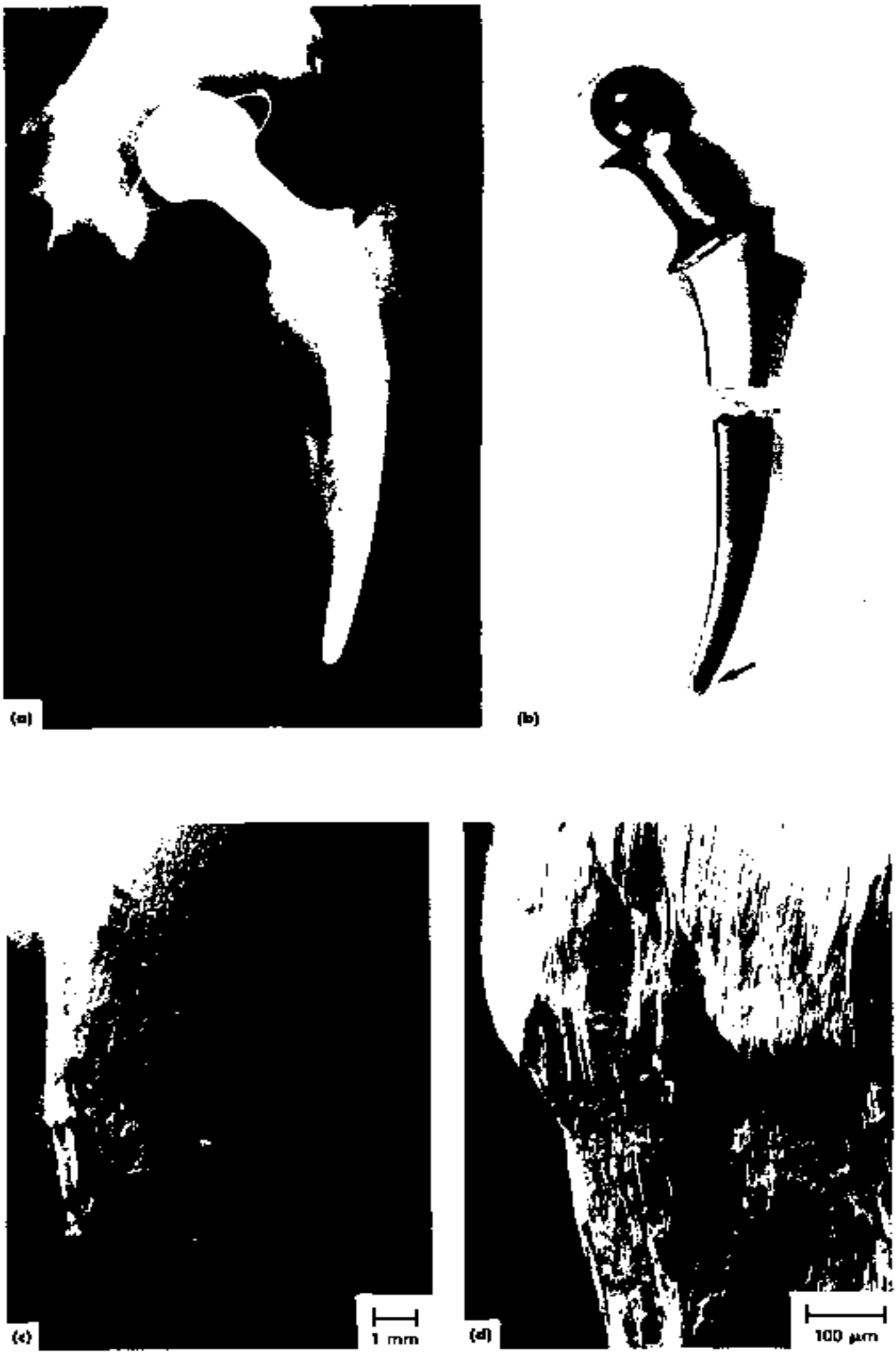

(a)
(b)
(c)
1 mm
(d)
100 μm

Fig. 33 Broken hip prosthesis of cast type ASTM F75 cobalt-chromium-molybdenum alloy. (a) Radiograph of total hip prosthesis. Circular wire marks acetabulum component made from plastics. Arrows (from top to bottom) indicate the area where the prosthesis stem is loosening at the collar, a stem fracture, and a fracture of bone cement at the end of the stem, respectively. (b) Fracture of prosthesis stem. Wear at end of stem (arrow) indicates stem movement due to loosening. (c) Wear at end of stem. (d) Close-up view of stem end showing material transfer and layering from wear. Corrosion signs are not observed. See also Figures 34, 35, and 36.

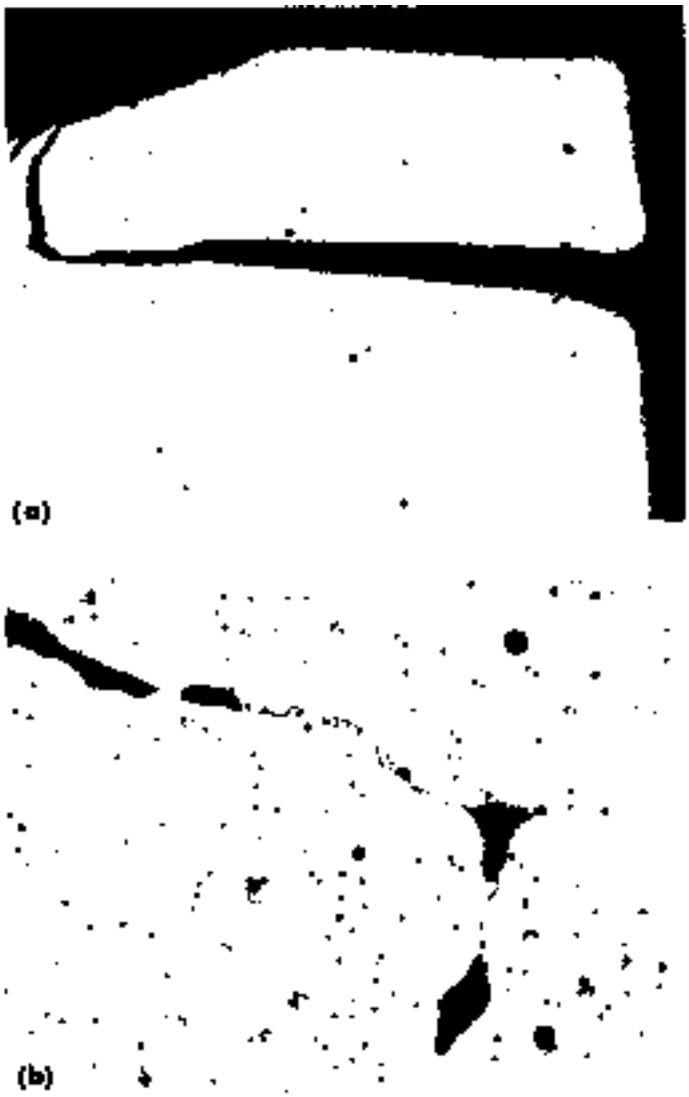


Fig. 34 Metallographic sections of failed hip prosthesis shown in Fig. 33. (a) Longitudinal section through fracture surface showing secondary fatigue crack parallel to fracture surface. 35×. (b) Cross section through prosthesis stem showing gas pores and second phase at grain boundaries and in grains. 105×

Fig. 35 Optical fracture analysis of the failed hip prosthesis shown in Fig. 33. (a) Fracture surface exhibiting large grains. The upper right grain is identical to grain A shown in Fig. 36(a). Crack originated at the lateral side of the stem. (b) Longitudinal section through fracture surface with glide systems activated by the fatigue process. Two cracks, one of which propagated through the grain, initiated along these glide systems. 195×. (c) Longitudinal section through fracture surface at the upper left area in Fig. 35(a) containing secondary cracks. Arrow indicates a gas pore. 270×

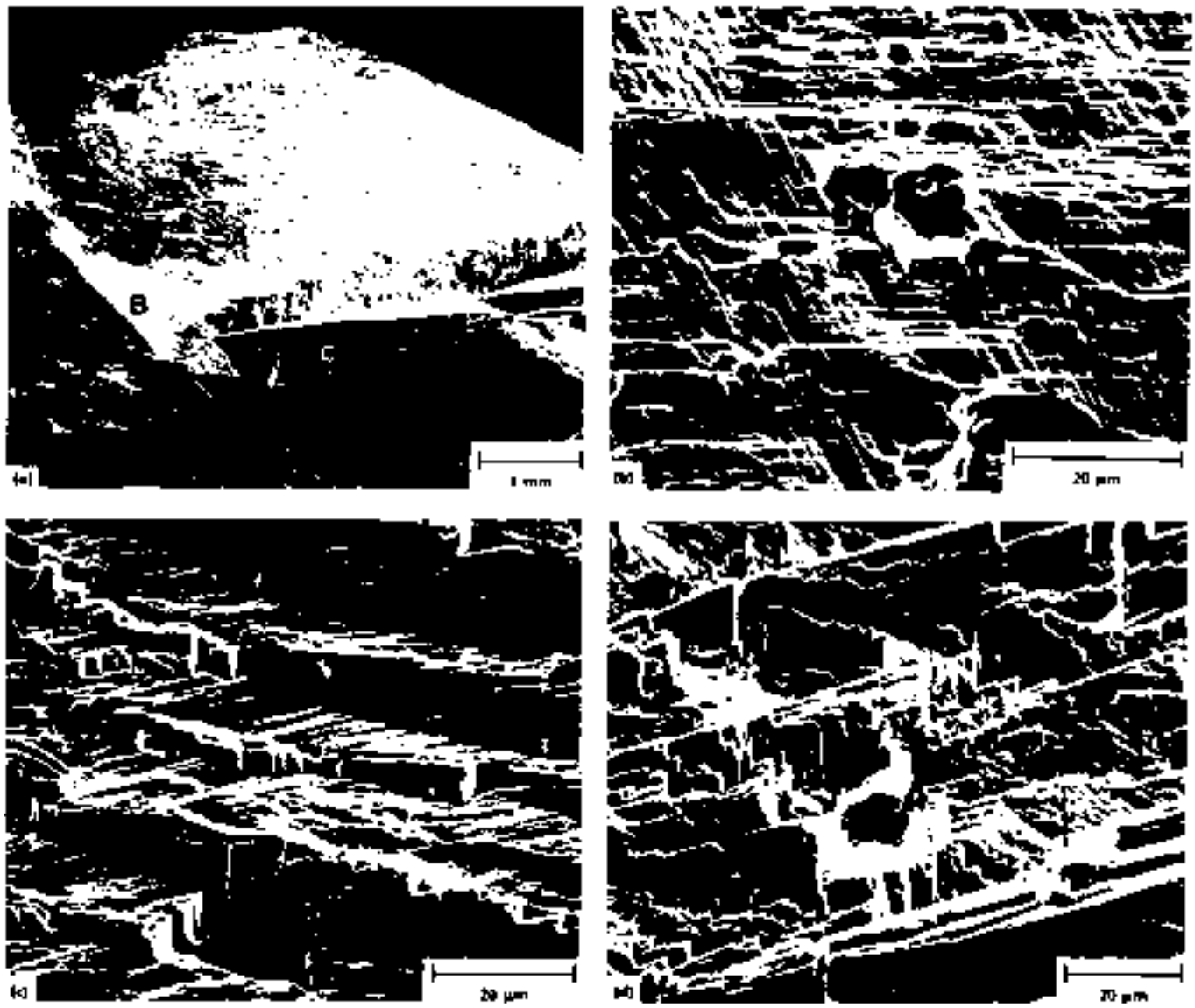


Fig. 36 SEM fracture-surface analysis of the failed hip prosthesis shown in Fig. 33. (a) Fracture surface showing three distinct grains labeled A, B, and C. (b) Grain A has a shallow crystallographically oriented fracture structure. (c) Grain B has a crystallographically oriented fracture structure with slip traces. Note vertical cracks running in the same direction as some of the fracture planes. (d) Grain C has a crystallographically oriented fracture structure and small vertical cracks similar to those shown in (c). A ruptured gas hole is visible at the center of the fractograph. 300×

Example 13: Broken Stem of Femoral Head Component of Total Hip Prosthesis Made From Cast Cobalt-Base Alloy. In a 65 year old male patient, radiotranslucency was visible around the collar of the femoral head prosthesis on radiographs taken 5 months after implantation. One month later, the bone cement was broken at the distal end of the prosthesis stem, and a small indentation on the lateral contour of the stem was visible where the stem had broken 2 weeks later.

Investigation. Figure 33(a) shows a radiograph illustrating the broken prosthesis. The dislocation of the fragment of the prosthesis indicated the degree of loosening and implant loading.

Under weight beating, the proximal prosthesis fragment was pushed toward lateral, and the loosening in the medio-lateral plane can be seen as a gap between the lateral stem contour and the bone or cement. The upper arrow in Fig. 33(a) indicates the loosening at the collar. The bottom arrow marks the crack in the bone cement plug at the end of the stem. Figure 33(b) shows an overview of the broken prosthesis component. Because of the loosening, the end of the stem was heavily worn in contact with the bone cement (Fig. 33c to d).

In analyses of broken prosthesis stems, such wear or fretting traces are good indicators of the presence of motion and loosening. In other investigated prostheses, flat fretting zones have also been found on other areas of the stem. On stainless steel prostheses, this wear appears occasionally as slight fretting corrosion attack. One may, however, distinguish wear that took place after fracture of the stem.

A metallographic specimen was taken parallel to the stem surface and perpendicular through the fracture surface of the distal fragment. This section revealed a secondary crack that had originated at the lateral aspect of the stem. The top contour in Fig. 34(a) represents the fracture edge, where two crack openings along slip planes can be observed. On a transverse metallographic section (Fig. 34b), gas pores are apparent in the grain and at the grain boundaries.

Discussion. Figure 35(a) shows an overview of the fracture surface. The grain size is considerable. The plateau in the upper right-hand corner represents a single grain, which is identical to grain A in Fig. 36(a). It is the same grain through which the secondary crack shown in Fig. 34(a) runs, and it is where the fracture originated. Figure 35(b) shows a higher-magnification view of the section through the fracture surface. The stresses created through the fatigue process activated glide systems that serve the formation of secondary cracks along glide planes. On the section through the fracture surface, another secondary crack formation is found (Fig. 35c). At these cracks, multiple slip bands also formed in connection with the fracture surface, and the cracks follow along these structures. One of the cracks propagated through a larger gas pore.

Figures 36(a) to (d) show details of the fracture morphology by SEM. In Fig. 36(a), three distinct grains are marked. These grains are oriented differently with respect to the propagating crack and therefore exhibit different crystallographic fracture patterns. Grain A (Fig. 36b) has a relatively flat texture, but also shows crystallographically oriented structures. Grain B (Fig. 36c) is characterized by a staircase pattern. Material separation occurred on preferred slip planes. The fine parallel line structures that run diagonally through the fractograph may be slip traces. Of special interest are the parallel microcracks that appear black and are superimposed on the other structures. On grain C (Fig. 36d), identical microcracks are found, although the grain is obviously oriented differently. A ruptured gas pore is visible in the center of Fig. 36(d). Typical fatigue striations were not found on the fracture surface. Glide systems appear to be activated at relatively low energies so that cracks can easily propagate along glide systems.

References cited in this section

27. M. Ungethüm, Technologische und biomechanische Aspekte der Hüft- und Kniealloarthroplastik, in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 9, Verlag Hans Huber, Bern Stuttgart Wien, 1978

28. P. Griss *et al.,* Ed., Findings on Total Hip Replacement for Ten Years, in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 21, Verlag Hans Huber, Bern Stuttgart Wien, 1982

29. R. Schneider, Die Totalprothese der Hüfte (Ein biomechanisches Konzept und seine Konsequenzen), in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 24, Verlag Hans Huber, Bern Stuttgart Wien, 1982

30. M. Semlitsch and B. Panic, Corrosion Fatigue Testing of Femoral Head Prostheses Made of Implant Alloys of Different Fatigue Resistance, in *Evaluation of Biomaterials,* G.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

31. M.F. Semlitsch, B. Panic, H. Weber, and R. Schoen, *Comparison of the Fatigue Strength of Femoral Prothesis Stems Made of Forged Ti-Al-V and Cobalt-Base Alloys,* STP 796, ASTM, Philadelphia, 1983

32. M. Semlitsch and B. Panic, Ten Years of Experience with Test Criteria for Fracture-Proof Anchorage Stems of Artificial Hip Joints, *Eng. Med.,* Vol 12, 1983, p 185

Failures of Metallic Orthopedic Implants

Ortrun E.M. Pohler, Institut Straumann AG

References

1. D.C. Mears, *Materials and Orthopaedic Surgery,* The Williams & Wilkins Co., Baltimore, MD, 1979

2. L.R. Rubin, Ed., Chapters 23–27, in *Biomaterials in Reconstructive Surgery,* The C.V. Mosby Co., St. Louis, MO, 1983

3. M.E. Müller, M. Allgöwer, R. Schneider, and H. Willenegger, *Manual of Internal Fixation,* 2nd ed., Springer Verlag, 1979

4. H. Gerber *et al.,* Quantitative Determination of Tissue Tolerance of Corrosion Products by Organ Culture, *Proceedings of the European Society for Artificial Organs,* Vol 1, 1974, p 29–34

5. H. Gerber and S.M. Perren, Evaluation of Tissue Compatibility of In Vitro Cultures of Embryonic Bone, in *Evaluation of Biomaterials,* G.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

6. V. Geret, B.A. Rahn, R. Mathys, F. Straumann, and S.M. Perren, A Method for Testing Tissue Tolerance for Improved Quantitative Evaluation, in *Evaluation of Biomaterials,* G.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

7. G.K. Smith, Systemic Aspects of Metallic Implant Degradation, in *Biomaterials in Reconstructive Surgery,* L.R. Rubin, Ed., The C.V. Mosby Co., St. Louis, MO, 1983

8. Medical Devices, *Annual Book of ASTM Standards,* Vol 13.01, ASTM, Philadelphia, 1985

9. A.C. Guyten, *Textbook of Medical Physiology,* W.B. Saunders Co., Philadelphia, PA

10. J.B. Park, *Biomaterials: An Introduction,* Plenum Press, 1979

11. P.G. Laing, Biocompatibility of Biomaterials, in *Orthopaedic Clinics of North America,* Vol 4, C.M. Evarts, Ed., 1973

12. F. Pauwels, *Der Schenkelhasbruch,* Enke, Stuttgart, 1935

13. R. Schenk, Fracture Repair—Overview, in *9th European Symposium on Calcified Tissues,* 1973

14. R. Schenk and H. Willenegger, Zur Histologie der primären Knochenheilung, *Unfallheilkunde,* Vol 80, 1977, p 155

15. S.M. Perren and M. Allgöwer, *Nova Acta Leopoldina,* Vol 44 (No. 223), 1976, p 61

16. S. Perren, R. Ganz, and A. Rüter, Mechanical Induction of Bone Resorption, in *4th International Osteol. Symposium,* Prague, 1972

17. H. Yamada, *Strength of Biological Materials,* The Williams & Wilkins Co., Baltimore, MD, 1970

18. L.E. Sloter and H.R. Piehler, Corrosion-Fatigue Performance of Stainless Steel Hip Nails—Jewett Type, in *STP 684,* ASTM, Philadelphia, 1979, p 173

19. O.E.M. Pohler and F. Straumann, Fatigue and Corrosion Fatigue Studies on Stainless Steel Implant Material, in *Evaluation of Biomaterials,* C.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

20. O.E.M. Pohler, "Study of the Initiation and Propagation Stages of Fatigue and Corrosion Fatigue of Orthopaedic Implant Materials," dissertation, Ohio State University, 1983

21. E. Powell, M.S. thesis, Ohio State University, 1983

22. R.W. Hertzberg, Deformation and Fracture Mechanics of Engineering Materials, John Wiley & Sons, 1976

23. O.E.M. Pohler, Degradation of Metallic Orthopaedic Implants, in *Biomaterials in Reconstructive Surgery,* L.R. Rubin, Ed., The C.V. Mosby Co., St. Louis, 1983

24. B.C. Syrett and A. Acharya, Ed., *Corrosion and Degradation of Implant Materials,* STP 684, ASTM, Philadelphia, 1979

25. A. Fraker and C.D. Griffin, Ed., *Corrosion and Degradation of Implant Materials,* STP 859, ASTM, Philadelphia, 1985

26. J.O. Galante and W. Rostoker, Wear in *Total Hip Prostheses, Acta Orthop. Scand. Suppl.,* Vol 145, 1973

27. M. Ungethüm, Technologische und biomechanische Aspekte der Hüft- und Kniealloarthroplastik, in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 9, Verlag Hans Huber, Bern Stuttgart Wien, 1978

28. P. Griss *et al.,* Ed., Findings on Total Hip Replacement for Ten Years, in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 21, Verlag Hans Huber, Bern Stuttgart Wien, 1982

29. R. Schneider, Die Totalprothese der Hüfte (Ein biomechanisches Konzept und seine Konsequenzen), in *Aktuelle Probleme in Chirurgie und Orthopaedie,* Band 24, Verlag Hans Huber, Bern Stuttgart Wien, 1982

30. M. Semlitsch and B. Panic, Corrosion Fatigue Testing of Femoral Head Prostheses Made of Implant Alloys of Different Fatigue Resistance, in *Evaluation of Biomaterials,* G.D. Winter *et al.,* Ed., John Wiley & Sons, 1980

31. M.F. Semlitsch, B. Panic, H. Weber, and R. Schoen, *Comparison of the Fatigue Strength of Femoral Prothesis Stems Made of Forged Ti-Al-V and Cobalt-Base Alloys,* STP 796, ASTM, Philadelphia, 1983

32. M. Semlitsch and B. Panic, Ten Years of Experience with Test Criteria for Fracture-Proof Anchorage Stems of Artificial Hip Joints, *Eng. Med.,* Vol 12, 1983, p 185

Failures of Pipelines

R.J. Eiber and J.F. Kiefner, Battelle Columbus Laboratories

Introduction

PIPELINES have established the enviable record of having the fewest fatalities of any of the various modes of transportation. Failures do occur, however, for a variety of reasons. In this article, the known failure causes and failure characteristics of high-pressure long-distance pipelines will be summarized. In addition, the procedures used to investigate pipeline failures will be described. Finally, uses of fracture mechanics in failure investigations and in developing remedial measures will be described.

Characteristics of High-Pressure Long-Distance Pipelines

Three general types of pipelines exist: gathering lines, transmission lines, and distribution lines. Gas or crude oil gathering lines exist between a well and a treatment plant or collection point. These are usually relatively small-diameter lines (51 to 203 mm, or 2 to 8 in., in diameter) and operate at a variety of pressures, usually from 3490 to 8376 kPa (500 to 1200 psi). Cross-country transmission pipelines of over 1600 km (1000 miles) in length transport natural gas, natural gas liquids, liquid petroleum products, anhydrous ammonia, or crude oil. These are continuous lengths of pipe interrupted only by valves at roughly 16-km (10-mile) spacings and compression stations (gas lines) or pumping stations (liquid lines) at 80- to 160-km (50- to 100-mile) spacings. High-pressure lines generally operate at pressures up to 6895 kPa (1000 psi) and are made of steel pipes welded or mechanically coupled together. Since the 1940s, all of the lines have been assembled by welding. The third type of pipeline is a gas-distribution line that mainly transports natural gas within cities at pressures that vary from several tens of pounds per square inch to a few inches of water where a line enters a home. This article will not discuss failures associated with these latter lines, but will address only the high-pressure gathering and transmission lines.

The steels that have been used in the United States for line pipe range in yield strength from American Petroleum Institute (API) grade A (207 MPa, or 30 ksi) to API 5L-X70 (483 MPa, or 70 ksi). Generally, API 5L Specification steels with yield strengths above 359 MPa (52 ksi) are microalloyed and controlled rolled to achieve the desired strength and fracture properties. Pipeline wall thicknesses are based on the pressure in the line and on the allowable hoop stress levels. The allowable stress levels for gas pipelines vary from 40 to 72% of the specified minimum yield strength based on the population density in the area of the pipeline and are regulated by the U.S. Department of Transportation (DOT) in Title 49 of the Code of Federal Regulations (CFR) Part 192. Liquid petroleum pipelines are regulated by DOT in Title 49 CFR Part 195. Their maximum allowable hoop stress level is 72% of the specified minimum yield strength.

One of the reasons for the excellent safety record of the pipeline industry is that pipelines are pressure tested before being put into service. It is common to test pipelines hydrostatically to hoop stress levels of 90 to 105% of their specified minimum yield strength. In addition, this pressure is maintained for 8 to 24 h. This ensures that the pipeline, at the time it is put into service, contains no defects that might induce failure at its operating stress level. The ability to preservice test a pipeline to a high stress level is a unique characteristic of a pipeline. The lines are designed such that they can withstand this type of test without undergoing plastic deformation that would induce residual stresses or damage the pipe.

One of the failure problems unique to gas pipelines was the potential length of a failure. Because the pressurized gas contains an enormous amount of stored energy and because the energy-release rate upon rupture is often slow relative to the speed of a propagating fracture, a fracture may propagate along the pipe axis for a considerable distance. The longest fracture on record was brittle in nature and propagated for a distance of 13.3 km (8.3 miles). Thus, pipelines are among the few structures for which the extent of fracture propagation is of concern.

R.J. Eiber and J.F. Kiefner, Battelle Columbus Laboratories

Pipeline Failure Investigation Procedures

Location of the Fracture Origin. The general procedures used to investigate pipeline failures are no different from those for other types of structures. The first step is to locate the fracture origin. This may be difficult when fractured pipe is scattered along the line for several hundred feet or more, but it is not impossible.

A study of the fracture surface is necessary to locate the origin. Features of the fracture surface make it possible to trace the fracture back to the origin. There are two basic modes of fracture propagation: brittle and ductile. Figure 1 shows a brittle, or cleavage, fracture. As shown in Fig. 1(a), brittle fractures in pipelines exhibit herringbone or chevron markings on the fracture surface. These markings are continuous and relatively uniform. This fracture pattern is produced because the fracture initiates first at the midwall of the pipe, then extends out to the surface, causing the V-shaped markings. As the result of stress-wave patterns established by the rupture event, brittle fractures in pipelines commonly exhibit a sinusoidal fracture path along the pipe axis. More than one fracture can often be observed due to the excessive energy available in the gas. If a brittle fracture propagates in a stress field that contains significant bending, the center of the chevron may move off-center (Fig. 1b). Brittle fractures that result from bending moments exhibit the appearance shown in Fig. 1(c).

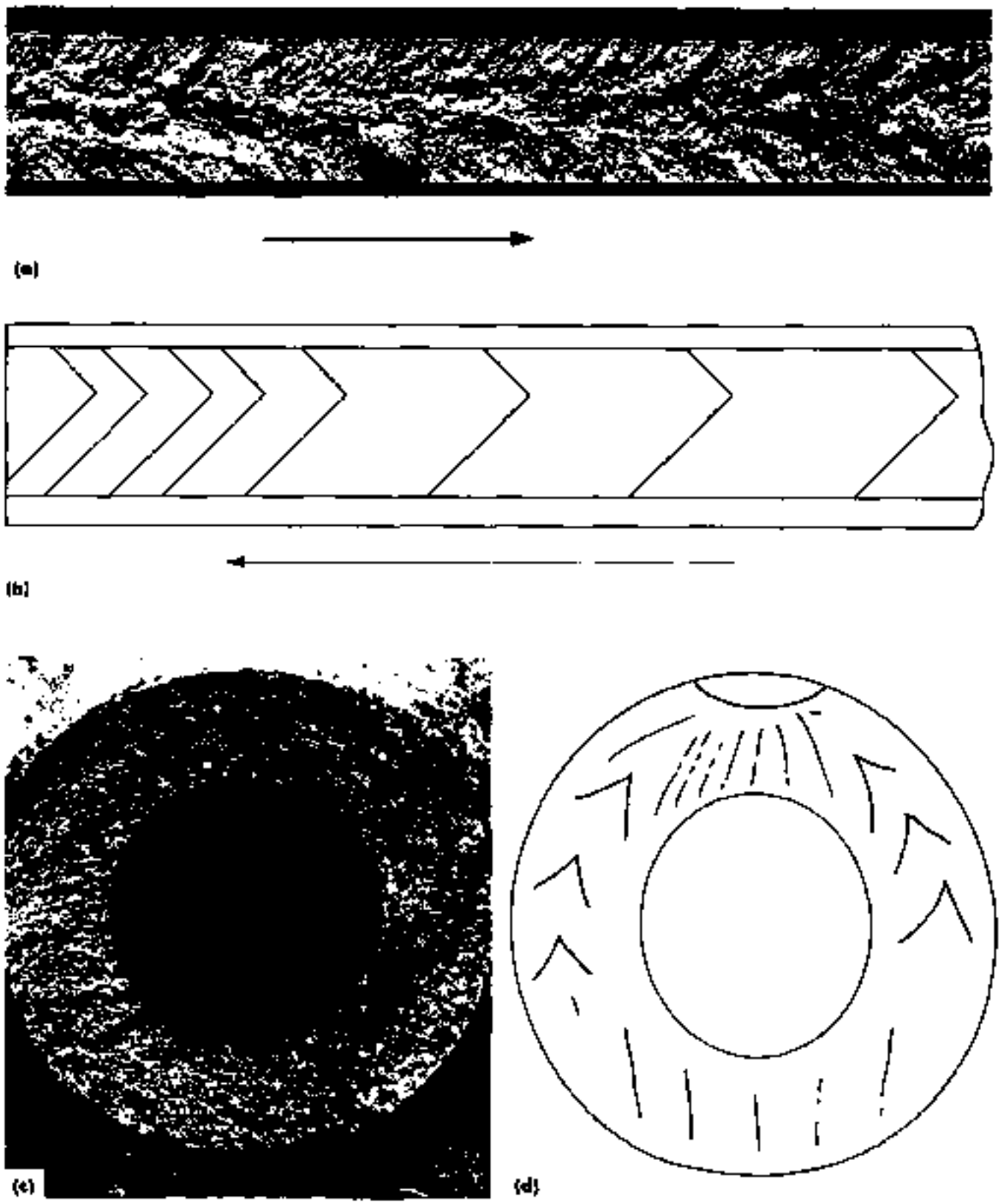


Fig. 1 Modes of propagation for brittle fractures. (a) Fracture resulting from primary tensile loading. (b) Fracture resulting from combined tensile and through-thickness bending load. The center of the chevron is offset from the pipe midwall. (c) Fracture from bending load with schematic of resulting chevron pattern. Arrows indicate fracture-propagation direction.

A propagating ductile-fracture surface is shown in Fig. 2(a). Under gaseous pressurization due to the excessive stored energy, these fractures can also propagate long distances. Under liquid pressurization, both brittle and ductile fractures are

usually short—of the order of a few feet. The fracture surface shown in Fig. 2(a) is typical of those found in API, grades B to X60, pipe manufactured by conventional rolling processes. The fracture surface is produced in a manner similar to a brittle fracture; that is, the fracture initiates first at midwall thickness and creates a deformation pattern on the fracture surface that can be likened to the chevron pattern of a brittle fracture but is not as distinct. This pseudo chevron pattern points toward the fracture origin and serves as a road map to locate the origin. Ductile fractures characteristically propagate axially along the pipeline, and in general, there is only one fracture.

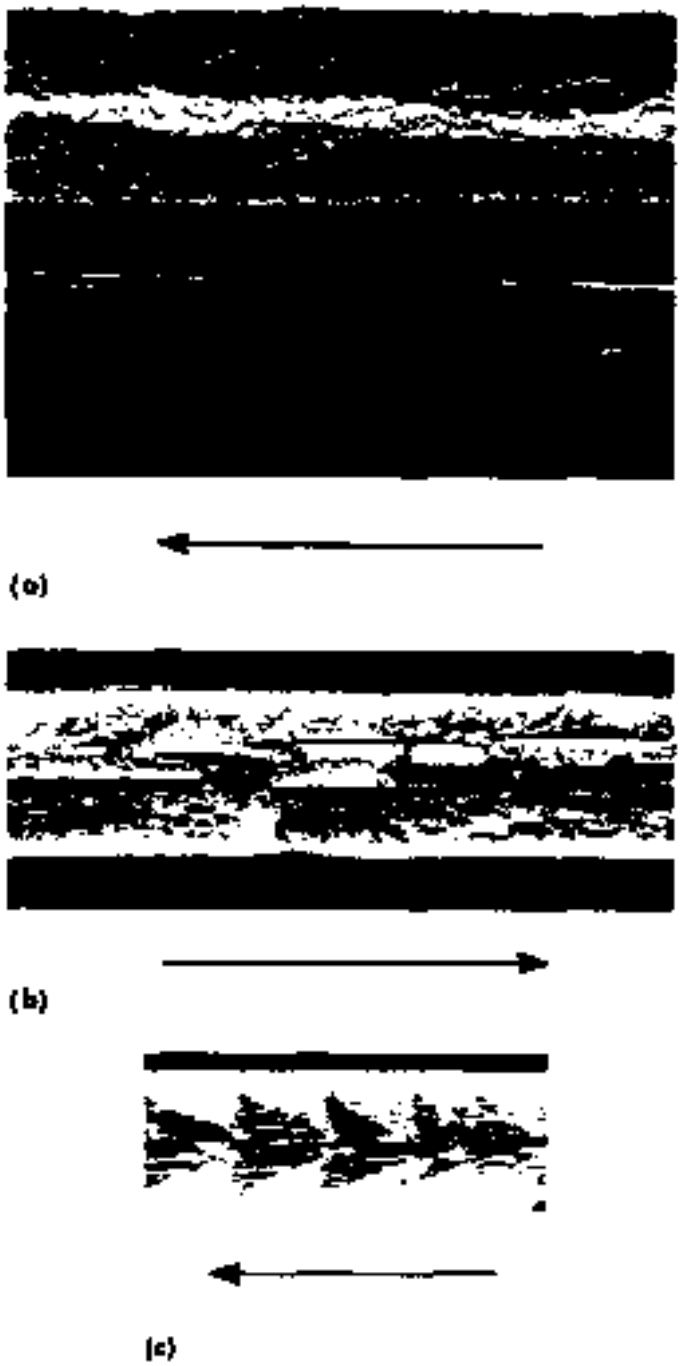


Fig. 2 Modes of propagation for ductile fractures. (a) Propagating ductile fracture. (b) Ductile fracture in separated material. (c) Ductile fracture with arrowheads, and illustration of chevron pattern. Arrows indicate fracture-propagation direction.

Other types of ductile-fracture surfaces may be encountered in a failure investigation. One type (Fig. 2b) exhibits separations. Separations have been observed in the newer controlled-rolled X65 to X80 pipe steels due to crystallographic texturing of the steel during processing. This causes the steel to exhibit secondary fractures in the plane of the plate in a brittle mode under the influence of triaxial stress created as the primary ductile fracture propagates. One characteristic of a separated fracture surface is that it is more difficult to determine the direction of crack propagation based on the fracture-surface appearance. Another propagating ductile-fracture surface is the so-called arrowhead fracture surface

shown in Fig. 2(c). The arrowheads on the fracture surface point away from the origin or in the direction of fracture propagation.

A third fracture appearance, called a tearing shear fracture, is found in conjunction with the arrest of ductile or brittle propagating fractures. Tearing shear fractures are pure shear fractures (Fig. 3), as opposed to the tensile fractures shown in Fig. 1 and 2. Tearing shear fractures occur at the arrest point, where an axially propagating fracture turns and spirals in a helical direction. Figure 4 shows an illustration of the location and configuration of tearing shear fractures. Tearing shear fractures occur because of the kinetic energy of the fracturing pieces of pipe and are therefore secondary to the main fracture.

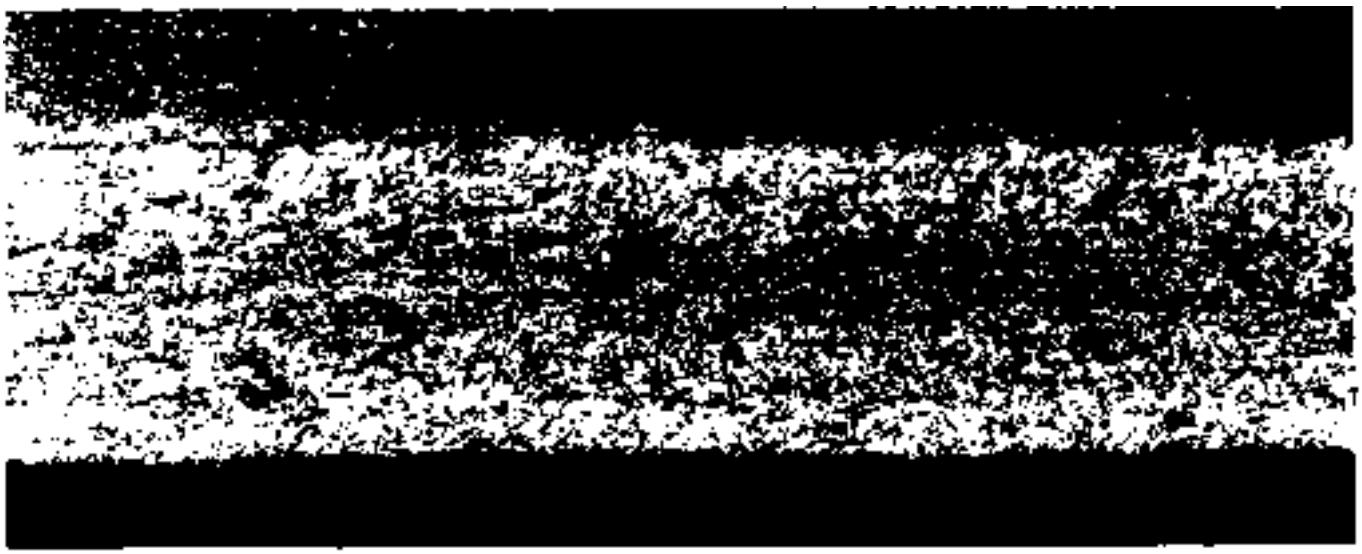


Fig. 3 Fracture surface of a tearing shear ductile fracture.

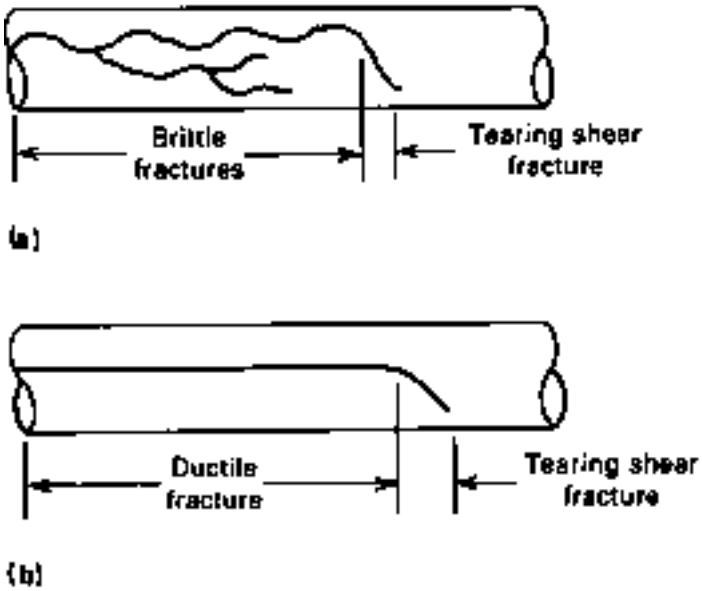


Fig. 4 Tearing shear fractures. (a) In brittle material. (b) In ductile material

Characteristics of an Origin. The origin of a fracture is a point at which there is generally only one fracture. The origin usually contains the defect that, in combination with the stress applied to the pipe steel, is responsible for the failure.

If the fracture propagation is brittle, it is possible that the fracture appearance will be brittle from the star-burst pattern at the origin to the tearing shear fracture at the arrest. Occasionally, the fracture at the origin will be ductile and then change to brittle as the fracture accelerates. The type of fracture that occurs is a function of the temperature of the pipe relative to the transition temperature of the steel.

Sample Selection. In failure investigations, a reasonable guess as to the cause of the accident can frequently be established in the field inspection. Based on the anticipated cause, it is possible to select samples logically for detailed examination. Normally, there are two objectives in selecting samples. The first is to select samples that help to define the cause of the failure, and the second is to select samples to investigate the mechanical and/or fracture-propagation properties of the pipe steel.

In selecting samples of the origin, large samples should be selected such that in flame cutting the heat of burning will not affect the fracture. Also, care should be taken to ensure that the fracture surface at the origin is not damaged during removal or shipment. Wire brushing of the origin fracture surface should not be permitted, because this removes the fine features of the fracture surfaces and eliminates the possibility of a fractographic examination in a scanning electron microscope. If the origin is associated with a weld, samples of the intact weld remote from the origin in the same length of pipe should be obtained to examine an unfractured section. If defects are associated with the weld, they may also be noted in the remote location. If an environmentally caused failure is suspected, consideration should be given to selecting and obtaining soil samples that are representative of the soil at pipe depth. Also, supplemental information regarding the cathodic-protection levels imposed on the pipe may be useful.

Samples selected for fracture-propagation property determination should be representative of the material that failed and should not be deformed or heated as a result of the failure. Samples measuring approximately 0.2 m^2 (2 ft^2) are usually removed for material-property determination.

Failures of Pipelines

R.J. Eiber and J.F. Kiefner, Battelle Columbus Laboratories

Causes of Pipeline Failures

One of the significant observations that have been made as a result of numerous failure investigations is that failures generally initiate at a defect, with the only exceptions being the few failures that have occurred from a gross overload, such as secondary stresses, internal combustion, or sabotage.

Table 1 lists the defects that have caused pipeline failures. The description of defects has been divided into two categories: those that cause failures in the preservice test and those that cause failures in service. The reason for the division is that the defects present in the pipe before the preservice test can generally be removed or detected by the preservice test if the test pressure is high enough.

Table 1 Types of defects that can cause pipeline failures

Causes of preservice test failures

- **Defects in the pipe body**
  - Mechanical damage
  - Fatigue cracks
  - Material defects

Longitudinal weld defects

- Submerged arc welds
  - Weld-area cracks
  - Incomplete fusion
  - Porosity
  - Slag inclusions
  - Inclusions at skelp edge
  - Off seam
  - Repair welds
  - Incomplete penetration
- Electric welds[(a)]
  - Upturned fiber cracks
  - Weld-line inclusions
  - Cold weld
  - Excessive trim
  - Contact burns

Field weld defects

Causes of service failures

- **Defects in the pipe body**
  - Mechanical damage
  - Environmental causes
    - Corrosion (external or internal)
    - Hydrogen-stress cracking
    - External stress corrosion cracking
    - Internal sulfide-stress cracking
    - Hydrogen blistering
  - Fatigue
  - Miscellaneous causes
    - Secondary loads
    - Weldments to pipe surface
    - Wrinkle bends
    - Internal combustion
    - Sabotage

Longitudinal weld defects

- High-hardness region[(b)]

(a) Electric flash welding, electric-resistance welding, or electric-induction welding without the addition of extraneous metals.
(b) These arc a problem in gathering lines carrying sour gas (gas containing H_2S).

Causes of Preservice Test Failures

As shown in Table 1, the defects that cause failures in line pipe in the preservice test can be broken down into three types, according to location: defects in the body of the pipe, longitudinal weld defects, or field girth weld defects.

Causes Associated With the Pipe Body. The three general causes of failure associated with the body of the pipe are mechanical damage, fatigue cracks, and material defects.

Mechanical damage consists of gouges or gouges and dents that have generally been produced by excavation or handling equipment during construction. The failure characteristics will be presented in the section “Causes of Service Failures” in this article because this is the leading cause of service failures.

Material defects may be laminations, laps, seams, holes, hot tears, and so on. A few failures have resulted from foreign material being rolled into the pipe skelp. This results in a local region where the wall thickness is reduced or completely penetrated, which is generally detected as a leak during the preservice test.

Investigation of failures caused by material defects normally consists of one or two metallographic sections through the defective area. These metallographic sections are examined in an optical (light) microscope to determine the nature of the defect and the general features of the microstructure. When inclusions or foreign particles are present, examination in the scanning electron microscope using an energy-dispersive x-ray analysis will usually identify the composition of the microstructural constituents. Frequently, a deep etch will show the flow pattern of the steel grains during hot rolling, which can help to identify the cause. Figure 5 shows a lamination that was inclined to the pipe wall, which effectively reduced the wall thickness and resulted in a slant fracture in approximately one-half of the wall thickness.

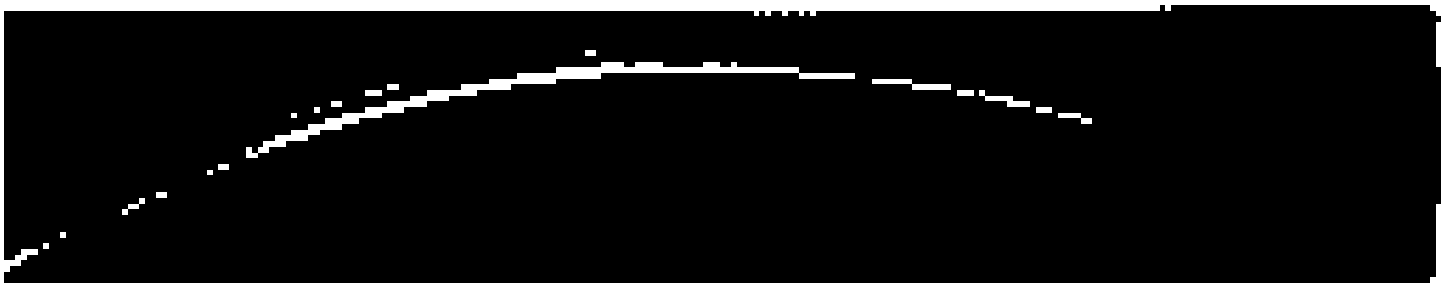


Fig. 5 Failure from lamination inclined with respect to the pipe wall.

These types of failures are generally related to (1) insufficient cropping of an ingot in the case of inclusions or a lamination in the steel, (2) foreign objects, such as parts of rolls or rolling equipment, being pushed into the steel during hot rolling, or (3) hot tears that form, particularly during piercing and rolling of relatively thin-wall large-diameter seamless pipe. All of these problems can be controlled by careful mill practice and the application of adequate nondestructive inspection.

Fatigue cracks have occurred in the body of the pipe and adjacent to longitudinal welds. Figure 6 shows an example of a fatigue crack that occurred alongside the longitudinal weld. In this case, the cracks initiated at both the inside and outside surfaces of the pipe at the edge of the longitudinal weld. These cracks were detected in the preservice test of the line. Fatigue cracks have also been observed in the base metal. In these cases, the cracks have initiated from the outside surface in a region of mechanical damage on the pipe surface, which is usually produced during rail shipment.

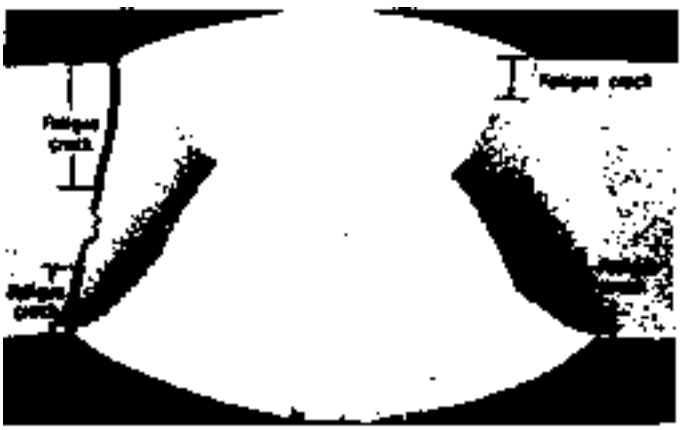


Fig. 6 Typical example of fatigue cracking adjacent to a longitudinal weld.

Because these cracks were detected in pipe that had not been in service, the shipping of the pipe was subjected to investigation. The investigations have usually revealed that the fatigue cracks were associated with rail shipment over distances of 1125 to 1600 km (700 to 1000 miles). In pipe with diameter-to-thickness ratios of 80 or greater, the cracks were generally observed to be within 3 m (10 ft) of the end of a pipe length.

The fatigue cracks had been produced during rail shipment of the pipe as a result of the combination of static stresses produced in the lower layers of stacked pipe by the static weight of the pipe above and cyclic stresses produced by the vertical acceleration of the rail car. When mechanical-damage regions were evident, the pipe had usually contacted the rail car, frequently on a rivet head, which served as a stress concentrator.

A recommended practice for the loading of pipe in rail cars was developed by the API (Ref 1). This publication defines loading practices for rail car shipment to protect pipe from this type of failure. Since the introduction of this recommended practice, failures of this type have been significantly reduced.

Causes of Failures in Longitudinal Welds. Five types of longitudinal welds have been used to produce pipe. These are single- and double-submerged arc welds, electric-resistance welds, furnace butt welds, flash welds, and lap welds. Currently, only double-submerged arc welds and electric-resistance welds are used. The following are brief descriptions of the welding processes and common causes of failures.

Double-Submerged Arc Welds. These welds are made with two passes (hence the term double): one on the inside and one on the outside. Either the inside or outside pass can be made first, depending on the producer. The weld passes can be made with one to four electrodes. The most common causes of failures in double-submerged arc welds are weld cracks, toe cracks, and lack of penetration.

Cracks may appear in either the weld metal or the heat-affected zone (HAZ). In the weld metal, the cracks are typically in the center of the weld metal parallel to the columnar structure of the weld metal, and they generally occur when the weld metal is subjected to a high stress or restraint during solidification. Figure 7 shows cracks in the HAZ; these cracks are usually referred to as toe cracks at the edge of the weld reinforcement. Generally, these result when the edges of the plate are not adequately curved when the plate is formed into pipe. As the pipe is pressurized or cold expanded during manufacture, high strains occur at the edge of the weld reinforcement, causing cracks similar to those shown in Fig. 7. An excessive number of inclusions in the steel at the edge of the weld reinforcement have been observed to assist the formation of toe cracks.

Fig. 7 Toe cracks in the HAZ of a double-submerged arc weld.

Another type of defect in double-submerged arc welds results from lack of penetration, in which the inside and outside beads are too shallow and do not overlap. This leaves a nonfused portion, thereby reducing the effective wall thickness.

Electric-Resistance Welds and Flash Welds. These welding processes locally heat the edges of the plate to a suitable forging temperature; the edges are then pushed together, upsetting the wall thickness and forming a bond. In an electric-resistance welded seam, the weld is formed as the pipe passes through the welding station, with the edges making a vee before welding. In the flash-welding process, a complete length of pipe is welded at one time as the edges of the plate are pushed together (flash welds are no longer used to produce pipe). In either type, after welding, the upset material is trimmed off.

Causes of failures common to both electric-resistance and flash welds are upturned fiber cracks, cold welds, and insufficient upset. Figure 8 shows the appearance of a metallographic section through an electric-resistance weld that contains a hook crack, or upturned fiber crack. Although it is somewhat difficult to see in Fig. 8, the steel has been upset such that the original plane of the pipe wall thickness extends through the thickness at the bond line. If inclusions (usually manganese sulfide inclusions) exist in the wall thickness before the upsetting, they may crack during the upsetting process, creating the upturned fiber defect. In this weld, it can also be observed that the upset removal on the inside of the weld was not uniform and that a notch was thereby created.

Fig. 8 Appearance of a hook crack in an electric-resistance weld. The thin white layer is a nickel coating that was applied to the fracture surface.

Figure 9 shows the origin fracture surface of one cause of failure in electric-resistance welded pipe. The smooth region is the result of inadequate upsetting and/or heating of the pipe edges (a cold weld). A stitched surface, or alternately bonded region, may also be observed at fracture origins associated with inadequate upsetting or too little heat.

Fig. 9 Appearance of an unwelded section in electric-resistance welded line pipe.

Lap Welds. These welds are made by forging a weld between overlapped plate edges. The weld is made using forging rolls that form a low-angle weld through the thickness. Causes of failure in this type of weld are oxides on the bond line and lack of fusion due to too low a temperature or too little upsetting pressure. These defects reduce the effective thickness of the weld.

Regardless of the welding process used, the investigation of weld failures usually involves use of low-power (2 to 30×) magnification in either the optical microscope or the scanning electron microscope to locate the origin. Advantages of the optical microscope are that the sample orientation is not lost and that a relatively long length of weld fracture can be examined. Once an origin is identified, scanning electron microscopy (SEM) examination may be helpful in identifying fracture features. Once the origin is identified, metallographic sections across the weld are prepared. In addition, sections across the intact weld from other locations in the same pipe length can be helpful. Intact weld sections can be used to:

- Identify whether repairs have been made to the weld
- Determine whether a heat treatment has been adequate at the origin
- Determine the general configuration of the weld and bond lines before the failure occurred

The cause of failure can generally be determined if the manufacturing process is understood and the origin is properly identified and examined.

Causes of Failure in Field Girth Welds. Relatively few girth welds have failed either in service or in the preservice hydrostatic test. One of the main reasons for this is that girth welds are not subjected to high stress due to internal pressure. Girth welds may, on occasion, experience high stresses from such sources as soil movement, subsidence, or severe thermal fluctuations. Except for such unusual sources, the longitudinal stress in a pipeline is about 30% of the hoop stress or about 22% of the specified minimum yield stress (SMYS) for a line operating at 72% SMYS.

Typical manual metal arc weld defects, such as lack of penetration, hollow bead, porosity, alignment problems (high-low), weld-metal cracks, and underbead cracks, have caused leaks or ruptures in welds. Because the failure modes and investigation procedures are similar for all cases, only underbead cracks will be discussed.

Underbead Cracks. One situation that can cause a failure is an underbead crack. Underbead cracks occur when regions of high hardness form because of rapid cooling of the HAZ. High-hardness regions may result from welding pipes of two

different wall thicknesses or a pipe and a heavy wall thickness fitting without adequate preheating. On occasion, due to fit-up or alignment problems, it is necessary to add a weld bead to the internal surface to provide a complete weld. Because this requires a welder inside the pipe, it is difficult to preheat the weld area. Weld beads deposited under these conditions with cellulosic electrodes may result in rapid cooling of the HAZ with a tendency to produce untempered martensite. The untempered martensite is highly susceptible to cracks from hydrogen-stress cracking, producing underbead cracks. The same condition can result if welds are made with cellulosic electrodes between pipes that are at low temperatures, providing high-hardness HAZs due to the rapid cooling.

Figure 10 shows a metallographic section through a girth weld between a pipe and a heavy flange. The backup bead on the inside surface produced an HAZ with a maximum Knoop hardness of 405. A semicircular crack (underbead crack) occurred roughly parallel to the weld bond line. This crack caused the weld to fail during pressure testing.

Fig. 10 Underbead cracks (hydrogen cracking) in a girth weld.

Causes of Service Failures

Table 1 also lists the causes of service failures. Each category will be discussed to identify its characteristics, and investigative steps will be presented for selected causes.

Defects in the Pipe Body. Mechanical damage is the leading cause of failure in pipelines, accounting for more than one-half of all service failures. Generally, mechanical damage occurs from excavation equipment that accidentally contacts the pipe. The damage usually consists of a gouge and dent in the pipe surface. After a failure, the dent is usually difficult to detect because the pressure in the pipe tends to remove it. Mechanical damage has resulted in penetration of the pipe wall thickness, producing an immediate leak, or in damage to the pipe, creating a potential for future failure as the dented area continues to deform due to creep from increased pressure. Frequently, the gouge appears to be relatively shallow. However, because of the strain imposed by the outward deflection of the dent, cracks may form in the cold-worked layer that are not visible on the gouged surface because of the smeared metal.

Figure 11 shows a photograph of a 406-mm (16-in.) diam API, grade X46, pipe that has been mechanically damaged. The damage is located in regions 1, 2, and 3. The chevron markings on the fracture surface locate the fracture origin between regions 2 and 3. The pipe surface was indented approximately 6 mm (0.25 in.) after the failure. In all probability, a much deeper dent existed at the time the damage was created.

Fig. 11 Mechanical damage in a 406-mm (16-in.) diam API, grade X52, pipe. S, shear- or ductile-fracture surface; C, cleavage- or brittle-fracture surface. See also Fig. 12, 13, and 14.

The presence of a gouged surface will usually immediately identify the primary cause. Metallographic sections through the mechanical damage are necessary to identify the extent of the cold-worked steel, the depth of the gouges, the extent of cracking (if any), the hardness of the cold-worked steel, and whether or not foreign particles are attached. The identification of foreign particles, if they exist on the gouged surface, may be useful in identifying the source of the mechanical damage through chemical analysis of the foreign material (the authors, however, have not been successful in identifying the specific damaging equipment). Examination of the fracture surface in a scanning electron microscope may be helpful in identifying the nature of the fracture, that is, transgranular versus intergranular and whether or not any evidence of fatigue can be observed. Most of the mechanical-damage failures have had transgranular fractures, but

intergranular fracture surfaces have been observed and are believed to have resulted from hydrogen-stress cracking, with the hydrogen resulting from the cathodic potential applied to the pipeline for corrosion control. A problem in attempting to use the scanning electron microscope is that the fracture surfaces are often too corroded to permit examination.

Figure 12 shows a macrograph of one side of the fracture at region 3. It shows the fracture edge and the outside pipe surface (cracked surface). It should be noted that the section has been nickel coated, which is the white coating. The overall fracture in this region is a slant or ductile fracture, as evidenced by the deformation along the fracture edge. The slant cracks evident on the outside surface are believed to have formed as a result of the cold-worked region on the surface having been subjected to high strains by the outward deformation of the dent from the pressure in the pipe. These cracks are very significant the depth of the gouge measured by the reduced wall thickness is approximately 0.2 mm (0.008 in.), while the maximum crack depth is 0.85 mm (0.033 in.). The crack that initiated the fracture was estimated to be 1.47 mm (0.058 in.) deep. Figure 13 shows a micrograph of the cold-worked outside pipe surface that illustrates the normal ferrite-pearlite microstructure of the carbon steel. Also shown is the cold-worked surface with a highly elongated grain structure, which has resulted from the mechanical damage.

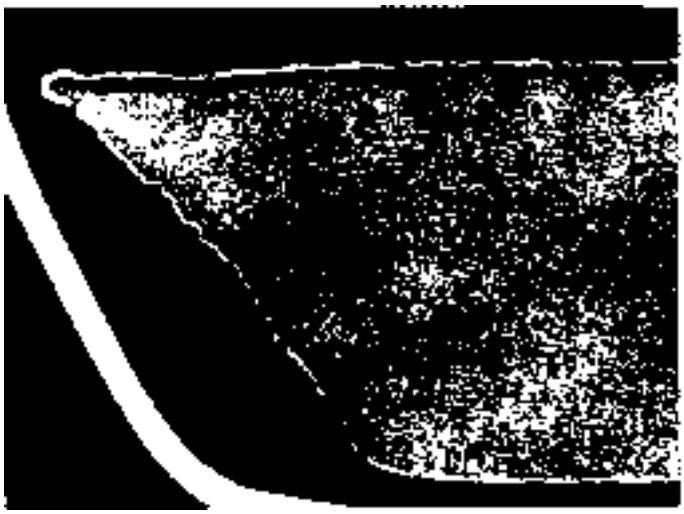


Fig. 12 Macrograph of region 3 from Fig. 11. 5×. See also Fig. 13 and 14.

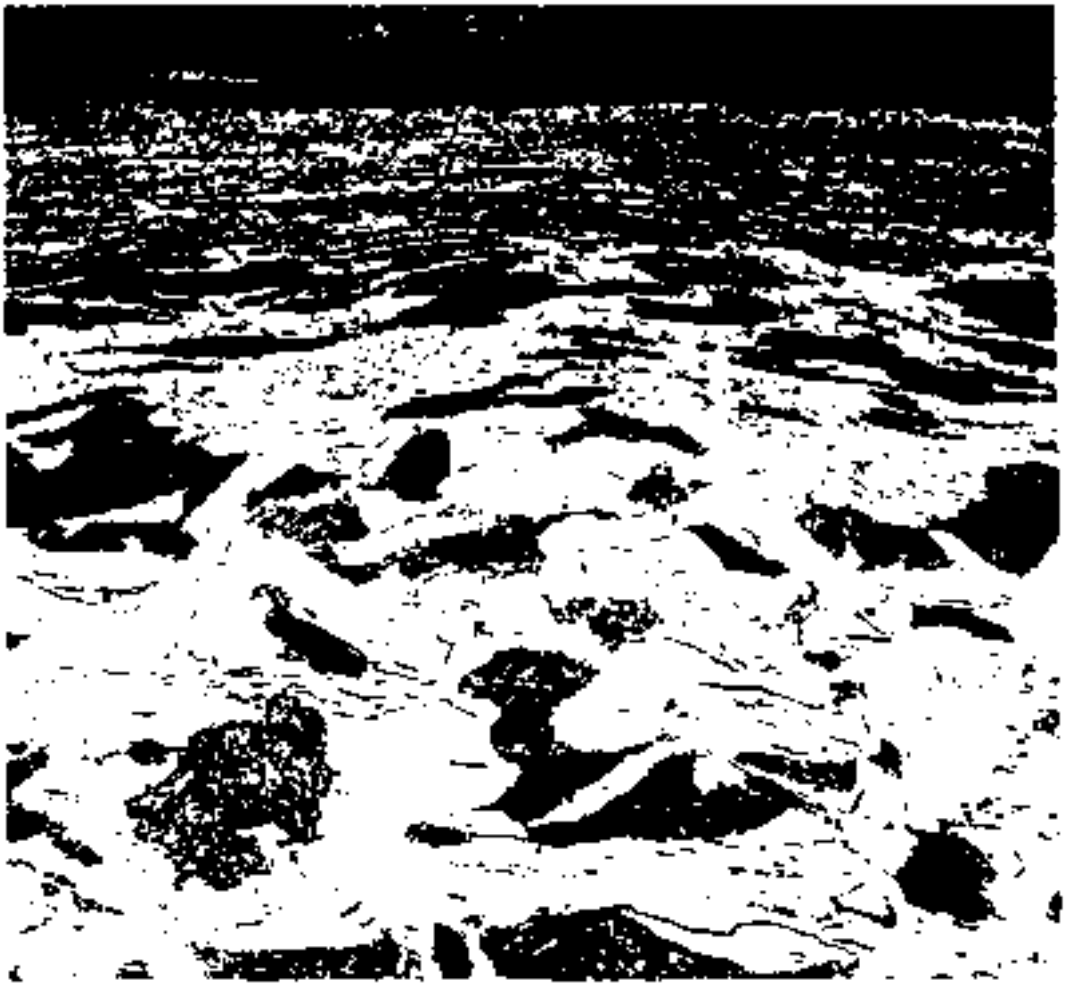


Fig. 13 Micrograph of cold-worked pipe surface at region 3 in Fig. 11. Etched with nital. 500×. See also Fig. 12 and 14.

Figure 14 illustrates another area of the surface in region 3 that shows a crack and the cold-worked structure. Near the surface in Fig. 14, there is evidence that, in addition to the cold work, the steel was heated to a temperature above the Ac_1 temperature (707 °C, or 1305 °F) and cooled at a fast rate. This layer of material has a metallurgical structure different from that of the severely cold-worked layer and different from that in the pipe. This structure is probably a bainitic structure, which is intermediate in hardness between ferrite and martensite. The heating was produced by friction between the pipe wall and the object that made the gouge. The rapid cooling is produced by the surrounding pipe steel. In some instances of mechanical damage, untempered martensite has been observed on the gouged surface.

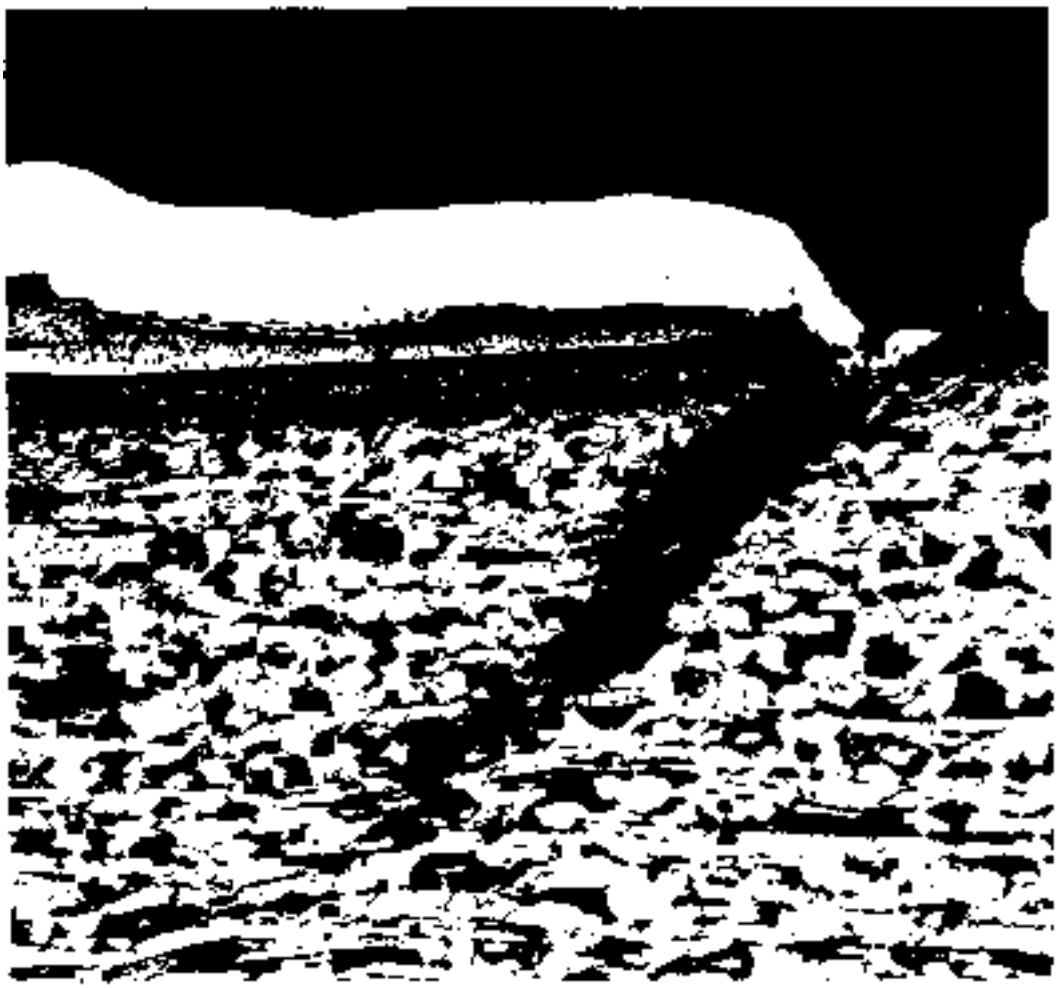


Fig. 14 Micrograph showing foreign material and secondary surface crack at region 3 in Fig. 11. Etched with nital. 100×. See also Fig. 12 and 13.

An additional feature that is sometimes observed can be seen in Fig. 14. The light gray area just beneath the white nickel plating to the left of the crack is an area of extremely hard steel that is foreign to the pipe wall. The hardness is 870 HK and probably came from the hard surfacing materials used for construction equipment bucket and blade edges. This apparently became friction welded to the gouged surface during the damage process.

Failure of mechanically damaged areas can occur from a number of causes and may or may not occur at the time damage is done. The possible causes of failure are:

- The damaged area may crack due to the high strains imposed by the dent
- The pressure fluctuations in the pipeline may cause a low-cycle high-stress fatigue failure
- The normal cathodic protection applied to the pipe to prevent corrosion may charge the highly stressed steel with sufficient hydrogen to cause a hydrogen-stress crack

When a region of mechanical damage is detected, action should be taken to remove it or to reinforce it through the application of a sleeve. If a sleeve is applied, the gap between the sleeve and pipe should be filled with a compound, such as epoxy or polyester, that will prevent further outward deformation of the dent.

Environmental Causes of Pipeline Failures. There are a number of environmental causes of pipeline failures. The major causes are general corrosion, hydrogen-stress cracking in hard areas, external stress-corrosion cracking (SCC), internal sulfide-stress cracking, and hydrogen blistering. The failures break down into two general classes: those occurring from the inside surface of the pipe and those occurring from the outside surface of the pipe. Generally, the failures initiating on the inside pipe surface are associated with gathering lines containing moisture, hydrogen sulfide, carbon dioxide, or other corrosive impurities. The causes associated with the inside pipe surface are internal corrosion, sulfide-stress cracking, and hydrogen blistering. The remaining causes are associated with the external pipe surface. Investigation of this general category of failures is similar; therefore, only selected examples of these failure causes will be described.

Hydrogen-Stress Cracking. Figure 15 shows the surface of a fracture that was initiated from a hydrogen-stress crack in line pipe. The features of the fracture surface are:

- Extremely brittle fracture surface with very fine grains
- Fracture origin generally at the outside pipe surface or subsurface
- No out-of-plane deformation of the pipe wall thickness at the origin

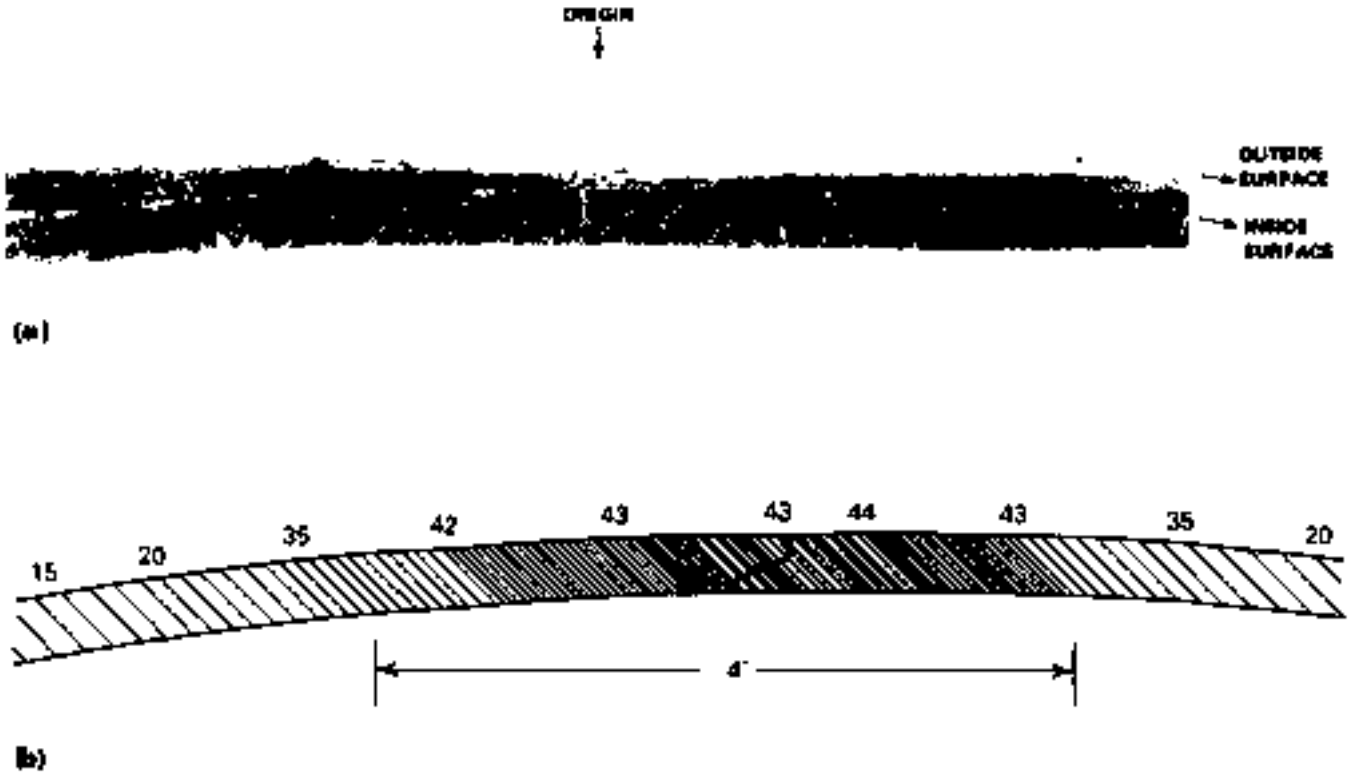


Fig. 15 Appearance of hydrogen-stress cracks in line pipe. (a) Fracture surface. 0.5×. (b) Results of Rockwell C hardness traverse of pipe

The investigation consists of examining the fracture at low-power magnification to locate the origin. It hydrogen-stress cracking is suspected, hardness impressions are made on the surface to determine if a region of high hardness is associated with the origin area. The surface-hardness impressions generally identify the extent of the high-hardness region, which assists in the selection of metallographic specimens. In one failure, no region of high hardness was found on the surface, but alloy segregation at midwall thickness created a region of high hardness. Usually, two metallographic sections are prepared: one section through the fracture origin and one outside of the high-hardness region. The section outside of the suspected region of high hardness is taken for comparison purposes to determine the normal microstructure of the pipe steel.

Examination of the origin metallographic section usually reveals that this is a region of the plate that contains transformation products that may range from bainite to untempered martensite. These local regions of high hardness are believed to have been quenched during hot rolling. They generally exhibit a hardness gradient from the ferritic-pearlitic structure to the region of high hardness. Microhardness readings should be taken to determine hardness level. Failures have occurred with hardnesses from 30 to 50 HRC in API, grade X52, steels. Generally, no secondary cracks have been observed on the fracture surfaces of these failures. Also, the fracture is usually partly intergranular and partly transgranular.

If hydrogen-stress cracking is suspected, information on the cathodic potential applied to the pipe for corrosion protection should be obtained as well as information on the coating condition. The atomic hydrogen necessary to cause the fracture is believed to result from the cathodic potential applied for corrosion control, which reaches the pipe surface through pores (commonly termed holidays in the pipeline industry) in the coating.

Hydrogen-stress cracks are caused by the combined factors of a very high-strength high-hardness steel, an applied hoop tensile stress from the pressure in the pipe, and atomic hydrogen in the steel from the cathodic-protection current applied to the pipe. Failures of this type may be prevented in several ways if the hard spots can be located. Success has been achieved by using internal nondestructive inspection devices based on the magnetic flux-leakage principle to locate hard spots. Once the hard spots are located, they can be removed, shielded to prevent the cathodic current from reaching them, or tempered to reduce the hardness. To date, failures have been observed only in areas with hardnesses exceeding 30 HRC. Also, hard spots are now prohibited by the API (Ref 2) and must be inspected for during the production of pipe. One of the best inspection techniques is to examine the pipe surface visually for flat spots.

Stress-Corrosion Cracking. Figure 16 shows the fracture surface of a stress-corrosion crack in line pipe. Failures of this type have primarily occurred on gas-transmission pipelines immediately downstream of compressor stations, where the pipe temperatures are the highest. The characteristic feature is a series of elliptical crack shapes, which initiated from the outside surface. Normally, a black oxide layer can be observed on the elliptical crack surfaces. Also, the cracks frequently overlap in the axial direction and link together. Figure 17 shows secondary cracks on the pipe surface. Information on the pipe-to-soil cathodic-potential level, pipe temperature, pressure, and pressure fluctuations in the pipeline should be obtained.

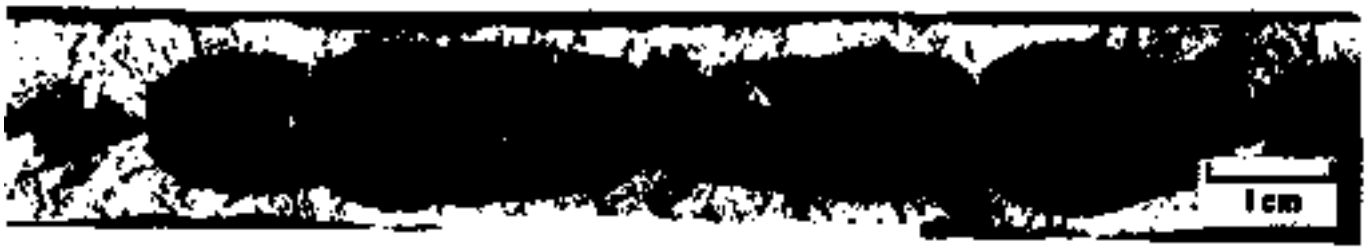


Fig. 16 Appearance of fracture surface of line pipe that failed by SCC. Actual size. See also Fig. 17, 18, and 19.

Fig. 17 Secondary cracking in pipe surface adjacent to stress-corrosion cracks shown in Fig. 16. Actual size. See also Fig. 18 and 19.

The investigation should begin with a low-power examination of the fracture surface to select locations for metallographic sections. The nature of the cracks can be determined by preparing several matched metallographic sections across the fracture surface. Also, analysis of any corrosion product on the elliptical crack surface using an energy-dispersive x-ray spectrometer may be helpful in defining the corrosive medium. In a high percentage of the failures, iron carbonate or bicarbonate has been detected on the crack surface (this environment is produced at a pore in the pipe coating as the hydroxide created by the cathodic potential is converted by the CO_2 in the soil to a carbonate-bicarbonate environment). Scanning electron microscopy examination of the fracture surface can also be helpful in identifying whether the fracture is transgranular or intergranular.

Usually, the fracture surfaces are primarily intergranular, and a number of branched secondary fractures are often evident in the metallographic sections (Fig. 18). Figure 19 shows an unetched section of a stress-corrosion crack that illustrates the branching and intergranular nature of the fracture.

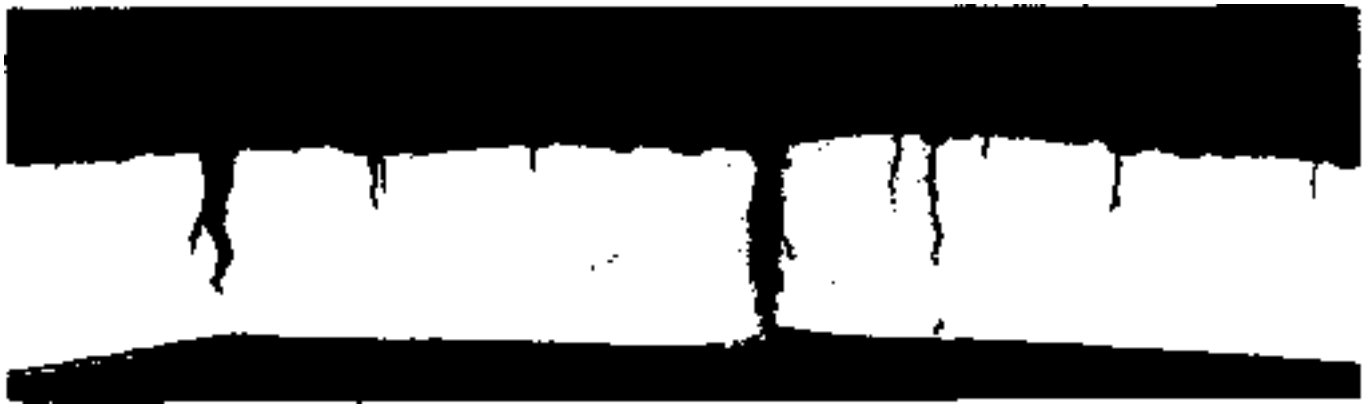


Fig. 18 Macrograph of unetched section through stress-corrosion cracks in line pipe. See also Fig. 16, 17, and 19.

Fig. 19 Polished section through stress-corrosion cracks showing branching and intergranular cracking. See also Fig. 16, 17, and 18.

The conditions that have been responsible for a high percentage of the stress-corrosion cracks in line pipe are the carbonate-bicarbonate environment, the pipe surface in the critical potential range for cracking (-0.6 to -0.7 V versus the saturated calomel electrode, SCE, for a carbonate-bicarbonate environment), and the applied tensile stress.

The carbonate-bicarbonate environment is believed to be formed by the cathodic potential at pores in the coating where the current penetrates to the pipe surface. The cathodic potential also creates the critical potential for SCC. The hoop stress in the circumferential direction of the pipe is created by the internal pressure in the pipe.

There is no simple control for external SCC, because one of the three conditions necessary for cracking must be prevented. This means keeping the cathodic potential outside of the critical range for cracking, which is difficult, if not impossible, at all points on a long pipeline.

Keeping the temperature as low as possible slows the reaction rate, which helps to prevent cracking. Many of the pipeline failures have been located in the first 16 km (10 miles) downstream of a gas compressor station, which is the high-temperature region on a gas pipeline. Applying a good coating to the pipe surface is also a deterrent, as is shot peening the pipe surface to remove mill scale and to provide a good surface for coating bonding. Small pressure fluctuations have been found to promote cracking. Thus, pressure control is helpful, but is difficult to achieve in most situations.

Corrosion failures can initiate from either the inside or the outside surface of the pipe. Corrosion on the inside is generally associated with water in the pipe, which most commonly occurs in gathering pipelines where H_2S and CO_2 may also be present. Corrosion frequently takes place in the form of nonuniform metal loss whether it is on the internal or the external surface of the pipe. Both external and internal corrosion tend to be concentrated on the pipe bottom. General corrosion results in loss of metal over large areas, and pitting corrosion results in pits, which may be completely isolated or in overlapping arrays.

At the origin of a corrosion failure, the goal is to determine the nature of the corrosion and to rule out other environmental causes. In this discussion, rather than focusing on the investigation of the corrosion (which is covered in other discussions), the emphasis will be on the use of fracture mechanics to estimate the failure pressure of the corrosion flaw to verify that the failure resulted from the observed corrosion defect. This is also helpful in evaluating corrosion-defect severity when corrosion is detected on an operating pipeline.

An example of a corrosion flaw in 610- × 9.5-mm (24- × 0.375-in.) API, grade B, pipe will be described. The measured yield strength of the pipe was 336 MPa (49 ksi).

The corrosion defect that caused the failure is shown in Fig. 20. This defect occurred on the outside bottom surface of the pipe. The hoop stress in the pipe at the time the corroded area failed was 173 MPa (25 ksi). The failure was a rupture, with the fracture propagating a total axial distance of 5 m (17 ft).

Fig. 20 Corroded area in 610-mm (24-in.) outside-diameter × 9.5-mm (0.375-in.) wall-thickness API, grade B, line pipe. See also Fig. 21 and 22.

The predicted failure pressure for the corroded area is based on treating the loss of metal as a part-through or surface flaw in the pipe (Ref 3). A grid is drawn on the corroded area, and remaining wall-thickness measurements are made from which a contour map of remaining wall thickness can be developed (Fig. 21). The next step is to develop a profile of the remaining wall thickness along the fracture edge (if this were a prediction being made for a non-failed corroded area, the minimum remaining thickness path through the corroded region would establish the flaw profile). Figure 22 shows the measured profile in an exaggerated view.

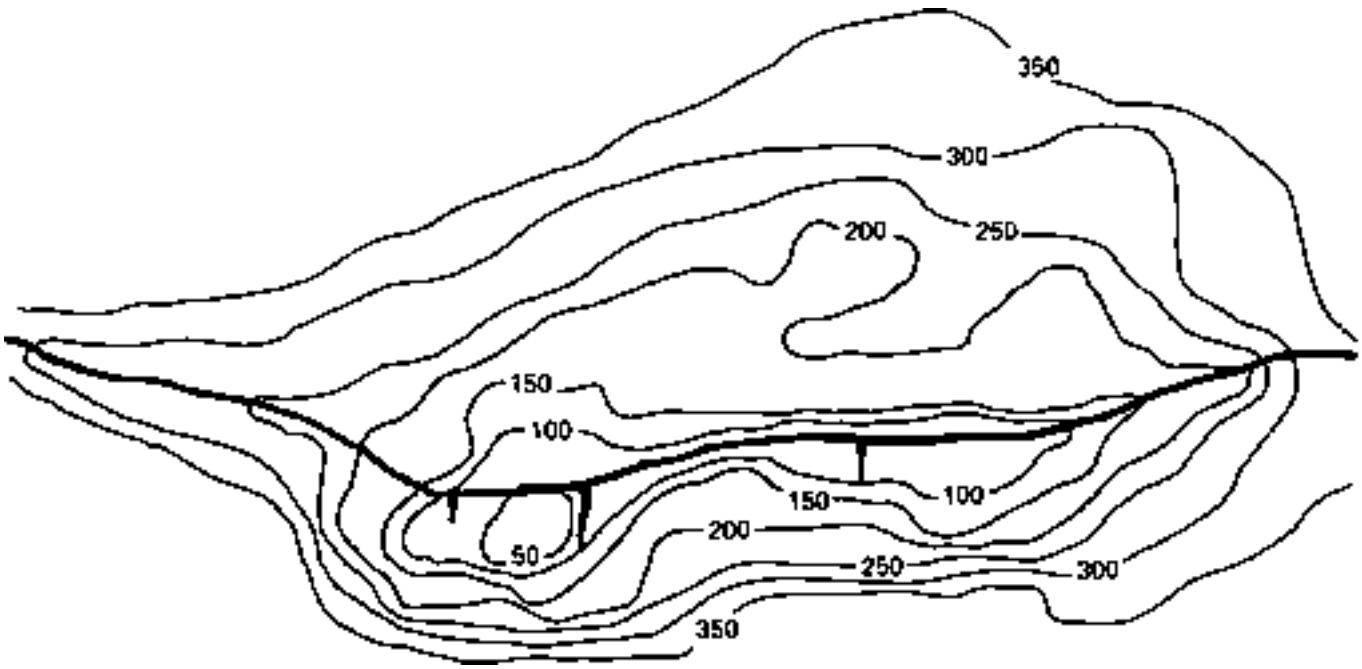


Fig. 21 Contour map of remaining wall thickness in corroded areas shown in Fig. 20. Dimensions are remaining wall thickness in mils. Fracture path is indicated by the heavy line. See also Fig. 22.

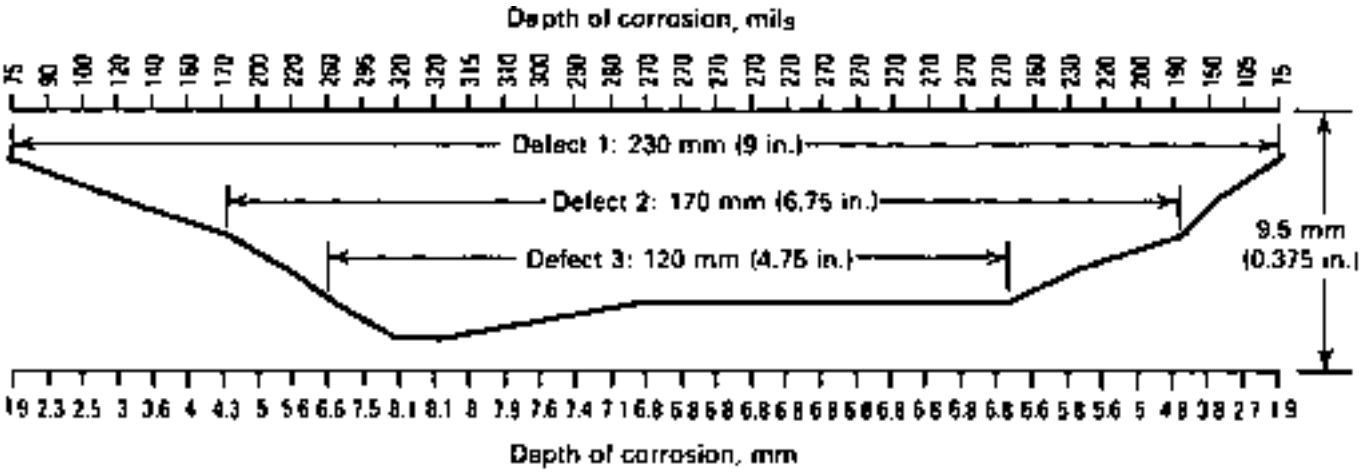


Fig. 22 Exaggerated depth profile of corroded area shown in Fig. 20 and 21.

Once the profile is available, the failure pressure can be predicted from Eq 1 (Ref 3):

$$s_{hp} = \bar{s}\,\frac{1 - A/A_o}{1 - (A/A_o)M_T^{-1}} = \bar{s}\,M_p^{-1} \qquad \text{(Eq 1)}$$

where σ_{hp} is the failure stress of the part-through flaw; $\bar{\sigma}$ is the flow stress of the steel, which has been taken as the yield strength plus 70 MPa (10 ksi); A is the area of the longitudinal profile of missing wall thickness; A_o is the area of full wall thickness for the length of the corrosion flaw; M_T is the bulging correction for an axial through-wall flaw in pressurized pipe; and M_p is the bulging factor for the part-through flaw.

Because corroded areas generally exhibit gradual transformation to full wall thickness along the critical axial profile, it is often difficult to preselect the axial length of the corroded region, which is termed the effective 2*c* length. By making several calculations of strength based on various trial values of crack length, one can usually obtain a minimum predicted failure stress level representing the effective crack length. This was done for the corroded region shown in Fig. 22. Using the trial lengths of 228.6, 171.5, and 120.65 mm (9, 6.75, and 4.75 in.), the values shown in Table 2 predict failure stresses for the three assumed flaw lengths.

Table 2 Controlling parameters and predicted failure modes for three assumed defect lengths

Defect 2*c* length		M_T^{-1} (a)	**A**		M_p^{-1} (b)	**σ_{hp}**		Predicted failure mode
mm	**in.**		**cm²**	**in.²**		**MPa**	**ksi**	
228.6	9.0	0.390	13.5	2.09	0.503	204	29.6	**Rupture**
171.5	6.75	0.485	11.2	1.74	0.456	185	26.8	**Leak**
120.65	**4.75**	**0.610**	**8.6**	**1.34**	**0.455**	**184**	**26.7**	**Leak**

(a) M_T, bulging correction for an axial through-wall flaw.
(b) M_p, bulging factor for part-through flaw

The minimum of these three values is a predicted failure stress level, σ_{hp}, of 184 MPa (26.7 ksi) corresponding to the assumed 120.65-mm (4.75-in.) defect length. This value is 6% higher than the failure hoop stress of 174 MPa (25.2 ksi).

The mode of failure predicted for the 120.65-mm (4.75-in.) flaw in the 9.5-mm (0.375-in.) wall-thickness pipe is a leak. A leak is predicted because $M_p^{-1} < M_T^{-1}$ (0.455 versus 0.610). These M factors can be thought of as stress-concentration factors. When a part-through flaw fails, it becomes a through-wall flaw. Thus, if the failure stress, which is proportional to M_p^{-1}, for a part-through flaw is higher than for that same length through-wall flaw, the part-through flaw would be

predicted to rupture. The reverse would predict a leak. Thus, if $M_p^{-1} > M_T^{-1}$, a rupture is predicted, and if the reverse, a leak. In the present case, the prediction is a leak, but a rupture occurred. In this case, however, the effective 120.65-mm (4.75-in.) flaw is in a region of much less than full wall thickness, as can be seen in Fig. 22. Computing the failure conditions for the 120.65-mm (4.75-in.) flaw length in a 5.1-mm (0.2-in.) average wall thickness results in an $M_T^{-1} = 0.8$ as contrasted with an $M_p^{-1} = 0.503$, which predicts a rupture.

The severity of a corrosion defect can be predicted using a fracture mechanics approach for line-pipe steels. This can be used to make sure the origin identified is the origin responsible for the failure, because the predicted failure pressure should be within ±15% of service pressure at the time of failure, and to help in assessing the remaining strength of corroded pipe before failure.

The chances of this type of failure can be minimized by monitoring loss of wall thickness through nondestructive inspections from the inside of the pipe, using magnetic flux leakage devices as a means of detecting corroded areas. Monitoring the pipe-to-soil potential level can also identify areas of probable corrosion as areas of low potential levels.

Fatigue cracks resulting from service loadings are unusual in gas-transmission pipelines remote from compressor stations. This is because there are few pressure cycles and no mechanical vibration that would produce them. Pressure cycles on gaseous pipelines are usually a maximum of approximately 10% of the line pressure and may occur once or twice a day. Thus, in 40 years, this would represent only 14,400 to 28,800 cycles.

At compressor stations on gas pipelines, fatigue cracks have developed due to mechanical vibrations of the piping system. These are typical of the problems associated with rotating equipment.

On liquid pipelines, problems have been experienced from the relatively large and frequent pressure cycles often associated with liquid pipelines. Large pressure cycles can occur a number of times each day, and smaller cycles can occur more frequently, depending on how the line is operated and where deliveries are being made. The most common fatigue-crack problem occurs at the longitudinal weld—especially if there is a weld reinforcement, such as at a double-submerged arc weld. The appearance of these fatigue cracks is similar to the fatigue cracks due to shipping shown in Fig. 6. On liquid lines, the pumping stations also experience fatigue cracks due to mechanical vibrations and high-frequency pressure cycles similar to the potential problems at gas compressor stations.

Miscellaneous Failure Causes. A number of unusual failures have occurred. Many of these have been observed only once or twice, but can cause problems. Other causes observed are described below.

Secondary Loads. Soil instabilities, land slides, and mine subsidence have caused failures from excessive overloads in bending or tension.

Weldments to the Pipe. A few failures have been attributed to attachments that were improperly welded to the pipe. Before 1940, weldments to the pipe were common. Currently, no attachments are made to the pipe without careful design to minimize stress concentrations.

Wrinkles, Bends, or Buckles. At one time, it was not possible to obtain high-bend angles without wrinkling the pipe wall on the compressive side of the bend. Many of these are in service, and a number have cracked after years of service due to a low-cycle fatigue mechanism. This has been largely solved by the use of bending machines with internal mandrels. Also, because problems have developed, bends are carefully inspected for wrinkles and are rejected if any are present.

Internal Combustion. Several failures have occurred in new gas pipelines being tested with gas that were not adequately purged of air. The ignition source is unknown but may be static electricity. The characteristic of the fracture is that a detonation wave propagates along the inside of the pipe. The wave builds up a pressure front that causes pipe ruptures intermittently along the pipe and at major changes in direction. In between the points of rupture, the pipe is frequently expanded.

Sabotage. Several failures have occurred where the integrity of the pipeline was purposely breached. The suspected method is through the use of an explosive charge. The characteristic feature of the cause is that the pipe is usually severely fragmented at the origin area. Also, severe inward denting or deformation of the pipe is common, and multiple randomly oriented origins have been observed.

Longitudinal Weld Defects. The causes in this category are similar to those described previously, with the exception of the following two causes. First, selective corrosion of the weld bond line has been observed in electric-resistance welds. This forms an axial V-shaped groove centered on the bond line. This generally occurs because the bond line is ferritic. The ferritic region becomes anodic to the surrounding pearlite, creating a local corrosion cell.

Second, in the late 1940s and early 1950s, a limited quantity of electric-resistance weld pipe was produced with a martensitic weld region. It was believed that this high-strength region would be less prone to failure because it was higher in strength. Unfortunately, this region is also much more susceptible to failure from atomic hydrogen created by the cathodic potential applied to pipelines for corrosion control. The fracture surface near the origin exhibits an extremely brittle appearance.

References cited in this section

1. “Recommended Practice for Railroad Transportation of Line Pipe,” API RP5L1, American Petroleum Institute, Dallas

2. “Specification for Line Pipe,” API 5L, 35th ed., American Petroleum Institute, Dallas

3. J. F. Kiefner and A.R. Duffy, “Summary of Research to Determine the Strength of Corroded Areas in Line Pipe,” American Gas Association, Arlington, VA, July 20, 1971

R.J. Eiber and J.F. Kiefner, Battelle Columbus Laboratories

Arrest of Pipeline Fractures

A propagating fracture in a gas pipeline can be extensive because of the relatively slow pressure decay associated with the compressed natural gas. Historically, in the early 1950s in the gas testing of pipelines, there were several cleavage fractures that propagated to distances of 914 m (3000 ft). Other than the length of the fracture, the unusual aspect of the failure was that there was no reduction in thickness at the fracture edge, and the fractures propagated along the pipe in a sinusoidal pattern.

A failure occurred in the early 1960s in which a cleavage fracture initiated from a fatigue crack alongside a longitudinal weld. The fatigue crack formed during rail shipment of the pipe. The pipe material was 762-mm (30-in.) diam × 9.5-mm (0.375-in.) wall-thickness API, grade X56, pipe. The fracture initiated when this section of line was being used to transfer gas to the next section to be tested. The pressure at the time of failure was equivalent to a hoop stress of 63% SMYS. This was reported to be the highest pressure that this section of pipe had experienced. The fracture extended from the origin as a cleavage fracture and propagated for a total distance of 13.3 km (8.3 miles). The origin was 5.1 km (3.2 miles) from one end. The fracture propagated in a sinusoidal pattern, with the number of simultaneous fractures varying from one to six, depending on the transition temperatures of the pipe steel. On one end, the fracture arrested as a ductile fracture in a pipe length with a low transition temperature, and on the other end, the fracture arrested at the transition from the 9.5-mm (0.375-in.) line pipe to a heavy-wall forged tee. The fracture did not penetrate the forged tee, but stopped at the girth weld.

Figure 23 shows Charpy V-notch impact curves from 19 random samples taken from pipe lengths through which the fracture propagated. The fracture appearance of the Charpy specimens at the failure temperature ranged from approximately 10 to 40% shear. Figure 24 shows the Charpy V-notch impact curve for the end in which the fracture turned to a shear fracture, spiraled around the pipe, and arrested. At the failure temperature, the Charpy V-notch shear appearance was 100% shear. Thus, in this failure, the fracture arrested on one end because it encountered a length of pipe with a low transition temperature, and it arrested on the other end because it encountered a reduced stress level.

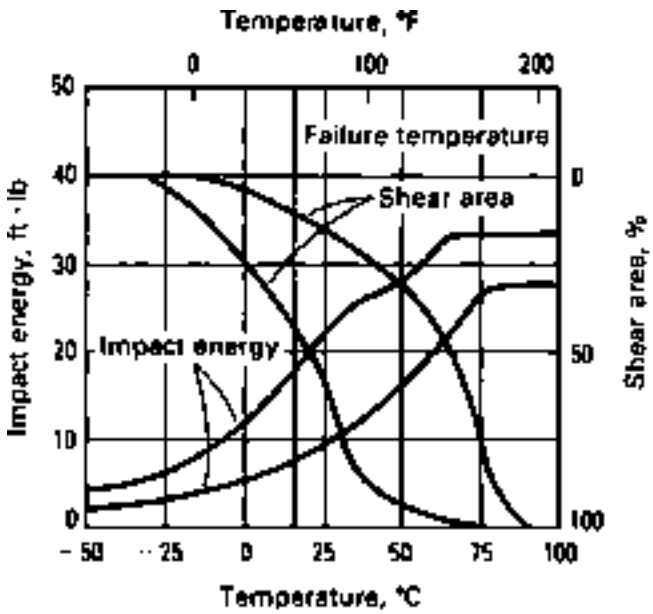


Fig. 23 Scatter bands for Charpy V-notch impact specimens from 19 random pipe lengths through which a fracture propagated in a 762-mm (30-in.) outside-diameter × 9.5-mm (0.375-in.) wall-thickness API, grade X56, pipe. Data are for two-thirds thickness transverse specimens.

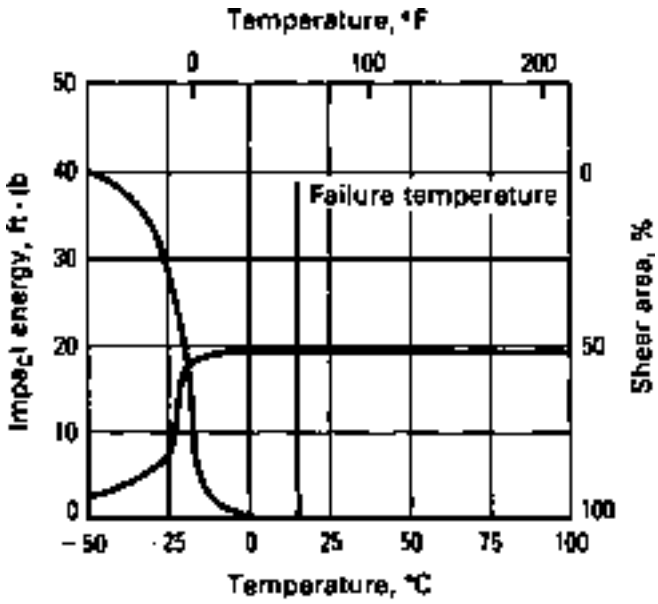


Fig. 24 Charpy V-notch curve for 762-mm (30-in.) outside-diameter × 9.5-mm (0.375-in.) wall-thickness API, grade X56, pipe in which a ductile fracture propagated and arrested. Data are for two-thirds thickness transverse specimens.

Another failure that occurred in 915-mm (36-in.) outside-diameter × 10-mm (0.406-in) wall-thickness API, grade X52, pipe was initiated by sabotage. The fracture propagated through 2.5 pipe lengths in one direction as a cleavage fracture and, upon crossing a girth weld, turned to shear and arrested. In the other direction, the fracture propagated through portions of three lengths as a cleavage fracture. It suddenly turned to tearing shear in the middle of a pipe length and arrested by spiraling around the circumference. Figure 25 shows the Charpy V-notch and drop-weight tear test (DWTT) curves for two pipe lengths through which cleavage fractures propagated and for the length in which a ductile fracture arrested. The DWTT results clearly show the arrest length as having an 85% shear-area transition temperature (SATT) below the failure temperatures thus, the shear fracture would be expected (SATT is used to identify the ductile-to-brittle transition temperature for pipeline steels). It should be noted that the transition temperature predicted by the DWTT (85% shear) is slightly above that predicted by the Charpy (85% shear) due to the 10-mm (0.406-in.) thickness of the DWTT. The two lengths through which the cleavage fracture propagated exhibited DWTT shear areas of 10 and 17%. This is consistent with the fracture appearance of 12 and 25% shear, respectively, for the two pipe lengths.

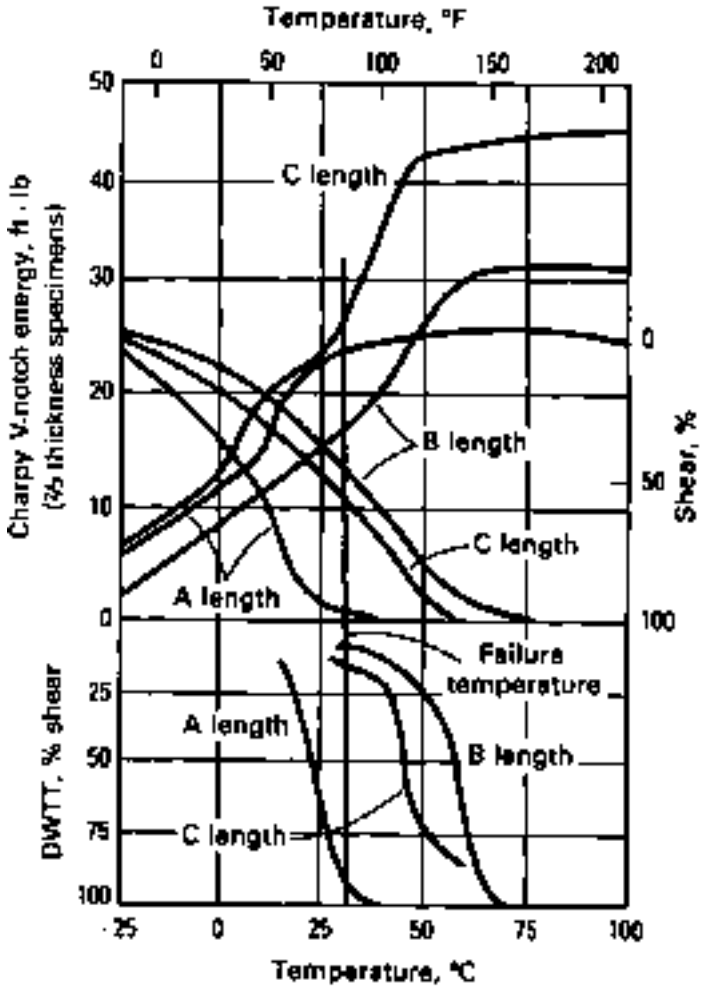


Fig. 25 Charpy V-notch impact energy and DWTT data for a failure in 915-mm (36-in.) outside-diameter × 10-mm (0.406-in.) wall-thickness API, grade X52, pipe. A length, ductile-fracture arrest; B and C lengths, cleavage-fracture propagation

In the opposite direction, the fracture propagated through parts of three lengths as a cleavage fracture and arrested in the middle of a pipe length as a tearing shear fracture. The Charpy and DWTT curves for the three pipe lengths are shown in Fig. 26. The DWTT predicts that the fracture appearance in all three lengths would be cleavage, as observed. The Charpy curves are in agreement, with the exception of length B, which exhibits an 85% SATT below the failure temperature. This discrepancy is probably due to thickness effects.

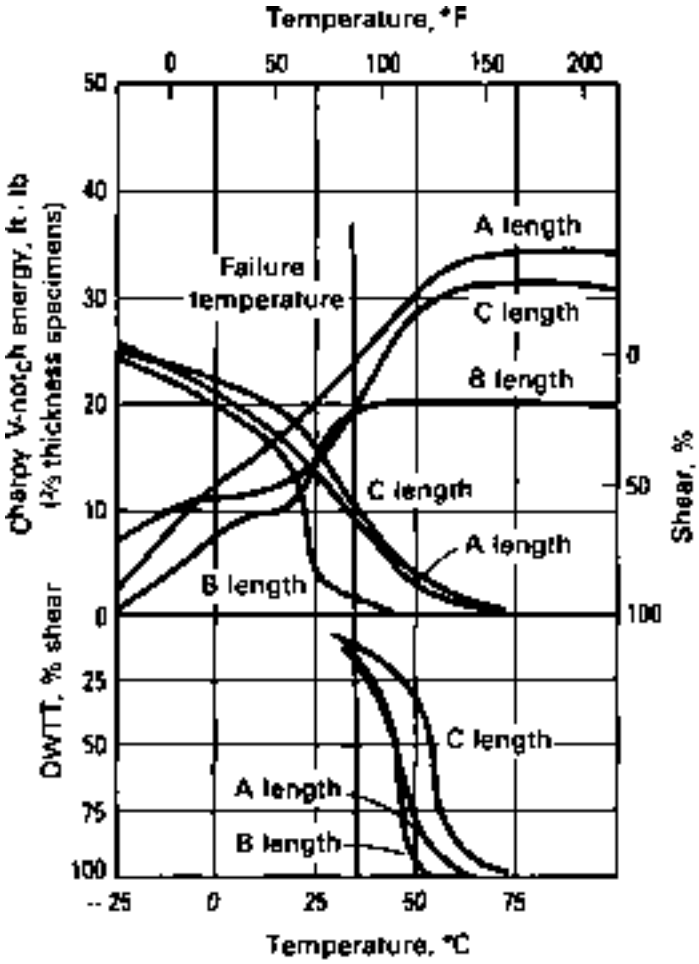


Fig. 26 Charpy V-notch and DWTT test data for failure in a 915-mm (36-in.) outside-diameter × 10-mm (0.406-in.) wall-thickness API, grade X52, pipe. A length, cleavage arrest; B and C lengths, cleavage-fracture propagation

Comparing the energy available to drive the cleavage fracture with the energy absorbed in the fracture process, it is apparent that the cleavage-fracture arrest is to be expected. The energy *U* available to drive the fracture in length A (the cleavage-fracture arrest length) is determined by (Ref 4):

$$U = \frac{s^2 p R}{E} = \frac{(33250)^2 p 1.5}{30 \text{ x } 10^6} = 174\, ft.lb/in^2 \qquad \text{(Eq 2)}$$

where σ = 750(18)/0.406 = 33.25 ksi hoop stress (229.3 MPa), $E = 3 \times 10^4$ ksi (2.07×10^5 MPa), and *R* is the pipe radius (457 mm, or 18 in.). The energy absorbed in the fracturing process can be determined from the Charpy energy at the shear area equivalent to the fracture appearance of the failed pipe (Ref 4). The average fracture appearance of the failed pipe was 22% shear. The Charpy energy at 22% shear is 19 J (14 ft · lb); converting this to Joules per square centimeter by dividing by the Charpy specimen area, the energy per unit area is 36 J/cm^2 (169 ft · lb/in.^{2}). Thus, the energy available to drive the cleavage fracture is very close to the energy absorbed in the fracture process, and the fracture arrest is highly probable.

Cleavage fractures have been observed to arrest in several ways. In pipe lengths with low transition temperatures, cleavage fractures turn to ductile fractures and arrest. Cleavage fractures have also arrested at an abrupt change in cross section, such as the transition to a forged tee or flange or at a road crossing casing, and by encountering a pipe length with sufficient energy-absorption capacity so that there is insufficient energy to drive the fracture.

A second mode of fracture propagation and arrest is ductile-fracture propagation and arrest. There have been several service failures in which ductile fractures have propagated up to 305-m (1000-ft) total length and then have arrested. In these failures, a single fracture propagated axially along the pipe. The fractures that have propagated unstably have generally propagated on the top of the pipe. The scatterband of the Charpy V-notch curves for one failure involving 914-mm (36-in.) diam pipe is shown in Fig. 27. The Charpy curves from the two lengths in which the fracture arrested are also shown. This fracture occurred at a hoop stress level of 307.6 MPa (44.6 ksi), and the total fracture length was 259 m (850 ft). All of the transition temperatures were below the operating temperature; therefore, the fractures were all ductile. Because the Charpy plateau energy was relatively low in the propagate lengths, unstable ductile-fracture propagation

occurred. The fracture arrested in the end lengths. The two-thirds thickness Charpy data indicate that ductile fractures propagated through lengths having 15- to 17-J (11- to 12.5-ft · lb) upper plateaus and arrested in pipe lengths with 19- and 30-J (14- and 22-ft · lb) plateaus. This type of failure indicates that material with a low transition temperature may not provide fracture control unless consideration is also given to the toughness level of the material. Equation 3 (Ref 5) provides an approximate relationship for estimating the Charpy V-notch plateau energy required for ductile-fracture arrest in various diameters, wall thicknesses, and stress levels between 60 and 80% of the specified minimum yield stress in gas-transmission pipelines:

$$CVN = 0.0108\ \sigma^2\ (Rt)^{1/3} \qquad \text{(Eq 3)}$$

where CVN is the Charpy V-notch plateau energy (in foot-pounds) for a 10- × 10-mm (0.4- × 0.4-in.) specimen, σ is the hoop stress (in kips per square inch), *R* is the pipe radius (in inches), and *t* is the pipe wall thickness (in inches).

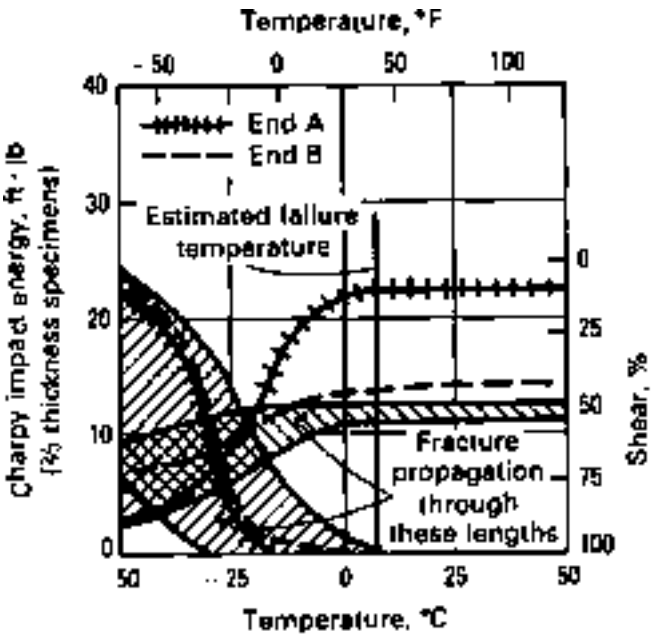


Fig. 27 Charpy V-notch impact data for pipe lengths in which ductile fracture propagated and arrested. End A, fracture arrest; End B, fracture arrest

References cited in this section

4. W.A. Maxey, J.F. Kiefner, and R.J. Eiber, “Brittle Fracture Arrest in Gas Pipelines,” Catalog L51436, American Gas Association, Arlington, VA, April 4, 1983

5. W.A. Maxey, J.F. Kiefner, and R.J. Eiber, “Ductile Fracture Arrest in Gas Pipelines,” Catalog L32176, American Gas Association, Arlington, VA, Dec 12, 1975

Failures of Pipelines

R.J. Eiber and J.F. Kiefner, Battelle Columbus Laboratories

References

1. “Recommended Practice for Railroad Transportation of Line Pipe,” API RP5L1, American Petroleum Institute, Dallas

2. “Specification for Line Pipe,” API 5L, 35th ed., American Petroleum Institute, Dallas

3. J. F. Kiefner and A.R. Duffy, “Summary of Research to Determine the Strength of Corroded Areas in Line Pipe,” American Gas Association, Arlington, V, July 20, 1971

4. W.A. Maxey, J.F. Kiefner, and R.J. Eiber, “Brittle Fracture Arrest in Gas Pipelines,” Catalog L51436, American Gas Association, Arlington, VA, April 4, 1983

5. W.A. Maxey, J.F. Kiefner, and R.J. Eiber, “Ductile Fracture Arrest in Gas Pipelines,” Catalog L32176, American Gas Association, Arlington, VA, Dec 12, 1975

Failures of Bridge Components

John W. Fisher, Fritz Engineering Laboratory, Lehigh University

Introduction

FROM 1960 TO 1986, a number of localized failures developed in steel bridge components due to fatigue-crack propagation. which resulted in some instances in brittle fracture from crack instability. More than 200 bridge sites have developed one or more types of cracking. Several types of cracking have often developed at a single site because of the existence of different details on several types of structures (Ref 1, 2).

The largest category of cracking is a result of out-of-plane distortion in a small gap, which is usually a segment of a girder web. When distortion-induced cracking develops in a bridge component, large numbers of cracks usually form nearly simultaneously in the structural system because the cyclic stress is high and the number of cycles needed to produce cracking is relatively small. Displacement-induced fatigue cracking has developed in a wide variety of bridge structures, including suspension, two-girder floor-beam, multiple-beam, tied-arch, and box-girder bridges. In general, the cracks formed in planes parallel to the stresses from loading and were not detrimental to the performance of the structure provided they were discovered and retrofitted before turning perpendicular to the applied stresses.

The next largest category of cracked members and components comprises large initial defects and cracks. In several cases, the defects in this category resulted from poor-quality welds that were produced before nondestructive test methods were well developed. However, the largest number of this type of crack resulted because the groove-welded component was considered a secondary member or attachment; consequently, weld-quality criteria were not established, and nondestructive test requirements were not imposed.

Most of the remaining cracks resulted from the use of low fatigue strength details that were not anticipated to have such low fatigue resistance at the time of the original design.

References cited in this section

1. J.W. Fisher, *Fatigue and Fracture in Steel Bridges*, Wiley-Interscience, 1984

2. J.W. Fisher and D.R. Mertz, Hundreds of Bridges—Thousands of Cracks, *Civil Eng.*, April 1985

Failures of Bridge Components

John W. Fisher, Fritz Engineering Laboratory, Lehigh University

Details and Defects

Low Fatigue Strength Details. The possibility of fatigue cracks forming at the ends of welded cover plates was demonstrated at the American Association of State Highway Officials (AASHO) Road Test in the 1960s (Ref 3). Multiple-beam bridges subjected to relatively high stress-range cycles (; 83 MPa, or 12 ksi) under controlled truck traffic experienced cracking after 500,000 vehicle crossings. In general, few cracked details were known to exist until a cracked beam was discovered in Span 11 of the Yellow Mill Pond Bridge in 1970 (Fig. 1). Between 1970 and 1981, the Yellow Mill Pond multibeam structures located at Bridgeport, CT, developed extensive numbers of fatigue cracks at the ends of cover plates (Ref 1). These cracks resulted from the large volume of truck traffic and the unanticipated low fatigue resistance of the large cover-plated beam members (Category E′) (Ref 4).

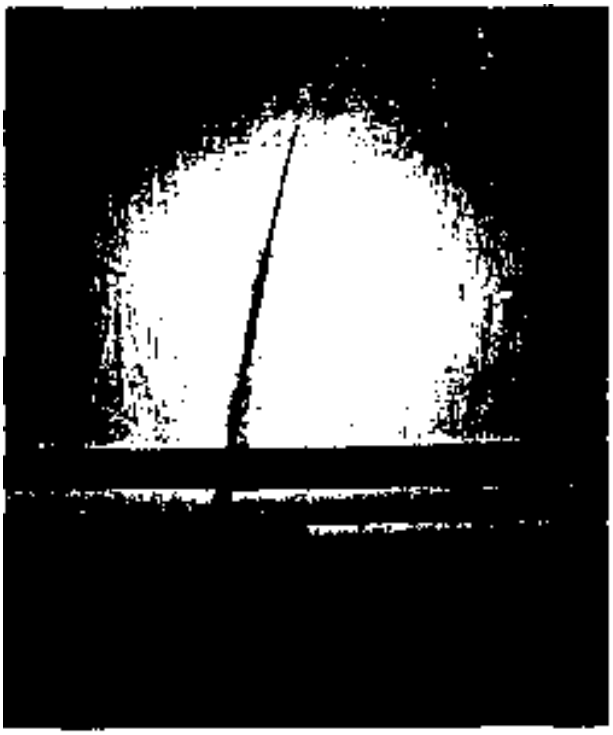


Fig. 1 Cracked girder at the end of the cover plate.

Several other Category E or E′ details have experienced fatigue cracking at weld termination (Ref 1). The crack shown in Fig. 2 typifies these structures and this type of cracking. It originated at the end of a longitudinal fillet weld that was used to attach a 0.9-m (3-ft) long lateral connection plate to the edge of a flange.

Fig. 2 Crack at the end of the lateral connection plate.

The Yellow Mill Pond Bridge was constructed in 1956–1957 and was opened to traffic in January 1958. In October-November 1970, the steel superstructure of the Yellow Mill Pond Bridge was inspected after cleaning and repainting. On November 2, 1970, during a routine inspection of the repainting, it was discovered that a crack had developed in the

eastbound roadway in Span 11. The crack started at the end of the cover-plated beam; it propagated through the flange and up to 400 mm (16 in.) into the web of one of the main girders (Fig. 1).
Additional cracks, such as that shown in Fig. 3, were detected between 1970 and 1981. The entire structural complex was retrofitted in 1981 by peening the weld toe at small cracks or uncracked details and by installing bolted splices at large cracks.

Fig. 3 Typical large crack at the weld toe at the end of the cover plate.

The eastbound and westbound bridges of Span 10 were selected for two major stress-history studies. The state of Connecticut and the Federal Highway Administration (FHWA) conducted both studies, the first in July 1971 (Ref 5) and the second between April 1973 and April 1974 (Ref 6). Figure 4 shows a typical stress histogram for the westbound lanes. The truck distribution and its position in the lanes were also concurrently recorded, and their weights were sampled. Some additional measurements were made in 1976 (Ref 7). The composition of the average daily truck traffic (ADTT) at the Yellow Mill Pond Bridge was about the same as the gross vehicle weight distribution developed from the 1970 FHWA nationwide Loadometer survey. From January 1958 to June 1976, approximately 259 million vehicles crossed the structure. Approximately 13.5% of this total flow was truck traffic. Therefore, approximately 35 million trucks crossed the eastbound and westbound bridges between 1958 and 1976.

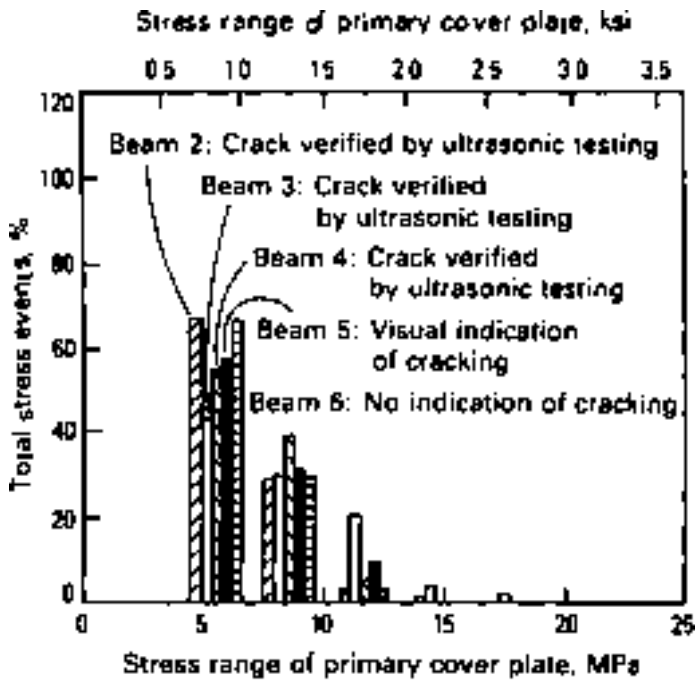


Fig. 4 Stress-range histograms.

The measurements acquired in June 1971, from April 1973 to April 1974, and in June 1976 were used to construct composite stress-range response spectra. Miner effective stress-range values S_{rMiner} were determined for the gage locations (Ref 7):

$$S_{\mathrm{rMiner}} = [\Sigma \alpha_i S_{ri}^3]^{1/3} = 13.6 \text{ MPa } (1.98 \text{ ksi})$$

where α_i is the frequency of occurrence of stress-range level s_{ri}. The measurements indicated that the effective stress range in other girders varied from 7.6 to 13.6 MPa (1.1 to 1.98 ksi). Those beams located under the outside lane tended to provide the higher stress-range values.

An examination of the stress response to the passage of a truck also indicated that more than one stress cycle occurred. The variable stress spectrum appeared to correspond to about 1.8 events per truck. This corresponded to 62.8×10^6 cycles if it were applied to each vehicle passage.

During 1981, cores were removed from several of the beams to examine the crack surfaces and the retrofit conditions conducted in 1976. Figure 5 shows the center of a polished-and-etched core, and the crack at the weld toe can be seen to extend more than halfway through the beam flange. The crack surface of the core was exposed and examined with the electron microscope. Because the surface was relatively clean and not extensively corroded, the crack-surface features were not destroyed. Figure 6 shows a transmission electron microscopy (TEM) fractograph at a magnification of 56,000× at the crack tip, which was 16 mm (0.63 in.) deep. Striationlike features are visible on the crack surface at various locations. The random variable loading produced striations of variable height and spacing.

Fig. 5 Polished-and-etched core showing a fatigue crack in the beam flange.

Fig. 6 TEM fractograph showing fatigue-crack growth striations. 39,000×

The cracks forming at the cover-plate weld toes were modeled as semielliptical surface cracks in the flange (Ref 1). The stress intensity K was defined as (Ref 7):

$$K = F_e F_s F_w F_g \sigma \sqrt{p\,a}$$

where

$$F_e = \frac{1}{E(k)}$$

$$E(k) = \int_0^{\pi/2} \left[1 - \left(\frac{c^2 - a^2}{c^2} \right) \sin^2 \theta \right]^{1/2} d\theta;$$

$$c = 5.46 a^{1.153} \text{(in.)}$$

$$F_s = 1.211 - 0.186 \sqrt{\frac{a}{c}}$$

$$F_g = \frac{K_T}{1 + 6.79 (a/t_f)^{0.44}}$$

$$K_T = -3.539 \ln \left(\frac{z}{t_f} \right) + 1.981 \ln \left(\frac{t_{cp}}{t_f} \right) + 5.798$$

$$F_w = \sqrt{\sec \frac{\pi a}{2 t_f}}$$

With z = 16 mm (0.63 in.), t_f = 32 mm (1.26 in.), t_{cp} = 32 mm (1.26 in.), a_i = 0.75 mm (0.03 in.), and $\Delta\sigma_{Miner}$ = 13 MPa (1.9 ksi), the number of cycles N required to grow a fatigue crack 25 min (1 in.) deep was estimated at:

$$N = \int_{a_{min}}^{1.0} \frac{da}{3.6 \times 10^{-10} \Delta K^3} = 36 \times 10^6 \text{ cycles}$$

This is in reasonable agreement with the larger fatigue cracks that formed between 1958 and 1976 when an estimated 35 million trucks crossed the bridges. It is also compatible with the experimental data on full-scale cover-plate beams (Ref 4, 8).

Large initial defects and cracks constitute the second largest category of cracked members and components. In several cases, the defects resulted because of poor-quality welds that were produced before nondestructive test methods were well developed. A larger number of these cracks resulted because the groove-welded component was considered a secondary member or attachment; therefore, weld-quality criteria were not established, and nondestructive test requirements were not imposed on the affected weldment. Splices in continuous longitudinal stiffeners represent a common condition that falls into this category. A related condition has occurred when backing bars were used to make a groove weld between transverse stiffeners and a lateral gusset plate. Lack of fusion often exists adjacent to the girder web in the transverse groove welds. If the transverse welds intersect with the longitudinal welds, they provide a path for a crack to enter the girder web.

Cracks were discovered at the vertical buttweld detail at haunch inserts (Fig. 7a). One of these cracks extended 1.1 m (44 in.) into the girder web along a diagonal line starting from the vertical butt-weld detail.

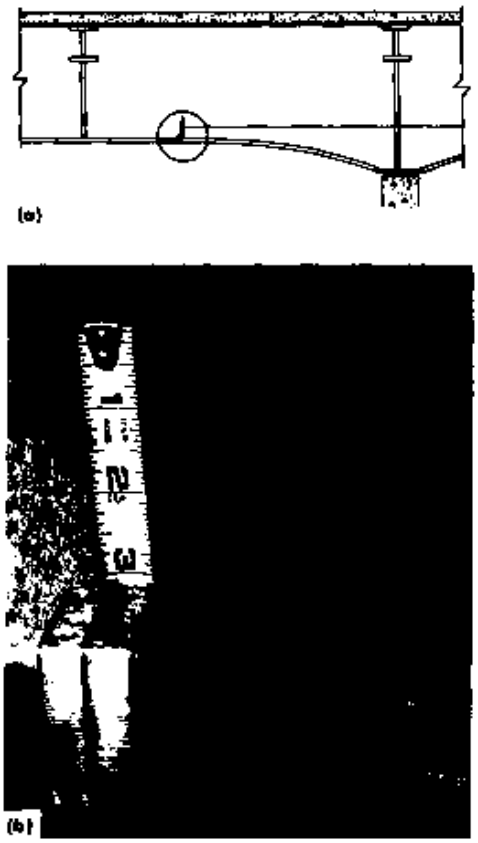


Fig. 7 Crack at a vertical butt weld. (a) Schematic showing insert and crack location. (b) Typical crack at a vertical groove weld. Source: Ref 9

The fatigue cracks that developed propagated from large initial weld imperfections or inclusions in the short transverse groove welds at the ends of the parabolic haunch inserts in the main girders (Fig. 7b). Of 24 welded details, 7 had cracked by 1973.

In November 1973, a large crack was discovered in the south facia girder of the suspended span in the center portion of the Quinnipiac River Bridge (Ref 10). The bridge had experienced approximately 9 years of service life before discovery of the crack. Figure 8(a) shows the crack that developed in the south facia girder web. The crack propagated approximately to mid-depth of the girder and penetrated into the bottom flange of the girder before it was discovered. A second crack was detected toward the midspan about 9 m (29 ft) from the cracked section. This crack severed the stiffener but did not propagate through the web. Figure 8(b) shows this crack at the time it was detected.

Fig. 8 Cracks in the Quinnipiac River Bridge after 9 years of service. (a) Crack in the girder web at the longitudinal stiffener groove weld. (b) Crack in the longitudinal stiffener weld

A detailed study of the fracture surface was made on pieces that were removed from the section (Ref 10). This indicated that the fracture initiated at the unfused butt weld in the longitudinal stiffener.

Cracks that have developed in the web at lateral connection plates have generally occurred as a result of intersecting welds. The lateral connection plate is often framed around a transverse stiffener. It was used to connect diaphragms and lateral bracing members to the longitudinal girders of bridges.

One of the first bridge structures to exhibit this cracking was the Lafayette Street Bridge over the Mississippi River at St. Paul, MN (Ref 11). Figure 9 shows the cracked girder at the detail attached to the web. The primary problem was the large defect in the weld attaching the lateral connection plate to the transverse stiffener. Because this weld was perpendicular to the cyclic stresses and because the weld intersected with the vertical welds attaching the stiffener to the web and the longitudinal welds of the connection plate, a path was provided into the girder web. A detailed study indicated that the crack originated in the weld between the gusset plate and the transverse stiffener as a consequence of a large lack-of-fusion discontinuity in the welded connection. Figure 10 shows a schematic of the crack surface. The intersecting welds at the corner permitted the transverse crack to penetrate into the girder web.

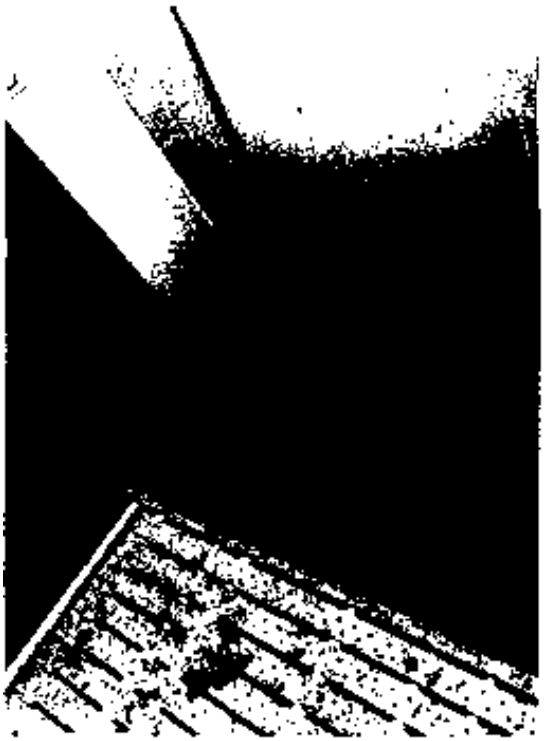


Fig. 9 Cracked girder at the Lafayette Street Bridge.

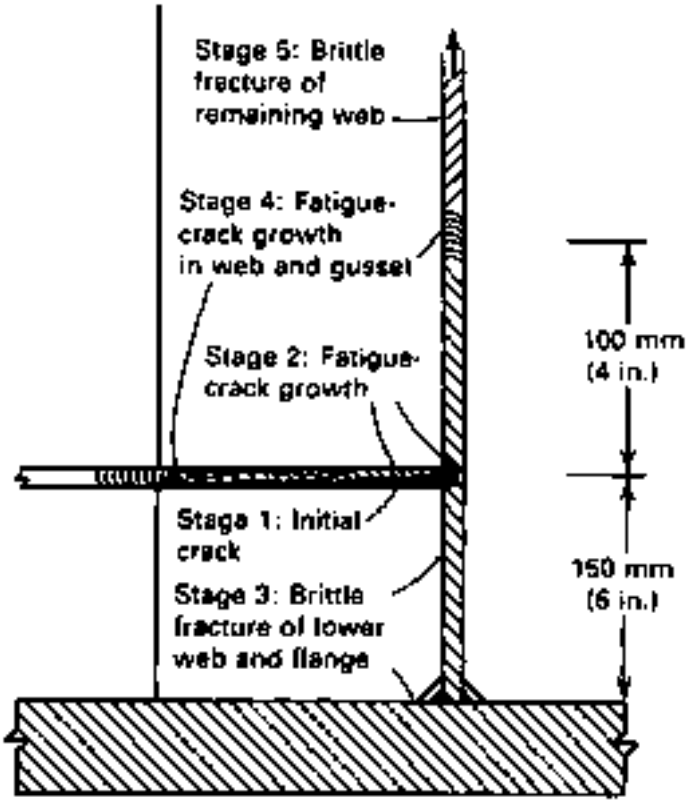


Fig. 10 Schematic of crack-growth stages.

Typical of the cracks that have developed in flange groove welds is the splice in the cover plate at thickness transitions (Fig. 11). Numerous cracks caused by lack of fusion have been detected in groove welds (Fig. 12). In general, these cracks formed because of lack of fusion at the root (land) of the groove weld. Regarding Fig. 11, it is probable that the single-U groove weld was made using the beam flange as a backup. Figures 12 and 13 show a double-V groove weld made in the same era with lack of fusion at the root.

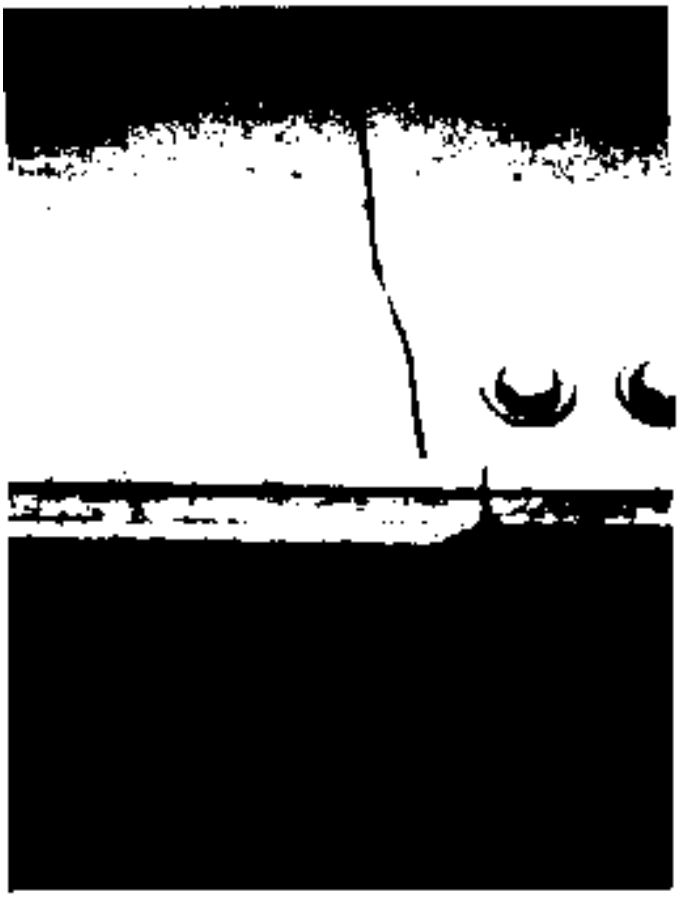


Fig. 11 Cracked groove weld in the cover plate.

Fig. 12 Lack-of-fusion defect in the flange groove.

Fig. 13 Polished-and-etched section showing lack of fusion.

The design stress range at the cracked groove welds like that shown in Fig. 11 varied between 58.6 and 78.6 MPa (8.5 and 11.4 ksi). Total truck traffic crossing the structure was estimated at 15.7 million vehicles.

The actual service stresses were not available. However, the effective stress range can be estimated by assuming that the daily truck traffic corresponds to the nationwide distribution (Ref 12). The effective stress range becomes:

An effective stress range of 14 to 21 MPa (2 to 3 ksi) is probable, because the factor γ_i is between 0.25 and 0.4. Assuming one stress cycle for each vehicle passage and no significant changes in total truck traffic between 1958 and 1980, the result would be 15.7×10^6 random variable stress cycles corresponding to the Miner effective stress range.

The probable size of the initial crack a_i is likely to be about 6 mm (0.25 in.). Because the cover plate is connected to the flange only along the edges, the applicable stress-intensity factor for an edge crack is:

$$K = 1.12\sigma\sqrt{\pi a}\sqrt{\sec\frac{\pi a}{2t_f}}$$

where t_f is the thickness of the thinner plate at the groove-weld splice. The predicted cycles of stress are given by:

$$N = \int_{a_i}^{a_f} \frac{da}{3.6 \times 10^{-10}\Delta K^{3}}$$

With a_i = 6 mm (0.25 in.), S_{rMiner} = 14 MPa (2 ksi), and a_f = 12.5 mm (0.5 in.), a predicted life of 16×10^6 cycles results.

Penetrations Through Box-Girder Webs. A related lack-of-fusion type of defect and crack among the most severe that has been encountered has occurred at details where a plate component was inserted through an opening cut into girder webs. The resulting detail was usually welded into place with either fillet or groove welds. In either case, large cracks resulted at the edge of the flange plate where short vertical weld lengths resulted with large unfused areas.

A large crack in one of the steel box bents (rigid frames) supporting the elevated track of Chicago's Mass Transit Dan Ryan Line was discovered on January 4, 1978 (Ref 13). Subsequent inspection showed that two adjacent bents were also cracked.

The bents have a box-shaped cross section consisting of two column legs and a horizontal box member (Fig. 14). The top flanges of the plate girders pass over the top flanges of the boxes. The bottom flanges pierce the boxes through flame-cut slots near the bottom of the box side plates. In the cracked bents, the bottom flanges were welded to the box-girder web.

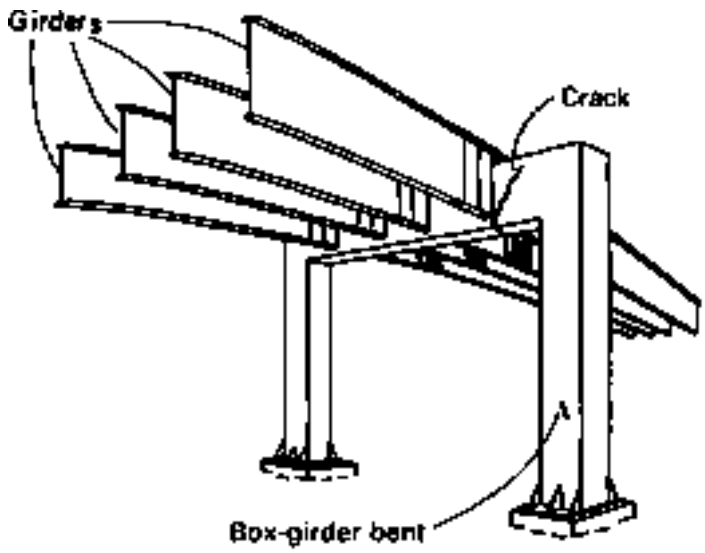


Fig. 14 Section of box-girder bent showing the location of the crack.

The initial field examination of the fracture indicated that all of the cracks started at the welded junction of the plate girder flange tip to the box side plate (Fig. 15). All three cracks completely severed the bottom flange of the box girders and the webs.

Fig. 15 Cracked box girder.

The crack surfaces of the side plates of the box girders adjacent to the flange tip were examined. Chevron markings found on the fracture surface indicated that the crack originated near the tip of the girder flange. Figure 16 shows a photograph of one fracture surface. The overall crack surface shows that the weldment connecting the flange tip to the box had a large lack-of-fusion area. Paint that had penetrated from the inside of the box was found in some of these areas. The largest lack-of-fusion area was about 18 × 83 mm (0.70 × 3.27 in.) and was approximately elliptical. Fatigue-crack propagation was detected at the inside and outside web surfaces (Ref 14).

Fig. 16 Fracture surface of the web plate.

Reaction diaphragm plates were also passed through box-girder webs (Fig. 17). Backup bars were used to make the groove welds between the girder webs and the insert plate (Fig. 18). Figure 19 shows a core removed from the box-girder web and reveals the crack that propagated from the initial flaw condition provided by the backup bar and slag at the root. Several bridges with these types of details have experienced fatigue-crack growth that was visible after 1 or 2 years of service.

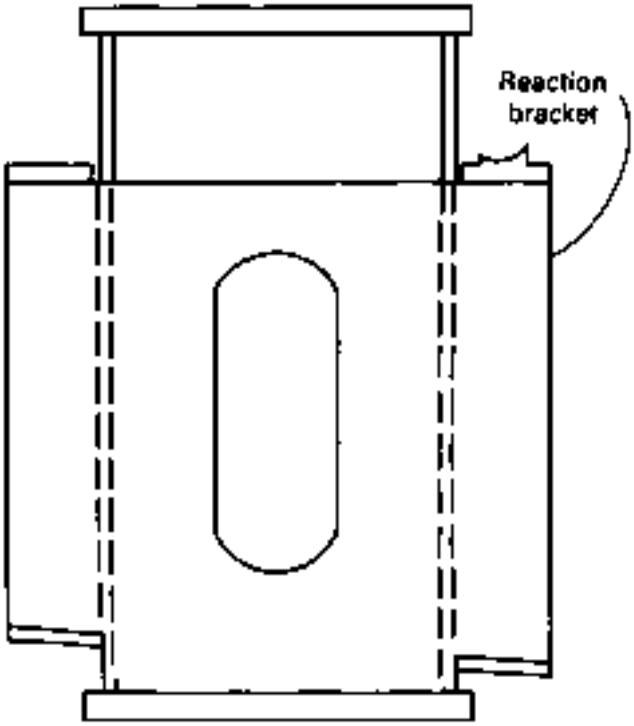


Fig. 17 Steel pier-cap section showing reaction brackets passing through the web.

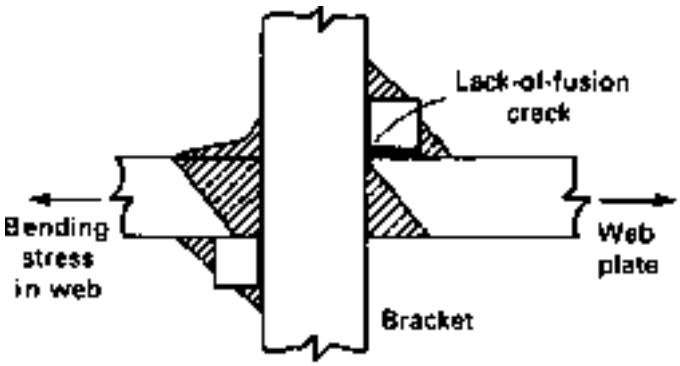


Fig. 18 Schematic of the lack-of-fusion crack from the backup bar.

Fig. 19 Investigation of cracking of a box-girder web. (a) Hole resulting from removal of the core. (b) Polished-and-etched surface showing crack, slag, and lack of fusion

---

References cited in this section

1. J.W. Fisher, *Fatigue and Fracture in Steel Bridges,* Wiley-Interscience, 1984

3. J.W. Fisher and I.M. Viest, *Fatigue Life of Bridge Beams Subjected to Controlled Truck Traffic,* Preliminary publications, 7th Congress, International Association for Bridge and Structural Engineering, Zurich, 1964

4. J.W. Fisher, H. Hausammann, M.D. Sullivan, and A.W. Pense, *Detection and Repair of Fatigue Damage in Welded Highway Bridges,* NCHRP report 206, Transportation Research Board, Washington, June 1979

5. D.G. Bowers, *Loading History Span 10 Yellow Mill Pond Bridge 195 Bridgeport, Connecticut,* Research project HPR 175-332, Connecticut Department of Transportation, Whethersfield, CT, 1972

6. R.L. Dickey and T.P. Severga, *Mechanical Strain Recorder on a Connecticut Bridge,* Report FCP 45G1-222, Federal Highway Administration Department of Transportation, Washington, 1974

7. J. W. Fisher, R.E. Slockbower, H. Hausammann, and A.W. Pense, Long Time Observation of a Fatigue Damaged Bridge, *J. Tech. Counc.,* Vol 107, (No. TCl), April 1981

8. J.W. Fisher, K.H. Frank, M.A. Hirt, and B.M. McNamee, *Effect of Weldments on the Fatigue Strength of Steel Beams,* NCHRP report 102, Transportation Research Board, Washington, 1970

9. J.P.C. King, P.F. Csagoly, and J.W. Fisher, Field Testing of Aquasabon River Bridge in Ontario, *Trans. Res. Rec.,* No. 579, 1976

10. J.W. Fisher, A.W. Pense, H. Hausammann, and G.R. Irwin, Analysis of Cracking in Quinnipiac River Bridge, *J. Struct. Div.,* Vol 106 (No. ST4), April 1980

11. J.W. Fisher, A.W. Pense, and R. Roberts, Evaluation of Fracture of Lafayette Street Bridge, *J. Struct. Div.,* Vol 103 (No. ST7), July 1977

12. J.W. Fisher, *Bridge Fatigue Guide—Design and Details,* Publication T112-11/77, American Institute of Steel Construction, Chicago, 1977

13. Engineers Investigate Cracked El, *Eng. News Rec.,* Vol 200 (No. 3), Jan 19, 1979

14. J.W. Fisher, J.M. Hanson, H. Hausammann, and A.E.N. Osborn, *Fracture and Retrofit of Dan Ryan Rapid Transit Structure,* Final report, 11th Congress, International Association for Bridge and Structural Engineering, Zurich, 1980

Failures of Bridge Components

John W. Fisher, Fritz Engineering Laboratory, Lehigh University

Out-Of-Plane Distortion

At least 100 bridge sites have developed fatigue cracks as a result of out-of-plane distortion in a small gap (Ref 1). Most often, this was a segment of a girder web. When distortion-induced cracking develops in a bridge, large numbers of cracks usually form before corrective action is taken because the cyclic stresses are often very high. As a result, many cracks form simultaneously in the structural system. Other types of cracking, such as a low fatigue strength detail or a large built-in detect, often result only in a single significant crack at a given bridge. Other potential crack locations can be identified and retrofitted before significant damage develops elsewhere.

Displacement-induced fatigue cracking has developed in such diverse structures as suspension bridges, two-girder floor-beam bridges, multiple-beam bridges, tied-arch bridges, and box-girder bridges. Cracks have formed mainly in planes parallel to the stresses from loading; therefore, these cracks have not been detrimental to the performance of the structure provided they were discovered and retrofitted before turning perpendicular to the applied stresses from loads. In some

structures, the cracks have arrested when located in low-stress areas, and the crack has served to relieve the restraint condition.

Floor-Beam-Girder Connection Plates. One of the earliest and most common sources of fatigue cracks from distortion is the cracking in the web gaps at the ends of floor-beam connection plates (Ref 15, 16). These cracks have occurred in the web gap near the end reactions (Fig. 20) when the floor-beam connection plate was not welded to the bottom tension flange. However, the most extensive cracking is in the negative-moment regions of continuous bridges where the connection plate was not welded to the top flange of the girder.

Fig. 20 Crack along the web-flange weld toe at the end reaction.

The cracking illustrated in Fig. 20 and 21 has occurred in skewed, curved, and right bridges. Generally, it is more severe and occurs earlier in skewed and curved structures.

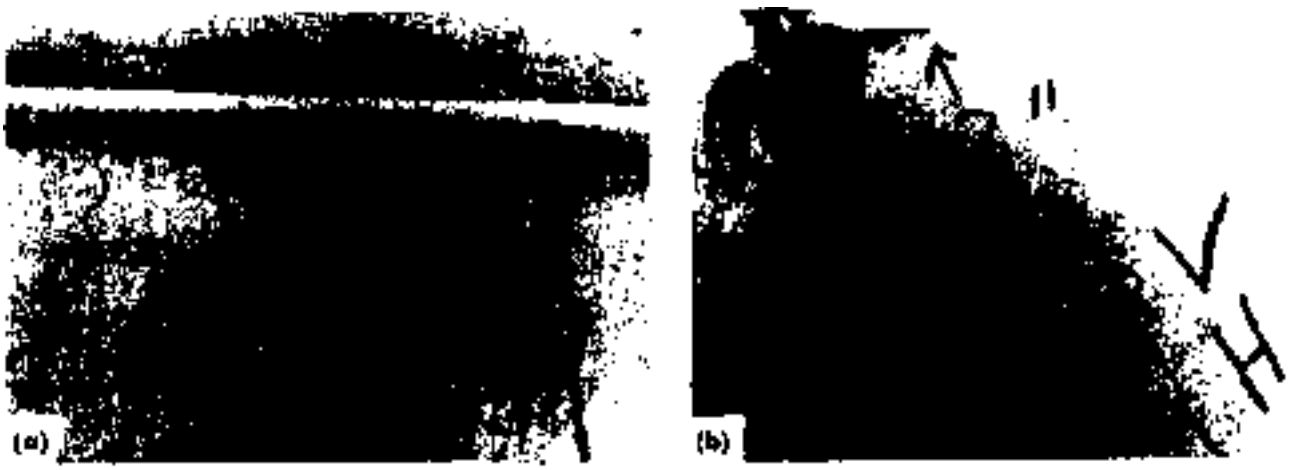


Fig. 21 Cracking in floor-beam-girder connection plates. (a) Cracks in the outside web surface. (b) Cracks from the inside web surface

Strain measurements have verified that the primary cause of the cracking is the out-of-plane movement at the end of the connection plate (Fig. 22). Figure 23 shows the stress gradient that was observed in one bridge structure (Ref 15, 16). The end rotation of the floor beam develops from bending of the floor beam and differential vertical movement between its two ends. The end rotation forces the web out-of-plane in the gap between the end of the connection plate and the web-flange weld.

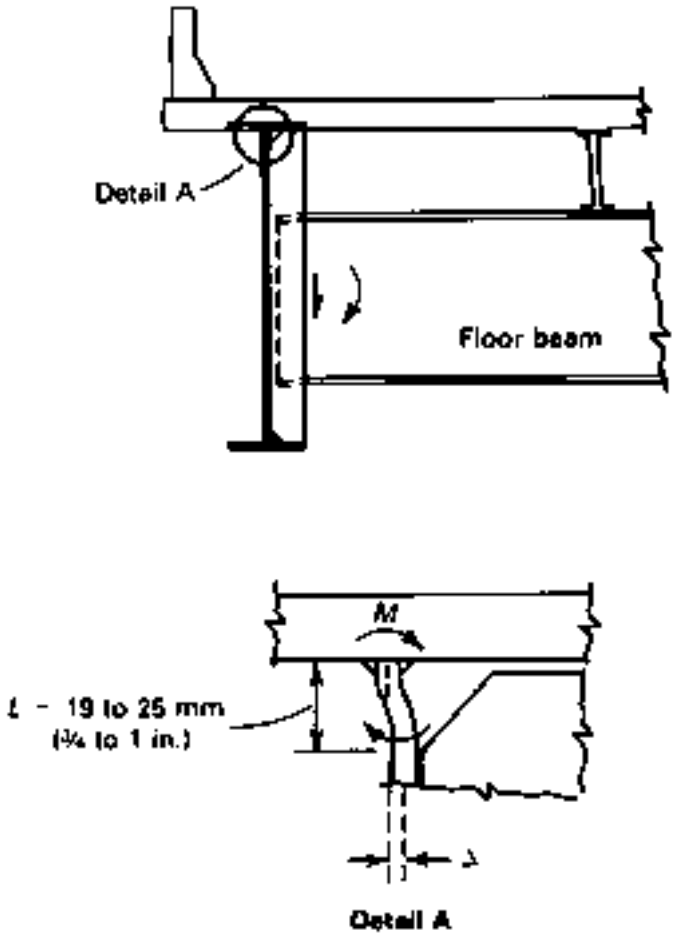


Fig. 22 Distortion in the web gap. *M*, bending moment; *L*, gap between surface of flange and weld to web

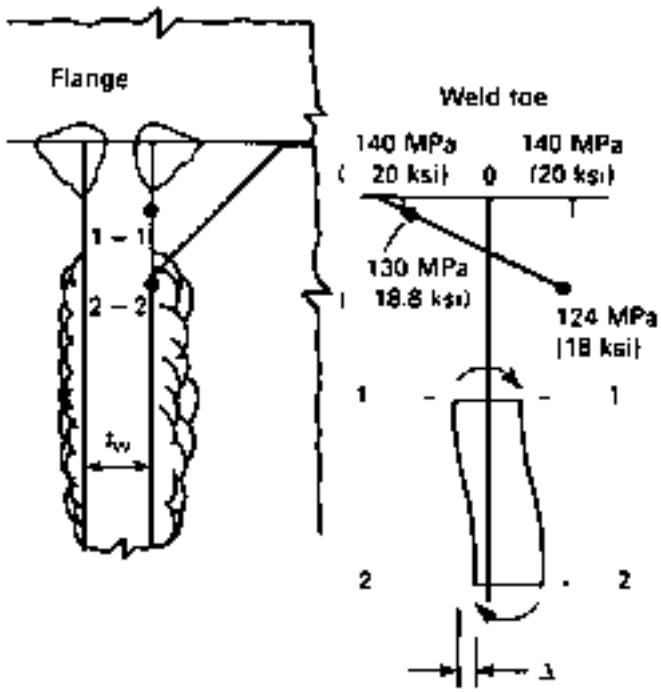


Fig. 23 Measured street gradient in the web gap.

Retrofit procedures for existing structures (Ref 1) have included removing a segment of the connection plate to increase the length of the web gap and welding or bolting the connection plate to the tension flange. Drilling holes to arrest the crack has provided only a temporary solution.

Multiple-Girder Diaphragm Connection Plates. Diaphragms and cross-frames in multiple-beam bridges often provide conditions at the transverse connection plates on the longitudinal girders to which they are connected that are similar to those of the floor-beam-girder connection plate. Generally, the diaphragms and cross-frames are connected to transverse stiffeners that are welded to the girder web. No connection has been provided between the stiffener and the girder tension flange. Sometimes, these stiffeners are not attached to either flange. Because adjacent beams deflect differing amounts, the differential vertical movement produces an out-of-plane deformation in the web gap at the stiffener ends that are not attached to the beam flange. The magnitude of this out-of-plane movement depends on the girder spacing, skew, and type of diaphragm or cross-frame. In general, skewed bridges undergo larger relative vertical movement of adjacent girders, which results in larger out-of-plane movement in the web gap.

Various types of diaphragms and girder spacings have resulted in this type of cracking. The diaphragms have ranged from simple X-bracing using angles to more rigid rolled sections.

Figure 24 shows a five-girder bridge with relatively stiff diaphragms composed of a blocked rolled beam with haunched ends. Cracks formed in the web gap along the web-flange weld toe (Fig. 25). In all but one highway bridge, these cracks have been confined to the negative-moment region where the top flange is in tension. The slab has provided greater fixity to the top flange, and the adjacent web has experienced more deformation and has cracked first.

Fig. 24 Diaphragms and girders in a continuous-span structure.

Fig. 25 Crack along the web-flange weld.

A more flexible diaphragm has been the X-type cross-frame shown in Fig. 26. Cracking has developed in the negative-moment region (top flange in tension) of continuous-span structures in the web gap (Fig. 27). The transverse stiffener to which the X-bracing was attached was not welded to the tension flange. This permitted the occurrence of out-of-plane displacement in the web gap and the formation of cracks along the web-flange weld and at the end of the connection plate.

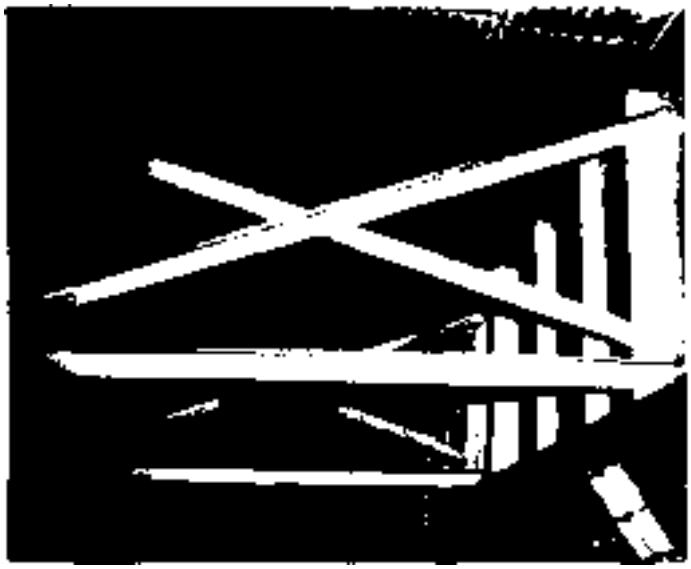


Fig. 26 View of bracing and girders.

Fig. 27 Fatigue cracks along the web-flange weld and at the end of the connection plate.

Several simple-span railroad and mass-transit bridges with relatively rigid diaphragms have developed cracks in the web gap adjacent to the bottom flange (Fig. 28). These structures were usually skewed, with the diaphragms placed perpendicular to the longitudinal girders. Cracking was initially observed at the ends of the transverse connection plate, which was cut short of the tension flange by 25 to 50 mm (1 to 2 in.). Cracking has since developed along the web-flange weld.

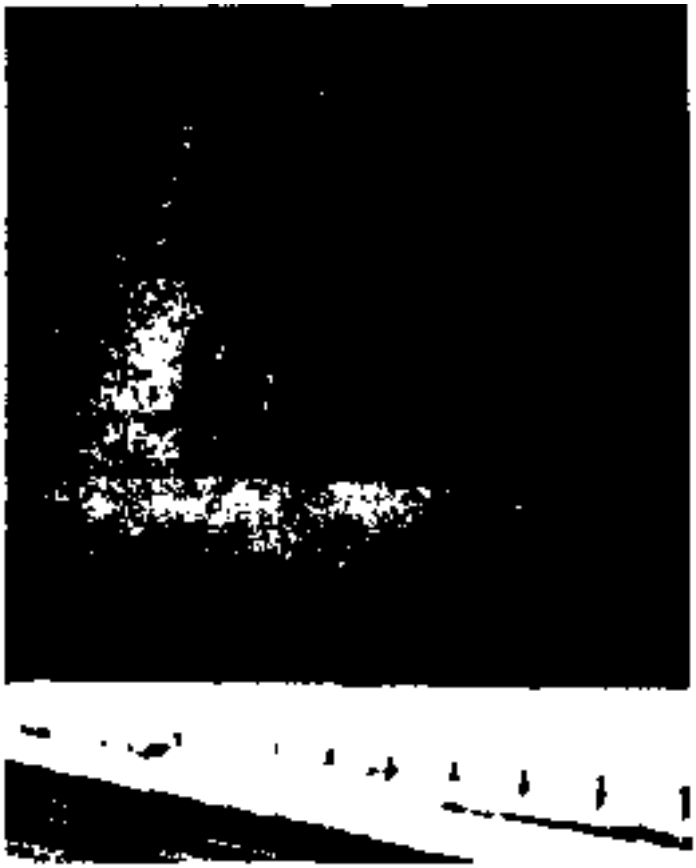


Fig. 28 Crack in the girder web at the bottom of the transverse connection plate.

Tied-Arch Floor Beams. A number of tied-arch structures have been built with the floor beams framed into the tie girder with web shear connection alone. The tie girder is a bending tie that is deeper than the floor beams. No direct connection was provided between the floor-beam flange and the tie girder. In several older structures, the floor-beam end connections were riveted double angles attached to the floor-beam web, with the outstanding legs riveted to the tie girder. More recently, constructed arches have utilized welded transverse connection plates on the tie-girder webs. The floor beams have been bolted to the welded transverse connection plate.

Cracks have formed in the floor beams of structures with either riveted end connections or welded transverse connection plates. The cracks have formed in the floor-beam web along the web-flange connection at the floor-beam web gap between the end of the connection angles or at the end of the welded connection plates. These cracks extended parallel to the floor-beam flange along the length of the web gap, then began to turn and propagate toward the bottom flange.

Figure 29 shows a crack that formed in the floor-beam web above the end of the riveted angle end connection. Visual observation showed that a horizontal displacement in the direction of the bridge developed between the tie girder and the top flange of the floor beam at the end connection. This relative movement produces out-of-plane deformation in the small web gap between the end of the floor-beam angle end connection and the top flange of the floor beam.

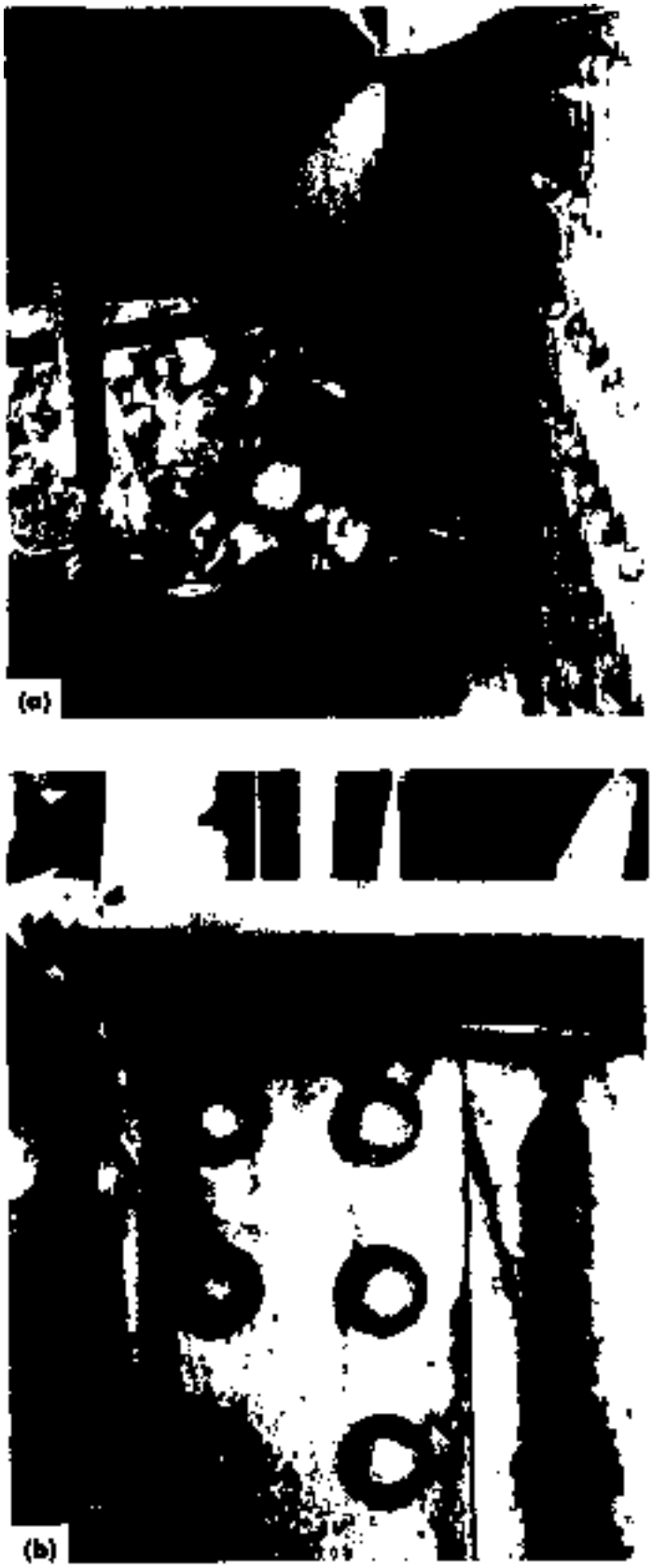


Fig. 29 Cracking in a floor-beam web above the end of the riveted angle end connection. (a) Floor beam/tie girder connection. (b) Crack along the web-flange weld above the end connection

At least eight tied-arch structures have experienced cracking in the floor beams along the web-flange connection of the top flange. In some of these structures, the stringers rested on the top of the floor-beam flange; in others, the stringers framed into the floor beam.

References cited in this section

1. J.W. Fisher, *Fatigue and Fracture in Steel Bridges*, Wiley-Interscience, 1984

15. J.W. Fisher, Fatigue Cracking in Bridges From Out-Of-Plane Displacements, *Can. J. Civil Eng.*, Vol 5 (No. 4), 1978

16. W. Hsiong, Repair of Poplar Street Complex Bridges in East St. Louis, *Trans. Res. Rec.*, No. 664, 1978

Failures of Bridge Components

John W. Fisher, Fritz Engineering Laboratory, Lehigh University

References

1. J.W. Fisher, *Fatigue and Fracture in Steel Bridges*, Wiley-Interscience, 1984

2. J.W. Fisher and D.R. Mertz, Hundreds of Bridges—Thousands of Cracks, *Civil Eng.*, April 1985

3. J.W. Fisher and I.M. Viest, *Fatigue Life of Bridge Beams Subjected to Controlled Truck Traffic*, Preliminary publications, 7th Congress, International Association for Bridge and Structural Engineering, Zurich, 1964

4. J.W. Fisher, H. Hausammann, M.D. Sullivan, and A.W. Pense, *Detection and Repair of Fatigue Damage in Welded Highway Bridges*, NCHRP report 206, Transportation Research Board, Washington, June 1979

5. D.G. Bowers, *Loading History Span 10 Yellow Mill Pond Bridge 195 Bridgeport, Connecticut*, Research project HPR 175-332, Connecticut Department of Transportation, Whethersfield, CT, 1972

6. R.L. Dickey and T.P. Severga, *Mechanical Strain Recorder on a Connecticut Bridge*, Report FCP 45G1-222, Federal Highway Administration Department of Transportation, Washington, 1974

7. J. W. Fisher, R.E. Slockbower, H. Hausammann, and A.W. Pense, Long Time Observation of a Fatigue Damaged Bridge, *J. Tech. Counc.*, Vol 107, (No. TCl), April 1981

8. J.W. Fisher, K.H. Frank, M.A. Hirt, and B.M. McNamee, *Effect of Weldments on the Fatigue Strength of Steel Beams*, NCHRP report 102, Transportation Research Board, Washington, 1970

9. J.P.C. King, P.F. Csagoly, and J.W. Fisher, Field Testing of Aquasabon River Bridge in Ontario, *Trans. Res. Rec.*, No. 579, 1976

10. J.W. Fisher, A.W. Pense, H. Hausammann, and G.R. Irwin, Analysis of Cracking in Quinnipiac River Bridge, *J. Struct. Div.*, Vol 106 (No. ST4), April 1980

11. J.W. Fisher, A.W. Pense, and R. Roberts, Evaluation of Fracture of Lafayette Street Bridge, *J. Struct. Div.*, Vol 103 (No. ST7), July 1977

12. J.W. Fisher, *Bridge Fatigue Guide—Design and Details*, Publication T112-11/77, American Institute of Steel Construction, Chicago, 1977

13. Engineers Investigate Cracked El, *Eng. News Rec.,* Vol 200 (No. 3), Jan 19, 1979

14. J.W. Fisher, J.M. Hanson, H. Hausammann, and A.E.N. Osborn, *Fracture and Retrofit of Dan Ryan Rapid Transit Structure,* Final report, 11th Congress, International Association for Bridge and Structural Engineering, Zurich, 1980

15. J.W. Fisher, Fatigue Cracking in Bridges From Out-Of-Plane Displacements, *Can. J. Civil Eng.,* Vol 5 (No. 4), 1978

16. W. Hsiong, Repair of Poplar Street Complex Bridges in East St. Louis, *Trans. Res. Rec.,* No. 664, 1978

Failures of Locomotive Axles

George F. Vander Voort, Carpenter Technology Corporation

Introduction

FAILURES OF LOCOMOTIVE AXLES caused by overheated traction-motor support bearings are discussed in this article. These failures are of interest because the analysis shows an example of what can be done when the fracture face and origin are destroyed during the failure incident. In most failure analyses of broken components, it is generally assumed that conclusive results cannot be obtained if the fracture and the fracture origin cannot be identified and examined. In many failures this is true. However, the failures described in this article possess some unique characteristics that permit successful analysis despite the lack of a preserved fracture face and origin.

Failures of locomotive axles due to overheated friction bearings are rather common in the railroad industry and have been observed for more than 100 years. Because of this, such failures are usually diagnosed merely by visual inspection of the damage. Comprehensive metallographic studies, therefore, are not done and have not been accurately documented in the open literature for many years. However, detailed analysis of such failures reveals a number of significant features.

Failures of Locomotive Axles

George F. Vander Voort, Carpenter Technology Corporation

Background

Friction bearings have been used for many years and are perfectly adequate if they are lubricated. The bearing is essentially a bronze cylinder lined with babbitt metal. Typically, the bronze alloy composition is close to Cu-16Pb-6Sn-3.5Zn, while the babbitt composition is usually Pb-3.5Sn-8-11.5Sb. A window is cut into the bearing and packed with cotton waste, which trails down to an oil reservoir. Oil is drawn up the wick during service to lubricate the contacting surfaces.

Friction-bearing failures due to overheating have been examined by the metallurgist for many years. Perhaps the earliest example, documented in 1914, involved a failed Krupp railroad axle (Ref 1). The study revealed a rather complex crack pattern, evidence of exposure to very high temperatures, and ruptures in the overheated region. Bronze bearing metal was observed in the cracked surface region located beneath the support bearing. The bearing metal was molten when it penetrated the axle. In addition, the railroad that submitted the failed

axle had tried to remove surface evidence of overheating. Also mentioned was earlier work that had been done on broken axles that were overheated. In each of these cases, small particles of bearing metal had penetrated the axle in the overheated region.

In 1944, a review was published of railroad-car axle failures due to the absorption of molten copper (Ref 2). Most axle-journal failures occurred near the wheel hub, an area of high stress and temperature. A broken axle was shown in which the fracture was not destroyed after breakage had occurred. From the surface inward, the fracture surface was rough, indicating the depth of copper penetration. The central portion was smooth, indicative of a fatigue fracture that ultimately led to failure. Color photomicrographs revealed a yellow grain-boundary copper phase. It was pointed out that adequate lubrication is required to keep the operating temperature of the contacting surface below the melting point of the bearing materials.

Two reports were issued—one in 1947, the other in 1954—concerning copper-penetration axle failures and the use of nondestructive testing to detect surface cracks in such failures (Ref 3, 4). The 1947 paper (Ref 3) discusses twist-off failures due to overheated bearings. This failure mode is referred to as a hot-box in railroad terminology. Almost all axle-journal failures were claimed to be attributed to intergranular embrittlement of the steel by molten brass or copper. The steel in contact with the bearing must be heated to a temperature above the melting point of the brass journal bearing and must be under a stressed condition. Experiments were performed with 13-mm ($\frac{1}{2}$-in.) diam medium-carbon steel, loaded as cantilever beams. Samples heated to 925 °C (1700 °F) ran for hours without failure at 1750 rpm. However, the instant they were wetted with molten brass, catastrophic failure occurred. These samples were loaded above the yield point of the steel at 925 °C (1700 °F). Samples were also treated in the same manner, but the load was removed before complete rupture occurred. Microscopic analysis showed that molten brass entered the steel in a narrow canyon at the surface, then spread out in a delta pattern.

The 1954 report (Ref 4) studied two failed railroad axles (from the Macon, Dublin & Savannah and the Chicago & Northwestern Railroads) with evidence of copper penetration. The axle from the Macon, Dublin & Savannah Railroad had gross cracking with visible copper-colored material in the cracks. Spectrographic analysis of samples from both axles containing copper penetration revealed that the major constituents were copper and lead with a minor amount of tin. These elements are the major constituents of the bronze friction bearing. The copper-penetration failure occurred in the following sequence:

- The bearing surface was heated by friction because of loss of lubrication
- The babbitt metal lining melts between about 240 and 315 °C (465 and 600 °F) and wets the surface, but penetration does not occur
- The babbitt metal is displaced, possibly by mechanical action or volatization
- The bronze backing is heated to its melting point (900 to 925 °C, or 1650 to 1700 °F) and penetrates the axle, causing failure

In 1959, the New York Central Railroad Company studied copper penetration in overheated journals (Ref 5). Two types of failures were observed. The first type, referred to as a burn-off, is indicative of a single continuous heating to failure due to penetration of bearing metals into journals at elevated temperatures. The second type, referred to as a cold break, also results from overheating, but is a two-stage failure process. These fractures exhibit an outer circumferential zone of irregular detail with evidence of thermal checks and intergranular separation and an inner fracture zone typical of a progressive-type fatigue fracture. In both cases, copper from the bronze-backed journal bearings was absorbed intergranularly into the hot steel journal. A surface analysis for copper indicated that the copper content exceeded the residual copper content of the steel to a depth of 1 mm (0.040 in.). A metallographic study confirmed the presence of grain-boundary copper penetration in the affected surface layer. This work was subsequently published (Ref 6).

An important source of information on copper-penetration failures is the reports of the Committee on Axle and Crank Pin Research, formed at the 1949 annual meeting of the Association of American Railroads (AAR). A major problem faced by this group concerned the decision to scrap or to recondition overheated axles. Overheating can result from loss of lubrication of either friction or roller bearings and possibly from other mechanical problems. Because the number of overheated axles per year was considerable, outright scrapping of all overheated axles represented a significant expense.

In the 1951 proceedings of the AAR, two types of overheating failures were described (Ref 7). One type of burn-off failure exhibited a necked-down, elongated fracture that generally displayed large surface ruptures. The diameter may be necked down to half its original size before the axle twists off. Copper penetration is not observed in these failures. In the second type of overheating failure, the fracture is flat across the section with no reduction in diameter. Copper penetration and thermal cracks are usually found to be present in this type of failure. The Union Pacific Railroad reported that when hot steel and copper are in contact, the steel is affected by the copper and breaks sharply without a reduction in diameter. The Atcheson, Topeka & Santa Fe Railroad reported that when an axle that is necked down and elongated is inspected, the brass is usually broken, and most of the journal brass is usually intact in the box.

In 1951 and in 1953, the Delaware, Lackawana & Western Railroad reported a revealing test (Ref 7, 8). A non-copper-containing (cast iron) bearing that was not lubricated was placed on one side of the axle, while a lubricated copper-base bearing was put on the other side. The car (a hopper car with a 63,500-kg, or 140,000-lb, load) was run for 108 km (67 miles) without burning off the journal, although it was red hot. The lubrication was then removed from the bronze bearing at the other end of the axle. The car traveled 55 km (34 miles) before the journal broke under the bronze bearing due to copper penetration (Ref 7). The journal under the nonlubricated cast iron bearing was still red hot, but had not broken.

The 1952 AAR proceedings discussed stress measurements made on a new 140- × 250-mm ($5\frac{1}{2}$ × 10-in.) standard black-collar freight-car axle fabricated from AAR M-126-49, grade F, steel (Ref 9). Strain gages were placed at five locations, and the journal loads were 69 to 138 MPa (10 to 20 ksi), with speeds from 65 to 135 km/h (40 to 84 mph). Results showed that the dynamic stresses on the journals were very low, in most cases less than one-half the level of stress at the wheel seat.

In this meeting, there was considerable discussion on the reuse of overheated axles. It was thought that if the bearing lining were melted out but surface cracks were not observed, the axle could be safely returned to service. However, when cracks were detected, it was thought that the axle surface could not be turned down to remove the crack without going below the minimum allowable diameter.

The Pennsylvania Railroad favored the scrapping of overheated axles when the babbitt metal was melted out. Experience indicated that about 90% of the overheated axles contained cracks that could not be turned out within the minimum diameter tolerance permitted. The expense of turning down all overheated axles when only 10% could be salvaged was concluded to be unjustifiable. The Southern Railway reported that 85% of its broken axles had been turned down previously, indicating that they had been overheated in earlier service. Removal of the cracking was concluded to be insufficient to guarantee that the axle was safe for additional service. The Atcheson, Topeka & Santa Fe Railroad reported on tests conducted using 19- to 25-mm ($\frac{3}{4}$ to 1-in.) diam miniature journals that were heated, loaded, and subjected to melted bearing metal. The tests resulted in the type of break that 99 out of 100 burn-off journals show, and 99% of the journals exhibited copper penetration.

The 1952 proceedings (Ref 10) contained additional information on the Laudig iron-backed journal bearing that was developed to overcome the copper-penetration problem encountered with bronze-backed bearings. The Laudig journal bearing was recommended as an alternative to the bronze-backed journal (AAR M-501-48). In 1950, six railroads applied 600 test bearings successfully.

In the 1956 AAR proceedings (Ref 11), the Baltimore & Ohio Railroad stated that at least three factors must exist to cause copper-penetration fractures:

- A bearing that contains a metal or metals that will diffuse into the axle material in the grain boundaries
- Heat of the proper range
- Stress

Wetting of the axle surface was considered to be a possible fourth prerequisite. A comprehensive, cooperative study was suggested to solve the problem of copper-penetration fractures. The Chesapeake & Ohio Railroad reported on an investigation that began in 1952 to determine what could be done to reduce or eliminate wrecks and hot-boxes due to defective journals. The railroad developed a journal-inspection car that used ultrasonic inspection. During preliminary testing of this device, a cracked axle was found under a loaded car that was ready to be dispatched. The axle was removed and broken open. The crack had penetrated about an inch from

the surface of the journal around approximately one-half of the axle circumference. Subsequent use of this inspection car showed that about one defective axle was found for every 580 cars inspected, based on inspection of about 70,000 cars. Testing of cars owned by other railroads revealed that about 40% of the axles tested were defective.

As a result of axle failures due to copper penetration, the AAR conducted a study to determine the effects of refined copper or alloyed copper on axles under hot-box conditions (Ref 12). The study attempted to answer the often raised question of whether the copper is a precursor of axle failure or is a post-fracture phenomenon, that is, does the molten copper flow into the grain boundaries and cause failure or does the molten copper flow into stress cracks after failure. In these tests, 19-mm ($\frac{3}{4}$-in.) diam centerless-ground AISI 1045 carbon steel samples were left bare or wrapped with 0.13-mm (0.005-in.) thick copper foil and held in place with fine copper wire. In some tests, wrapping was done with 70Cu-30Zn. Specimens that fractured in the presence of brass failed suddenly and had sharp fracture faces; that is, no neckdown resulted. Copper penetration was found in these samples. The temperature above which failures occurred decreased as the applied bending stress increased. As the temperature increased, the probability of fracture increased.

Two distinct fracture modes were observed in these tests. When the failed samples exhibited reduced diameters (necking), evidence of twisting and the presence of cavernous cracks and fissures were due to high-temperature fatigue in an oxidizing atmosphere. When a brittle failure occurred, that is, a sharp, well-defined fracture, microscopic examination showed copper or brass in the grain boundaries. This type of fracture occurred at temperatures above the melting point of the penetrating metal. The penetration followed grain-boundary paths, as in stress-corrosion cracking (SCC). Branching of the penetration was found in the grain structure of several specimens. Copper penetration was as deep as 10 mm (0.4 in.) below the surface.

There is a considerable body of published information that shows that steels will fail by grain-boundary penetration of molten copper. The following conditions are required:

- Presence of applied tensile stress
- Wetting of the steel surface by molten copper
- Heating of the steel substrate to temperatures high enough to austenitize the steel

The failure of railroad axles due to the overheating of bronze-backed bearings has been well documented. It is well recognized that the steel used in these axles is sensitive to liquid-metal embrittlement (LME) by molten copper when the steel surface is heated above the austenitizing temperature by frictional heat due to loss of lubrication. Stress is present in the axle at this location due to the weight of the traction motor, but the stress level is less than one-half that at the gear seat.

Overheating of friction bearings results in LME and failure of the axle. Liquid-metal embrittlement is a complex phenomenon; the example described in this article is only one form of LME. Because of the different forms of LME that can occur, a brief review of the nature of LME is presented.

Liquid-metal embrittlement is a phenomenon in which the ductility or fracture stress of a solid metal is reduced by exposure of the surface to a liquid metal (Ref 13). The subject has been extensively reviewed (Ref 13, 14, 15, 16, 17, 18), and specific aspects have been reported in numerous scientific papers. The first recorded recognition of the problem was made in 1914 by Huntington, who observed the embrittlement of α-brass by mercury (Ref 19). Since then, LME has been identified in numerous failure analyses that include some notable cases: damage to aircraft by liquid-gallium embrittlement of aluminum and the embrittlement of stainless steel piping by molten zinc.

References 13, 14, 15, 16, 17, 18, and 19 have helped to eliminate some of the confusion in the literature regarding LME. This confusion is partly due to the existence of several different forms of LME. Four distinct forms have been identified (Ref 13):

- Instantaneous fracture of a particular metal under an applied or residual tensile stress when in contact with particular liquid metals. This is the most common type of LME
- Delayed failure of a particular metal in contact with a specific liquid metal after a certain time interval at a static load below the ultimate tensile stress of the metal. This form of LME is related to grain-boundary penetration by the liquid metal and is less common than the first form

- Grain-boundary penetration of a particular solid metal by a particular liquid metal such that the solid metal eventually disintegrates. Stress is not a prerequisite of this form of LME in all observed cases
- High-temperature corrosion of a solid metal by a liquid metal, causing embrittlement. This problem is entirely different from the others

In some respects, LME is similar to SCC. However, SCC always requires a measurable incubation period before fracture occurs. Hence, SCC is more similar to the second form of LME than to any of the other types. Like SCC, LME is a rather perplexing subject since embrittlement is restricted to only certain combinations of solid and liquid metals.
The following observations have been made regarding liquid metal/solid metal interactions (Ref 15):

- No apparent interaction
- Simple dissolution of the solid metal into the liquid metal
- Simple diffusion of the liquid metal into the solid metal
- Formation of intermetallic compounds at the liquid/solid interface
- Intergranular penetration of the liquid metal into the solid metal without the presence of applied or residual stress
- Brittle, premature failure of the solid metal due to intergranular penetration of the liquid metal into the solid metal under the influence of applied or residual tensile stress. Failure occurs after a finite incubation period
- Brittle, instantaneous failure of a stressed metal when a particular liquid metal is applied. Grain-boundary penetration is not necessarily observed

The last three types of interactions are the prime forms of LME. Most studies of LME discuss the last type, the instantaneous form of embrittlement, which is the most common form. Several authors refer to this form as adsorption-induced embrittlement. Other authors prefer to consider LME as merely a special type of brittle fracture and view the subject from a fracture mechanics approach.
A large number of metals and alloys can fail in this manner. Only specific liquid metals are known to embrittle each specific metal or alloy. Indeed, this is one of the more perplexing aspects of LME, and some investigators refer to this as specificity. Many of the studies of LME couples have involved tests consisting of only one temperature and fixed conditions regarding stress, grain size, and so on (Ref 19). Until all possible couples have been thoroughly studied using a wider range of test conditions, the specificity feature is of questionable validity (Ref 20). In addition, the LME tests have generally been “go, no-go” type tests rather than quantitative measurements of the degree of embrittlement.
The following conditions are required but are not necessarily sufficient to cause embrittlement (Ref 13):

- Wetting of the solid metal substrate by a liquid metal must be adequate
- An applied or residual tensile stress must be present in the solid metal
- A barrier to plastic flow must exist in the solid metal at some point in contact with the liquid metal

Obviously, for failure to occur the liquid metal must be of the correct composition to cause LME of the particular solid metal.
If the solid metal being embrittled is not notch sensitive, as in the case of face-centered cubic (fcc) metals, the crack will propagate only when the liquid metal feeds the crack. In a notch-sensitive metal—for example, in a body-centered cubic (bcc) metal, such as iron—the nucleated crack may become unstable and propagate ahead of the liquid metal.
In cases of LME, crack-propagation rates can be extremely rapid. Indeed, crack-propagation rates between 50 and 500 cm/s^2 (20 and 200 $in./s^2$) have been measured. In most cases, the crack paths are intergranular; however, instances of transgranular crack paths have been observed.
Factors influencing the fracture stress and ductility of metals subject to LME include:

- Composition of the solid metal
- Composition of the liquid metal
- Temperature

- Strain rate
- Grain size
- Thermal-mechanical history of the solid metal

Studies of specific LME couples have empirically identified certain trends that are usually, but not always, obeyed. For example, most embrittlement couples have very little mutual solid solubility. Solid metals that are highly soluble in the liquid metal and solid/liquid metal combinations that form intermetallic compounds are usually immune to LME. These studies have also demonstrated that good wetting of the substrate is required in LME. Interestingly, the factors that promote the lowest interfacial energy and therefore the best wetting are high mutual solid solubility or intermetallic-compound formation, both of which are generally related to a low degree of LME.

The presence of notches or stress concentrators increases the severity of LME, and many current investigators have used a fracture mechanics approach to understand LME. Since stress raisers are known to influence brittle failure detrimentally under ordinary (dry) situations, the effect on LME is understandable.

In the previously discussed adsorption-induced mode of LME, grain-boundary penetration is not a prerequisite. There are a few well-known cases in which LME occurs by the rapid penetration of a particular liquid metal along grain boundaries. Examples of this form of LME are much less common and not as well studied as the previously described form. Known examples include the wetting of aluminum by gallium (Ref 15), carbon steel by copper (Ref 15), and copper by bismuth (Ref 21).

Failures Caused by LME of Steel. The first reported case of LME of steel, in this instance by molten brass, was documented in 1927 (Ref 22). In 1931, tests were conducted on the influence of certain liquid metals on plain carbon steel, silicon steel, and chromium steel (Ref 23). These steels were embrittled at 1000 to 1200 °C (1830 to 2190 °F) by liquid tin, zinc, antimony, copper, 5% tin-bronze, and 10% zinc-brass. Liquid bismuth, cadmium, lead, and silver caused little or no embrittlement. In 1935, the bend strength of steel coated with a lead-tin solder at 250 °C (480 °F) and with a bearing metal at 350 °C (660 °F) was determined (Ref 24). Penetration of the liquid metals was observed on the tension side only of the bend specimens. In a similar study (Ref 25), the influence of liquid zinc at 425 to 500 °C (795 to 930 °F) on various steels was tested using bend, tensile, and creep-rupture specimens. Grain-boundary penetration of the zinc into the steels was observed. Another study investigated intercrystalline failures of a nickel-chromium steel airplane axle (Ref 26). The fracture occurred in the middle of the tubular axle where a brass number plate was attached by soft solder. When 50Pb-50Sn solder was applied to stressed rings cut from the axle, the specimen ruptured rapidly. The intergranular crack paths exhibited some solder penetration. The same results were observed when liquid tin, lead, zinc, cadmium, or Lipowitz alloy (a tin-lead-bismuth-cadmium alloy sometimes containing indium) was used.

In 1968, studies were conducted on the influence of cold work on the LME of pure iron and Fe-2Si (Ref 27). Sheet samples were tested in the annealed state and after 10, 25, 50, and 75% reduction in thickness. The degree of embrittlement increased with a small amount of cold work, but further working reduced the degree of embrittlement. For pure iron, maximum embrittlement occurred with the sample reduced 10% in thickness, while LME did not occur with the sample reduced 75%. For the Fe-2Si alloy, maximum embrittlement occurred with 25% reduction in thickness, and LME did not occur for either 50 or 75% reduction. The lack of LME at high levels of cold work was attributed to the fragmentation of the grain boundaries.

From 1970 to 1974, a series of studies was performed on the LME of iron (Ref 28, 29, 30, 31). These tests were made under nonoxidizing conditions using notched tensile-creep test specimens. A thin copper coating (8.5 μm thick) was plated onto the notched surface using a cyanide bath. The amount of copper was increased by wrapping pure copper wire (0.18 mm, or 0.007 in., in diameter) around the notch. Testing was done at 1100 and 1130 °C (2010 and 2065 °F) under stresses of 8.3 and 11 MPa (1.2 and 1.6 ksi). Copper significantly altered the creep behavior of pure iron and caused premature failure. An increase in applied stress or temperature reduced the time for LME failure. Embrittlement was of the delayed type and occurred by diffusion-controlled grain-boundary penetration of copper. The liquid copper appeared to be ahead of the growing surface cracks. The depth of surface cracking was found to be controlled by the depth of copper penetration. Dihedral-angle measurements (Ref 30) for liquid copper in steel at 1100 and 1130 °C (2010 and 2065 °F) revealed that the most frequent angle was 34° for both temperatures. The liquid copper/austenite interfacial energy was calculated as 444 ergs/cm^2 at 1100 and 1130 °C (2010 and 2065 °F). Sessile drop

measurements of the contact angle between liquid copper and steel revealed similar results: 35° at 1100 °C (2010 °F) and 28° at 1130 °C (2065 °F).

Five earlier studies regarding the dihedral angle are significant. In the first, the dihedral angle of copper in steel was measured at 30° (Ref 32). In the second, a minimum value was observed for the dihedral angle in the region of the melting point of copper (Ref 33). The third study claimed that stress had little effect on the dihedral angle; however, stress was observed to promote spreading of the liquid metal along the grain boundaries (Ref 34). The fourth study maintained that the dihedral angle decreased with increasing applied stress (Ref 35). In the last study, it was found that in the presence of molten copper, the surface energy that must be applied to drive a crack along austenite grain boundaries is reduced by about a factor of 100, and the corresponding stress required is about one-tenth that required without the presence of copper (Ref 36).

There have been numerous studies of LME by copper occurring during welding and joining processes applied to steel. One such study involved the influence of copper deposition by welding of austenitic and ferritic stainless steels (Ref 37). In stainless steels with mixed ferritic-austenitic structures, the presence of the stable ferrite phase in austenite reduced the penetration of copper. When the stable ferrite content was greater than 30%, penetration of copper was not observed. When a wholly ferritic stainless steel was deposited with copper, cracking did not occur. The wettability of molten copper on austenitic and ferritic stainless steel was measured. The contact angle was 92 to 100° for the ferritic stainless (no wetting) and 22 to 28° for the austenitic stainless steel (wetting) at 1100 °C (2010 °F).

A series of reports has been published regarding the embrittlement of iron-base alloys by copper during welding (Ref 38, 39, 40, 41). In a study of HY-80 steel welded with or without copper-nickel filter metal (Ref 38), infiltration of the grain boundaries by the copper-nickel deposit under an applied strain field was observed. The attack occurred in the heat-affected zone (HAZ) while the steel in this area was austenitic. In heated areas that remained bcc, penetration did not occur.

In another study involving copper deposition using a gas metal welding arc (Ref 39), alloy steels, such as AISI 4340 and 4140, and type 304 stainless steel were found to be extremely sensitive to copper penetration. In comparison, the penetration depths of AISI 1340, 1050, Armco iron, and carburized Armco iron were only about one-third those of the heat-treated alloy steels. A ferritic stainless steel, AISI 430, was almost completely immune to copper penetration. When stresses are absent, the ease of penetration by molten copper was concluded to be a function of the alloy content of the steel. When grain-boundary copper penetration occurs under an applied stress, partially filled or open cracks are formed in the steel. Notch-sensitive heat-treated steels lose much of their strength and ductility when copper penetration occurs. Steels that remain ferritic at the melting point of copper do not lose strength as a result of exposure to molten copper.

In a subsequent study (Ref 41), the Gleeble high strain rate hot-tensile test machine was used to determine the influence of temperature, atmosphere, stress, grain size, strain rate, and amount of copper on LME of iron- and cobalt-base superalloys. A copper contamination of only 0.08-mm (0.003-in.) thickness was found to be sufficient to cause hot cracking. Susceptibility to LME decreased with increasing temperature above some transition temperature in the HAZ. The intergranular LME cracks were oriented perpendicular to the direction of the principal stress. The hot cracks grew at high velocities after a critical amount of plastic strain was introduced in the HAZ.

It is well recognized that copper, as well as certain other elements in steel, can detrimentally influence hot workability (Ref 42, 43, 44, 45, 46). These studies have generally employed hot-torsion, bend, or cupping tests to assess the influence of the concentration of certain elements on hot workability. Much of the early work on the influence of copper on hot workability can be summarized (Ref 42):

- Because the solubility of copper in iron is low, the concentration of copper appears as a layer of nearly pure copper between the metal and the scale
- If the rolling or forging temperature is above the melting point of copper, the molten copper on the surface of the steel will penetrate the steel along the grain boundaries

Elements such as tin, arsenic, and antimony decrease the melting point of copper and increase the sensitivity of steel to surface cracking. Tin reduces the solubility of copper in austenite by a factor as high as three (Ref 43). Hence, tin additions can cause the precipitation of a molten copper phase at a much lower level of copper enrichment than in the absence of tin.

It is quite clear that molten copper, or copper alloys, will embrittle steels under the proper conditions. The steel must be austenitic with tensile stresses present. Carbon steels contacting liquid copper will be heated into the fully austenitic range. Liquid copper will penetrate the austenitic grain boundaries. The speed of penetration depends on the magnitude of the applied tensile stresses. The depth of cracking depends on the depth of copper penetration. Thus, if the elements present in the friction-bearing sleeve are observed in the prior-austenite grain boundaries in failed axles, it can be safely concluded that the axle was intact and under normal loading when the bearing over-heated and that frictional heating due to loss of lubrication heated the axle and the bearing to a temperature at which the axle surface was austenitic and the bearing surface was molten. If the axle was broken before the bearing overheated, it would not be stressed in tension at the failure location, and penetration of bearing elements would not occur, irrespective of the crystallographic structure of the axle.

References cited in this section

1. B. Straub, Mikroskopische Stahluntersuchung, *Stahl Eisen,* Vol 34 (No. 50), Dec 1914, p 1814–1820

2. F.H. Williams, Car Axle Failures Traced to Absorption of Bearing Metal, *Prod. Eng.,* Vol 15, Aug 1944, p 505–508

3. "The Mechanism of Copper Penetration Failure," Magnaflux Corp., Chicago, 1947

4. "Copper Penetration—New and Revealing Information Concerning the Analysis of the Penetrating Material Itself," Research Laboratory Technical Memo No. 220, Project 125, Magnaflux Corp., Chicago, Feb 1954

5. E.A. Unstich and E.J. Palinkas, "A Study of Copper Penetration in Overheated Journals," Report 69222, New York Central Railroad Company, New York, July 1959

6. New York Central Research Shows … Turning Removes Copper Penetration, *Rail. Locom. Cars,* Feb 1961, p 34–35

7. Report of the Committee on Axle and Crank Pin Research, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1951, p 158–166

8. J. J. Laudig, Hot-Box Research Results … in the Iron-Back Journal Bearing, *Rail. Age,* Vol 134 (No. 12), March 1953, p 65–69

9. Report of the Committee on Axle and Crank Pin Research, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1952, p 116–129

10. Laudig Iron Back Journal Bearing, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1952, p 358–359

11. Report of Committee on Axles, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1956, p 94–104

12. "The Failure of Railroad Axle-Journal Steel Including the Effects of Copper-Bearing Metals in Contact with Stressed Steel Rods at Elevated Temperatures," Report MR-434, Association of American Railroads, Research Department, Washington, Dec 1963

13. N.S. Stoloff, Liquid Metal Embrittlement, in *Surfaces and Interfaces, II,* Proceedings of 14th Sagamore Army Materials Research Conference, Syracuse University Press, Syracuse, NY, 1968, p 157–182

14. A. R. C. Westwood *et al.,* Adsorption-Induced Brittle Fracture in Liquid-Metal Environments, in *Fracture,* Vol III, Academic Press, 1971, p 589–644

15. W. Rostoker *et al.*, *Embrittlement by Liquid Metals,* Reinhold, 1960

16. M.H. Kamdar, Liquid Metal *Embrittlement, in Embrittlement of Engineering Alloys,* Vol 25, *Treatise on Materials Science and Technology,* Academic Press, 1983, p 361–459

17. M.G. Nicholas and C.F. Old, Review: Liquid Metal Embrittlement, *J. Mater. Sci.,* Vol 14, Jan 1979, p 1–18

18. M. H. Kamdar, Ed., *Embrittlement by Liquid and Solid Metals,* American Institute of Mining, Metallurgical, and Petroleum Engineers, Warrendale, PA, 1984

19. A.R. Huntington, discussion of paper by C.H. Desch, The Solidification of Metals From the Liquid State, *J. Inst. Met.,* Vol 11, 1914, p 108–109

20. F.A. Shunk and W.R. Warke, Specificity as an Aspect of Liquid Metal Embrittlement, *Scr. Metall.,* Vol 8, 1974, p 519–526

21. E. Scheil and K.E. Schiessl, The Diffusion of Liquid Bismuth Along Grain Boundaries in Copper, *Z. Naturforsch,* Vol 4a, 1949, p 524–526

22. R. Genders, The Penetration of Mild Steel by Brazing Solder and Other Metals, *J. Inst. Met.,* Vol 37, 1927, p 215–221

23. H. Schottky *et al.,* The Red-Shortening of Steels by Metals, *Arch. Eisenhüttenwes.,* Vol 4, 1931, p 541–547

24. W. E. Goodrich, The Penetration of Molten White Metals into Stressed Steels, *J. Iron Steel Inst.,* Vol 132, 1935, p 43–66

25. W. Radeker, The Production of Stress Cracks in Steel by Molten Zinc, *Stahl Eisen,* Vol 73 (No. 10), May 1953, p 654–658

26. L.J.G. van Ewijk, The Penetration of Steel by Soft Solder and Other Molten Metals at Temperatures up to 400 °C, *J. Inst. Met.,* Vol 56 (No. 1), 1935, p 241–256

27. J.V. Rinnovatore and J.D. Corrie, The Effect of Cold Work on the Liquid Metal Embrittlement of Iron and Iron-Silicon, *Scr. Metall.,* Vol 2 (No. 8), Aug 1968, p 467–470

28. R.R. Hough and R. Rolls, The High-Temperature Tensile Creep Behavior of Notched, Pure Iron Embrittled by Liquid Copper, *Scr. Metall.,* Vol 4 (No. 1), 1970, p 17–24

29. R.R. Hough and R. Rolls, Creep Fracture Phenomena in Iron Embrittled by Liquid Copper, *J. Mater. Sci.,* Vol 6, 1971, p 493–1498

30. R.R. Hough and R. Rolls, Copper Diffusion in Iron During High-Temperature Tensile Creep, *Met. Trans.,* Vol 2, Sept 1971, p 2471–2475

31. R.R. Hough and R. Rolls, Some Factors Influencing the Effects of Liquid Copper on the Creep-Rupture Properties of Iron, *Scr. Metall.,* Vol 8 (No. 1), Jan 1974, p 39–44

32. L.H. VanVlack, Intergranular Energy of Iron and Some Iron Alloys, *Trans. AIME,* Vol 191, March 1951, p 251–259

33. W.J.M. Salter, Surface Hot Shortness in Mild Steel, *J. Iron Steel Inst.,* Vol 200, Sept 1962, p 750–751

34. R.B. Waterhouse and D. Grubb, The Embrittlement of Alpha-Brasses by Liquid Metals, *J. Inst. Met.*, Vol 91, 1962–1963, p 216–219

35. C.A. Stickles and E.E. Hucke, The Effect of Stress on the Dihedral Angle in Leaded Nickel, *J. Inst. Met.*, Vol 92, 1963–1964, p 234–237

36. R. Eborall and P. Gregory, The Mechanism of Embrittlement by Liquid Phase, *J. Inst. Met.*, Vol 84, 1955, p 88–90

37. E.A. Asnis and V. M. Prokhorenko, Mechanism of Cracking During the Welding or Depositing of Copper on to Steel, *Weld. Prod.*, Vol 12 (No. 11), 1965, p 15–17

38. S.J. Matthews and W.F. Savage, Heat-Affected Zone Infiltration by Dissimilar Liquid Weld Metal, *Weld. J.*, Vol 50 (No. 4), 1971, p 174s–182s

39. W. F. Savage *et al.*, Intergranular Attack of Steel by Molten Copper, *Weld. J.*, Vol 57 (No. 1), 1978, p 9s–16s

40. W.F. Savage *et al.*, Copper-Contamination Cracking in the Weld Heat-Affected Zone, *Weld. J.*, Vol 57 (No. 5), May 1978, p 145s–152s

41. W.F. Savage *et al.*, Liquid-Metal Embrittlement of the Heat-Affected Zone by Copper Contamination, *Weld. J.*, Vol 57 (No. 8), Aug 1978, p 237s–245s

42. J.W. Halley, Residual Elements in Steel, in *Effect of Residual Elements on the Properties of Metals*, American Society for Metals, 1957, p 71–87

43. D.A. Melford, Surface Hot Shortness in Mild Steel, *J. Iron Steel Inst.*, Vol 200, April 1962, p 290–299

44. K. Born, Surface Defects in the Hot Working of Steel Resulting from Residual Copper and Tin, *Stahl Eisen*, Vol 73, HB 3255, 1953, p 1268–1280

45. G.G. Foster and J.K. Gilchrist, The Influence of Copper, Nickel and Tin on the Hot Working Properties of Mild Steel, *Metall.*, May 1952, p 225, 228, 230

46. W.J.M. Salter, Effects of Alloying Elements on Solubility and Surface Energy of Copper in Mild Steel, *J. Iron Steel Inst.*, Vol 204, May 1966, p 478, 487

Failures of Locomotive Axles

George F. Vander Voort, Carpenter Technology Corporation

Results of Axle Studies

Three axles were examined. Two, coded 1611 and 2028, were provided by the Seaboard Coast Line Railroad. Visual examination of the failures indicated that overheating of the traction-motor support bearings, caused penetration of the bronze bearing material and failure. The third axle was the cause of a train derailment (Ref 47); a limited study was performed on this specimen.

Views of the fracture faces of axles 1611 and 2028 are shown in Fig. 1, 2, and 3. Both axles exhibit little, if any, necking in agreement with the general fractures of such failures. Axle 1611 was used in a yard engine, and little damage was done to the fracture after rupture. Axle 2028 experienced considerable deformation after fracture. Figures 1 and 2 of axle 1611

show that the outer portion of the fracture face experienced considerable tearing and rupture, while the central portion exhibits hot torsional rupture (note the swirl pattern of the fracture). This suggests that LME occurred to a depth of nearly 25 mm (1 in.). The axle, which was heated to at least 925 °C (1700 °F), then ruptured to failure by twisting under the applied loads. The fracture features of axle 2028 are less distinct because of the rubbing action after fracture (Fig. 3). The outside-diameter surface of each of these axles in the area in contact with the traction-motor support bearing exhibited scale, and some of the friction-bearing material was present, as identified by x-ray fluorescence and diffraction.

Fig. 1 Fracture surface at the drive-wheel side of axle 1611.

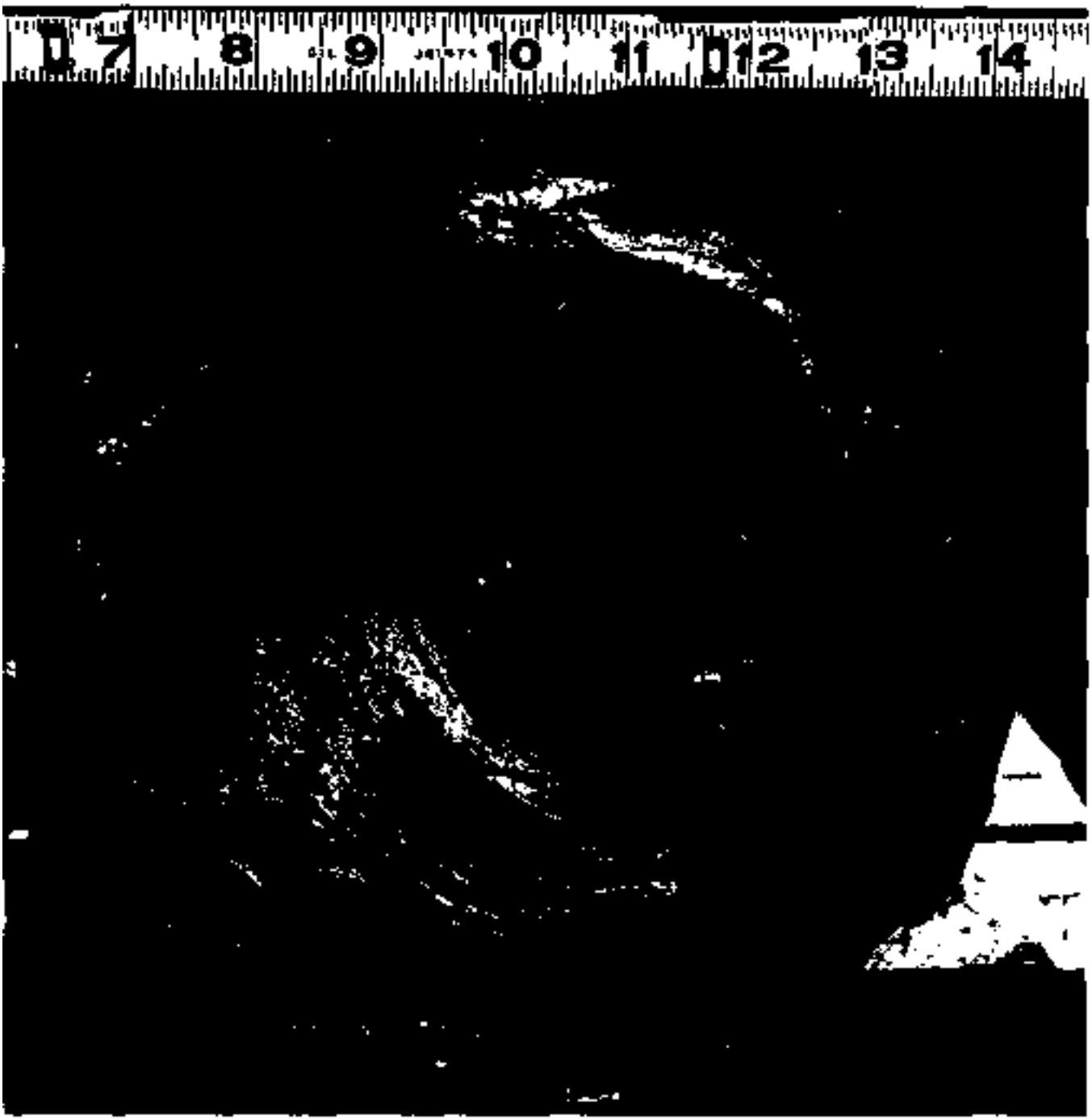


Fig. 2 Fracture surface at the commutator side of axle 1611.

Fig. 3 Fracture surface at the commutator side of axle 2028.

Figures 4, 5, 6, and 7 show longitudinal sections cut from each side of the axle fractures after hot-acid etching (1:1 HCl in water at 70 °C, or 160 °F). This etching revealed a number of secondary cracks propagating from the outside-diameter surface inward along the portion of the axle that was in contact with the overheated bearing. The regions near the centerline of each axle behind the fracture face reveal evidence of the high temperatures and the metal flow from twisting to rupture.

Fig. 4 Hot acid etched longitudinal centerline section from the drive-wheel side of axle 1611.

Fig. 5 Hot acid etched longitudinal centerline section from the commutator side of axle 1611.

Fig. 6 Hot acid etched longitudinal centerline section from the drive-wheel side of axle 2028.

Fig. 7 Hot acid etched longitudinal centerline section from the commutator side of axle 2028.

Examination of the microstructure along the surface of the axle from the fracture face outward reveals extensive penetration of the bearing elements. Macrographs of two of the microsamples are shown in Fig. 8, which illustrates some of the gross crack patterns. Figure 9 shows electron microprobe views of three typical regions from axle 1611. Specimen current images and scans for copper and tin are illustrated. The three indicated areas were analyzed quantitatively; results appear under the appropriate dot maps. Minor amounts of lead, zinc, and antimony were also detected, but were not analyzed quantitatively.

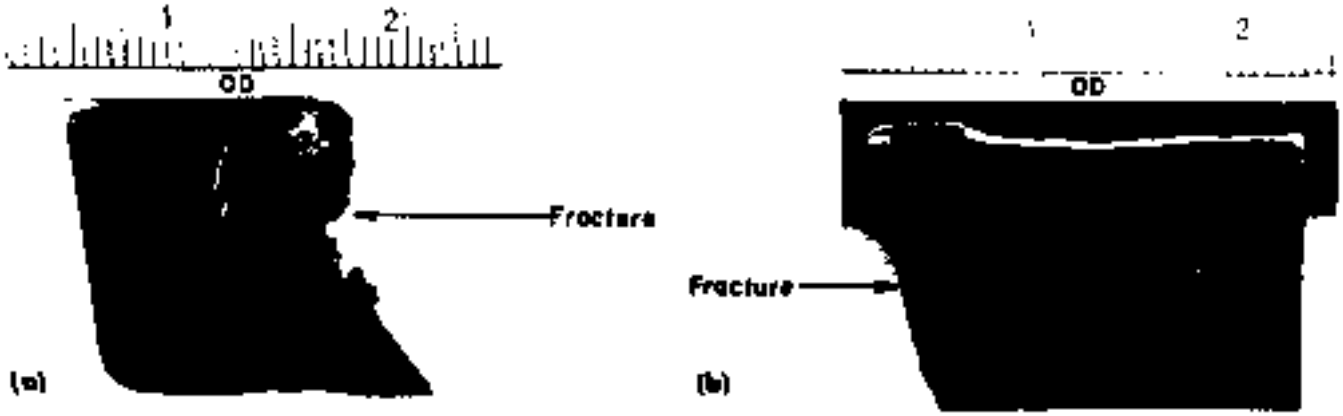


Fig. 8 Macrographs of two polished sections from the failed axles. (a) Axle 1611. (b) Axle 2028. Each was positioned along the fracture at the outside-diameter surface.

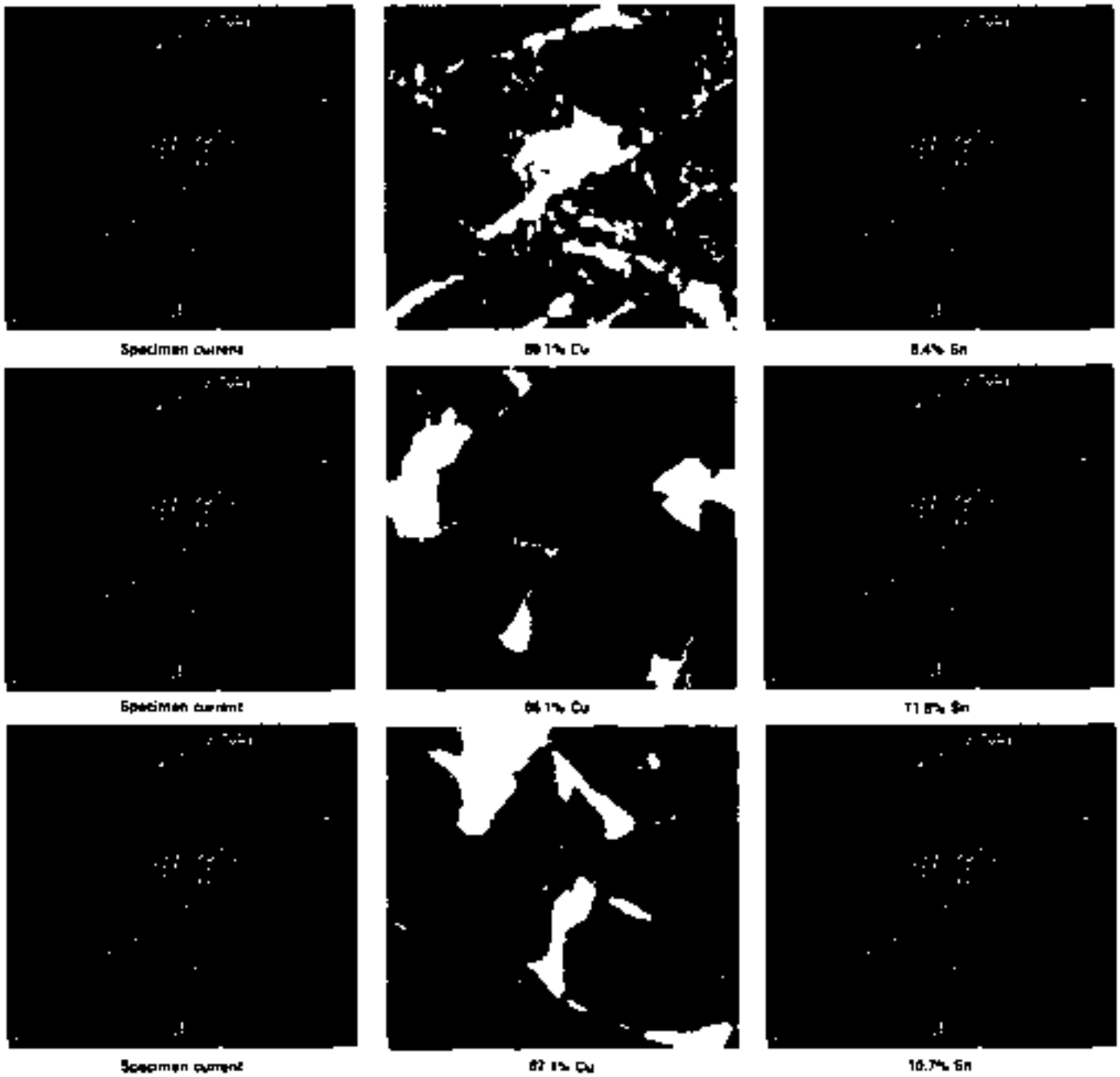
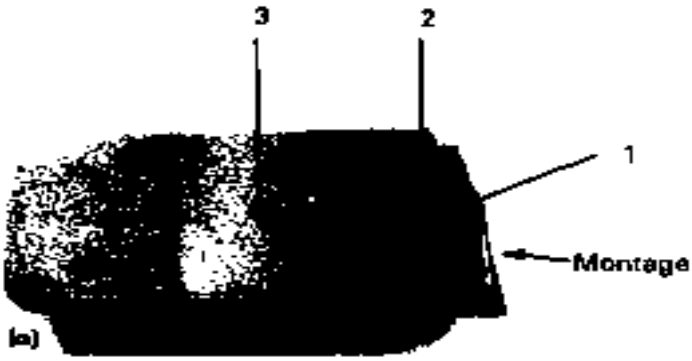


Fig. 9 X-ray elemental dot maps for copper and tin taken at three typical areas exhibiting penetration of the bearing elements. The three regions shown in the specimen current

images (left) were quantitatively analyzed for copper and tin; results are given under the concentration maps. All 320×

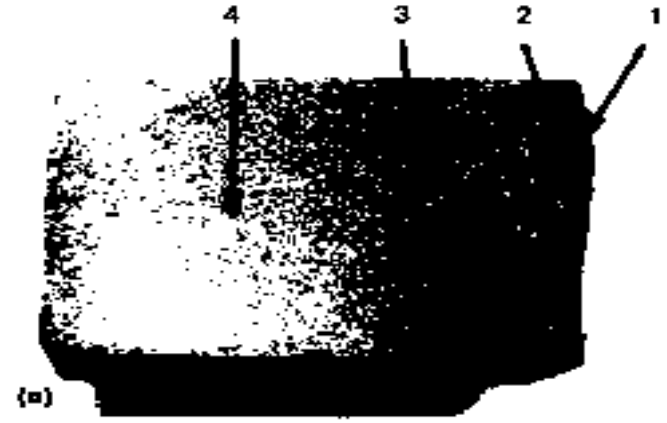
Figure 10 shows an etched section from the central region of axle 1611 and a montage of the microstructure. The temperature in regions 1 and 2 was high enough to reaustenitize the steel. Region 3 was heated into the two-phase region, while the lower portion of the sample, below region 3, was not heated above the lower critical temperature. The montage shows that the original longitudinal fiber of the forged axis was removed by the high temperature. The twisting action produced a metal-flow pattern perpendicular to the original hot-working axis and subsequent rupture. Around the larger crack, recrystallization of the grain structure occurred.


3
2
1
Montage
(a)

(b)

Fig. 10 Structure of a specimen cut from the area near the center of the fracture of axle 1611. (a) Macrograph of specimen. Actual size. (b) Montage showing the microstructure from the fracture surface inward to just above region 3.9×

Figure 11 shows a similar microsample cut from along the centerline of axle 2028 at the fracture face. The microstructure demonstrates the influence of the very high temperature and torsional rupture metal flow.

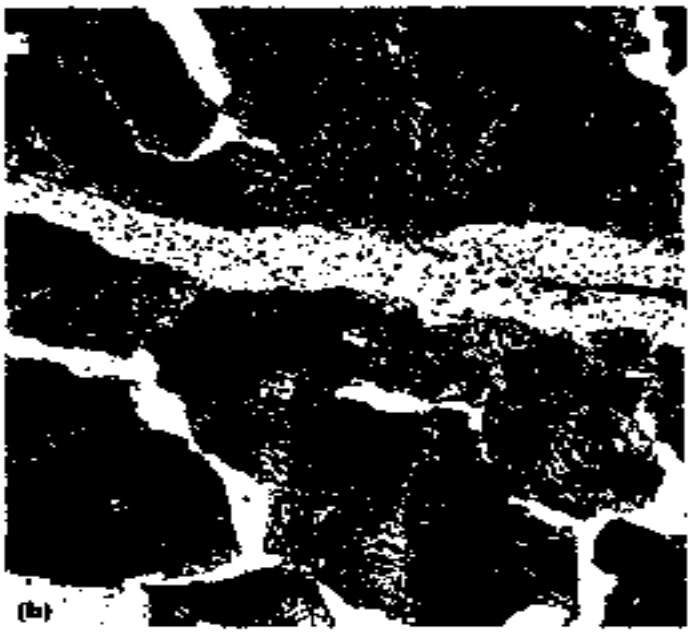

4
3
2
1
(a)

(b)

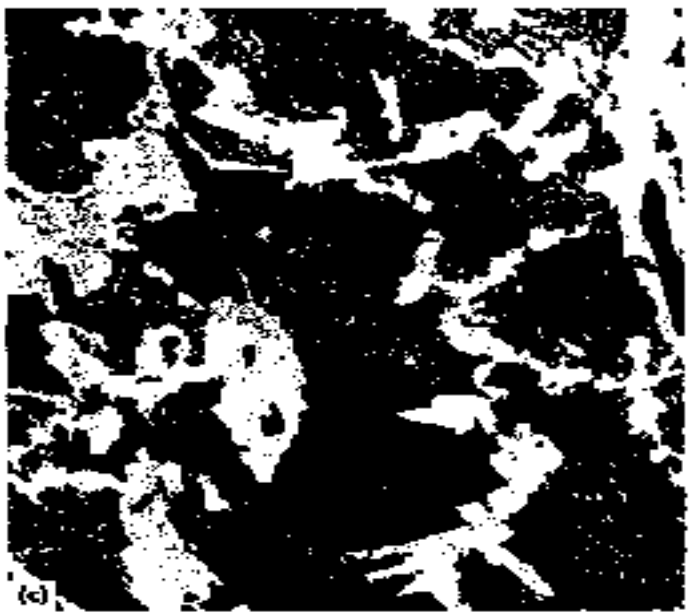
Fig. 11 Structure of a specimen cut from the area near the center of the fracture of axle 2028. (a) Macrograph of specimen. Actual size. (b) Montage showing the microstructure in regions 1 and 2. 14×

The influence of the high temperatures near the fracture and the rotational forces on axle 1611 are further illustrated in Fig. 12, which shows views of the microstructure taken at three locations with respect to the fracture along the centerline. All are oriented in the same manner; that is, the longitudinal axis is vertical in each illustration. At a distance of 38 mm ($1\frac{1}{2}$ in.) from the fracture face, the sulfide inclusions are parallel to the hot-working axis, as normally encountered. However, at 6.4 mm (0.25 in.) from the fracture, the sulfides are perpendicular to the longitudinal axis. In addition, they are much smaller, and many fine globular sulfides are observed. This indicates that the temperature was high enough at this location to dissolve most of the sulfides, which precipitated upon cooling as small globules. Near the edge of the fracture, only very fine spherical sulfides are observed, and considerable decarburization is evident.

(a)



(b)



(c)

Fig. 12 Three views of the microstructure of axle 1611. The specimens were taken from the area near the center at different distances from the fracture face. (a) Near the fracture surface. (b) 6.4 mm (0.25 in.) from the fracture face. (c) 38 mm (1.5 in.) from the fracture face. The longitudinal axle direction is vertically oriented in each micrograph. All 500×

Figure 13 shows similar micrographs from axle 2028 at regions 1 to 4 shown in Fig. 11. The microstructure in region D shows that the temperature was in the correct range to spheroidize the pearlitic structure to some extent (near the lower critical temperature). This same type of behavior was observed in samples from the third axle studied. Figure 14 shows measurements of the average angle of the sulfides with respect to the longitudinal axis as a function of the distance from the fracture face.

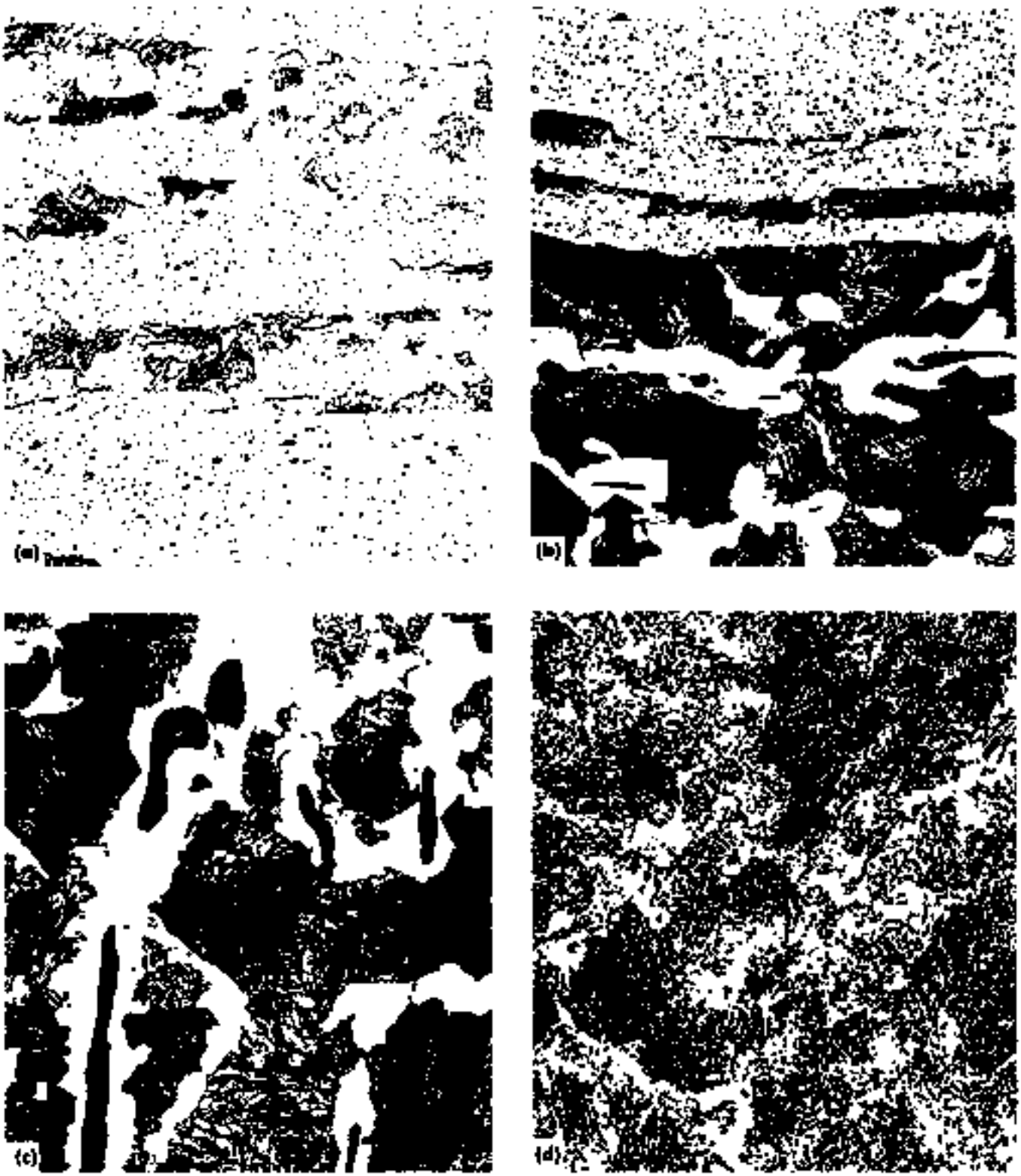


Fig. 13 Four views of the microstructure of axle 2028. The specimens were taken from the area near the center at different distances from the fracture face. (a) At the fracture surface. (b) At the boundary between the edge structure and the heat-affected structure. (c) At the heat-affected area. Arrows in (b) and (c) show sulfide inclusions. (d) 25 mm (1 in.) from the fracture surface

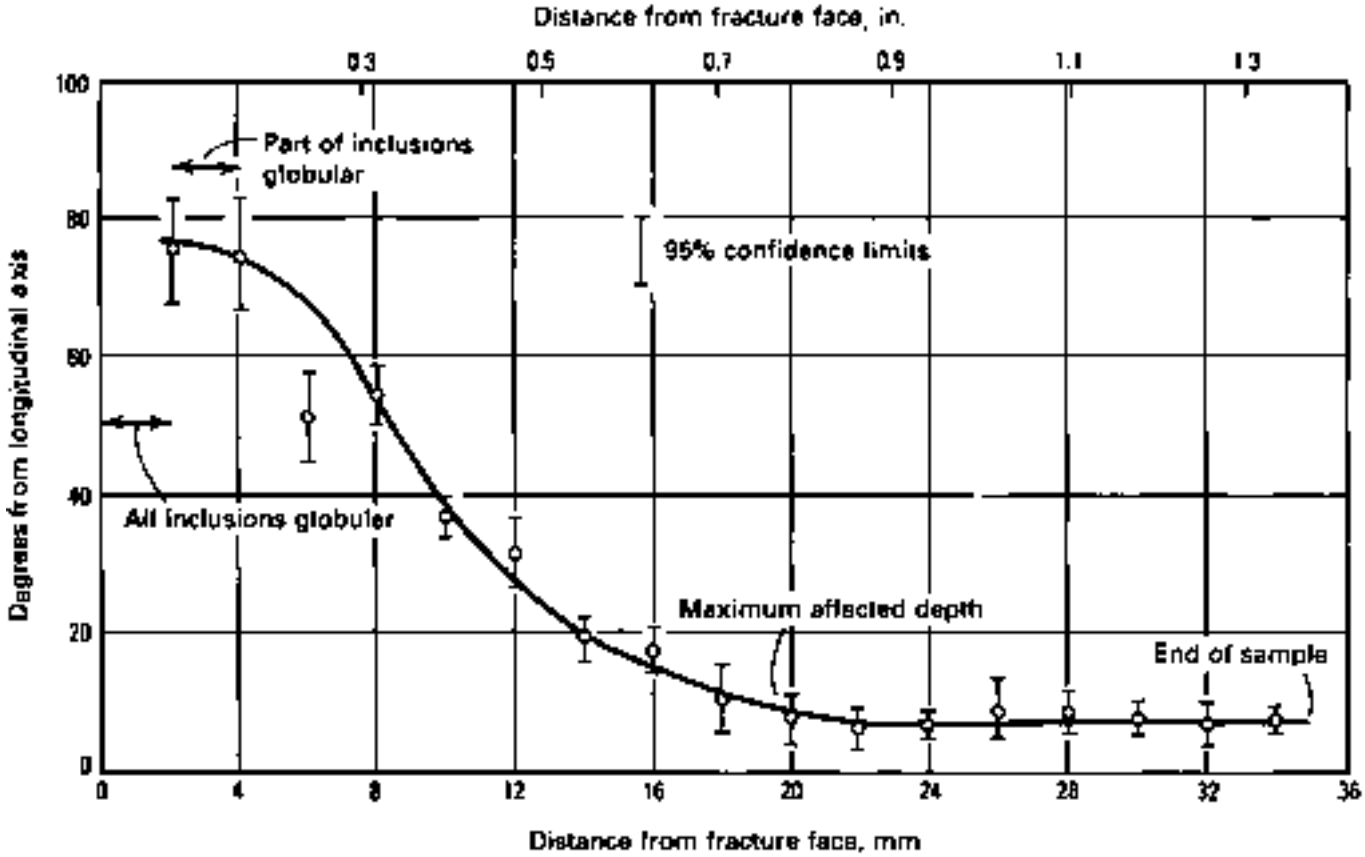


Fig. 14 Average angle of sulfide inclusions relative to the longitudinal axle axis as a function of distance from the fracture face. Specimen was taken from the area near the centerline at the fracture face. Compare to the qualitative examples in Fig. 12 and 13.

Table 1 shows the results of chemical analysis of axles 1611 and 2028 taken in unaffected regions compared with the specified range for M-126, grade F, axle steel. Figure 15 shows the normal microstructure of axles 1611 and 2028, taken well away from the affected regions. These axles are double normalized and tempered. The microstructure is fine lamellar pearlite with grain-boundary ferrite.

Table 1 Chemical analysis of axles 1611 and 2028[a]

Axle	Element, %															
	C[b]	Mn[c]	P[d]	S[b]	Si[e]	Ni[f]	Cr[f]	Mo[f]	Al[a]	Sn[g]	Cu[h]	V[f]	Nb[g]	Ti[g]	N[i]	O[j]
2028	0.57	0.70	0.012	0.042	0.16	0.01	0.08	0.004	0.004	0.002	0.057	0.003	0.01	0.007	0.0045	0.0057
1611	0.54	0.79	0.012	0.026	0.19	0.01	0.09	0.005	0.004	0.002	0.023	0.003	0.01	0.007	0.0058	0.0056
M-126-F	0.45–0.59	0.60–0.90	0.045	0.050	>0.15	…	…	…	…	…	…	…	…	…	…	…

(a) Weight percent.
(b) Analysis by high-temperature combustion; infrared detection
(c) Peroxydisulfate-arsenite titrimetric analysis.
(d) Alkalimetric analysis.
(e) Gravimetric analysis.
(f) Atomic absorption spectrometric analysis
(g) Plate spectrographic analysis
(h) Neocuproine photometric analysis
(i) Analysis by inert gas fusion; thermal conductivity defection.
(j) Vacuum fusion analysis

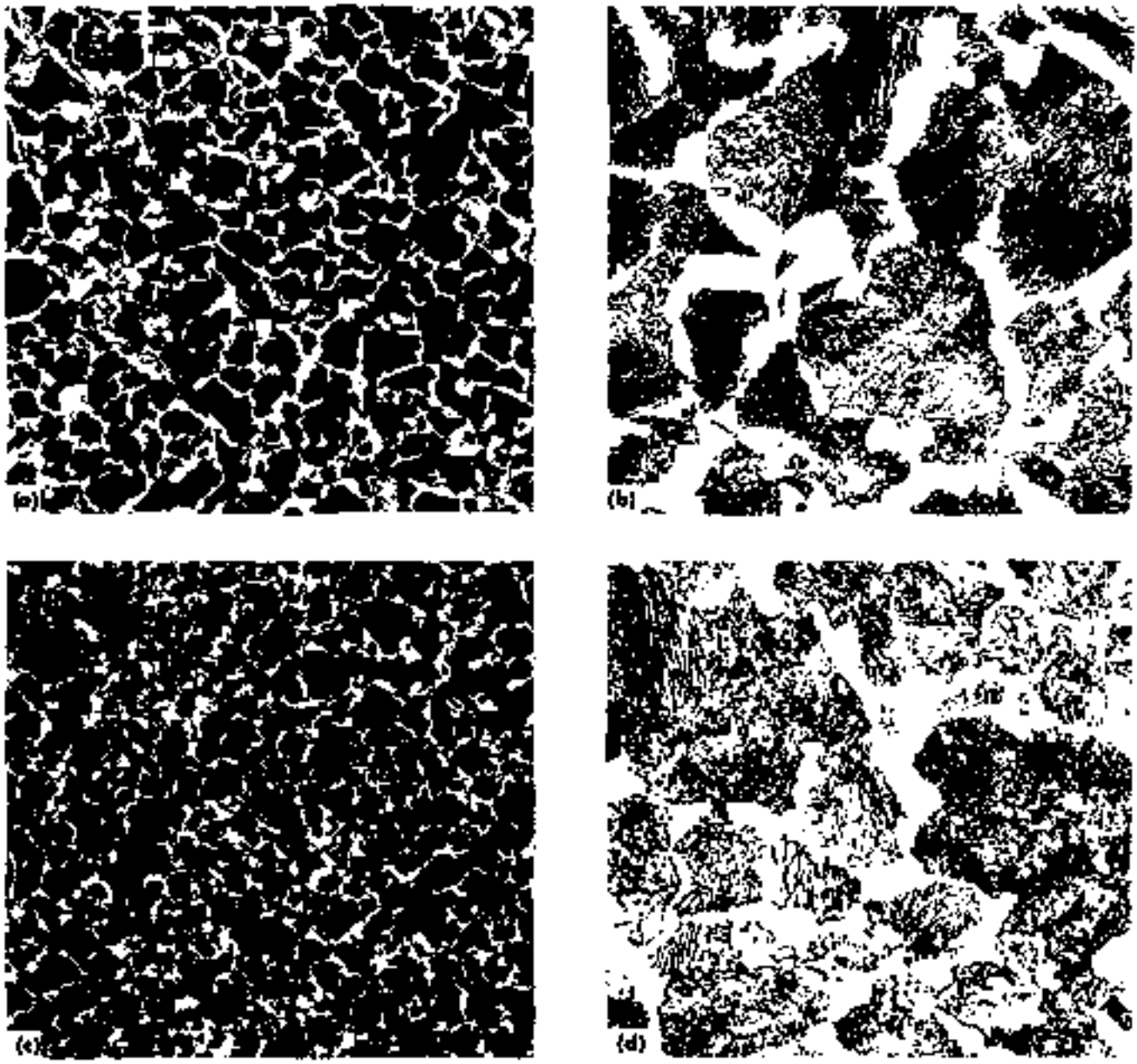


Fig. 15 Examples of the normal microstructure of the axles. Specimens were taken well away from the fracture surface. (a) Axle 2028. Etched with 2% nital. 100×. (b) Axle 2028. Etched with 4% picral. 500×. (c) Axle 1611. Same etchant and magnification as (a). (d) Axle 1611. Same etchant and magnification as (b)

Penetration of bearing elements in the third axle is shown in Fig. 16. Color metallography reveals the characteristic copper color of the penetrating material. Figure 17 shows scanning electron microscope views of the area shown in Fig. 16, further confirming the presence of copper. More extensive elemental mapping of penetrating bearing material is shown in Fig. 18. Again, all of the typical bearing elements are observed.

Fig. 16 Optical micrograph of copper penetration (arrows) near the outside-diameter surface of the third axle. Etched with 2% nital. 55×

Fig. 17 SEM micrographs of the area from the third axle shown in Fig. 16. (a) Secondary electron image. (b) Backscattered electron image. (c) Copper x-ray dot map. All 55×

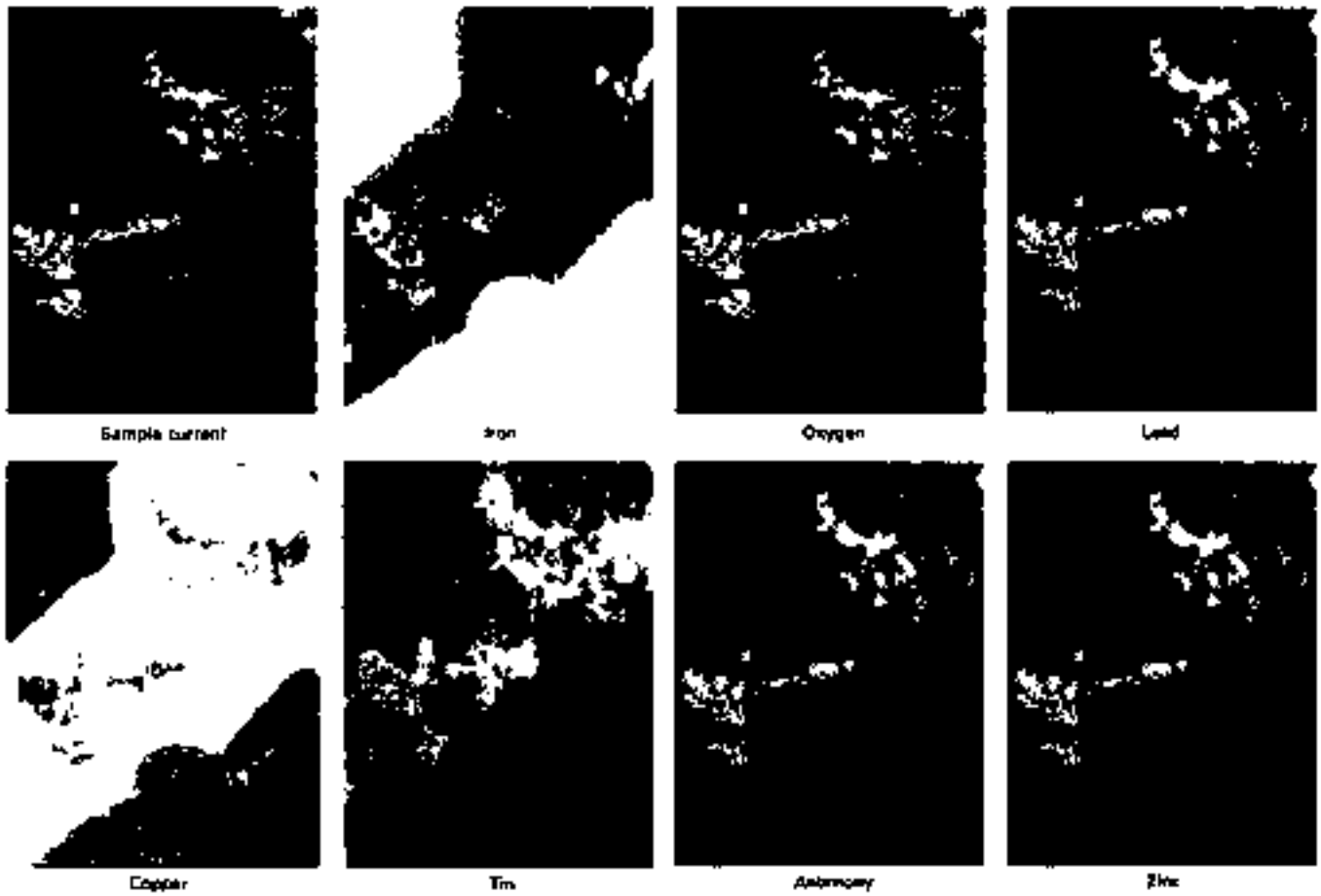


Fig. 18 X-ray elemental composition maps made with the electron microprobe using a region of copper penetration shown in Fig. 16 and 17. All 270×

Simulation of LME Mechanism

To provide additional information on the embrittlement mechanism, laboratory tests were conducted using material from one of the axles. Tensile specimens measuring 12.8 mm (0.505 in.) in diameter × 250 mm (10 in.) long were prepared. A 60° V-notch was machined into the periphery at midlength of each sample. The diameter at the root of the notch was 9.1 mm (0.357 in.). The notch surface, as well as some of the bar surface on each side, was electroplated with copper. The V-notch region was then wound with fine copper wire. One tensile specimen was left bare to determine the tensile strength of the steel at 1100 °C (2010 °F) without the influence of copper. All tests were performed using a Gleeble. The uncoated, notched sample was heated to 1100 °C (2010 °F) and pulled in tension with a crosshead speed of 2.5 mm/s (6 in./min). A maximum load of 259 kg (570 lb) was observed; that is, the ultimate tensile strength was 39 MPa (5.7 ksi).

Samples with copper in the notch were then heated to 1100 °C (2010 °F), high enough to melt pure copper, and held at constant loads of 32, 65, 78, 97, and 129 kg (71, 143, 171, 214, and 285 lb), that is, 12.5, 25, 30, 37, and 50% of the 1100 °C (2010 °F) tensile strength without copper. If the liquid copper had no influence on the steel, the tensile samples should have stayed intact for a long time.

The tests showed, however, that fracture occurred after 18 s when a 129-kg (285-lb) load was applied; this corresponds to a crack-growth rate of 907 mm/h (35.7 in./h). At a load of 97 kg (214 lb, or 37% of the normal tensile strength), complete rupture occurred in 24 s; this is a crack-growth rate of 680 mm/h (26.8 in./h). At 30% of the normal tensile strength (78 kg, or 171 lb), rupture occurred in 79 s. A number of tests were conducted at 32 kg (71 lb) and 65 kg (143 lb). These samples were held for different lengths of time, cooled to room temperature, sectioned, and examined to measure the crack depths. All of the test data are plotted in Fig. 19, which reveals a consistent relationship between the test load and crack-growth rate.

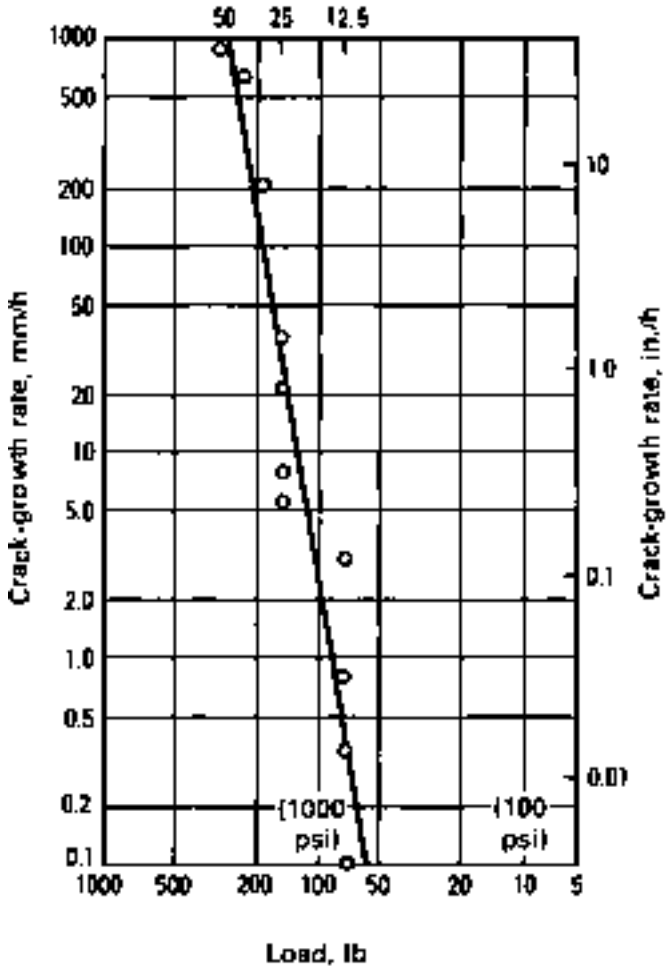


Fig. 19 Graph showing the influence of load on the LME crack-growth rate for the experimental conditions described in text. Copper-free tensile strength at 1100 °C (2010 °F), %

The microstructure of the sample that was not coated with copper when tested at 1100 °C (2010 °F) is shown in Fig. 20. The spherical cavities found at the grain boundaries are indicative of ductile overload rupture. Figure 21 shows examples of the fracture surface of a specimen tested to rupture when copper was present at 1100 °C (2010 °F). The intergranular fracture pattern is indicative of LME. Two views of the microstructure of a partially broken specimen tested at 65-kg (143-lb) load (25% of normal tensile strength) are also shown. The white grain-boundary films are copper, showing the intergranular nature of the penetration, also shown by the crack path.

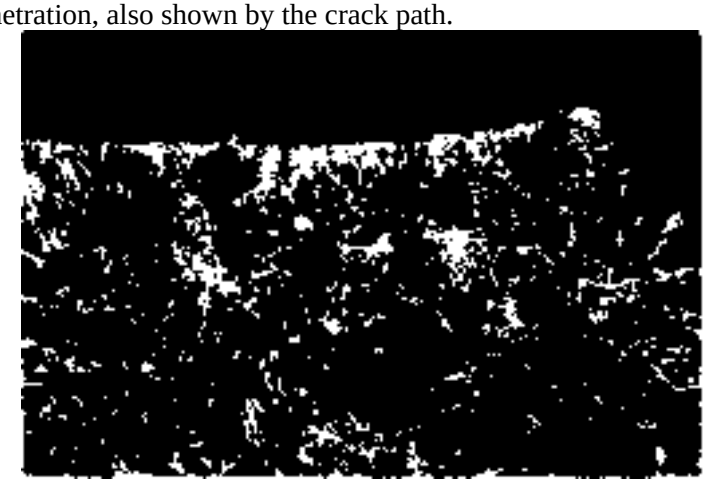
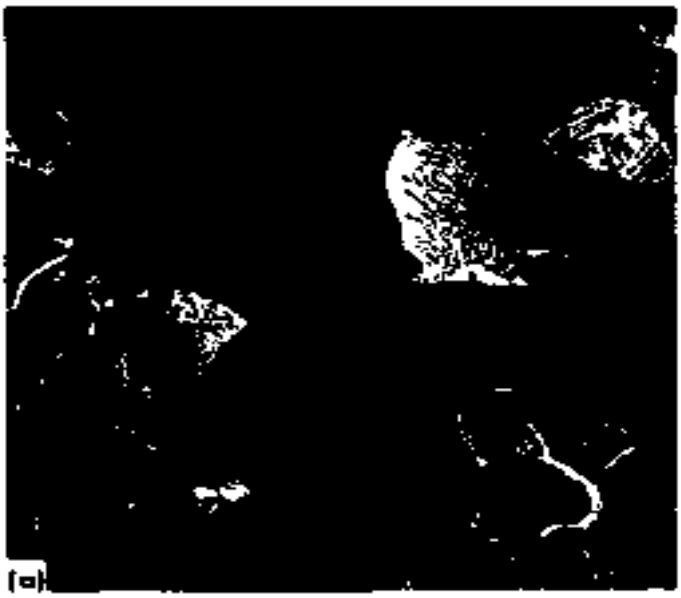


Fig. 20 Micrograph of the fractured end of the specimen tested with the Gleeble at 1100 °C (2010 °F) without copper present. The fracture is transgranular. Etched with 4% picral. 55×



(a)

(b)

(c)

Fig. 21 Examples of axle steel tensile specimens tested at 1100 °C (2010 °F) with the Gleeble, with copper present. The liquid copper has penetrated prior-austenite grain boundaries, producing intergranular fracture. (a) Bottom of notch of a specimen held for 10 min at temperature at 65 kg, or 143 lb, load (25% of the normal tensile strength). 70×. (b) Below crack in the same specimen as (a). 70×. (c) Fracture surface of a specimen tested to rupture. 35×

For comparison, a similar copper-filled V-notch specimen was heated to 1100 °C (2010 °F) in a laboratory furnace without any applied load. Figure 22 shows that no cracking occurred at the root of the V-notch. The high-magnification micrograph shows some dark-etching grain boundaries at the extreme surface where copper diffused into the specimen. As predicted by the literature, no evidence of LME was detected because there was no applied stress.

Fig. 22 Two views of a notched specimen that was electroplated with copper, wrapped with copper wire, and tested for 3 h at 1100 °C (2010 °F) without application of a load. No grain-boundary penetration of copper or intergranular cracking is evident, but there is evidence of minor bulk and grain-boundary diffusion [dark-etching lines in (b)]. (a) Root of notch. 55×. (b) Edge of specimen etched with 2% nital. 275×

Conclusions

The carbon steel railroad axles examined failed because of LME due to friction-bearing overheating. After the babbitt metal had melted away, the bronze sleeve rubbed against the steel surface. Both overheated because of the frictional stresses, to a point where the bearing melted and the contacting steel surface became austenitic. In this region the axle is under stress because of the weight of the traction motor; hence, the liquid bearing material penetrated the axle. Further heating of the axle in the contact region continued while the copper penetrated the axle. Cracking around the periphery reduced the effective cross-sectional area of the axle. Because the axle was quite hot, its strength decreased substantially, and twisting began under the action of the rotational forces. Thus, fracture was initiated by LME and propagated to final rupture by torsion. This is consistent with the observed depth of copper penetration and the observed metal flow in the central region of the axle.

Although the fracture face and origins were destroyed, there was evidence of LME along the entire surface in contact with the bearing surface. Fracture occurred in the axle under the approximate center of the bearing, where temperatures and stresses were highest.

Reference cited in this section

47. “Railroad Accident Report—Derailment of Auto-Train No. 4 on Seaboard Coast Line Railroad, Florence, South Carolina, February 24, 1978,” Report NTSB/RAR-84/1, National Transportation and Safety Board, Washington, June 1984

References

1. B. Straub, Mikroskopische Stahluntersuchung, *Stahl Eisen,* Vol 34 (No. 50), Dec 1914, p 1814–1820

2. F.H. Williams, Car Axle Failures Traced to Absorption of Bearing Metal, *Prod. Eng.,* Vol 15, Aug 1944, p 505–508

3. "The Mechanism of Copper Penetration Failure," Magnaflux Corp., Chicago, 1947

4. "Copper Penetration—New and Revealing Information Concerning the Analysis of the Penetrating Material Itself," Research Laboratory Technical Memo No. 220, Project 125, Magnaflux Corp., Chicago, Feb 1954

5. E.A. Unstich and E.J. Palinkas, "A Study of Copper Penetration in Overheated Journals," Report 69222, New York Central Railroad Company, New York, July 1959

6. New York Central Research Shows … Turning Removes Copper Penetration, *Rail. Locom. Cars,* Feb 1961, p 34–35

7. Report of the Committee on Axle and Crank Pin Research, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1951, p 158–166

8. J. J. Laudig, Hot-Box Research Results … in the Iron-Back Journal Bearing, *Rail. Age,* Vol 134 (No. 12), March 1953, p 65–69

9. Report of the Committee on Axle and Crank Pin Research, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1952, p 116–129

10. Laudig Iron Back Journal Bearing, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1952, p 358–359

11. Report of Committee on Axles, in *Proceedings of the Association of American Railroads, Mechanical Division,* 1956, p 94–104

12. "The Failure of Railroad Axle-Journal Steel Including the Effects of Copper-Bearing Metals in Contact with Stressed Steel Rods at Elevated Temperatures," Report MR-434, Association of American Railroads, Research Department, Washington, Dec 1963

13. N.S. Stoloff, Liquid Metal Embrittlement, in *Surfaces and Interfaces, II,* Proceedings of 14th Sagamore Army Materials Research Conference, Syracuse University Press, Syracuse, NY, 1968, p 157–182

14. A. R. C. Westwood *et al.,* Adsorption-Induced Brittle Fracture in Liquid-Metal Environments, in *Fracture,* Vol III, Academic Press, 1971, p 589–644

15. W. Rostoker *et al., Embrittlement by Liquid Metals,* Reinhold, 1960

16. M.H. Kamdar, Liquid Metal *Embrittlement, in Embrittlement of Engineering Alloys,* Vol 25, *Treatise on Materials Science and Technology,* Academic Press, 1983, p 361–459

17. M.G. Nicholas and C.F. Old, Review: Liquid Metal Embrittlement, *J. Mater. Sci.,* Vol 14, Jan 1979, p 1–18

18. M. H. Kamdar, Ed., *Embrittlement by Liquid and Solid Metals,* American Institute of Mining, Metallurgical, and Petroleum Engineers, Warrendale, PA, 1984

19. A.R. Huntington, discussion of paper by C.H. Desch, The Solidification of Metals From the Liquid State, *J. Inst. Met.,* Vol 11, 1914, p 108–109

20. F.A. Shunk and W.R. Warke, Specificity as an Aspect of Liquid Metal Embrittlement, *Scr. Metall.,* Vol 8, 1974, p 519–526

21. E. Scheil and K.E. Schiessl, The Diffusion of Liquid Bismuth Along Grain Boundaries in Copper, *Z. Naturforsch,* Vol 4a, 1949, p 524–526

22. R. Genders, The Penetration of Mild Steel by Brazing Solder and Other Metals, *J. Inst. Met.,* Vol 37, 1927, p 215–221

23. H. Schottky *et al.,* The Red-Shortening of Steels by Metals, *Arch. Eisenhüttenwes.,* Vol 4, 1931, p 541–547

24. W. E. Goodrich, The Penetration of Molten White Metals into Stressed Steels, *J. Iron Steel Inst.,* Vol 132, 1935, p 43–66

25. W. Radeker, The Production of Stress Cracks in Steel by Molten Zinc, *Stahl Eisen,* Vol 73 (No. 10), May 1953, p 654–658

26. L.J.G. van Ewijk, The Penetration of Steel by Soft Solder and Other Molten Metals at Temperatures up to 400 °C, *J. Inst. Met.,* Vol 56 (No. 1), 1935, p 241–256

27. J.V. Rinnovatore and J.D. Corrie, The Effect of Cold Work on the Liquid Metal Embrittlement of Iron and Iron-Silicon, *Scr. Metall.,* Vol 2 (No. 8), Aug 1968, p 467–470

28. R.R. Hough and R. Rolls, The High-Temperature Tensile Creep Behavior of Notched, Pure Iron Embrittled by Liquid Copper, *Scr. Metall.,* Vol 4 (No. 1), 1970, p 17–24

29. R.R. Hough and R. Rolls, Creep Fracture Phenomena in Iron Embrittled by Liquid Copper, *J. Mater. Sci.,* Vol 6, 1971, p 493–1498

30. R.R. Hough and R. Rolls, Copper Diffusion in Iron During High-Temperature Tensile Creep, *Met. Trans.,* Vol 2, Sept 1971, p 2471–2475

31. R.R. Hough and R. Rolls, Some Factors Influencing the Effects of Liquid Copper on the Creep-Rupture Properties of Iron, *Scr. Metall.,* Vol 8 (No. 1), Jan 1974, p 39–44

32. L.H. VanVlack, Intergranular Energy of Iron and Some Iron Alloys, *Trans. AIME,* Vol 191, March 1951, p 251–259

33. W.J.M. Salter, Surface Hot Shortness in Mild Steel, *J. Iron Steel Inst.,* Vol 200, Sept 1962, p 750–751

34. R.B. Waterhouse and D. Grubb, The Embrittlement of Alpha-Brasses by Liquid Metals, *J. Inst. Met.,* Vol 91, 1962–1963, p 216–219

35. C.A. Stickles and E.E. Hucke, The Effect of Stress on the Dihedral Angle in Leaded Nickel, *J. Inst. Met.*, Vol 92, 1963–1964, p 234–237

36. R. Eborall and P. Gregory, The Mechanism of Embrittlement by Liquid Phase, *J. Inst. Met.*, Vol 84, 1955, p 88–90

37. E.A. Asnis and V. M. Prokhorenko, Mechanism of Cracking During the Welding or Depositing of Copper on to Steel, *Weld. Prod.*, Vol 12 (No. 11), 1965, p 15–17

38. S.J. Matthews and W.F. Savage, Heat-Affected Zone Infiltration by Dissimilar Liquid Weld Metal, *Weld. J.*, Vol 50 (No. 4), 1971, p 174s–182s

39. W. F. Savage *et al.*, Intergranular Attack of Steel by Molten Copper, *Weld. J.*, Vol 57 (No. 1), 1978, p 9s–16s

40. W.F. Savage *et al.*, Copper-Contamination Cracking in the Weld Heat-Affected Zone, *Weld. J.*, Vol 57 (No. 5), May 1978, p 145s–152s

41. W.F. Savage *et al.*, Liquid-Metal Embrittlement of the Heat-Affected Zone by Copper Contamination, *Weld. J.*, Vol 57 (No. 8), Aug 1978, p 237s–245s

42. J.W. Halley, Residual Elements in Steel, in *Effect of Residual Elements on the Properties of Metals*, American Society for Metals, 1957, p 71–87

43. D.A. Melford, Surface Hot Shortness in Mild Steel, *J. Iron Steel Inst.*, Vol 200, April 1962, p 290–299

44. K. Born, Surface Defects in the Hot Working of Steel Resulting from Residual Copper and Tin, *Stahl Eisen*, Vol 73, HB 3255, 1953, p 1268–1280

45. G.G. Foster and J.K. Gilchrist, The Influence of Copper, Nickel and Tin on the Hot Working Properties of Mild Steel, *Metall.*, May 1952, p 225, 228, 230

46. W.J.M. Salter, Effects of Alloying Elements on Solubility and Surface Energy of Copper in Mild Steel, *J. Iron Steel Inst.*, Vol 204, May 1966, p 478, 487

47. "Railroad Accident Report—Derailment of Auto-Train No. 4 on Seaboard Coast Line Railroad, Florence, South Carolina, February 24, 1978," Report NTSB/RAR-84/1, National Transportation and Safety Board, Washington, June 1984

Glossary

Compiled and revised by Debbie Aliya, Aliya Analytical, W.T. Becker, Consultant (University of Tennessee, Retired), and Roch Shipley, Packer Engineering Inc.

COMMUNICATION is often the key issue in practical failure analysis. Communication is required among specific parties who are trying to solve a problem. It is recognized that the exact definition of a particular term may be less important than conveying the concept that solves the problem. However, it is typical in failure analysis and problem solving to consult the literature (including Handbooks such as this one) or other people to find information on similar problems. Without clear communication and consistent terminology, one cannot know if the problems are really similar.

In failure analysis work, it is often not possible to pin down a particular damage sequence or root cause with complete certainty. However, if the work is performed in a professional manner, there is a good probability of identifying some helpful solutions to at least reduce the likelihood of recurrence. This glossary is intended to help promote clear thinking and useful failure analysis. The process of defining terms is useful to the parties involved in solving a failure analysis problem, in that it increases their own comprehension of the physical facts and also facilitates communication with the other parties who need to understand the results of the investigation.

Strict use of proper terminology is critical to clear communication. Several organizations have compiled glossaries of terms used in failure analysis, including ASTM, SAE, and ASM International. In some cases (e.g., ASTM), accepted definitions of terms is by consensus approval of a committee. In other instances (e.g., ASM), there is no formal approval procedure nor is there a committee responsible for terminology. As a result, there may be more than one common use of a specific term, and there may be more than one definition of a term.

The definitions used in the following glossary are not always consistent with those for the same terms found in the *ASM Materials Engineering Dictionary* (J.R. Davis, editor, ASM International, 1992) or in glossaries in other *ASM Handbook* volumes. The use of a term (e.g., corrosion fatigue) may change with time as understanding evolves. The definitions presented here are those used in this Volume and reflect common and modern understanding of these terms as used in the literature and in reports by practicing failure analysts. However, as in all communication, if the way in which a particular term is being used is not clear, the speaker or author should be asked for clarification.

Several terms used in discussions of failure analysis can be confusing or interpreted variously by different persons. Cases in point include mode, stress cracking, ductile, brittle, and cleavage. An example of the evolution of definitions, as previously noted, is the term quasi-cleavage, which is truly a form of cleavage and a microscale fracture surface appearance of a quenched and tempered steel. The fracture surface is typically dominated by cleavage, but there are typically small patches of microvoid coalescence (MVC) present or thin ribbons of MVC contained in the fracture surface. As the patches of MVC increase, the fracture surface is more accurately described as (microscale) mixed cleavage and MVC. In a matrix dominated by cleavage, there is nothing “quasi” about the cleavage, because the term *quasi* implies something different than cleavage.

Sometimes a modifying adjective may considerably improve the clarity of the sentence in which it is used. For example, *ductile* can refer to both a process/mechanism at the microscale or an appearance at the macroscale. The term *mode* is used in multiple ways: It is defined in ASTM E 1823 in terms of surface displacement of the crack tip created by various loading conditions. It is defined within *ASME Failure Methods and Effects Analysis* terminology, and there are some eight additional definitions in *Webster's Unabridged Dictionary.* There are still more uses of the term, as in *fracture mode, failure mode,* and so on. These latter two terms are not in the *ASM Materials Engineering Dictionary,* nor are they in an ASTM glossary. Consequently, their use has no underlying, commonly accepted definition and should be avoided if possible.

Other examples of terms that may be misinterpreted include *static fatigue* and *stress cracking.* The compilers also discourage the use of terms that have been removed from glossaries created by consensus approval. An example is the term *endurance limit.* The endurance limit presumably indicates a threshold stress for infinite

life in cyclic loading. It appears in *S-N* data for materials that strain age and is therefore associated with a dislocation-interstitial pinning mechanism. If some event occurs to cause depinning (say, an increased stress), a new saturation stress is obtained; that is, the endurance limit is changed. The term endurance limit is no longer contained in a standard glossary of terms used in fracture and fatigue (ASTM E 1823).

There are a group of terms used to describe deviations from ideality and that are in some instances the technical root cause for failure. They include *imperfection, discontinuity,* defect and *root cause.* In some instances these terms are used interchangeably by some writers, and in other cases clear distinctions are made. In some instances, specific terms have been defined by legal jurisdictions. The failure analyst should consider the legal implication of his terminology. As noted, the courts have defined the meanings of certain words, and these meanings may not correspond to what the analyst intends. Furthermore, various courts may not be consistent. For example, some metallurgists have examined a fracture surface and identified a defect at the fracture origin. By this they mean nothing more than a discontinuity and perhaps a minute discontinuity. However, within a legal context, the identification of a defect suggests a defective component, legal liability for the manufacturer, and a possible need for a recall or other corrective action.

The definition of defect herein is based on deviation from a specification or a component being unfit for its intended purpose. An actual or perceived failure does not automatically mean there is a defect. Even if there is a defect, that condition may or may not relate to the failure. It is believed this framework provides a rational method for deciding if there is, in fact, a defect. However, it does not eliminate all controversy as specifications may or may not be clear, appropriate, and up-to-date. There may also be disagreement regarding the intended purpose of a component. A discussion on this matter from a legal perspective is in two references: (a) Asperger, J.J., "Legal Definition of a Product Failure: What the Law Requires of the Designer and the Manufacturer," *Failure Prevention Through Education, Getting to the Root Cause, Conference Proceedings,* ASM International, May 2000, and (b) "Product Liability and Design" in this Volume of the *ASM Handbook.*

475 °C embrittlement

Embrittlement of ferritic and semiferritic steels containing more than 13% Cr that occurs when they are held or slowly cooled through the 400 to 500 °C (750 to 930 °F) temperature range. Embrittlement is presumed to be due to the grain-boundary precipitation of a chromium-rich phase together with the creation of a precipitate-free zone (PFZ) adjacent to the grain boundary. Because the precipitation creates a PFZ, fracture is likely to be intergranular.

500 to 700 °C embrittlement

Embrittlement that occurs when high-alloy steels (e.g., hot work steels and high-speed steels) are cooled to form martensite and are subsequently tempered or slowly cooled through the 500 to 700 °C (930 to 1290 °F) temperature range. Tempering results in a fine dispersion of carbides. A small increase in hardness is associated with the embrittlement.

A

abrasion

The process of grinding or wearing away through the use of abrasives; a roughening or scratching of a surface due to abrasive wear.

abrasive wear

The removal of material from a surface when hard particles slide or roll across the surface under pressure. The particles may be loose or may be part of another surface in contact with the surface being abraded. Compare with adhesive wear.

adhesive wear

The removal or displacement of material from a surface by the welding together and subsequent shearing of minute areas of two surfaces that slide across each other under pressure. Compare with abrasive wear.

adiabatic

Occurring without loss or gain of heat in the system, which may result in a local temperature increase or decrease. Adiabatic conditions differ from isothermal conditions (under which the temperature remains constant).

adiabatic shear bands

show that the outer portion of the fracture face experienced considerable tearing and rupture, while the central portion exhibits hot torsional rupture (note the swirl pattern of the fracture). This suggests that LME occurred to a depth of nearly 25 mm (1 in.). The axle, which was heated to at least 925 °C (1700 °F), then ruptured to failure by twisting under the applied loads. The fracture features of axle 2028 are less distinct because of the rubbing action after fracture (Fig. 3). The outside-diameter surface of each of these axles in the area in contact with the traction-motor support bearing exhibited scale, and some of the friction-bearing material was present, as identified by x-ray fluorescence and diffraction.

Fig. 1 Fracture surface at the drive-wheel side of axle 1611.

where τ and σ are shear and normal stress, respectively, and μ and τ_i are material constants.

crack

(1) A pair of surfaces and the associated discontinuity created by separation under stress of atoms previously bonded. This does not include internal free surfaces created by corrosion processes or solidification imperfections. (2) The process by which a solid object becomes fragmented due to separation of atomic bonds under stress. Cracking is often accompanied by or preceded by deformation.

crack bifurcation

The spliting of a crack into two paths, or the intersection of a crack with a preexisting crack.

crack extension

An incremental increase in crack size. See also crack length (a) or depth, effective crack size, and physical crack size.

crack extension force

The elastic energy per unit of new separation area that is made available at the front of an ideal crack in an elastic solid during a virtual increment of forward crack extension. Also described as the strain energy release rate.

crack length (*a*) or depth

In fracture-mechanics-based analyses, the physical crack size used to determine the crack growth rate and the stress-intensity factor. For a compact-type specimen, crack length is measured from the line connecting the bearing points of load application. For a center-crack tension specimen, crack length is measured from the perpendicular bisector of the central crack.

crack mouth opening displacement (CMOD)

See crack tip opening displacement (CTOD).

crack opening displacement (COD)

See crack tip opening displacement (CTOD)

crack plane orientation and growth direction

A three-letter identification system used to identify the longitudinal (L), transverse (T), and short transverse (S) direction of the applied stress normal to the plane on which the crack propagates and the direction of crack growth. Used for both prismatic and cylindrical sections. For example, a L-T orientation indicates a stress in the longitudinal (or rolling) direction, and crack propagation is in the wide transverse direction. See ASTM E 399 for a complete description.

crack size (*a*)

A lineal measure of a principal planar dimension of a crack. This measure is commonly used in the calculation of quantities descriptive of the stress and displacement fields. In practice, the value of crack size is obtained from procedures for measurement of physical crack size, original crack size, or effective crack size, as appropriate to the situation under consideration. See also crack length (a) or depth.

crack tip opening displacement (CTOD)

The crack displacement resulting from the total deformation (elastic plus plastic) at variously defined locations near the original crack tip.

crack-tip plane strain

A stress-strain field near a crack tip that approaches plane strain to the degree required by an empirical criterion.

creep

Time-dependent strain occurring under stress. The creep strain occurring at a diminishing rate is called primary or transient creep (stage I); that occurring at a minimum and almost constant rate, secondary or steady-rate creep (stage II); and that occurring at an accelerating rate, tertiary creep (stage III).

creep rate

The slope of the creep-time curve at a given time. See also minimum creep rate.

creep-rupture strength

The stress that causes fracture in a creep test at a given time in a specified constant environment. Also known as stress-rupture strength.

creep strain

The time-dependent total strain (including the initial instantaneous elastic strain) produced by applied stress during a creep test.

creep strength

The stress that causes a given creep strain in a creep test at a given time in a specified constant environment.

creep stress

The constant load divided by the original cross-sectional area of the specimen.

crevice corrosion

Localized corrosion of a metal surface at, or immediately adjacent to, an area that is shielded from full exposure to the environment because of close proximity between the metal and the surface of another material.

cross direction

See transverse direction.

crush

(1) An indentation in a casting surface due to displacement of sand into the mold cavity when the mold is closed. (2) Permanent localized compressive deformation. (3) The spalling of (usually thin) surface-treated cases on gears due to overload; also known as case crushing.

cumulative damage

In fatigue loading, a measure of damage, as expressed as a fraction of life to failure, as it accumulates with the number of cycles. When the stress level is changed, the life is changed. Therefore, damage accumulates according to the stress level and number of cycles spent at a given stress. The most common (but not only) model for linear cumulative damage is the Palmgren-Miner cumulative damage model. In the following equation, n_i is the number of load cycles spent at a stress for which the life is N_i. Failure is assumed to occur when the damage summation becomes unity:

$$D = \sum \frac{n_i}{N_i}$$

cup fracture (cup-and-cone fracture)

A mixed-mode fracture, often seen in cylindrical-shaped components or tension test specimens of a ductile material, where the core of the necked region undergoes fracture by microvoid coalescence, creating the cup bottom. As the crack grows from this central region, the degree of constraint is reduced, and fracture continues in the remaining outer ring under conditions of plane stress, creating the slanted cup walls. Closely related fracture appearance also is seen in small width-to-thickness-ratio prismatic tensile specimens. One of the mating fracture surfaces looks similar to a miniature cup, having a central depressed flat-face region surrounded by a shear lip. The other fracture surface looks similar to a miniature truncated cone.

cupping

The condition sometimes occurring in heavily cold-worked rods and wires in which the outside fibers are still intact and the central zone has failed in a series of cup-and-cone fractures.

cut

A raised, rough surface on a casting due to erosion by the metal stream of part of the sand mold or core.

cyclic loading

(1) Repetitive loading, as with regularly recurring stresses on a part, that sometimes leads to fatigue fracture. (2) Loads that change value by following a regular or irregular repeating sequence of change.

D

dealloying

The selective corrosion of one or more components of a solid-solution alloy. Also called parting or selective leaching. See also decarburization, denickelification, dezincification, and graphitic corrosion.

decarburization

Loss of carbon from a free surface (component surface or surface of a crack intersecting the surface) of a carbon-containing alloy due to reaction with one or more chemical substances in a medium that contacts the surface.

decohesion

The process of creating a pair of free surfaces where a grain boundary or second-phase boundary existed previously.

defect

(1) An imperfection (deviation from perfection) that can be shown to cause failure by a quantitative analysis and that would not have occurred in the absence of the imperfection. (2) (Legal) Manufacturing defect: (a) a failure to conform to stated specifications; (b) nonsatisfaction of user requirements; (c) deviation from the norm; (d) when a product leaves the assembly line in substandard condition, differs from the manufacturer's intended result, or differs from other ostensibly identical units of the same product line. (3) (Legal) Design defect: (a) less safe than expected by the ordinary consumer; (b) excessive preventable danger. (4) (Legal) Marketing defect: failure to warn or inadequate warning of hazard and risk involved with use of a product. Note: Legal definitions may vary from jurisdiction to jurisdiction. Those legal definitions cited here are from the article "Product Liability and Design" in this Volume and also in *Materials Selection and Design,* Volume 20, *ASM Handbook,* page 147. An example is a deviation from specification. It can be the cause of a failure that makes the part unsuitable for its intended purpose. However, just because a part fails does not imply that it contained a defect, and not all defects are a cause for failure. (2) Laws of various jurisdictions may also define what constitutes a defect of defective product.

deformation

A change in the shape of a body due to stress, thermal change, change in moisture, or other causes.

deformation bands

Bands produced within individual grains during cold working that differ in orientation from the matrix.

deformation twin

See mechanical twin (deformation twin) and Neumann bands.

dendrite

A crystal with a treelike branching pattern. Dendrites are most evident in cast metals slowly cooled through the solidification range.

denickelification

Corrosion in which nickel is selectively leached from nickel-containing alloys. Most commonly observed in copper-nickel alloys after extended service in fresh water. See also selective leaching.

depletion

Selective removal of one component of an alloy, usually from the surface or preferentially from grain-boundary regions. See also selective leaching.

deposit attack or corrosion

See poultice corrosion.

deviatoric stress

The nonhydrostatic component of the state of stress on a body. It is the deviatoric component that causes shape change (plastic deformation).

dezincification

Corrosion in which zinc is selectively leached from zinc-containing alloys. Most commonly found in copper-zinc alloys containing less than 85% Cu after extended service in water containing dissolved oxygen. See also selective leaching.

diamond pyramid hardness test

See Vickers hardness test.

diffuse necking

The nonuniform strain distribution along the length of a member loaded in tension that develops at the maximum load. The term "diffuse" is used because the neck develops slowly, with little change in load, but an increase in axial strain near the load maximum. Contrast to local necking.

dimpled rupture fracture

Ductile fracture that occurs through the formation and coalescence of microvoids (dimples) along the fracture path. The fracture surface of such a ductile fracture appears dimpled when observed at high magnification and usually is most clearly resolved when viewed in a scanning electron microscope.

distortion

Any deviation from an original size, shape, or contour that occurs because of the application of stress or the release of residual stress and may be associated with exposure to high or low temperature.

ductile

Capable of being plastically deformed before fracturing.

ductile-brittle transition temperature (DBTT)

Temperature at which a marked change occurs in the fracture resistance of body-centered cubic and hexagonal close-packed metals from ductile behavior to brittle behavior. The transition occurs in those metals in which the yield strength increases sharply with decreasing temperature and which are capable of fracturing by cleavage or an intergranular mode with very little accompanying plastic deformation. The transition temperature also depends on strain rate.

ductile crack propagation

Slow crack propagation that is accompanied by noticeable plastic deformation and requires energy to be supplied from outside the body. Contrast with brittle crack propagation.

ductile erosion behavior

Erosion behavior having characteristic properties (such as considerable plastic deformation) that can be associated with ductile fracture of the exposed solid surface. A characteristic ripple pattern forms on the exposed surface at low values of angle of attack. Contrast with brittle erosion behavior.

ductile fracture

Macroscale and microscale term. Used at the macroscale to indicate visible plastic deformation associated with fracture (often implies high energy absorption). Used at the microscale to describe plastic deformation associated with fracture, typically by microvoid coalescence. Contrast with brittle fracture.

ductility

The ability of a material to deform plastically before fracturing. Measured by elongation or reduction of area in a tension test, by height of the cup formed, in a cupping test, or by the radius or angle of bend in a bend test. Contrast with brittleness; see also plastic deformation.

dynamic

Moving, or having high velocity. Frequently used with high strain-rate ($\geq 0.1\ s^{-1}$) testing of metal specimens.

E

effective crack size, a_e

The physical crack size augmented for the effects of crack-tip plastic deformation. Sometimes the effective crack size is calculated from a measured value of a physical crack size plus a calculated value of a plastic-zone adjustment. A preferred method for calculation of effective crack size compares compliance from the secant of a load-deflection trace with the elastic compliance from a calibration for the type of specimen.

effective stress

A calculated parameter used in a mathematical expression that predicts the onset of an event (typically yield and fatigue failure) when a critical value of the effective stress is obtained. The most common effective stress is the von Mises stress.

elastic constants

The factors of proportionality that relate elastic displacement of a material to applied forces. See also bulk modulus of elasticity, modulus of elasticity, Poisson's ratio, and shear modulus.

elastic deformation

Shape change and volume change due to an applied stress that is low enough so that the material returns to its original shape and volume when the stress is removed.

elasticity

The property of a material by virtue of which deformation caused by stress disappears on removal of the stress. A perfectly elastic body completely recovers its original shape and dimensions after release of stress.

elastic limit

The maximum stress a material is capable of sustaining without any permanent shape change remaining on complete release of the stress. See also proportional limit.

elastic strain

The mathematical value that quantifies elastic deformation.

elongation

A term used in mechanical testing to describe the amount of extension of a test piece when stressed. Expressed in units of length or as strain (percent change in length). Quantitative value describing the length increase of the gage length of a tensile-test bar due to the deformation up to and including the fracture process. Note that elongation (expressed as length or strain) is not an inherent material property but depends on the dimension of the specimen tested as well as whether it is measured under load, such as with a laser extensometer, or after fracture, generally after manually remating the fracture surfaces in a fixture.

embrittlement

The severe loss of ductility and/or toughness of a material.

end grain

Grain flow lines that intersect with the exposed surface of the ends of bar stock, tubing, or the parting lines of forgings. A long, narrow test specimen sectioned so that the grain is parallel to the longitudinal axis of the specimen has no exposed end grain, except at the extreme ends. In contrast, a corresponding specimen cut in the transverse direction has end-grain exposure at all points along its length. End grain is especially pronounced in the short-transverse direction on die forgings designed with a flash line.

end-grain attack

Preferential corrosive attack of grains exposed by cutting through the cross section or at the parting lines of forgings.

endurance limit

Obsolete term used to describe a characteristic in components subject to fatigue cracking. The use of this term is now discouraged, because it does not reflect the current understanding of how fatigue cracks initiate and grow. It referred to the maximum stress below which a material can presumably endure an infinite number of stress cycles. If the stress is not completely reversed, the value of the mean stress, the minimum stress, or the stress ratio also should be stated. Compare with fatigue limit.

erosion

Destruction of materials by the abrasive action of moving fluids, usually accelerated by the presence of solid particles carried with the fluid. See also erosion-corrosion.

erosion-corrosion

A conjoint action involving corrosion and erosion in the presence of a moving corrosive fluid, leading to the accelerated loss of material.

etch pits

(1) Localized corrosion attack at the microscopic scale. Typically seen on a polished and etched metallographic specimen. Etch pits typically have recognizable, simple geometric shape (square, rectangle, triangle) and therefore reveal, in a qualitative way, the orientation of a grain. (2) A conjoint action involving corrosion and erosion in the presence of a moving corrosive fluid, leading to the accelerated loss of material.

exfoliation

Corrosion that proceeds from the sites of initiation along planes parallel to the surface, generally along grain boundaries, forming bulky corrosion products that create generally undesired free internal surfaces and give rise to a layered appearance.

F

failure

(1) A general term used to imply that a part in service (a) has become completely inoperable, (b) is still operable but is incapable of satisfactorily performing its intended function, or (c) has deteriorated seriously, to the point that it has become unreliable or unsafe for continued use. (2) Also commonly applied to manufacturing processes that produce components that do not meet specifications.

false Brinelling

Damage to a solid bearing surface characterized by indentations not caused by plastic deformation resulting from overload but thought to be due to other causes such as fretting corrosion. See also Brinelling.

fatigue

Failures of Pipelines

R.J. Eiber and J.F. Kiefner, Battelle Columbus Laboratories

Introduction

PIPELINES have established the enviable record of having the fewest fatalities of any of the various modes of transportation. Failures do occur, however, for a variety of reasons. In this article, the known failure causes and failure characteristics of high-pressure long-distance pipelines will be summarized. In addition, the procedures used to investigate pipeline failures will be described. Finally, uses of fracture mechanics in failure investigations and in developing remedial measures will be described.

Characteristics of High-Pressure Long-Distance Pipelines

Three general types of pipelines exist: gathering lines, transmission lines, and distribution lines. Gas or crude oil gathering lines exist between a well and a treatment plant or collection point. These are usually relatively small-diameter lines (51 to 203 mm, or 2 to 8 in., in diameter) and operate at a variety of pressures, usually from 3490 to 8376 kPa (500 to 1200 psi). Cross-country transmission pipelines of over 1600 km (1000 miles) in length transport natural gas, natural gas liquids, liquid petroleum products, anhydrous ammonia, or crude oil. These are continuous lengths of pipe interrupted only by valves at roughly 16-km (10-mile) spacings and compression stations (gas lines) or pumping stations (liquid lines) at 80- to 160-km (50- to 100-mile) spacings. High-pressure lines generally operate at pressures up to 6895 kPa (1000 psi) and are made of steel pipes welded or mechanically coupled together. Since the 1940s, all of the lines have been assembled by welding. The third type of pipeline is a gas-distribution line that mainly transports natural gas within cities at pressures that vary from several tens of pounds per square inch to a few inches of water where a line enters a home. This article will not discuss failures associated with these latter lines, but will address only the high-pressure gathering and transmission lines.

The steels that have been used in the United States for line pipe range in yield strength from American Petroleum Institute (API) grade A (207 MPa, or 30 ksi) to API 5L-X70 (483 MPa, or 70 ksi). Generally, API 5L Specification steels with yield strengths above 359 MPa (52 ksi) are microalloyed and controlled rolled to achieve the desired strength and fracture properties. Pipeline wall thicknesses are based on the pressure in the line and on the allowable hoop stress levels. The allowable stress levels for gas pipelines vary from 40 to 72% of the specified minimum yield strength based on the population density in the area of the pipeline and are regulated by the U.S. Department of Transportation (DOT) in Title 49 of the Code of Federal Regulations (CFR) Part 192. Liquid petroleum pipelines are regulated by DOT in Title 49 CFR Part 195. Their maximum allowable hoop stress level is 72% of the specified minimum yield strength.

One of the reasons for the excellent safety record of the pipeline industry is that pipelines are pressure tested before being put into service. It is common to test pipelines hydrostatically to hoop stress levels of 90 to 105% of their specified minimum yield strength. In addition, this pressure is maintained for 8 to 24 h. This ensures that the pipeline, at the time it is put into service, contains no defects that might induce failure at its operating stress level. The ability to preservice test a pipeline to a high stress level is a unique characteristic of a pipeline. The lines are designed such that they can withstand this type of test without undergoing plastic deformation that would induce residual stresses or damage the pipe.

One of the failure problems unique to gas pipelines was the potential length of a failure. Because the pressurized gas contains an enormous amount of stored energy and because the energy-release rate upon rupture is often slow relative to the speed of a propagating fracture, a fracture may propagate along the pipe axis for a considerable distance. The longest fracture on record was brittle in nature and propagated for a distance of 13.3 km (8.3 miles). Thus, pipelines are among the few structures for which the extent of fracture propagation is of concern.